INTERACTIVE EXPLANATION OF HIGH-DIMENSIONAL DATA PROJECTIONS

JULIAN THIJSSEN

Project Supervisor:

PROF. DR. ALEX TELEA

Second Examiner:

DR. MICHAEL BEHRISCH

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

Utrecht University

July 13, 2022

UTRECHT UNIVERSITY

STUDENT ID: 6987664

ABSTRACT

Companies, institutions and researchers around the world are collecting enormous sets of high-dimensional data at breakneck speed. However, our understanding of the collected data is not nearly keeping up. One of the main approaches to understanding these datasets has been to reduce the data to a low-dimensional representation, called a projection, that can subsequently be visualised.

Seeing visible patterns in these projections indicates there are relationships between the dimensions of the high-dimensional data. However, it does not tell us anything about what those relationships are. Several efforts have previously been made to explain the patterns in the projection in terms of their original dimensions. However, they tend to fall short in adequately explaining them, or the techniques don't scale well to a higher number of dimensions. Therefore, this thesis aims to answer the question how to adequately explain these patterns in projections of high-dimensional data, while simultaneously scaling better than previous techniques in the number of data dimensions.

We extend the variance-based explanations of previous work with a value-based explanation, that gives insight into, not only why the patterns are there, but what they represent. Furthermore, we introduce a user-driven exploration mechanism that provides significantly more detailed explanations of regions in the projection. In addition, these explanations are augmented by a number of tools that support their function. We integrate all of the above elements into a visualisation solution for exploring high-dimensional data projections.

We assess the visualisation system using an evaluation study asking a mix of 23 experts and non-experts to analyze several datasets of increasing dimensionality (12, 31, 58) using the proposed solution, as well as their opinion on the usefulness of each of the elements of the visualisation solution.

Participants rated each of the elements of the visualisation system highly in terms of their usefulness. In addition, with minimal training and by overwhelming majority, participants answered correctly to a series of twelve control questions meant to test whether they understood how to read the explanations generated by the visualisation system. On a series of nine more complex analysis questions, where participants had to use the system themselves, the majority gave answers that strongly aligned with our analysis. This indicates use of the system results in consistent insights about the data with only minor training or expertise required.

Overall, the evaluation study indicates that our visualisation solution is capable of providing detailed and consistent explanations of patterns in data projections, even as the dimensionality of the data gets higher.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

1 INTRODUCTION

1.1 INTRODUCTION

Many domains such as business, medicine, communications and research are becoming increasingly data-driven [15, 25]. Massive amounts of data are collected in these fields every day [2, 15]. This data is then analysed in order to gather insights and make informed decisions about the future [23]. However, such analyses are not always straightforward.

Generally, gathered data consists of a number of observations, or data points, that describe events, objects, people, or some other phenomena, and for every observation several measurable properties are recorded. These measurable properties, sometimes called features, variables, attributes or dimensions, describe a characteristic of the observation. For example, an observation might the test score on an exam of a particular student, and possible variables that influenced it may have been the number of hours they studied, how long they slept the night before, and so on. In a more complicated setting, an observation might be an earthquake, and its measured attributes, the magnitude, location, duration, time of origin, and so on.

Datasets collected nowadays are often huge, consisting of thousands to millions of data points. Each data point, in turn, may consist of multiple dimensions, such datasets are called *multidimensional* datasets. Moreover, complex (non-linear) relationships between the dimensions of the data may exist [1, 27] that are not easily brought forward by common analysis techniques [16].

If the number of dimensions associated with each data point is higher than in typical multidimensional datasets, the dataset can be considered as *high-dimensional*. Several of these dimensions may be completely irrelevant to the observation, for example, the population of Mexico on the observation of an earthquake happening in Japan. Other dimensions may be profoundly relevant to the observation, such as the current atmospheric pressure on the observation that it rains today.

When analysing such high-dimensional data, we are looking for patterns in the data. These patterns are interesting because they can tell us about hidden relations in the data. For example, which dimensions correlate with other dimensions, in which observations dimension values are out of the ordinary, even which observations themselves are out of the ordinary. As an example, we might find that the combination of certain food ingredients are harmful, learn about which abnormalities in the human genome cause a genetic disorder, or which factors caused stock prices to plummet.

However, obtaining such insights from looking at the raw data is practically impossible for a human. For low-dimensional data, such insights can be obtained from visualising the data directly. For high-

dimensional data however, this is not a feasible approach as visualisation of more than three dimensions is completely unintuitive to humans.

A common approach to deal with this problem of high-dimensionality is to employ the use of a *dimensionality reduction* technique [27]. Such a technique transforms the high-dimensional data points to a low-dimensional representation, called a projection, while maintaining certain properties of the high-dimensional space as much as possible.

A lot of dimensionality reduction techniques exist, but what many of them have in common is that they try to preserve the distance relationship between the high-dimensional points, in the reduced low-dimensional space. Data points that have similar values for many of their dimensions are close together in the high-dimensional space, whereas points with very different values are far apart. Therefore, in the projection we similarly hope to see patterns such as, groups of points that are close together, separation between these groups as well as points that are far away from other points (outliers). In essence, we are trying to find the patterns that may exist in the high-dimensional structure of the data, by looking at a low-dimensional representation.

However, the projection itself does not explain anything about the structures we may see. We do not know *why* these structures are there, and *what* they contain. That is, we do not know what dimensions are most important in causing data points to be similar to each other, or the opposite. Multiple techniques exist for explaining these local point structures [4, 10, 11, 28], however they do not relate back to the original dimensions in the data. Whereas, the techniques that do explain the structures in terms of their original dimensions, tend to fall short in their explanations as the dimensionality of the dataset increases [6].

1.2 RESEARCH QUESTION & CONTRIBUTIONS

Given these limitations in the previous work, in this thesis, we investigate how we can explain the patterns in projections of multidimensional data in enough detail while simultaneously allowing for better scaling in terms of dimensions than the current state of the art.

More formally, the research question and focus of this thesis is:

How can local point patterns in projection embeddings of highly multidimensional data be explained in terms of their original dimensions?

We present a visualisation solution addressing these limitations and the research question. In short, we extend the work of da Silva et al. [6] and present the following new contributions:

1. An additional global explanation metric based on dimension values
2. Interactive user-guided detailed local explanations
3. Differential analysis tool between regions in the projection
4. On-the-fly exclusion of dimensions in all explanations

We combine both the global and local explanation approaches into a visualisation solution augmented with the latter two contributions that support analysis of the data. In order to test the effectiveness of this visualisation solution, we run an evaluation study asking participants to use the various explanatory mechanisms to solve analysis tasks on datasets of increasing dimensionality. As far as we are aware, this is the first evaluation study of its kind, and gives us insight into the ability of our solution to explain local point patterns in projections, its ease of use in doing so, and its ability to scale to datasets of higher dimensionality.

Our visualisation solution is implemented in the HDPS [7] software application for exploring high-dimensional data. The source code and binaries of our implementation are available on GitHub at `https://github.com/JulianThijssen/ProjectionExplorer` [13].

1.3 THESIS STRUCTURE

We next provide an overview of the thesis, and briefly highlight the content of each chapter.

Chapter 2 explains the background of why our research is relevant, it discusses multiple other techniques for understanding multidimensional data and makes the argument for why we focus specifically on projections.

Chapter 3 gives an overview of previous work done in explaining multidimensional data projections, identifies current limitations and lays out the research question and aim of this thesis.

Chapter 4 describes our extensions to the previous work in providing global explanations of a projection, which serve as an entry-point for the user-guided exploration explained in the following chapter.

Chapter 5 introduces a user-guided exploration that is capable of providing detailed explanations of multi-scale regions of the projection.

Chapter 6 describes the design and results of the evaluation study that was performed for testing the ability of our solution to explain local point patterns in projections, its ease of use in doing so, and its ability to scale to datasets of higher dimensionality.

Chapter 7 discusses the contributions in light of the research question, results of the evaluation, limitations of the proposed solution and potential future work.

2 BACKGROUND

In this chapter the groundwork is laid for the problem that we are trying to solve in this thesis. Starting from the definition of multidimensional data, we survey several visualisation techniques that have been used over the years to attempt to explain these datasets. We explain why this thesis is primarily focused on one of these techniques, i.e. projections, and why it is that it needs additional explanation. Related work that focuses on explaining projections is explored in detail in Chapter 3. While the visualisation techniques mentioned in this chapter are also related work to the topic of this thesis, their mention, and the focus of this chapter, is primarily to motivate why projections are the visualisation technique of concern in this work.

2.1 WHAT IS MULTIDIMENSIONAL DATA

Real-world phenomena are often complicated. Events generally happen for many reasons and many factors play a role. When analysing why a company has made a loss in the past year, when the next earthquake will happen in Japan, what the weather will be like in three days, test scores for a university course, single factor analysis is rarely appropriate. The institutions doing the analysis on this type of data have gathered enormous datasets that can typically be laid out in a data matrix or table. The rows of this data matrix contain *observations* such as years in which the company has made a loss, earthquakes in Japan, the weather on previous days, the test scores of students on previous exams. The columns, on the other hand, describe certain factors (also called variables, attributes or dimensions) that constitute or have potentially influenced the observation.

Taking the example of the test scores for a university course, each row of the data matrix could convey a particular student who has taken the test. Each column could then convey an attribute of the test-taking student, e.g., their test score, how many hours they spent on learning, how much previous knowledge they had on the subject, how many other courses the student was taking at the same time, their age, how many hours they slept the night before, and so on. We would be interested to find out what if any effect each of these variables has on a student's performance on the test. In other words, we are interested in finding patterns and correlations between the variables. Understanding what factors contribute to a high test score could help other students to get higher test scores. For example, through analysis one might find that a small group of students got significantly lower scores than others and precisely all of those students had a poor night's sleep before the exam. Similarly, if one student got a very high score but also had a poor night's

sleep, we would then be interested in finding out what other variables could have contributed to this score.

How we can visualise such multidimensional data such that we are able to analyse it is a field of study called *Visual Analytics*. This field allows us to tackle analysis problems which require human interpretation and are not easily performed by a machine.

2.2 MULTIDIMENSIONAL DATA VISUALISATIONS

Multiple visualisations have been developed over the years that can be used for analysis of multidimensional data. We will mention a couple here to introduce the topic.

2.2.1 *Tables / Table Lenses*

One of the simplest visualisations is displaying the data in table (see Figure 2.1) where each row in the table represents an observation and each column in such a row contains the value of a certain dimension of that observation. While every detail of the data is available for the user to see, it is not at all straightforward to see patterns in the data. Manual comparison of the dimension values of potentially thousands of observations must be done, for every dimension of the data.

	A	B	C	D	E	F	G	H	I	J
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak
2	63	1	3	145	233	1	0	150	0	2.3
3	37	1	2	130	250	0	1	187	0	3.5
4	41	0	1	130	204	0	0	172	0	1.4
5	56	1	1	120	236	0	1	178	0	0.8
6	57	0	0	120	354	0	1	163	1	0.6
7	57	1	0	140	192	0	1	148	0	0.4
8	56	0	1	140	294	0	0	153	0	1.3
9	44	1	1	120	263	0	1	173	0	0
10	52	1	2	172	199	1	1	162	0	0.5
11	57	1	2	150	168	0	1	174	0	1.6
12	54	1	0	140	239	0	1	160	0	1.2
13	48	0	2	130	275	0	1	139	0	0.2
14	49	1	1	130	266	0	1	171	0	0.6
15	64	1	3	110	211	0	0	144	1	1.8
16	58	0	3	150	283	1	0	162	0	1
17	50	0	2	120	219	0	1	158	0	1.6
18	58	0	2	120	340	0	1	172	0	0
19	66	0	3	150	226	0	1	114	0	2.6
20	43	1	0	150	247	0	1	171	0	1.5
21	69	0	3	140	239	0	1	151	0	1.8
22	59	1	0	135	234	0	1	161	0	0.5
23	44	1	2	130	233	0	1	179	1	0.4
24	42	1	0	140	226	0	1	178	0	0
25	61	1	2	150	243	1	1	137	1	1
26	40	1	3	140	199	0	1	178	1	1.4
27	71	0	1	160	302	0	1	162	0	0.4
28	59	1	2	150	212	1	1	157	0	1.6
29	51	1	2	110	175	0	1	123	0	0.6
30	65	0	2	140	417	1	0	157	0	0.8
31	53	1	2	130	197	1	0	152	0	1.2

Figure 2.1: Table visualisation of a multidimensional heart disease dataset [8]. The rows list observations (patients), and the columns are the dimension values of those observations. The last column is coloured according to the value displayed in the cell. Low values are coloured lighter and high values are darker. This makes it easier to spot patterns in the data. (Image produced with Microsoft Excel [20])

A relatively simple addition to make this comparison process easier is to colour the cells of the table according to some heatmap based on the values of the dimensions. Low values would be on one end of the

colourmap, while high values would be on the other end. The comparison process then becomes a matter of seeing colour patterns. Still, such comparison may not be easy to do as we generally have limitations on the real-estate (i.e. screen space, or paper size) to show these tables. If the data consists of many observations then a lot of scrolling might be required to see every part of the data, during which other parts are out of view again.

By colouring the cells it becomes possible to compare based purely on the heatmap colours assigned to each cell. Therefore, what size we display each cell in is largely irrelevant. *Table lenses* make use of this to mitigate the space limitations to some extent by reducing the height of each table cell to a single or just a few screen pixels (see Figure 2.2). This allows for comparing vastly more observations at a time than before, meaning it becomes much easier to spot patterns and trends.

Figure 2.2: Table lens visualisation of the heart disease dataset. Dimension values are encoded as bars (longer bars mean a higher value) and bars at the same height correspond to different dimensions of the same observation. The bars are coloured according to the gender of the patient (purple = female, blue = male). The *chol* column specifies the levels of cholesterol of a patient. It appears that in this dataset women are more often associated with higher levels of cholesterol, as there is a higher concentration of purple near the top of the cholesterol values. (Image produced with High-D software [18])

2.2.2 *Scatterplots*

A visualisation that allows seeing all observations at once is a *two-dimensional scatter plot*. In a 2D scatter plot observations are typically drawn as dots in a two-dimensional coordinate system. The coordinates of the points in the plot can be derived from two user-selected dimensions of the input data, each of which is assigned to an axis of the coordinate system (see Figure 2.3). Similarly, three-dimensional scatterplots are constructed in an analogous way and are also used relatively frequently, but come with a few additional difficulties such as picking an appropriate viewpoint to view the data from, and difficulties in assessing distances and depths of the points.

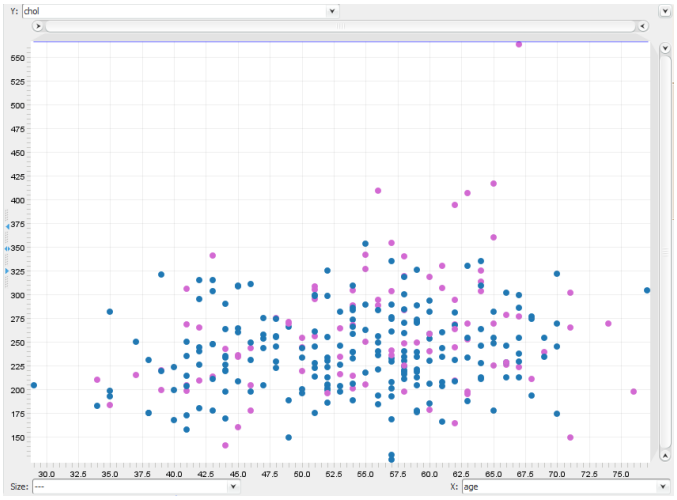


Figure 2.3: Scatterplot visualisation of the heart disease dataset. Observations (patients) are plotted in the scatterplot as dots, their locations determined by their values over the dimensions mapped to the axes. In this plot the X-axis is mapped to the age of the patients and the Y-axis to their level of cholesterol. The women with very high cholesterol values seen in Figure 2.2 are seen here as outlier values on the top half of the plot. (Image produced with High-D software [18])

In a two-dimensional scatterplot, patterns in the data can be found by looking at the distribution of the points in the plot. Points that lie close to each other are similar in the two user-selected dimensions. Likewise, points that lie along a horizontal or vertical line are similar in at least one of the dimensions. When a point is isolated from the rest of the points in the plot, this means it is different from other points in both dimensions and is called an *outlier*.

As a visualisation, scatterplots are easy to construct and interpret. However, there are some downsides as well. Drawn data points with very similar values in the dimensions used in the scatterplot can often overlap and obscure each other. This makes it hard to get an idea of the density of points at a certain location. One point drawn at a certain coordinate may hide several or many other points drawn at the same coordinate. One way to alleviate this problem is by drawing the points in a partially transparent manner using additive alpha blending. In this way, regions of high density will become more opaque than regions of lower density. Still, the degree of transparency with which the point is

drawn must be chosen carefully based on visual feedback. A different solution would be to render a density map in the plot space using, e.g., kernel density estimation, in order to make clear which regions in the plot have a higher or lower density of points.

Furthermore, when looking at a dataset with just two or three dimensions, the points in the dataset are simply visualised by drawing them in a two, or three-dimensional plot respectively and assigning one dimension to each plot axis. However, when the data has more than three dimensions, the original dimensions can not all be simultaneously mapped to the plot axes. Analysis of the data may then be performed by assigning a subset of two or three dimensions to the axes and analysing the data considering just these dimensions. The dimensions may then be switched out for a different subset and again assigned to the axes and analysed. This process can help explain the multidimensional data in terms of its various original dimensions, but is obviously impractical and intractable for data with many dimensions as it would require the user to switch back and forth between an inordinate amount of dimension subsets.

2.2.3 *Scatterplot Matrices*

The process of having to switch out and assign different subsets of dimensions to the scatterplot axes can be solved to some extent by utilising a *small multiples* [26] visualisation. Such a visualisation draws a series of charts in a spatial layout, where each chart shows different parts of the data. Applying this concept to scatterplots, we get a matrix of scatterplots also called a *SPLOM (Scatterplot Matrix)* [3]. Each cell of the matrix then shows a scatterplot using the data dimensions as axes that correspond to that cells matrix coordinates (see Figure 2.4). In this manner, one can see the scatterplots for more than two or three dimensions simultaneously.

Figure 2.4: Scatterplot Matrix visualisation of the heart disease dataset. (Image produced with High-D software [18])

While this visualisation allows for analysis over many dimensions, it is not always easy to do such analyses. Finding similar observations comes

down to finding which points are close together over multiple scatterplots. Without some identifying feature of which are the same observations on each scatterplot this would be impossible. Such a feature could be implemented by a linked selection, where selecting an observation or group of observations in one scatterplot selects it in all others. Nevertheless, finding patterns in the data using this method is very labour intensive and requires a lot of memorisation. Moreover, when the data has a large number of dimensions, the number of visualisations required would grow quadratically and be nearly impossible to deal with.

2.2.4 *Projections*

Rather than scaling the number of visualisations to the number of data dimensions, it is instead possible to scale down the number of dimensions. *Feature projection*, a type of *dimensionality reduction*, attempts to transform data from a high-dimensional space to a low-dimensional representation (called a *projection* or an *embedding*), while trying to retain some properties of the data such as distances between points in a local neighbourhood. In visual analytics, the high-dimensional space typically gets projected to two or three dimensions such that it can be easily visualised. The resulting low-dimensional projection can be plotted in a scatterplot as normal.

If any relationships between observations or dimensions of the input data exist, then these manifest themselves as structure in the high-dimensional space. As the dimensionality algorithm tends to attempt to retain the local structure of the high-dimensional data in the low-dimensional representation, patterns in the original data are expected to show up as patterns in the projection. This means that by looking at the projection, relationships in the data can be visually spotted in an easy and intuitive manner.

In theory, there is no limit to the number of observations or dimensions that can be reduced in this manner. Therefore, a big benefit of this technique is its scalability. However, in general the more dimensions and observations the data has, the harder it will be for the projection to faithfully represent the structures present in the high-dimensional space.

How well the dimensionality reduction managed to maintain the local high-dimensional structures is not immediately clear from the resulting projection. A lot of research has focused on ways to measure and/or visualise the quality of a projection [17, 19, 21, 22]. However, this is not the focus of this thesis, and we assume that the projection has a reasonable enough quality such that analysing it makes sense.

Figure 2.5 shows a projection produced by dimensionality reduction visualised in a scatterplot. All observations are captured in a single visualisation. Moreover, every observation can potentially be drawn using the most minimal representation possible on a display, a pixel. There-

fore, projections are not only algorithmically scalable in the number of observations and dimensions, but also visually very scalable.

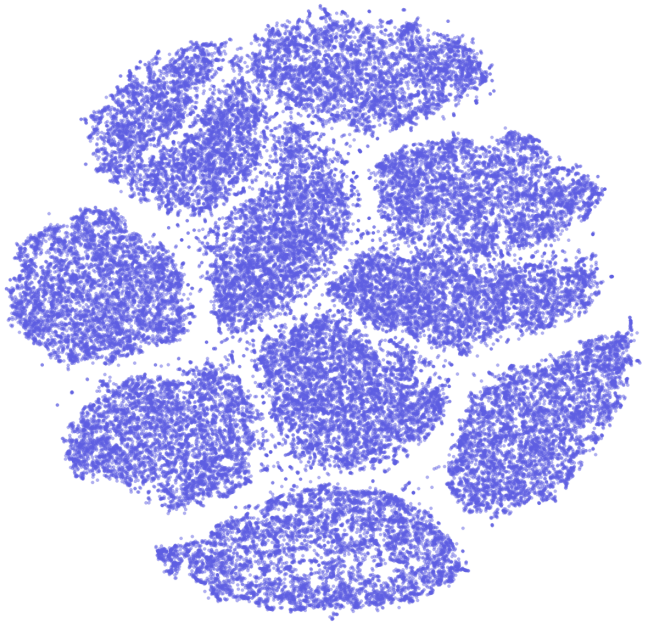


Figure 2.5: Scatterplot of a dimensionality reduction embedding. There are visible clusters of points meaning that the observations of those clusters are similar among a common subset of dimensions. (Image produced with HDPS software [7])

2.3 CONCLUSION

Several visualisations for multidimensional data have been discussed in this chapter, including their benefits and drawbacks. Projections stand out owing to their excellent scalability both algorithmically and visually, as well as how easy and intuitive it is to spot patterns in the data. For these reasons, the focus of this thesis lies in improving projection-based visualisation techniques.

What projections fail to show is why the patterns that can be seen have formed and what relationships these patterns indicate. In the next chapter we will therefore look at previous works that try to explain what is shown in the projection. We explore the taxonomy of work that has focused on explaining projections and discuss their strengths and limitations.

3 PROJECTION EXPLANATIONS

In the previous chapter we have discussed various visualisation techniques and their limitations. We have seen why projections are a valuable tool for rapidly discovering relationships between observations in multi-dimensional datasets. In this chapter we look at projections more closely and discuss a major challenge in understanding them, which forms the context of this thesis. We explore the related work that has focused on solving this challenge and discuss in what ways they succeed and what their limitations are.

3.1 UNDERSTANDING PROJECTIONS

In the figure showing a projection (see Figure 2.5) in the last chapter, it is possible to see clearly defined groups of points. What is not clear, is *why* these points seem to cluster together. That is, over which dimensions are points within the cluster similar to each other, and over which dimensions do they differ from points in other clusters. Thus, while projections allow us to quickly spot groups of similar and dissimilar observations, it is unclear what dimensions cause them to relate to each other in this manner. Understanding which dimensions contribute most strongly towards forming these groups of points is key in the analysis of projection. After all, we want to understand exactly what factors contribute to e.g. high stock prices, earthquake formation, developing a disease. Therefore, in order to be able to understand a projection in terms of its factors, it needs to be augmented with an *explanatory mechanism*. Such explanatory mechanisms should assist in understanding the relations between observations in terms of their dimensions.

It is worth mentioning that in a projection, like the one in Figure 2.5, which exhibits clearly defined groups of points, it would be possible to segment the projection into different clusters and attempt to assign a meaning to each of the segmented groups of points. After all, points within the same cluster must share some common values over their dimensions. Explaining the cluster by this group of common values can provide a lot of insight into the semantic meaning of such a cluster.

However, by far not all projections have such clearly delineated clusters. As an example the projection in Figure 3.1 shows regions of higher concentrations of points, but most of the projection is not clearly segmentable in any reasonable way. Therefore, an explanatory mechanism that is independent of the projection or projection technique used to generate it, must not rely on being able to perform segmentation.

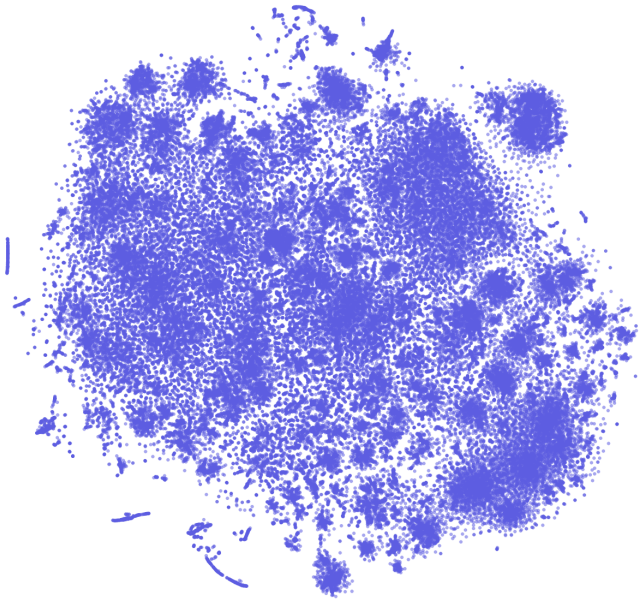


Figure 3.1: Scatterplot of a dimensionality reduction embedding. There are locations with a higher concentration of points and locations with a more sparse concentration, but there is no clear delineation between most clusters of points. (Image produced with HDPS software [7])

3.2 SOLUTION REQUIREMENTS

Before discussing the body of work that focuses on the issue of explaining projections, we define several requirements we consider to be crucial in being able to explain a projection adequately. This will give us a framework with which to evaluate the various explanatory mechanisms and identify their limitations.

We consider the following requirements key to a satisfactory explanatory mechanism, it must:

1. **Explain local point patterns in the projection in terms of their original dimensions**
   Projections can be explained in many different ways. The explanations may focus on, e.g., showing projection quality or local dimensionality. However, in order to understand the relationships between observations, we must understand them in terms of the original dimensions of the data.

2. **Be easy and intuitive in use**
   Performing analyses with the explanations must be as easy and intuitive for the user as possible. If it is not it will either be too hard to gather insights from them or nobody will want to use them.

3. **Scale well in the number of observations and dimensions**
   Multidimensional datasets oftentimes have a significant number of observations (sometimes running into the millions), and sometimes also have a large number of dimensions. Therefore, it is important

that the explanations are able to handle this in terms of staying effective at bigger scales.

4. **Scale well in terms of computational cost** Similar to the explanations scaling in terms of staying effective, they should scale computationally or they become impractical to use on larger datasets.

5. **Be applicable to projections in a black-box manner** The explanations should function without knowledge of how the projection was produced or access to the internal parameters of the projection technique. If it is specific to a projection technique then whenever a projection is generated using a different technique the explanations no longer apply, which is undesirable.

We believe these requirements cover most of what makes a satisfactory explanatory mechanism. The following sections will dive deeper into the existing work, their benefits and drawbacks, and how they relate back to the requirements laid out in this section.

3.3 TAXONOMY

We now explore some of the previous work in this field and how it relates to the requirements listed in section 3.2. We order and classify the collection of relevant previous work into several categories.

SPACE VS POINT-BASED EXPLANATIONS First of all, we differentiate between whether the explanatory mechanism explains the individual points in the projection by their original dimensions or properties, or whether it attempts to assign a meaning to the projection space. For example, in a dataset considering hobbies of people, one might find that people with hobbies generally performed by younger people tend cluster at one side of an axis, while people with hobbies performed by older people cluster on the other side. The axis along which these clusters are separated may then be assigned the semantic meaning of describing the age of the plotted individuals.

Other techniques produce a (low-dimensional) embedding of the data where the axes merely have a spatial meaning. That is, the axes merely convey the coordinates of the projected points. Here other mechanisms must be used to explain the plotted observations.

GLOBAL VS LOCAL In addition, some mechanisms attempt to explain the whole projection at once which we will refer to as *global explanations*. Others aim to explain only a subset of the projection at a time, often requiring some user-interaction such as hovering over or selecting individual points or clusters of points and only displaying the explanation for those points, which we will refer to as *local explanations*.

EXPLANATORY ELEMENTS The work is further classified according to what properties of the data are used in the explanation. Apart from

explanations based on the individual dimensions of the data, there are certain properties of the data which can give useful information about the visible structures. Examples of explanatory elements are the intrinsic dimensionality of the data in a particular structure, correlation between different dimensions or quality of the projection in a certain region.

3.4 RELATED WORK

We discuss the collection of related work in terms of the categories laid out in the previous section. It is practical to divide the works into two gross divisions, global approaches and local approaches. In general, mechanisms which attempt to explain the projection space tend to be global, and so the global approaches section will be subdivided between space and point-based explanations. Elements used in the explanation will be mentioned where applicable.

3.4.1 *Local Approaches*

Local approaches do not explain the whole projection in one go, but rather a subset of the projected points. Usually, this is paired with some user interaction where they select the points they are interested in, and the explanation for those points is shown.

BRUSHING WITH A TOOLTIP Explaining a subset of points by brushing over them and showing a tooltip is a very simple interactive way of explaining a projection. The user can move their cursor over, or select a particular point or close group of points (brushing) whereupon a little window, called a *tooltip*, will pop up to give additional information about those points. In the most basic case, it may simply show the dimension values associated with a single selected point. By then moving the cursor over points in a local neighbourhood, one can get an idea of which dimensions have similar values over all those points.

This technique is easy and intuitive to use and works regardless of the projection technique that was used to generate the embedding. It explains local point patterns, although the effort that a user must put into reaching the explanation depends strongly on what will be displayed in the tooltip. Only being able to select a single point at a time clearly does not scale well in the number of observations, and showing all dimension values in the tooltip does not scale well in the number of dimensions. Therefore, whether the technique scales in terms of both of these and additionally computational cost depends on the ability to select multiple points and get some form of aggregate explanation, where the explanations of multiple points are combined.

COLOURING POINTS BY DIMENSION Another most basic way to explain a projection by its dimensions is simply to pick a dimension and assign colours to the projected points based on their value in this

dimension. Commonly, this is done using a 1D-colourmap where the range of values the dimension takes on is mapped to the extents of the colourmap. Visually, one can then spot points or neighbourhoods where the dimension is highly expressed, and where it is not. While this can give valuable insights into the data it fails to capture why certain point patterns form.

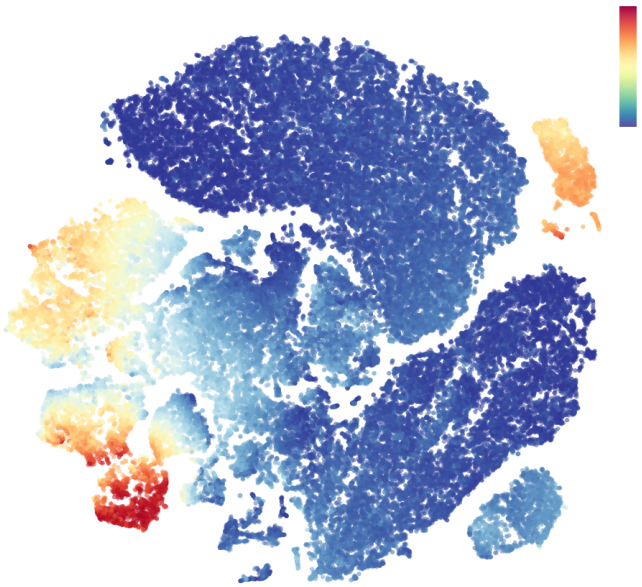


Figure 3.2: Projection visualisation where points are coloured according to their values over a single dimension. Colours are picked from a color gradient heat map, where high values correspond to red and low values to blue. (Image produced with HDPS software [7])

Going through all the dimensions present in the original multidimensional data and visualising them as colours can give a more complete picture of which dimensions cause a certain structure to form. However, it would require keeping keeping a mental map of which points showed high expression for a certain dimension and which did not. As the number of dimensions grows, this process simply becomes undoable. The problem can be alleviated somewhat by making use of a SPLOM-type visualisation, where each scatterplot can show the colours of a different dimension. However, this comes with all the drawbacks, in terms of visual scalability and difficulty in putting all the information together, as discussed in Chapter 2.

Still, the technique is easy and intuitive to use, works regardless of the projection technique, scales well in the number of observations and has almost no computational cost. These factors could explain why it is still commonly used in visual analytics.

FORCESPIRE The ForceSPIRE tool [10] takes a more exploratory approach to the interaction. Users can select points in the embedding and highlight certain dimension values of the point, which in turn will pull observations that share those values closer towards each other. In addition, points can be dragged to a different location in the embedding,

to which the rest of the points respond by adjusting their own locations. In this manner, the relationship with other points can be explored simply by how they respond in the embedding. Using these exploratory interaction mechanisms a user can get some idea of which factors are contributing to a local structure in the embedding. However, dragging around points in the embedding in order to understand the relations between points is quite a disruptive operation which may influence the analysis of other structures.

3.4.2 *Global Approaches*

Global approaches attempt to explain all of the observations in the projection at once. This means in general, that little interaction is required in order to understand why structures in the embedding are there. However, it is possible to augment such global explanations with interaction in order to provide more local information or guide the exploration. We subdivide this categorisation into explanations that attach meaning to the projection space, and explanations that attach meaning to the points or point-groups in the embedding.

3.4.2.1 *Explanations of the projection space*

BIPLOTS Biplots are a generalisation of a scatterplot of observations over two dimensions. Rather than only displaying the observations in the plot, they also display the dimensions. Often these dimensions are drawn as vectors which originate from a common point and spread out radially. Their direction can correspond to their relative contribution to the variance of some principal component eigenvectors computed for the plot axes. Or alternatively, when constructing a biplot by projection, the dimension vectors will point in the direction of maximal variation of that dimension. When these direction vectors are labeled with the corresponding dimension names, it is possible to get a general idea of why points-groups or outliers are formed by their location in the plot.

Biplots can work regardless of the projection technique used, however how the biplot dimension vectors are computed determines whether they are useful for finding patterns in the data. Even if computed in an optimal manner for explaining these patterns, the explanations themselves are quite broad and undetailed. Moreover, some experience with biplots and how they are computed is also required to understand what exactly is being conveyed. This knowledge might not be commonplace in all the fields where multidimensional data needs to be analysed. Biplots generally scale as well as scatterplots in the number of observations, but when the number of dimensions grows very large, or many dimensions contribute equally to the variance in the data, it may become hard to get any useful insight from the visualisation.

AXIS LEGENDS Broeksema et al. [4] propose adding several visual explanations to a biplot of projected multidimensional data. Firstly, they assign each dimension a colour from a categorical colourmap. Next, they

draw a Voronoi partitioning of the dimension plot, with the projected values as seeds (see Figure 3.3). The Voronoi cells are then coloured by their respective dimension colour and labeled with the dimension value. Inspection of this plot explanation can quickly reveal which dimensions or values are often seen together in the observations, or the contrary. Similarly, one might find that groups of values on one side of the plot indicate a certain property not described by the data when compared with values on the other side of the plot. This would give an implicit meaning to the axes found by the projection technique. This implicit meaning is reinforced by several bar charts or axis legends that show the contribution of original dimensions in the data to the respective axis.

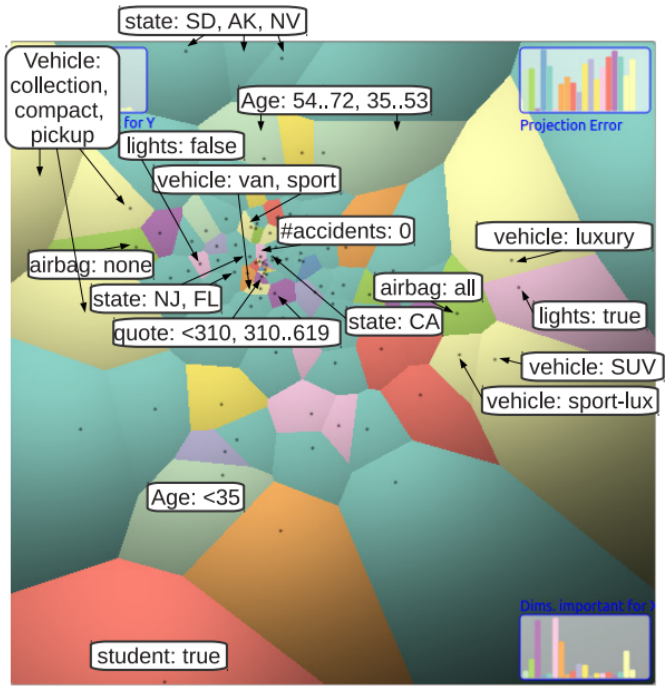


Figure 3.3: Axis legends visualisation of a projection. Voronoi cell colours represent their corresponding categorical dimension. Bar charts show the important dimensions along each spatial plot axis. Labels are added manually for illustration purposes. (Source: Paper by Broeksema et al. [4])

Similar to biplots, axis legends broadly explain why points are located in certain positions in the embedding, but the exact contributing dimensions that form groups of points remain unknown until deeper exploration using multiple visualisations. These visualisations are not immediately intuitive and need some explanation and experience to get comfortable with. The Voronoi cells depicting the dimension values of the data generally don't scale very well with many dimensions, but this has been somewhat alleviated by allowing cells to be merged for a more broad overview.

3.4.2.2 *Explanations of the projected points*

DIMENSION-BASED VISUAL EXPLANATIONS Da Silva et al. [6] propose a visual explanation based on computing a ranking of dimensions over a given local neighbourhood of projected points, where the lower the rank of a dimension the more it explains the similarity of points in the neighbourhood. The ranking is computed based on either the contribution of a particular dimension to the distances in the neighbourhood, or on the ratio of local variance of a given dimension in the neighbourhood to the global variance of that dimension over the whole projection. A number of top-ranking dimensions for most of the projection are selected and mapped to colours via a categorical colourmap. Points in the projection are then coloured by their top-ranking dimension and a legend is provided to link the colours to the dimension IDs (see Figure 3.4).

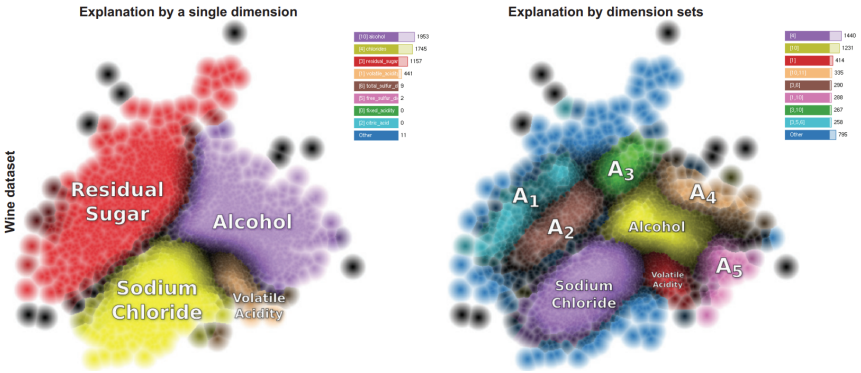


Figure 3.4: Dimension-based explanations of projections by da Silva et al. Points in the left projection are assigned a colour from a categorical colourmap for their most important dimension. Points in the right projection are assigned a colour for their unique set of important dimensions. Names of the dimensions or dimension sets are plotted on top of the groups of points with the same explanation. (Source: Paper by da Silva et al. [6])

For datasets with a moderate amount of specifically named dimensions, this approach can provide a lot of insight. Potential clusters in the projection are explained by their most important dimension, and it is clear and intuitive for the user to understand. The downside of this approach is that there are only a limited number of colours that can reasonably be used in a categorical colourmap. This restricts the number of important dimensions that are possible to visualise at once to a certain degree. Moreover, if a certain local point pattern is explained by multiple important dimensions, only one will be displayed to the user. In a dataset where many dimensions contribute roughly equally to a local point pattern such a visualisation may be particularly misleading. Similarly, when the dimensions are more abstract and not linked to a particular semantic identity, then showing the identity of such a dimension is usually not very helpful.

Another approach was introduced by van Driel et al. [28] where instead of computing the most important original dimensions, they computed several dimensional properties of the projection, such as

intrinsic dimensionality and correlation over a local neighbourhood. Different dimensionalities are then again mapped to different colours and the points are coloured accordingly. A legend is provided that links the colours with a certain dimensionality.

In contrast to the method by da Silva et al. [6] these explanations scale excellently in the number of dimensions of the data. However, merely seeing the intrinsic dimensionality of the local neighbourhoods doesn't convey a lot of information or insight. Especially in high-dimensional datasets where the intrinsic dimensionality of a local neighbourhood may be quite high. Furthermore, in the analysis of multidimensional datasets, one is generally interested in the contribution of original dimensions to the forming of local patterns, rather than some set of abstract intrinsic dimensions.

3.5 CONCLUSION

The explanatory mechanisms introduced in this chapter each satisfy a number of the desirable properties of a solution. Brushing with tooltips and colouring points by their dimension are easy in use, but do not scale in the number of observations or dimensions respectively. ForceSPIRE, biplots and axis legends tend to scale a bit better in this regard, but lack some ease of use. However, critically, none of the aforementioned techniques are able to satisfy the primary criterion of adequately explaining local point patterns. The approaches by da Silva et al. and van Driel et al. are specifically aimed at explaining these patterns and succeed in doing so to some extent.

The method proposed by da Silva et al. provides insight into why groups of points have clustered together and why outliers are different from other points. However, it does not explain anything about the semantic content of the clusters, i.e. what type of observations are contained in these clusters. Their method attempts to explain why points are similar or dissimilar by visualisation of the least varying dimension or particular subset of dimensions. In both cases, not all of the dimensions of the dataset are involved in the explanation and therefore important information may be missed. Moreover, the number of dimensions or dimension sets that can be involved in the explanation is inherently limited due to the fact that they need to be assigned visually distinct colours from a categorical colourmap. All this results in an inability to adequately explain the projection as the intrinsic dimensionality of the dataset grows. That is, as more and more dimensions are involved in contributing to the variance of the data.

On the other hand, the local dimensionality explanation proposed by van Driel et al. scales excellently with the number of dimensions, but the explanations do not relate back to the original dimensions of the data. The correlation explanation does use the original dimensions to explain local point patterns, but again does not scale well in the number of dimensions.

As there are many fields where the multidimensional data has a large number of dimensions, and analysis of point patterns is generally concerned with the contribution of the data dimensions on those patterns, there is clearly a lack of explanatory techniques that can handle such data. Therefore, we investigate how to provide local point pattern explanations on projections of data with a number of dimensions higher than the above techniques can handle.

More formally we define the follow main research question:

How can local point patterns in projection embeddings of highly multidimensional data be explained in terms of their original dimensions?

In order to answer the research question formulated, we aim to produce a solution that ideally satisfies all the requirements laid out in Section 3.2. The first thing to note is that the technique proposed by da Silva et al. explains local point patterns very clearly and intuitively for data with a low intrinsic dimensionality. The main limiting factor on the number of dimensions able to be explained in their method is that they attempt to globally explain all point neighbourhoods in the projection. It is common for the various neighbourhoods to be explained by different subsets of dimensions. However, as every neighbourhood explained uniquely by a certain subset of dimensions is assigned its own colour from a categorical colourmap, it is clear that one quickly runs out of visually distinct colours. The main research direction of this thesis will therefore be to investigate how to provide similar explanations for data with a higher intrinsic dimensionality.

In chapters 4 and 5, we propose several additional explanatory mechanisms that allow for both a global and local exploration of high-dimensional data embeddings. The local explanations scale to a significantly higher number of dataset dimensions, allowing for more complex data to be analysed. Finally, we implement all those mechanisms in a comprehensive projection exploration system.

We test the functionality and practical applicability of this system by running an evaluation study. In the study we ask experts and non-experts in the field of high-dimensional data analysis to analyse several datasets of increasing dimensionality using the proposed system. For each dataset, we ask several questions that serve to establish whether participants reach consistent insights using the various elements of the system. The set up of the evaluation study and its results are laid out in Chapter 6.

4 EXTENDING GLOBAL EXPLANATIONS

In Chapter 3 we presented the work of da Silva et al. [6] which proposes a global approach to explaining why points in a projection are close together or far apart. They introduced a colouring of the projected points according to the dimension with least variance in a local neighbourhood.

We discussed that this method has a problem with scaling to datasets with a higher local dimensionality. However, we believe that for many datasets these global explanations can still serve very well as an indication of where transitions in the dimension profiles of groups of points happen. That is, borders between differently coloured groups of points often seem to indicate a different set of dimensions is prominent on either side of the border. Such a visualisation can therefore be an excellent entry point into a more detailed analysis of the projection.

Taking this into account, we incorporated the work of da Silva et al. as a base for the more detailed local explanations that we explain in Chapter 5. In this chapter we describe the partial work of da Silva et al. that we include in our proposed answer to the research question, as well as proposing an additional global explanation method.

4.1 WINE DATASET

One of the datasets used for demonstration in da Silva et al. [6] is a dataset of Portuguese wine samples. In order to validate our implementation of the techniques described there and to serve as an example dataset for our proposed explanations, we briefly introduce it here.

The dataset [5] consists of roughly 6500 samples of wine from Portugal. For each of these wine samples 11 physicochemical properties of the wine were measured, and one quality attribute was assigned through sensory evaluation by human experts that graded the wine on a scale from 0 (very bad) to 10 (excellent), for a total of 12 attributes.

A projection of the dataset was computed using Local Affine Multidimensional Projection (LAMP) [12] and is shown in Figure (4.1).

4.2 VARIANCE RANKING

We make use of the global colouring scheme based on variance ranking introduced by da Silva et al. Therefore, we briefly reiterate their idea here.

LOCAL NEIGHBOURHOOD As the method tries to provide an explanation of why points are close together in the projection, it makes sense to define a local neighbourhood over which this explanation applies. That is, how many points are considered in the explanation. To this

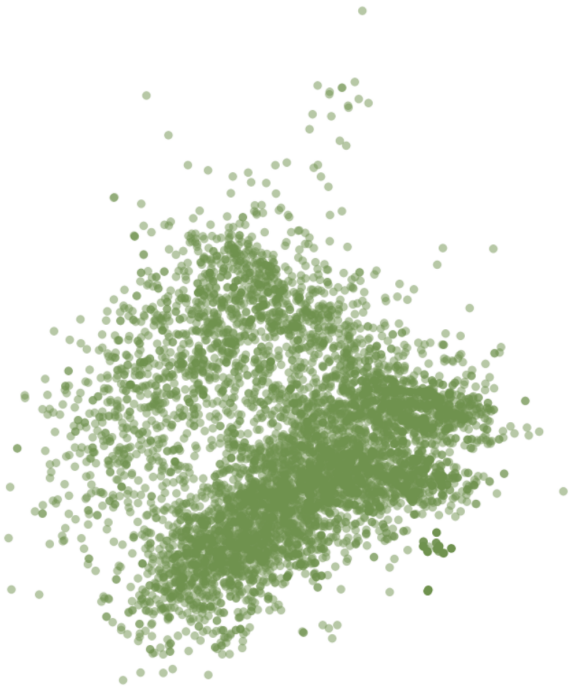


Figure 4.1: Projection of wine dataset computed using LAMP.

end, da Silva et al. define a 2-dimensional local neighbourhood around each point in the projection, where all projection points closer to $\boldsymbol{q}_i$ than some radius ρ are part of its neighbourhood (see Equation 4.1).

$$\boldsymbol{v}_i^P = \{\boldsymbol{q} \in D^P \mid \quad \|\boldsymbol{q} - \boldsymbol{q}_i\| \leq \rho\} \tag{4.1}$$

As each of the projected points corresponds to a point in the high-dimensional space, this local neighbourhood in the projection is equivalent to a neighbourhood in nD given by:

$$\boldsymbol{v}_i = \{\boldsymbol{p} \in D \mid \quad P(\boldsymbol{p}) \in \boldsymbol{v}_i^P\} \tag{4.2}$$

where $\boldsymbol{p}$ is some nD point part of the high-dimensional dataset D which is part of the neighbourhood defined in the projection if the corresponding projected point $P(\boldsymbol{p})$ is part of the local neighbourhood v_i^P in 2D.

DIMENSION RANKING Every point in the projection defines its own local neighbourhood and therefore has its own explanation. Which dimension best explains why points in this neighbourhood are similar is computed through ranking the dimensions based on their variance in the neighbourhood. The dimension ranking $\boldsymbol{\xi}_i = (\xi_i^1, \ldots, \xi_i^n) \in \mathbb{R}^n$ for a particular point consists of a list of ranks ξ_i^d for every dimension d in $\mathbf{p}_i$. Ranks with a lower value mean that this dimension has lower variance in the local neighbourhood and therefore has a bigger impact on the similarity of the points contained in this neighbourhood.

Ranks are computed based on the variance in the local neighbourhood which is defined as

$$LV_i^d = \frac{1}{|v_i|} \sum_{\boldsymbol{p} \in \boldsymbol{v}_i} (\boldsymbol{p}^d - \boldsymbol{\mu}_i^d)^2 \quad \text{with} \quad \boldsymbol{\mu}_i^d = \frac{1}{|v_i|} \sum_{\boldsymbol{p} \in \boldsymbol{v}_i} \boldsymbol{p}^d \tag{4.3}$$

where LV_i^d is the variance of the local neighbourhood around $\boldsymbol{p}_i$ for dimension d, calculated by summing the squared differences between a point $\boldsymbol{p}^d$ in the local nD neighbourhood v_i and the mean over that neighbourhood $\boldsymbol{\mu}_i^d$ for dimension d and dividing over the number of points in the neighbourhood.

It may be the case that a particular dimension has a very low variance over the dataset as a whole. These dimensions would not be a good explanation of why a cluster of points has been grouped together, as it has very similar values in the rest of the dataset. Therefore, da Silva et al. give more weight to dimensions that have local variances different from their global variance, defined as

$$GV^d = \frac{1}{|D|} \sum_{\boldsymbol{p} \in D} (\boldsymbol{p}^d - \boldsymbol{\mu}^d)^2 \quad \text{with} \quad \boldsymbol{\mu}^d = \frac{1}{|D|} \sum_{\boldsymbol{p} \in D} \boldsymbol{p}^d \tag{4.4}$$

where GV^d is the variance of the whole dataset for dimension d, calculated by summing the squared differences between the value of dimension d for all points $\boldsymbol{p}^d$ in the dataset D, and the mean $\boldsymbol{\mu}^d$ of that dimension over the dataset, and dividing over the number of points in the dataset.

The ranking ξ_i^d of a particular dimension d for a particular point i is then calculated as the ratio of local variance for point i and dimension d to the global variance of dimension d and normalised to indicate relative importance (see Equation 4.5).

$$\xi_i^d = \frac{LV_i^d / GV^d}{\sum_{j=1}^{n} (LV_i^j / GV^j)} \tag{4.5}$$

Dimensions that have a lower rank value for some point $\boldsymbol{p}_i$, are seen as more important to the points in the neighbourhood around $\boldsymbol{p}_i$ being positioned close to each other in the projection.

COLOURING Dimension rankings are computed for every point in the projection. The dimension that has the lowest rank value will be used to colour that particular point in the projection. Colours are assigned to dimensions from a categorical colourmap. Which and how many dimensions are assigned colours is discussed later on in this chapter.

A projection of the wine dataset with the points coloured according to their variance ranking results in Figure 4.2.

The projection is roughly divided into four differently coloured clusters of points. For the points coloured in light-blue, the `residual sugar`

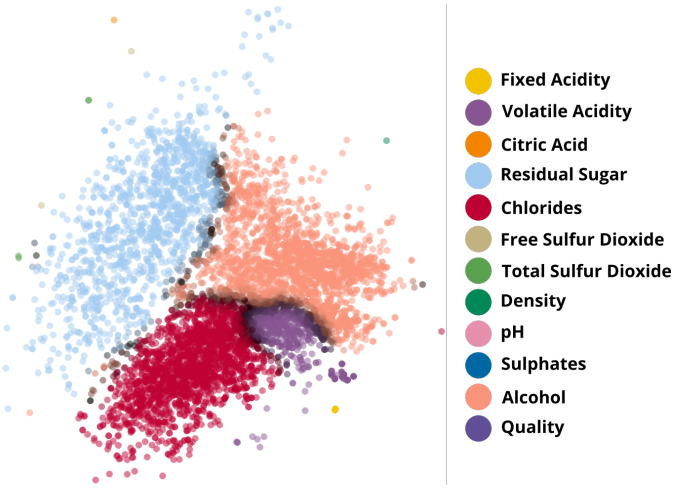


Figure 4.2: A projection of the wine dataset with points coloured according to the least varying dimension in a region around the point.

dimension has an unusually low variance. For the red, purple and salmon-coloured points, respectively the `chlorides`, `quality` and `alcohol` dimensions have unusually low variances.

4.3 VALUE RANKING

The variance ranking is useful for explaining *why* points in the projection are close together, but it does not explain *what* those points represent. In order to be able to get insights into this, we extend the global explanations with a colouring of the projection based on the values of the dimensions.

The value ranking uses the same definitions of a local neighbourhood and dimension ranking as the variance ranking. However, the way the ranks are computed is different.

DIMENSION RANKING As in the variance ranking, the value dimension ranking $\boldsymbol{\xi}_i = (\xi_i^1, \ldots, \xi_i^n) \in \mathbb{R}^n$ for a particular point consists of a list of ranks ξ_i^d for every dimension d in $\mathbf{p}_i$. However, different to the variance ranking, here ranks with a higher value indicate that this dimension has higher values in the local neighbourhood, while lower values indicate the dimension has lower values in the local neighbourhood.

A rank ξ_i^d is computed based on the average value dimension d takes on in a local neighbourhood around point $\boldsymbol{p}_i$, which is defined as

$$LA_i^d = \boldsymbol{\mu}_i^d = \frac{1}{|v_i|} \sum_{\boldsymbol{p} \in \boldsymbol{v}_i} \boldsymbol{p}^d \tag{4.6}$$

where LA_i^d is the mean average value of dimension d over the local neighbourhood around point $\boldsymbol{p}_i$. This definition is exactly the same as the one given for $\boldsymbol{\mu}_i^d$ in Equation 4.3.

As before it may be the case that a particular dimension has the same or very similar values for all the points in the dataset. In this case the dimension would not provide useful information about what makes the local neighbourhood of points special. Therefore, we instead compare the average value of the dimension over the local neighbourhood with the average value over the whole dataset defined as

$$GA_d = \boldsymbol{\mu}^d = \frac{1}{|D|} \sum_{\boldsymbol{p} \in D} \boldsymbol{p}^d \tag{4.7}$$

where GA_d is the mean average value of dimension d over the whole dataset. This definition is exactly the same as the one given for $\boldsymbol{\mu}^d$ in Equation 4.4.

The value ranking ξ_i^d of dimension d for a particular point i is then calculated as the difference of the mean average value of dimension d over the local neighbourhood v_i and its mean average value over the whole dataset D. As different attributes of a dataset often have very different ranges, we normalise the difference between the averages by dividing over the data range of the dimension over the whole dataset. As in the variance ranking, the value ranks are then also normalised to indicate relative differences (see Equation 4.8).

$$\begin{aligned} \xi_i^d &= \frac{1}{range(d)} \frac{LA_i^d - GA^d}{\sum_{j=1}^{n} \frac{1}{range(j)} |LA_i^j - GA^j|} \quad \text{with} \\ range(d) &= max(\boldsymbol{p}_1^d, \ldots, \boldsymbol{p}_{|D|}^d) - min(\boldsymbol{p}_1^d, \ldots, \boldsymbol{p}_{|D|}^d) \end{aligned} \tag{4.8}$$

Dimensions that have a positive rank value can be considered to be *unusually* high in the local neighbourhood around some point $\boldsymbol{p}_i$, whereas dimensions with a negative rank value can be considered to be unusually low. The higher or lower the rank value the more unusual the dimension values in that neighbourhood are. These rankings give an insight into what makes the points in that local neighbourhood special.

COLOURING Value rankings are computed for every point in the projection. In many datasets especially high values are more interesting than especially low values, therefore we colour points in the projection according to the dimension with the highest value rank. Although, whether the highest ranks or the lowest ranks are used for colouring could be easily adjusted. In some cases it may be even be useful to use the rank with the highest absolute value, resulting in a colouring that mixes both unusually high and unusually low values. Colours are assigned to dimensions from a categorical colourmap and the assignment will be discussed later in the chapter.

The projection of the wine dataset with the points coloured according to the highest value ranking results in Figure 4.3.

The projection is roughly divided into seven differently coloured clusters of points. Each of these coloured clusters corresponds to a

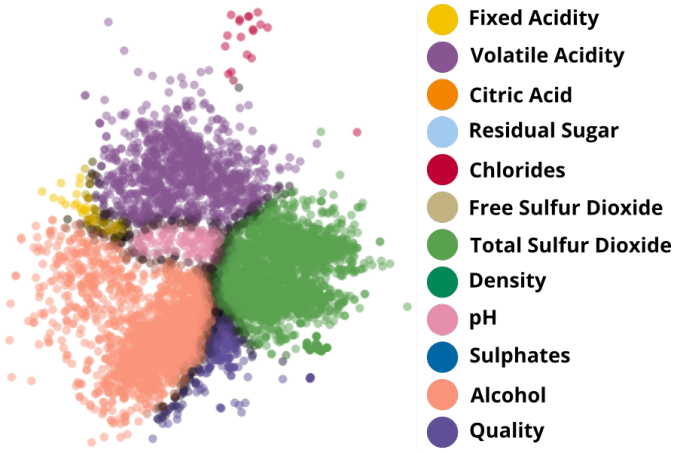


Figure 4.3: A projection of the wine dataset with points coloured according to the dimension with the most unusually high values in a region around the point.

dimension that has a particularly high value in the local neighbourhoods around points in that cluster. For example, wines in the salmon-coloured cluster appear to have an unusually high alcohol percentage (the salmon colour corresponds to the alcohol dimension, see legend).

4.4 CONFIDENCE

The projection may contain regions where several of the top-ranked dimensions have very similar rank values and therefore the top-ranked dimension for one point might not be the same as for the point next to it, resulting in a noisy colour allocation to these points. This frequently occurs on the borders between two differently coloured clusters as these mark the transition from one top-ranked dimension to another. In these regions we can not say with confidence that the top-ranked dimension is more important than the next.

4.4.1 *Rank-based Confidence*

Therefore, da Silva et al. proposed computing a measure of our *confidence* in the chosen top-ranked dimension. This measure is calculated for each point by looking at a local neighbourhood v_c^P centered at the projected point $\boldsymbol{q}_i$, and defined in the same way as v^P but with a smaller radius $\rho_c < \rho$. For every point in this neighbourhood that has the same top-ranked dimension as point $\boldsymbol{q}_i$, the rank value is summed up. This is then divided by the rank of that dimension for all the points in the same neighbourhood. The result is a fraction that indicates the extent to which this top-ranked dimension is ubiquitous in the local neighbourhood v_c^P (see Equation 4.9).

$$C_i^d = \frac{1}{\sum_{\boldsymbol{q} \in v_c^P} \xi^d(\boldsymbol{q})} \sum_{\boldsymbol{q} \in v_c^P} \begin{cases} \xi^d(\boldsymbol{q}) & \text{if } d \text{ is top ranked for } \boldsymbol{q} \\ 0 & \text{otherwise} \end{cases} \tag{4.9}$$

A high confidence C_i^d indicates the top-ranked dimension belonging to $\boldsymbol{q}_i$ is also commonly the top-ranked dimension for the other points in v_c^P. A low confidence means the top-ranked dimension for $\boldsymbol{q}_i$ is not commonly the top-ranked dimension in the other points.

4.4.2 *Simplified Confidence*

A problem with this measure of confidence is that when the ranks of a top-dimension in a particular neighbourhood are zero (which could easily happen with variance ranking), then there may be a division by zero. In fact, this equation for confidence in essence measures the ratio of occurences of a given top-ranked dimensions among the points in the local neighbourhood. However, due to the nature of summing the rank values, confidences are skewed in different directions based on the exact distribution of rank values in the neighbourhood.

We adopt a very similar measure of confidence that is less complex to understand by doing away with the summing of rank values. It implements confidence purely as the ratio of occurences of a given top-ranked dimension among the points in the local neighbourhood (see Equation 4.10). It has the added benefit of avoiding the division by zero issue.

$$C_i^d = \frac{1}{|v_c^P|} \sum_{\boldsymbol{q} \in v_c^P} \begin{cases} 1 & \text{if } d \text{ is top ranked for } \boldsymbol{q} \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

4.4.3 *Colour Encoding*

This confidence value is encoded in the brightness of the colours assigned to every point in the projection. Projected points that have a high confidence in their top-ranked dimension are drawn with a bright colour, whereas points with a low confidence value are drawn less bright. This effect is achieved by multiplying each RGB-component of the colour by the confidence value (which has a range of 0 through 1).

4.4.4 *Relative Confidence*

The two confidence measures introduced in the previous sections compute an absolute measure of confidence. This is a practical measure if there are both points with low confidence and high confidence. However, if all points are either low or high confidence, then the differences in brightness of projected points will be barely noticable.

Understanding which points have a lower or higher confidence in relation to the rest of the projection can still give practical insights. To address this, a relative measure of confidence can be computed from the absolute confidences. This is achieved by normalising the absolute confidence values to fall between 0 and 1.

In both Figure 4.2 and Figure 4.3 a projection of the wine dataset is shown. Along the borders between points that have different top-ranked dimensions the colours fade to black indicating low relative confidence.

4.5 COLOUR ALLOCATION

How colours are assigned to particular dimensions is particularly important when dealing with datasets that have more than roughly 15 dimensions as it becomes difficult to find enough visually distinct colours to assign a colour to every dimension.

4.5.1 *Colour Assignment*

The approach taken by da Silva et al. is to assign the colours from their categorical colourmap to dimensions based on their frequency of being picked as top-ranked dimension. That is, the dimension that is top-ranked for most of the points is assigned the first colour. The dimension that is top-ranked for the most points after that is given the second colour, and so on until all colours have been assigned to a dimension. If at any point during the process there is a tie in the frequencies of a dimension being top-ranked, the colour is assigned randomly to one of the tied dimensions. Dimensions that did not get assigned a colour in this process are assigned a neutral grey colour at the end. This manner of colour allocation works very well and we employ the same mechanism. As a colourmap we make use of Kenneth Kelly's 22 colours of maximum contrast [14], however we exclude white and black from this map as they would not contrast with our white projection view background and dark explanation widget background, leaving us with a total of 20 colours.

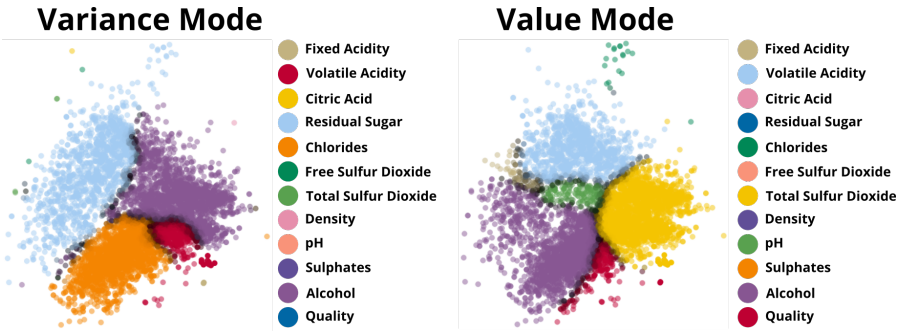
4.5.2 *Consistency*

During the exploration process, there are several factors that affect the colouring of the projection. Primarily, switching between different ranking metrics results in rank recomputation and therefore points need to be assigned updated colours. Moreover, the size of the neighbourhood used in the computation of the ranks may also be changed to match better to the projection, in which case ranks must be similarly recomputed and colours updated.

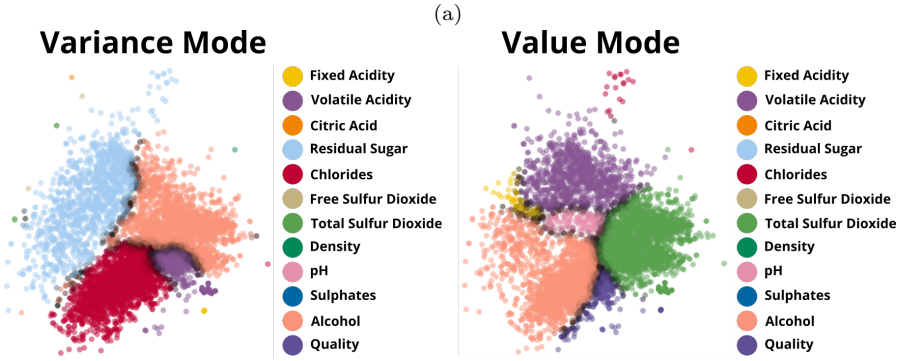
It does not suffice to let the dimensions keep their assigned colours, because the top-ranked dimensions in variance ranking may be a substantially different set to the ones in value ranking. In addition, changing the neighbourhood size can cause dimensions that weren't top-ranked before to suddenly become so, or vice versa. Because of this, triggering a recomputation of the ranks through the above mechanisms, could result in a significantly different colour allocation. Dimensions being associated with colours that change every once in a while is very confusing to deal with and disturbs the exploration process.

Therefore, we attempt to keep the colour allocation as consistent as possible throughout these changes. At the start of the exploration an initial colour allocation is computed based on the ranking mode that is in effect (variance or value). Whenever the exploration process triggers an update of the dimension ranks, a new colour allocation is computed. Dimensions that are part of both the new and the previous allocation, get assigned the same color. The other colours are distributed to the rest of the dimensions in the new allocation based on their frequency of being top-ranked as before.

In Figure 4.4, we show how the colour allocation changes when switching between variance and value mode. If not kept consistent, each mode has a completely different mapping of dimensions to colours because colours are assigned by ranking of the dimensions. With keeping color consistency, dimensions which appear in both the previous mapping and the new one are assigned the same colours. For the wine dataset this results in completely consistent dimension colours in variance and value mode.



(a)



(b)

Figure 4.4: Colours assigned to the wine projection in variance and value mode without keeping colours consistent between modes (a) and with keeping them consistent (b)

5 ADDING LOCAL EXPLANATIONS

In the previous chapter we looked at extending the global explanations proposed by da Silva et al. with additional ranking metrics, an altered confidence metric and consistent colour allocation.

However, global explanations are limited in the insights they can provide about local patterns in the projection. When attempting to explain all of the points in the projection at the same time in terms of their dimensions, space quickly runs out for visualising these explanations.

Consider two clusters of points that are distinct from each other in the projection. Barring any projection errors, this generally means that the dimension profiles, i.e. the values that dimensions take on in those clusters, are sufficiently different from each other, otherwise their points would form a single cluster. As there are more distinct clusters or outliers in the projection, the number of distinct dimension profiles increases. In order to fully explain the projection, all these distinct dimension profiles would need to be visualised at the same time. In fact, single clusters might have inter-cluster variability for several of their dimensions, for example, a dimension might have low values at one end of the cluster and high values at the other. Even more, a point with quite different values in one dimension might be located next to a point with small differences over all dimensions, simply because their distance to a third point is the same. All this shows, that there is a granularity to the level at which a projection can be explained. Global explanations at all these levels of detail are simply not feasible as there is not enough visual space to show the explanations. For that reason, we let go of trying to explain the full projection at once in this chapter, and focus on providing more detailed explanations for regions of the projection.

5.1 INTERACTIVE EXPLORATION

In addition to the global methods outlined in the last chapter, we take a more local approach in order to analyse parts of the projection in detail. This detail is possible owing to the fact that instead of attempting to explain all of the projected points at once, we explain only a subset. Through interactive exploration different parts of the projection can be highlighted and explained on demand.

5.1.1 *Lens Brushing*

There are many different ways in which points can be selected in a scatter plot. Some of the most common selection mechanisms include: box selection, polygon selection, lasso selection, brush selection. However,

when being presented with a projection of a complex dataset, exploration with these tools can quickly become cumbersome as there are usually many areas of the projection that are interesting for the analysis. Having to select areas over and over at different positions and scales is an unwieldy way to explore. Moreover, these selection tools demarcate artificial boundaries in the projection. A projection is a representation of the continuous high-dimensional space and therefore it is not appropriate to explain it as if consisting of a number of clearly separated clusters.

We provide a brushing tool that is more ideally suited to exploring such projections in a continuous and user-friendly manner. In much the same way as a looking glass, the user selects a circular region of points in the projection by dragging a circular brushing tool over it. Local explanations that give more detailed insight into the selected points are generated on the fly, akin to getting a more detailed look at something through the lens of a looking glass. The radius of the *lens brush* can be scaled up or down interactively, allowing for exploration at multiple levels of detail.

While the lens brushing mechanism still creates an artificial boundary of selected points, the key point is that it can be moved fluidly over the projection on multiple scales to get an integrated understanding of particular regions. A visual example of points selected with the lens brushing tool is shown in Figure 5.1.

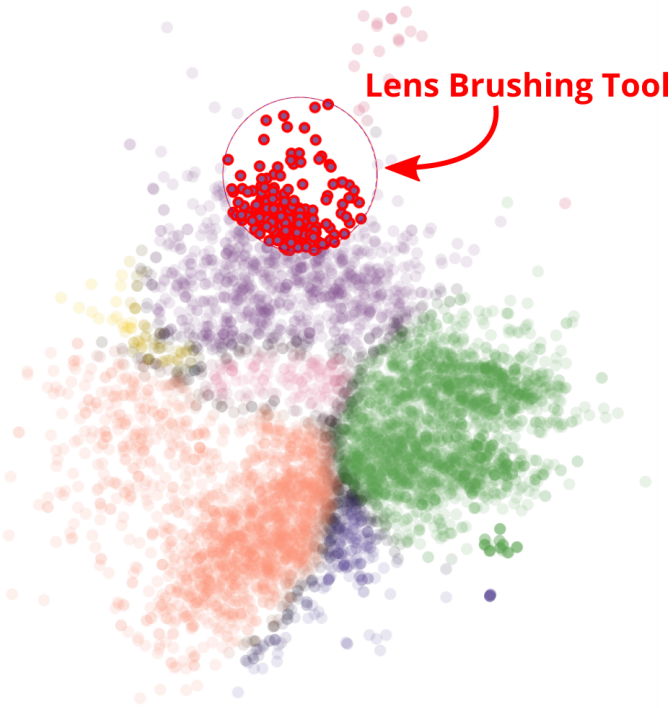


Figure 5.1: A projection of the wine dataset. The lens brushing tool is displayed as a red circle at the top and points within this circle are selected.

5.2 LOCAL RANKING

Exploring the projection using the lens brushing tool selects a subset of points in the projection that can then be explained in more detail than would be possible in a global approach. Such local explanations are generated in much the same way as in the global ranking schemes outlined in Chapter 4. The integration of both the global and local explanations result in a comprehensive visualisation system for understanding high-dimensional projections at any level of detail.

The system supports two modes for exploration of the projection. In *variance mode* the global and local explanations can be used to understand *why* points are close together in the projection. While, in *value mode* the global and local explanations can be used to understand *what* those points represent.

5.2.1 *Variance Mode*

In variance mode, colours are assigned to the projected points based on the global variance ranking explained in Chapter 4.2. Similarly to those global variance rankings, we compute a ranking of dimensions based on variance for the local selection of points. The variance rank ξ^d of a dimension is computed in the same way as in 4.5, except that it is not computed per point, but as an aggregate over the selected points $\mathcal{S}$ as follows

$$LV^d = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{p}\in\mathcal{S}}(\boldsymbol{p}^d - \boldsymbol{\mu}^d)^2 \quad \text{with} \quad \boldsymbol{\mu}^d = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{p}\in\mathcal{S}}\boldsymbol{p}^d \tag{5.1}$$

$$\xi^d = \frac{LV^d/GV^d}{\sum_{j=1}^{n}(LV^j/GV^j)} \tag{5.2}$$

Ranks computed in this way are displayed for all dimensions in a separate widget beside the projection.

5.2.2 *Value Mode*

In value mode, colours are assigned to the projected points based on the global value ranking explained in Section 4.3. The value rank ξ^d of a dimension is computed in the same way as in 4.8, except that it is computed over the selected points $\mathcal{S}$ as a whole as follows

$$LA^d = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{p}\in\mathcal{S}}\boldsymbol{p}^d \tag{5.3}$$

$$\xi^d = \frac{1}{range(d)} \frac{LA^d - GA^d}{\sum_{j=1}^{n} \frac{1}{range(j)} |LA^j - GA^j|} \quad \text{with}$$
$$range(d) = max(\boldsymbol{p}_1^d, \ldots, \boldsymbol{p}_{|D|}^d) - min(\boldsymbol{p}_1^d, \ldots, \boldsymbol{p}_{|D|}^d) \tag{5.4}$$

5.3 LOCAL ANALYSIS WIDGET

When points are selected in the projection, local explanations of the selected points are generated and shown in a *local analysis widget*. This widget is situated beside the projection and all dimensions of the data are listed there with their associated colour, as computed according to Section 4.5. The order in which dimensions are listed is discussed in 5.3.1. Statistics about the dimension values associated with the selected points are shown next to each of the listed dimensions. The details of this are discussed in 5.3.2. A legend explaining the details of the local analysis is shown in a collapsible box underneath the local analysis widget. Figure 5.2 shows an annotated example of the visualisation system layout featuring the projection of three faces of an axis-aligned three-dimensional cube.

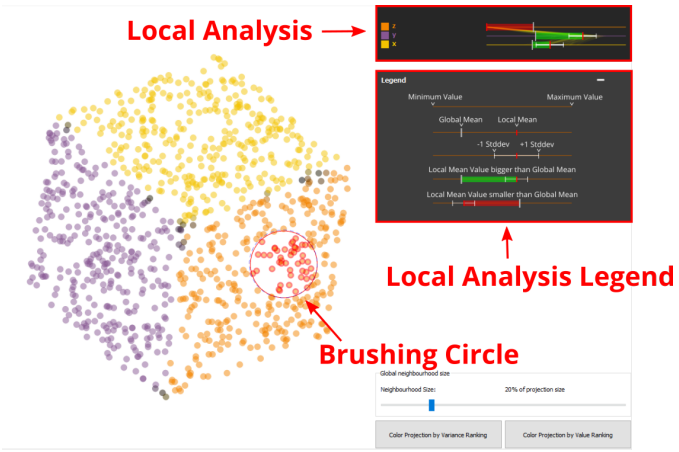


Figure 5.2: Layout of the visualisation system showing a projection, the lens brushing tool, and the generated explanations in the local analysis widget for the selected points.

5.3.1 *Dimension Sorting*

Points selected in the projection, cause ranks to be computed according to the current mode of the visualisation system, that is, variance mode or value mode. The order in which dimensions are listed in the local analysis widget is the result of sorting the dimensions according to their rank.

In variance mode, dimensions are sorted from lowest rank (lowest ratio of variance in the selected points versus the whole projection) at the top, to highest rank (highest ratio of variance) at the bottom. This

means that instead of only knowing the least varying dimension in a particular region of the projection, we get a ranking of all dimensions based on their variance in that region.

In value mode, dimensions are sorted from highest rank (highest mean average value in the selected points compared to the mean value over the whole projection) at the top, to lowest rank (lowest mean average value) at the bottom.

The sorting of dimensions in the different modes is shown in Figure 5.3.

Figure 5.3: Sorting of the dimensions according to variance (on the left) and according to value (on the right).

5.3.2 *Local Statistics*

Seeing the full ranking of the dimensions based on variance or value gives more insight into the relations between dimensions in the selected region in the projection, but does not say very much about the dimension variance or values themselves. As an example, a dimension listed at the top of the value ranking may have a relatively high value, or it may have a low value, as long as all the other dimensions have even lower values. Therefore it is useful to get a quantitative idea of the values of the dimensions for the selected points.

To this end, next to every dimension listed in the local analysis widget, a visualisation shows aggregate statistics of the selected points (see Figure 5.4).

As, during analysis, we are commonly interested in atypical dimension values, we chose to display the average value of the dimension over the selected points and the average value of the dimension over the whole dataset. Together they convey whether the selected points have unusually high or low values compared to their average value over the whole dataset.

In addition, we display the standard deviation of the dimension over the selected points. This conveys whether the values are unusually spread out or close together.

All in all, the displayed statistics are constructed from the following elements into an integrated view of the dimension values.

RANGE LINE A horizontal line is drawn that represents the full range of values that dimension takes on in the data. The left endpoint of the line represents the minimum value the dimension takes on, and the right endpoint represents the maximum value.

GLOBAL MEAN Along this range line, a vertical grey line is drawn to represent the average value of the dimension over the whole dataset. We call this the *global mean*. Since the range line represents the full range of values of the dimension over the dataset, the global mean line intersects the range line at the point where it would be situated along the dimension range. A low global mean would result in the line being drawn closer to the left endpoint (being the minimum) and a high global mean would result in being drawn closer to the right endpoint (the maximum).

LOCAL MEAN In addition to the global mean, a vertical red line is drawn along the range line to represent the average value of the dimension over the selected points in the projection. We call this the *local mean*. Whenever the local mean is greater than than the global mean (further to the right along the range line), a green bar is drawn between the two means. This allows for quick visual recognition for which dimensions the selected points have higher than usual values. Whenever the local mean is less than the global mean (further left along the right line), a red bar is drawn between the two means, indicating an unusually low value for this dimension.

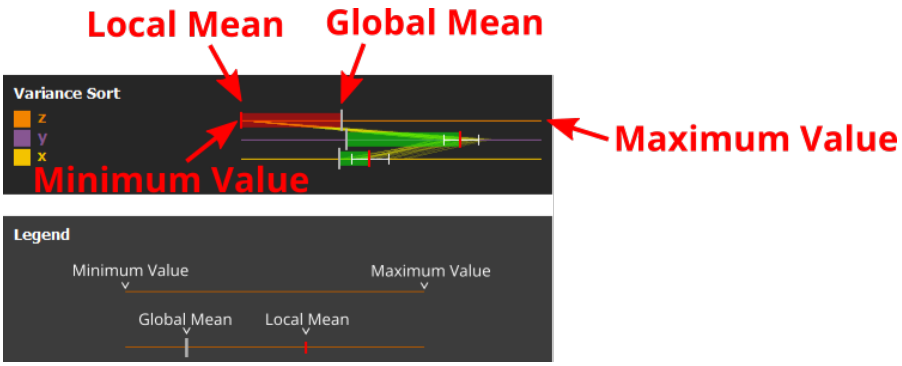


Figure 5.4: Local statistics of points selected in the projection. For each dimension a horizontal line is drawn representing the full range of values that dimension takes on in the data. Along this line a grey vertical stripe is drawn that indicates the average value of the dimension over all points in the dataset, and a red vertical stripe that indicates the average value over the selected points in the projection.

STANDARD DEVIATION The previous elements deal with the average value of the dimension, but say nothing about how the values are spread. However, the spread of dimension values are an important feature, and indicate whether the dimension had a big influence on the points being close together in the projection. Dimensions with a low variance for a number of points result in those points having shorter

distances in the high-dimensional space and therefore hopefully also having short distances in the low-dimensional embedding.

We therefore render white whiskers around the local mean of the selected points. Where the ends of the whiskers indicate one standard deviation to either side of the mean. Longer whiskers therefore represent dimensions with more spread in their values, whereas shorter whiskers indicate values that lie closer together.

As a sidenote, we want to stress that even though the visual elements of the resulting visualisation are similar to a boxplot, they express very different concepts. Firstly, the whiskers indicate a standard deviation here, and not the minimum or maximum values or quartiles. And secondly, the box drawn between the vertical lines represents the difference between the global and local means of the dimension.

5.3.3 *Parallel Coordinates Plot*

All of the statistics discussed in the previous sections are aggregate statistics over the selected points. However, sometimes such aggregate statistics can be deceiving. Dimensions that have the same local mean average over the selected points, might have wildly different distributions of values on a per-observation level. Therefore, purely understanding the selected region in the projection on an aggregate level is not enough, and can lead to wrong interpretations.

In order to avoid drawing erroneous conclusions based on just the average values, we draw all the selected points in a parallel coordinates plot (PCP). The range lines of each dimension form the parallel axes of this plot, and for every selected point, a polyline is drawn vertically through the dimension axes with its vertices intersecting each dimension at its respective value along the range line. This visualisation allows for understanding the true distribution of values over the selected points.

As an example of the problem that the parallel coordinates plot solves, see Figure 5.5a. Here we show two dimensions with exactly the same local mean, but it is totally unclear whether the actual distribution of the per-observation values for the two dimensions are the same. Figure 5.5b adds the PCP visualisation and it becomes clear that while the local means of the two dimensions were the same, their distributions of values were wildly different.

The standard deviation whiskers similarly show such differences in distributions, however they operate at an aggregated level and therefore cannot convey skewed distributions or distributions with discrete clusters of values.

5.4 DIFFERENTIAL ANALYSIS

With the tools discussed so far, it is possible to interactively explore a projection and explain points patterns at multiple levels of detail. However, there are several issues to address.

(a)

(b)

Figure 5.5: Several points are selected in the projection. The two dimensions shown in the analysis widget show that, for the selected points, the local mean for both dimensions is exactly the same. In figure (a) we can see that it is completely unclear whether the per-observation values are the same or different between the dimensions. In figure (b) the yellow parallel coordinates plot lines show that their per-observation values are wildly different. The selected points have values close to the local mean for dimension 1 (the lines intersect close to the mean), while for dimension 2 the values are far away from the mean.

First of all, the main advantage of having a global explanation is that the whole projection is explained at once, and all the information is on screen at the same time. As we are taking a more local interactive approach, we lose this benefit, meaning that the information is presented in a details-on-demand manner. Inherent to this approach is the fact that at any moment only a fraction of the total explanation is shown to the user. In order to obtain a total explanation of the projection, the user needs to fully explore it and combine the partial explanations in their mind. Therefore, memory plays a big role in gaining a full understanding of the data.

One common task where the user would have to intensely use their memory is in doing a differential analysis on the data. That is, finding out what the differences are between different subsets of the data. In order to find this out, the user would have to select several points in the projection, look at the local analysis widget, remember the visualisation shown there for many potential dimensions, then select several other points, mentally integrate both explanations and work out the differences.

To alleviate the burden on the memory of the user in this task, we provide a tool that automatically computes the difference between two sets of selected points. With the tool, the user selects their first set of points in the projection, holds a key on the keyboard, and selects a different set of points in the projection. The statistics that are normally

shown in the local analysis widget are now replaced by statistics showing the difference between the two sets of observations.

Figure 5.6 shows an example of a differential analysis on the wine dataset, where the original selection was done on the green points and then compared to the points selected in the salmon region. In the local analysis widget we see that wines in the second selection have a much higher alcohol content and perceptual quality score, as well as a low density and total amount of sulfur dioxide, among other smaller changes.

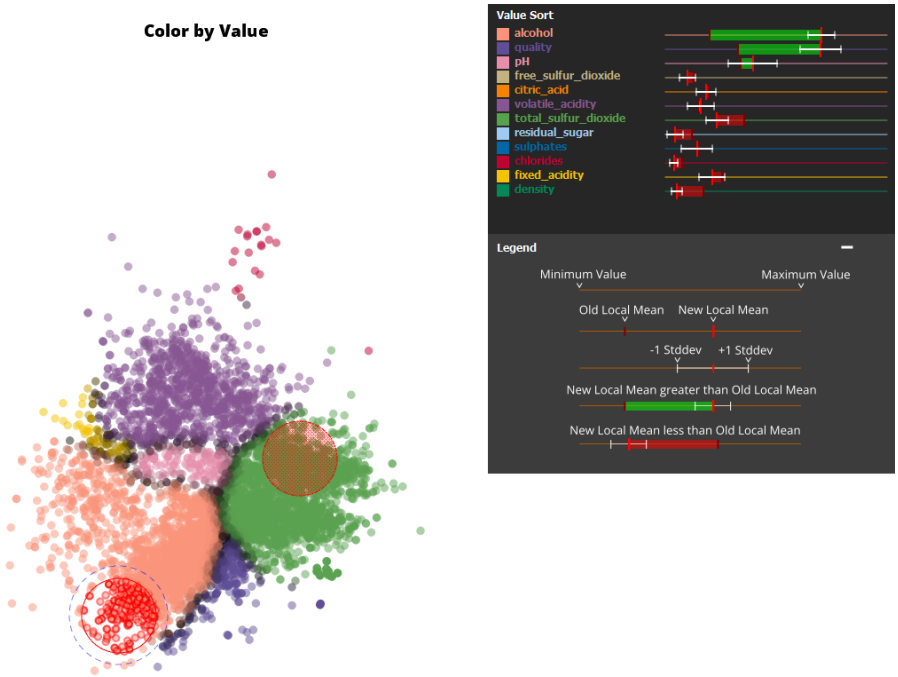


Figure 5.6: Example of a differential analysis on the wine dataset. The green points were initially selected and are then compared with the points selected in the salmon region. The local analysis widget shows wines in the salmon region have significantly higher alcohol percentages and perceived quality than the green points.

5.5 DIMENSION EXCLUSION

A second issue is that sometimes a dataset contains a number of dimensions that do not convey very much information and are simply unimportant or obstructive to the analysis. These dimensions can take up valuable colours in dimension colour assignment, as well as conceal other dimensions in the global colouring, and push more important dimensions away in the local explanation widget.

These dimensions can be excluded from the dataset as a whole with data preprocessing. However, doing this may be undesirable, as it requires regenerating the projection. Moreover, the dimensions might be desirable for generating the projection, but undesirable in the analysis of the data.

Therefore, in the local analysis widget, users can click on dimensions in order to temporarily exclude them from the generated explanations. Doing so reassigns colours to the remaining dimensions and instantly

regenerates the global and local explanations. When a dimension is desired again it may be reenabled by clicking it once more, and the explanations are again regenerated. Multiple dimensions can be selectively excluded from or included in the analysis like this in an interactive manner.

As an example, the left of Figure 5.7 shows a projection where nearly half of the points are dominated by the colour of one dimension. If we for any reason do not want to see this dimension anymore, it can be disabled by clicking on it. The dimension will turn white to indicate it is disabled, move to the bottom of the sorting and the explanations are regenerated without considering this dimension. As a result on the right of Figure 5.7, instead of seeing a big yellow blob, we can now see several groups of differently coloured points that can guide further exploration. If the, now disabled, dimension is clicked again it will be reenabled and included again in the generated explanations.

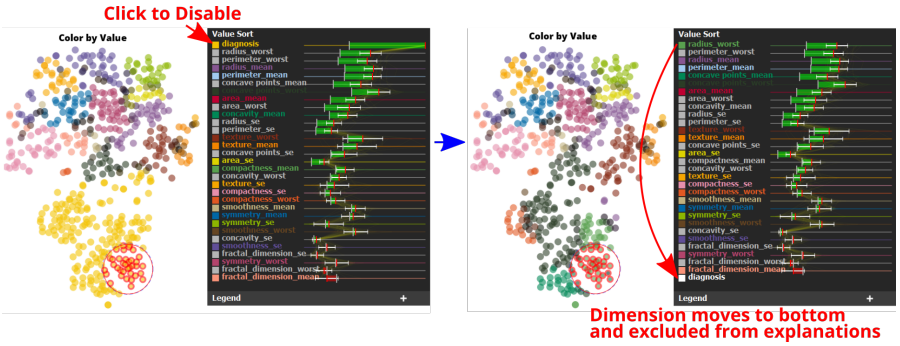


Figure 5.7: In the left figure a big chunk of the projection is dominated by the colour of just one dimension. If we are uninterested in this dimension, or done with its analysis, or want to know where to explore based on other dimensions, it can be disabled by clicking on it. This will exclude it from all explanations, turn it white and move it to the bottom. The right figure shows the same view after disabling the dimension. Several groups of points characterised by different dimensions are revealed and can be used to guide further exploration.

5.6 SCALABILITY

Now that we have introduced all of the elements of the proposed visualisation system, we will demonstrate how these tools work in practice on two, more complex, state-of-the-art biological datasets. The purpose of showing this is merely to show the results of the previous and current chapter, and to demonstrate both the visual, and computation scalability in the number of observations and dimensions.

SCALING IN OBSERVATIONS The first dataset contains a tissue sample from the cortex of a brain for which the expression of various genes has been measured. The dataset consists of 22 images of the same tissue patch, where each image is associated with a particular gene, and where in the tissue the gene is expressed is encoded in each pixel as the brightness. Each pixel in the image is treated as an observation and the values of that pixel over the different images treated as the dimensions

of the pixel. In total, the dataset has roughly 115,000 observations and 22 dimensions. We compute a projection of the data using t-SNE [24] and examine it in the proposed visualisation system (see Figure 5.8).

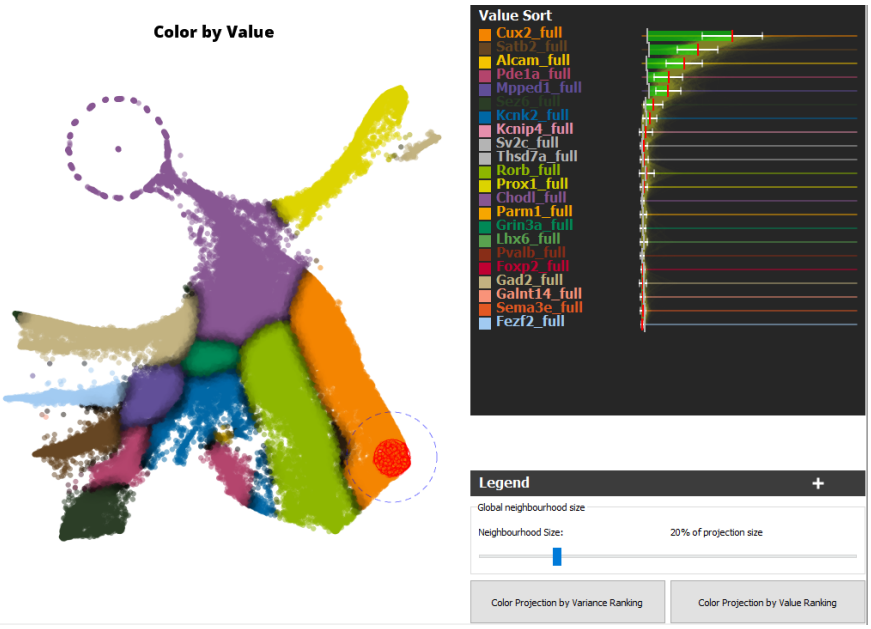


Figure 5.8: Projection of the first cortex dataset with over 100,000 observations, showing the explanations visually scale well in the number of observations.

We first of all see global value explanation, which divides the projection into very clear and distinct separated groups of points with similar explanations. In fact, the high number of observations makes it extremely clear where interesting transitions take place, because there is barely any white space between the points. This division can then be used to guide further local exploration of why those points have clustered together and what is contained in them. Using the lens brushing tool, we highlighted several points in the orange region of the projection, which corresponds to the `Cux2` gene. In the generated local explanations we see that indeed the `Cux2` gene is unusually highly expressed in these points, as well as a number of other genes to a lesser extent.

SCALING IN DIMENSIONS The second dataset [29] again contains a tissue sample from the cortex of a brain for which the expressions of various genes has been imputed. The dataset consists of roughly 2400 observations (cells from this region), and 314 dimensions (genes). We examine the spatial layout of these cells in our visualisation solution (see Figure 5.9).

Looking at the scatter plot we see that even though the dataset contains hundreds of dimensions, the global value explanation is still able to assign colours to each of the displayed data points, again guiding us where to explore further. Using the lens brushing tool, we highlighted several points in the purple region of the projection, we see from the legend that this corresponds to the `Foxp2` gene. In the generated local explanations, we can explore which other genes have high expressions,

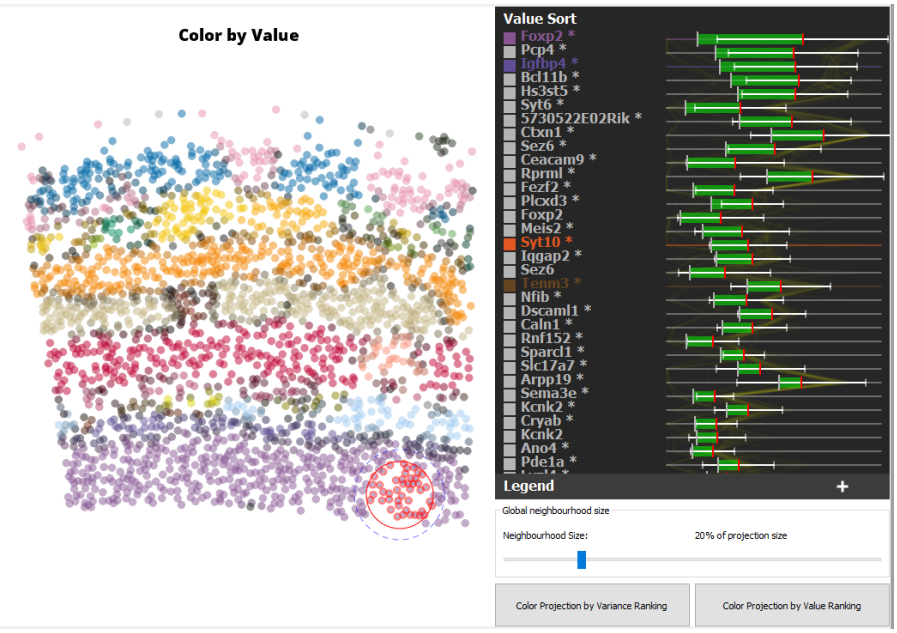


Figure 5.9: Projection of the second cortex dataset with over 300 dimensions, showing the explanations scale well in the number of dimensions.

how highly expressed they are, and how much the expression varies for the highlighted cells.

5.7 CONCLUSION

We presented an interactive user-guided local explanation mechanism that allows for detailed explanations of user selected regions in the projection. The points are selected using a circular brushing tool that can be resized in order to explain the projection at various levels of detail. This mechanism of local exploration and explanation is supported by an explicit way to compare multiple regions in the projection and do a differential analysis on their explanations. If any of the dimensions is obstructing the global or local analyses then it may be freely disabled and reenabled.

We combine the global explanations, local explanations, differential analysis and dimension exclusion into an integrated visualisation system and implement it as a software application. In addition, we tentatively showed that this visualisation system scales very well both visually and computationally in terms of the number of observations and dimensions of the dataset.

In the next chapter we describe our method of testing whether the designed system is capable meeting the requirements laid out in Section 3.2 and present the results.

6 EVALUATION

The various elements of the visualisation system described in the previous chapters are intended to allow for analysing projections of high-dimensional datasets on a deeper level. The combination of all these elements should be able to provide useful insights into the data even as the dimensionality of a dataset gets higher.

There are, however, three potential issues with using such a visualisation system. Firstly, the explanations generated by the system may just not be capable enough to help users in solving analysis problems and understanding the data. Secondly, if users are able to gather useful insights from the system, this capability may break down as the dimensionality of the analysed dataset increases. And lastly, if the provided explanations are not clear and intuitive, there is a risk that users of the system do not understand, or misinterpret the explanations and come to wrong insights.

In order to investigate all three of these potential issues, we performed a preliminary evaluation study. In this study we asked participants to solve analysis problems on several real-world datasets of increasing dimensionality. This study allows us to look into whether people using the same tool, on the same dataset, trying to solve the same analysis questions, arrive at similar conclusions. If they do, this means that either they all arrive at the right conclusion, or the wrong one. However, considering that we designed the questions, know the visualisation system, and have done a thorough analysis of the data ourselves, we have a good sense of which answers are likely to be right and which not. Therefore, if participants are able to consistently come to the right conclusions for analysis problems on datasets of increasing dimensionality, this means that users interpret the generated explanations of the visualisation system correctly, and that it is capable of helping users to solve these analysis problems, even as the datasets become more high-dimensional.

In addition, as we are introducing many new explanatory mechanisms and visual encodings, a more qualitative aspect of the study is to find out how people perceived the various elements of the visualisation system in terms of practicality and usefulness.

6.1 EVALUATION SETUP

Participants of the study were asked to analyse several datasets of increasing dimensionality, both by looking at screen captures of the corresponding software application and by using the application on their own machine. In addition, participants of the study were asked to appraise the various elements of the system in order to understand the benefits and drawbacks each of them have. Apart from when users

were asked to use the application on their machines, the evaluation was conducted fully within Google Forms and structured to take roughly 50-60 minutes to complete, however participation in the study was not timed. The full evaluation form is included in Appendix A.

6.1.1 *Invited Participants*

Participants to the study were gathered on an invitation basis and were in large part invited based on their familiarity with analysis of multidimensional datasets. Experience of the invited participants in this field ranged from no experience to decades of experience. Which participants responded to the invitation was not monitored and participation in the study was anonymous.

6.1.2 *Installation*

The application was compiled for both Microsoft Windows 10 and Debian-based Linux systems. Distributables were offered in .zip and .tar.gz form respectively, allowing for easy installation. Participants were asked to download the software and unpack it on their system in order to run it.

6.1.3 *Tutorial*

Before starting the evaluation proper, participants were made familiar with the functionality of the system by means of a tutorial (see Appendix A pages 4-11). The tutorial consists of six major parts and preceded asking any questions of the participants. In it, the following basic concepts of the system were introduced one-by-one:

1. How to load an example projection into the application (page 5)
2. How to display the example projection in a scatter plot (page 6)
3. Background on what the example projection is generated from (page 7)
4. Exploration using the lens brushing (page 8)
5. The local analysis widget statistics (excluding the PCP) (page 9)
6. How to change between variance and value ranking and their meanings (page 10)

After going through the parts of the tutorial, the participants were asked to proceed with the question part of the evaluation if they felt comfortable with the various elements described in the tutorial. In addition, they were made aware a copy of the tutorial was available to them during the evaluation, both packaged with the application and online via a separate link.

6.2 EVALUATION

The part of the evaluation where participants are asked questions consists of analysis of three different datasets plus a questionnaire at the end. For each dataset we ask participants to answer four *control* questions and three *live exploration* questions. The control questions serve to understand whether the participants understand how to read the explanations generated by the system, and to reinforce the concepts taught in the tutorial. The live exploration questions test whether the participants come to similar insights using the system, when tasked with solving a multidimensional analysis problem. At the end of the evaluation, participants are asked to provide feedback on each of the elements of the system, the evaluation study, and to state their experience with multidimensional data analysis. The full evaluation structure is portrayed in Figure 6.1.

Figure 6.1: Structure of the evaluation study. Participants were first asked to read a roughly 15 minute tutorial on the basic elements of the system. Then they were asked a series of 4 control questions, followed by 3 live exploration questions for each of the three datasets used in the study. Finally, they were asked to provide qualitative feedback in a questionnaire.

We briefly introduce the three datasets before discussing the different question types and questionnaire in more detail.

6.2.1 *Datasets*

Participants were asked to look at three projections of datasets during the study. All datasets used in the study were taken from the UCI machine learning repository [9]. Datasets were picked based on their suitability for analysis using projections and their number of dimensions. The dimensionality of the datasets are roughly 10, 30 and 60 respectively, giving us an idea of how the system operates on multiple ranges of dimensionality.

PORTUGUESE WINES The first dataset consists of roughly 6500 samples of white and red wines from Portugal (see Appendix A page 13). For each of these wine samples, 12 attributes are recorded (e.g., pH, alcohol %, total sulphur dioxide). One of the attributes (*quality*) is a dependent variable and represents the perceived quality of the wine (evaluated by a panel of testers) on a scale from 0 to 10 (0 = worst quality, 10 = best quality). A projection (see Figure 6.2a) of the dataset was computed using Local Affine Multidimensional Projection (LAMP) [12] as in [6].

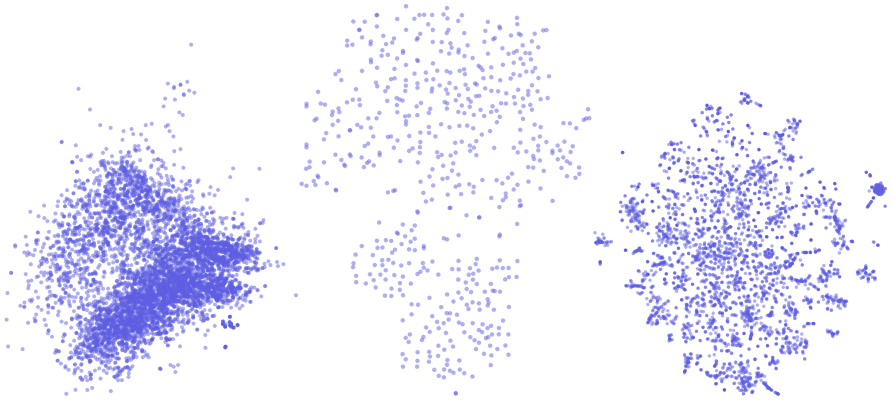
BREAST CANCER The Wisconsin breast cancer dataset from the UCI repository (see Appendix A page 25) examines roughly 570 slices taken from tissue samples. Each slice (each data point) contains a number of cells and 10 attributes are computed that describe the size, shape and texture of their cell nuclei. For each of these attributes the mean value (mean), the largest value (worst) and the standard error (se) are found. In addition, one dimension encodes whether the cells are considered malignant (1) or benign (0), resulting in a total of 31 attributes per slice.

The data was standardised and a projection (see Figure 6.2b) of the standardised data was computed using t-SNE [24].

SPAM E-MAILS The Spambase dataset (see Appendix A page 36) examines roughly 4600 e-mails, of which a number have been classified as spam (an unsolicited commercial e-mail). For each e-mail, the frequencies of 48 words and 6 characters have been counted (higher numbers mean the word or character occurred more often in the mail), as well as the run length of capital letters in the mail (how many sequences of capital letters there are on average, the longest run, and the total length of all sequences). Lastly, one dimension encodes whether the mails are considered spam (1) or not spam (0). Therefore, in total for each e-mail 58 attributes are recorded.

The data was standardised and a projection (see Figure 6.2c) of the standardised data was computed using t-SNE [24].

Participants were asked a series of control and live exploration questions about each of the datasets in the above order.



(a) Wine Projection (b) Breast Cancer Projection (c) Spam Projection

Figure 6.2: A scatter plot of the projections computed for (a) the wine dataset (b) the breast cancer dataset and (c) the spam dataset.

6.2.2 *Control Questions*

For each dataset a series of four multiple-choice single-answer *control* questions were asked. The purpose of these control questions was to establish whether the participant understood how to read the various elements of the system in order to come to the correct conclusion. In addition, as questions generally briefly reiterated the concepts necessary for answering it, if participants were not yet comfortable with a particular element these questions served to reinforce the concepts for them.

As participants were presented with a snapshot of the application where no interaction is possible, the generated explanations were the same for every participant, and therefore the analysis was more controlled. The questions were constructed in a way that there was an unambiguously correct answer and the other answers were clearly incorrect. Furthermore, we knew what the correct answers were, and so we could evaluate the understanding of the participants based on their given answers.

Each of the questions presented the participants with a snapshot of the dataset projection in the application. However, in each question different points in the projection were selected using the lens selection accompanied by both the global and local explanations generated by the system. The participants were then asked to make use of the explanations provided by the system in the snapshot to answer a question about the selected region in the projection.

An example question is shown in Figure 6.3 and can be found in Appendix A on page 15. Here we presented participants with the projection of the wine dataset in variance mode. This means that the dimensions of the selected points are sorted from least variance at the top to most variance at the bottom in the local analysis widget. Several points were preselected and participants were asked to answer which of the given options were true for the pH dimension. The first option is incorrect because the variance is not zero for the pH dimension (see size of the

whiskers). The second option is also incorrect because the local mean is slightly lower than the global mean. The third option is the correct answer as the pH dimension is all the way on the bottom of the sorting, meaning it has the most variance out of all dimensions for the selected points. Because the third option is correct, the fourth option cannot be correct.

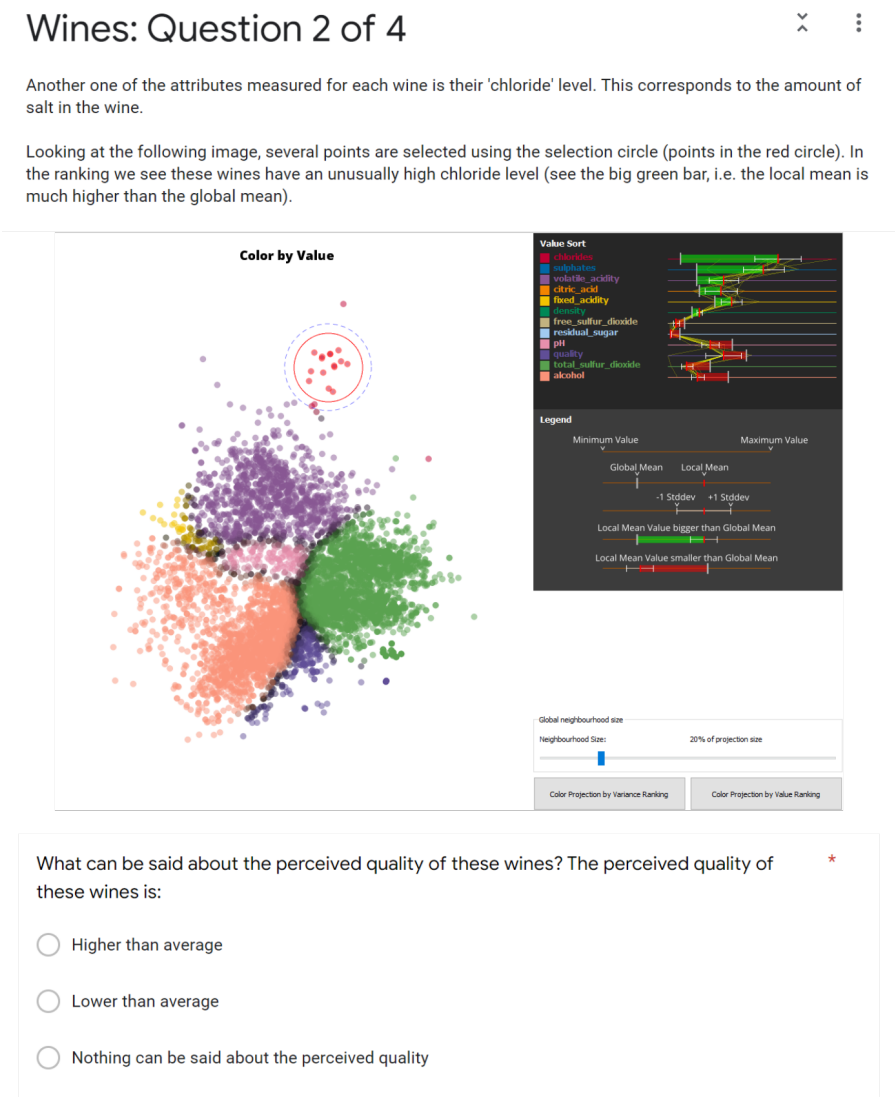


Figure 6.3: One of the control questions used in the evaluation study. The participants were presented with a snapshot of a projection of the wine dataset, and were asked to read the generated explanations in order to answer an analysis problem.

6.2.3 *Live Exploration Questions*

As the developed system uses interaction to explore projections, it is crucial for participants to be able to explore the datasets by using the system themselves. Therefore, for every dataset, the set of four control questions were followed up by a series of three multiple-choice multiple-

answer *live exploration* questions. For these questions participants were not provided with a snapshot of the application, but rather needed to explore the projection with the application themselves (on their machine) in order to answer the question.

In order to form their answer, participants were free to explore the projection however they desired. That means we had no knowledge of what view they were looking at, what points they brushed over, at what radius they set their lens tool and which visual elements they considered in coming to their conclusion. This made the evaluation especially challenging, but it is exactly what we wanted to test. We wanted to see if users of our system, when presented with an analysis problem in an uncontrolled environment, would come to the same or similar insights. Meaning, that similar to the control questions, we were looking for consistency in the answers.

In contrast to the control questions however, the live exploration questions were of a nature where there were no strictly right or wrong answers. Rather, because we performed our own extensive analysis of the data, we considered some answers to be more in line with our findings of what the explanations show than others.

An example question is shown in Figure 6.4 and can be found in Appendix A on page 21.

6.2.4 *Feedback*

The last part of the evaluation consisted of a series of qualitative feedback questions on each of the elements of the system. Here participants could rate how useful they found each element, provide additional commentary and leave comments on the evaluation and system in general.

Participants were asked to rate the variance ranking mode, value ranking mode, differential analysis and exclusion of dimensions on a usefulness scale from 1 to 7, where 1 represents the notion that the element was not very useful and 7 means they found it very useful.

In addition, three multiple choice questions were asked that allowed participants to tick several predefined assessments of the variance and value ranking mode and the parallel coordinates plot. Apart from the predefined assessment, they could write down their own feedback in a free form text field.

Lastly, participants were asked to indicate their experience with multidimensional data analysis in a multiple-choice question and note down any comments on the evaluation or system in two free-form text fields.

For details on the structure of the questionnaire see Appendix A page 45.

6.3 RESULTS

We now present the results of the evaluation study. We split up the results into three sections, firstly the answers to the control questions,

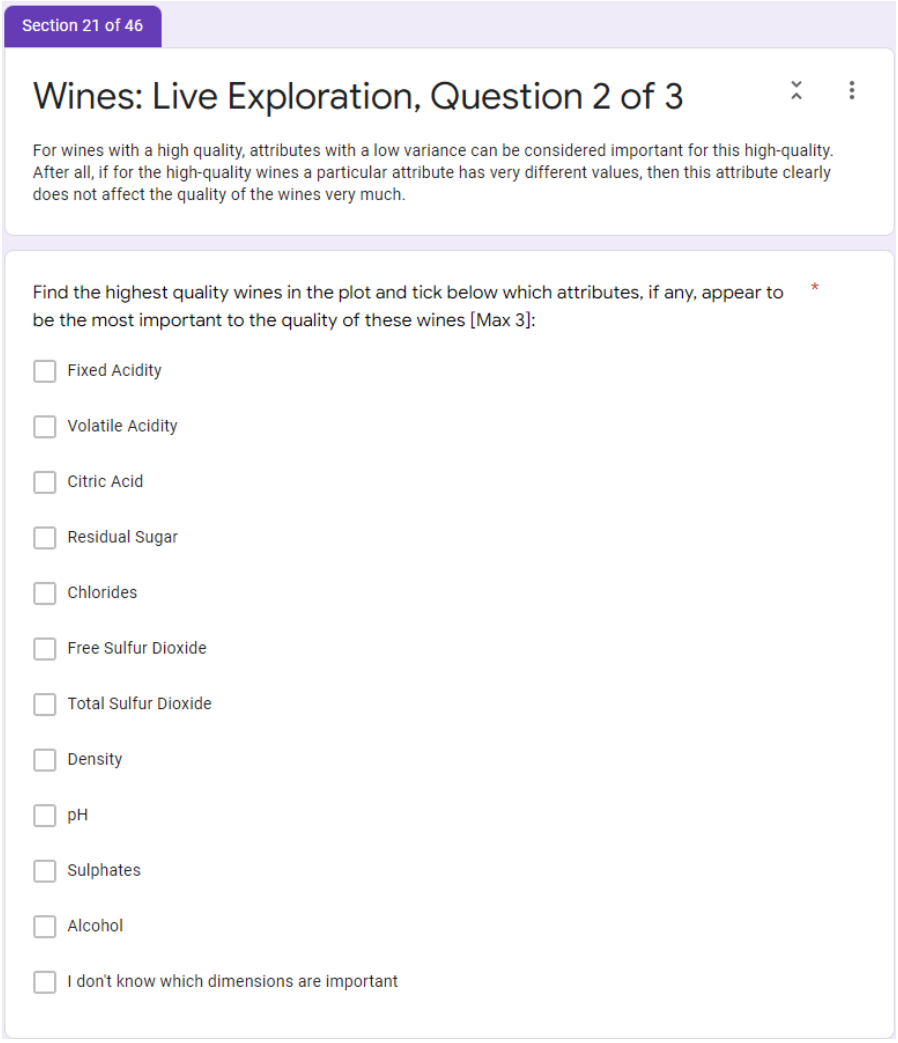


Figure 6.4: One of the live exploration questions used in the evaluation study. The participants were asked to answer an analysis problem by using the software application on their local machine.

then the answers to the live exploration questions and finally the answers to the questionnaire.

6.3.1 *Participants*

In total, 23 participants submitted responses to the evaluation. Their experience with multidimensional data analysis ranged anywhere from no experience to more than five years and was fairly uniformly distributed (see Figure 6.5). Therefore, further results from the evaluation can be considered to come from a broad range of expertise.

6.3.2 *Control Questions*

Responses to the total of 12 control questions were nearly unanimous in all cases apart from one question (wine dataset question ¾), with

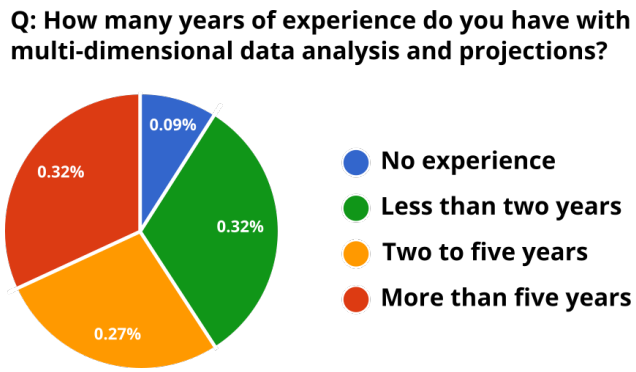


Figure 6.5: Participants reported a broad spectrum of experience levels with multi-dimensional data analysis and projections.

correct answers being chosen by, on average, 92% of the participants. The percentage of correct answers to all control questions are summarised in Figure 6.6. All of the recorded answers to control questions can be found in Appendix B.

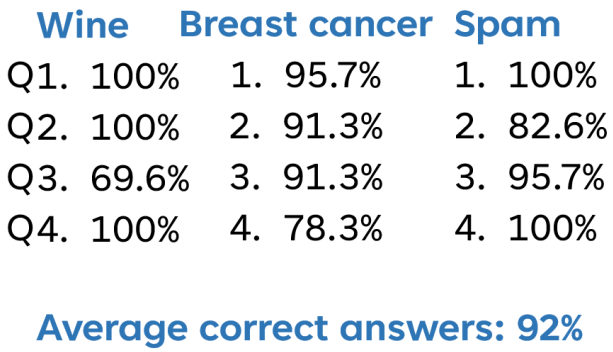


Figure 6.6: Percentage of correct answers to the control questions of every dataset.

6.3.3 *Live Exploration Questions*

Live exploration questions were, with one exception, all questions where participants could tick multiple options. These question didn't have a factually correct answer, as it depends very much on at what location and scale a region in the projection is analysed. As such, we evaluated the answers given by participants based on two factors. Firstly, their similarity to the answers found through our analysis. As we spent significant time with the system and the datasets, this is a measure of how easy the system is to learn and come to the same insights. Secondly, the spread of the distribution of the answers given, which indicates whether the explanations provided by the system are consistent and result in similar insights through the interpretation of different users.

ANSWER EVALUATION We roughly categorise the possible answers to each question according to how likely we would pick those options as our answer, based on our analysis. In order to visually indicate which

options we would likely pick and which not, participants responses are coloured according to the colourmap in Figure 6.7. The lighter the colour the more likely it is we would tick that option based on what we found using the system, vice versa, the darker the colour the less likely we would tick that option. We would therefore like to see answers chosen in the majority by participants to line up with the lighter colours, and answers rarely chosen by participants to line up with the darker colours.

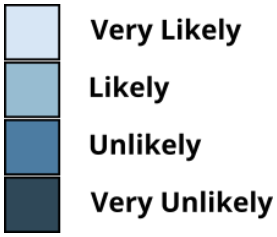


Figure 6.7: Colourmap indicating our likelihood of picking a particular answer to live exploration questions based on our extensive analysis, used for evaluation of the participants' answers.

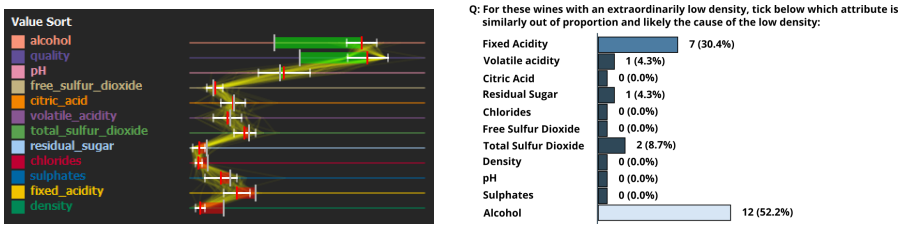
We next analyse the different questions and answers given based on their themes and the skills and system elements required to solve them.

SINGLE CLUSTER ANALYSIS The simplest type of analysis possible with the system is selecting a certain region in the projection with the lens selection and interpreting the statistics and sorting of the dimensions in that region.

Question ⅓ of the wine dataset asked participants to find the region in the projection containing wines with the lowest average density value (see Appendix A) and to answer with an attribute that is similarly out of proportion, likely causing the low density. A possible selection of the lowest density region in the projection results in the statistics displayed in Figure 6.8a. Through our analysis we found that the alcohol dimension, like density, has a local mean that is out of proportion (unusually high) and likely the cause of the low density.

Responses recorded by participants are summarised in Figure 6.8b. The colours of the bars indicate our likeliness in picking the same answer according to our own research (see Figure 6.7). The majority of respondents (52.2%) indeed answered `alcohol`, however a substantial number of participants (30.4%) answered `fixed acidity`. This is potentially due to ambiguous phrasing of the question, which could be interpreted as having to find a dimension which deviates from the global mean in the same proportion as the density dimension. Looking at the generated explanation, the most likely answer would indeed then be `fixed acidity`.

MULTIPLE CLUSTER ANALYSIS In both questions ⅓ and ⅔ of the spam dataset, users were asked to perform an analysis over more than one cluster. Specifically, they were asked to find out, for first non-spam and then spam e-mails respectively, which words occurred more often than unusual in the e-mails. This involved interactive exploration of the

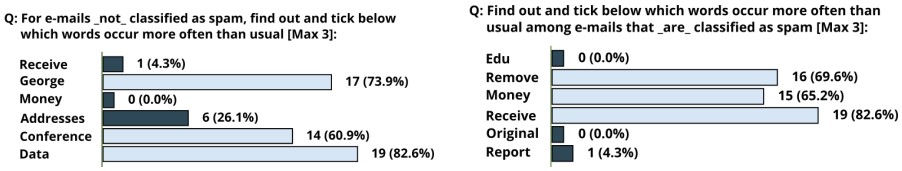


(a) (b)

Figure 6.8: (a) Possible generated explanation for low density region in projection. Alcohol has an unusually high value, while fixed acidity has a deviation from the global average similar to density. (b) Answers given by participants.

different regions in the projection while looking at the explanations to find regions where any of the six words listed as options to the question had unusually high values.

Responses recorded by participants are summarised in Figure 6.9a for question 1 and Figure 6.9b for question 2. The colours of the bars indicate our likeliness in picking the same answer according to our own research (see Figure 6.7). Participants were extremely close to unanimous in their answers, and answers with majority votes correspond exactly with our answers. On question 1, one answer (addresses) also has several votes, this is potentially due to confusion caused by both the words `addresses` and `address` being dimensions in the dataset, the latter of which indeed has unusually high values in the non-spam e-mails, whereas the first does not.



(a) (b)

Figure 6.9: Response summary of question 1 (a) and question 2 (b) from the spam dataset, participants' answers line up almost unanimously with our analysis.

MULTIPLE ATTRIBUTE ANALYSIS In addition to a multiple cluster analysis, question ⅔ of the breast cancer dataset required analysing multiple attributes per cluster. The question asked participants to explore all observations with a malignant diagnosis and evaluate which of the listed statements were likely true. This involved interactive exploration and comparison between regions in the projection with high or low values for a given attribute, and then analysing several other attributes for these regions. This was likely the most involved question in the evaluation.

Responses recorded by participants are summarised in Figure 6.10. The colours of the bars indicate our likeliness of picking the same answer according to our own research (see Figure 6.7). A large majority of the participants ticked the middle two boxes which is exactly in agreement with our analysis.

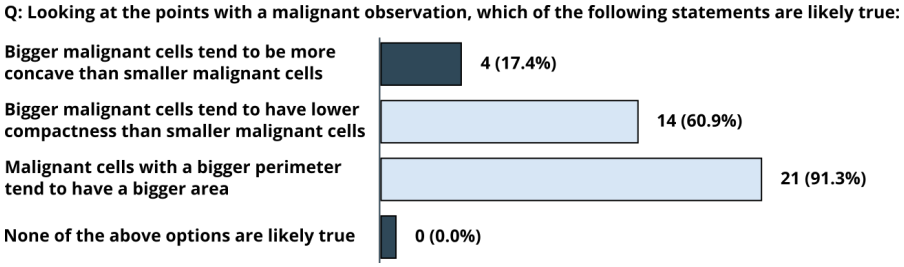
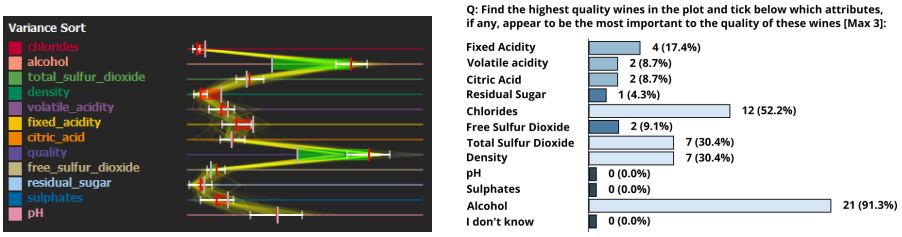


Figure 6.10: Response summary of question 3 of the breast cancer dataset, participants' answers line up very closely with our analysis.

IMPORTANCE A common scenario in the analysis of real-world datasets is trying to find out which variables influence a dependent variable, that is, which dimensions have an *important* contribution to the value of the dependent variable. Question ⅔ of the wine dataset asks the participant to perform such an analysis by finding the region in the projection with the highest quality wines and ticking which dimensions they believe to be important, based on all the generated explanations available to them.

Responses recorded by participants are summarised in Figure 6.11b. Colours of the bars indicate how likely it is we would also pick that answer (lighter colour means more likely, see Figure 6.7). A possible generated explanation is displayed in Figure 6.11a. From the ranking of the dimensions, we see that the `chlorides` dimension has the least variance for the selected high-quality wines (it is displayed as the top dimension) indicating that having this particular value of chlorides may be important for the high quality of the wines. It is followed by the `alcohol` dimension and then either `total sulphur dioxide` or `density` depending on where exactly the selection is made. These four dimensions are also given the most votes by participants. However, `alcohol` was found much more often than `chlorides`.



(a) (b)

Figure 6.11: (a) Possible generated explanation for high-quality wines region in projection. The chlorides dimension has the least variance, followed by alcohol, total sulphur dioxide, and density (b) Answers given by participants to question 2 of the wine dataset.

As this question is complicated and very much up to interpretation and personal judgement, we asked a follow-up question in order to find out how participants used elements of the system to reach their conclusion. Participants could indicate they used any of six predefined manners in which they came to their answer, or additionally indicate a different manner in a free text option. However, no different manners were

submitted outside of the predefined ones. Responses are summarised in Figure 6.12.

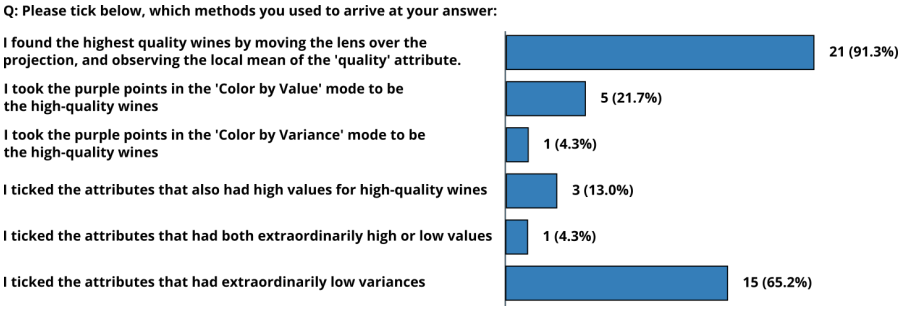


Figure 6.12: Methods, indicated by participants, they used to arrive at their answer for question 2 of the wine dataset.

We see the majority of users found the highest quality wines by moving the lens over the projection and keeping track of the local mean displayed for the `quality` dimension. Once they found some high-quality wines, the majority then subsequently indicated that they ticked the dimensions that had extraordinarily low variances. In our analysis we followed the same procedure.

DIFFERENTIAL ANALYSIS Questions ⅔ of the wine dataset and ⅔ of the spam dataset asked the participants to perform a differential analysis. Participants were briefed on how to perform a differential analysis on an example dataset before the first question.

On question 3 of the wine dataset, participants were instructed to tick a maximum of four attributes that are most different between red and white wines. This involved doing a differential analysis between the region in the projection representing red wines and the region representing white wines.

Responses recorded by participants are summarised in Figure 6.13b The colours of the bars indicate our likeliness of picking the same answer according to our own research (see Figure 6.7). Participants most commonly answered `volatile acidity` (82.6%), `total sulfur dioxide` (78.3%) and to a lesser degree `fixed acidity` (47.8%) and `pH` (26.1%).

Figure 6.13a displays a possible generated explanation for a differential analysis done between the red and white wines. We see that, both `volatile acidity` and `total sulfur dioxide` have the biggest differences followed by `fixed acidity` and `pH`. These results completely align with the responses of the participants.

A similar task was given to the participants in question 3 of the spam dataset. Their responses are summarised in Figure 6.13c. In our analysis we found the word frequencies of the words `your`, `receive` and `business` to be significantly different between the non-spam and spam emails. This analysis corresponds strongly to the answers given by participants.

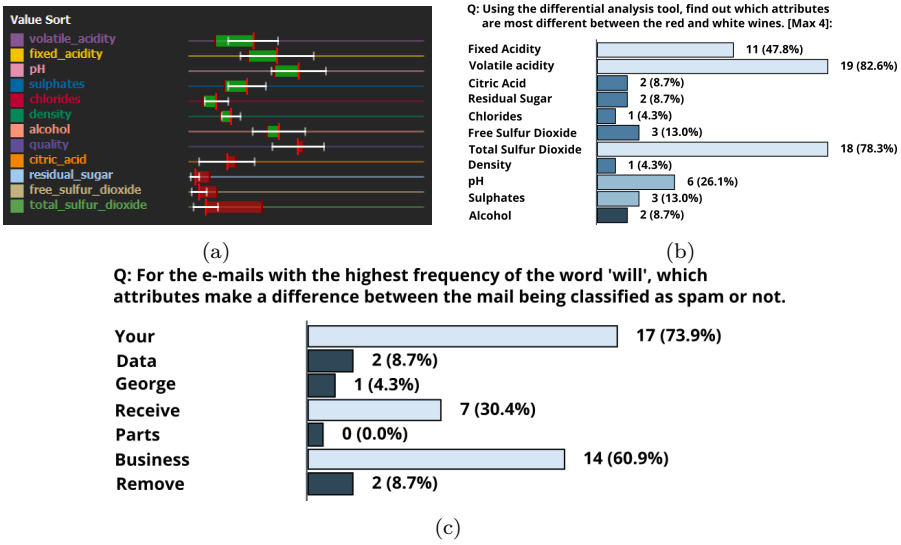


Figure 6.13: (a) Possible generated explanation for the difference between red and white wines. The biggest differences occur in the total sulphur dioxide, volatile acidity and fixed acidity dimensions (b) Answers given by participants to question 3 of the wine dataset. (c) Answers given by participants to question 3 of the spam dataset.

DIMENSION DISABLING Questions ⅓ and ⅔ of the breast cancer dataset asked the participants to find clusters of points in the projection where particular attributes had higher values than all other attributes, and to note down which attributes these were. However, these clusters had to be found in points that were completely dominated by a malignant diagnosis (high value) in the `diagnosis` dimension, meaning all points were assigned the same colour (of the diagnosis dimension, see Figure 6.14a).

In our analysis we found there were three major distinct subclusters within the cluster of points with a malignant diagnosis. They were characterised by high values of the `radius`, `concave points` and `compactness` dimensions.

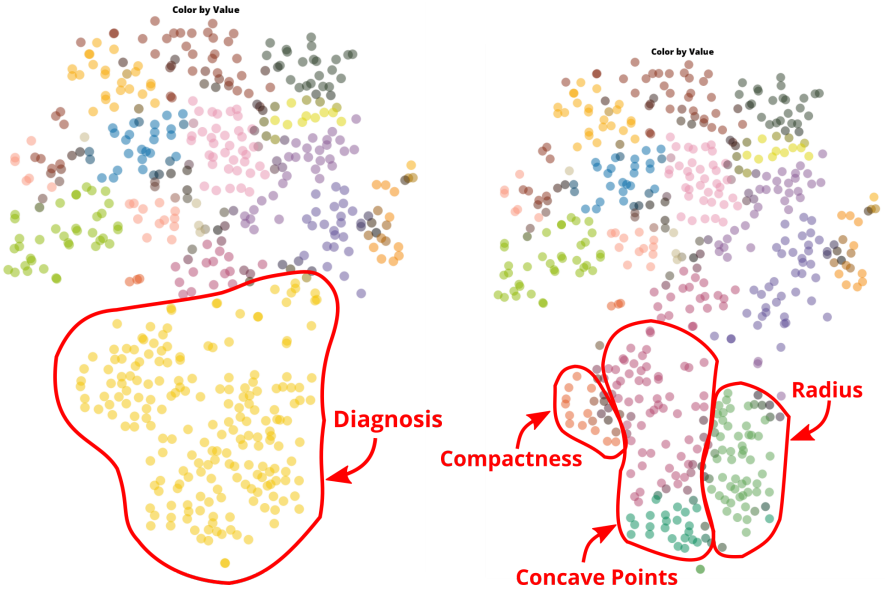
Responses to question 1 recorded by participants are summarised in Figure 6.15a. The colours of the bars indicate our likeliness in picking the same answer according to our own research (see Figure 6.7). Participants most commonly answered `concave points` (87.0%), `radius` (78.3%) followed by `compactness` (43.5%) which corresponds with our analysis.

Before proceeding to question 2, participants were briefed on how they can disable and re-enable dimensions and were instructed to disable the `diagnosis` dimension, thereby uncovering the colours of the subclusters (see Figure 6.14b). Here we can see that the compactness cluster is quite small and was therefore harder to find in the first question.

Question 2 then subsequently asked the participant to repeat the task of question 1 with the newly revealed colour groups. In this second task, we expected participants to have an easier time finding the specific clusters as the assigned point colours are indicating them. Given the relative small size of the compactness cluster, making it hard to find in the first task without being able to see the colours, we expected it to be

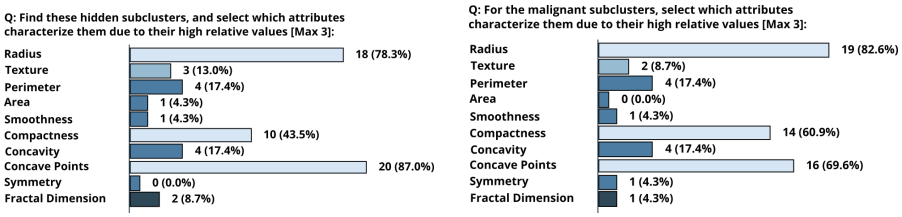
found much more often in the second task, as well as a lesser increase in the other cluster attributes.

Responses recorded by participants are summarised in Figure 6.15b. We see that the `compactness` dimension increased from being ticked by 43.5% of participants to 60.9%. However, the `concave points` dimension decreased from 87.0% to 69.6%.



(a) (b)

Figure 6.14: The value mode view of the breast cancer projection before disabling the diagnosis dimension (a) and after (b). Three major subclusters are revealed.



(a) (b)

Figure 6.15: Answers given by participants to question 1 (a) and question 2 (b) of the breast cancer dataset.

6.3.4 *Questionnaire*

In the questionnaire, we asked participants to provide feedback on each of the system elements.

VARIANCE AND VALUE MODE For the variance and value mode we asked participants to both provide a usefulness rating (1 = Not very useful through to 7 = Very useful) and their opinion on what it is useful for.

Participants rated the usefulness of the variance ranking mode with a mean score of 4.83 (SD = 1.63) and the value ranking mode with a 6.52 (SD = 0.77).

For the variance ranking mode, opinion was roughly equally divided between it helping to find important dimensions (52.2%) or identifying clusters to explore (39.1%), with 13% indicating they found it to provide no additional value. Comments mentioned it providing structure to the projection and making it interesting to explore (13%), not feeling they had to use it in the live exploration part (4.3%), being difficult to interpret (4.3%), and more useful combined with the value view (4.3%).

For the value ranking mode, participants found it helps for identifying clusters to explore (95.7%), helps identify which values are extraordinarily high (87%), and extraordinarily low (52.2%) and helps find important dimensions (65.2%), no participants indicated they felt it provided no additional value.

PARALLEL COORDINATES PLOT For the parallel coordinates plot in the local analysis widget, participants found it provides additional explanatory value (60.9%) and helps them understand the distribution of values (47.8%). Whereas some participants found it makes the local analysis widget more confusing (17.4%) or provides no additional value (8.7%). Additional comments mentioned using it mainly for outlier detection but it adding more clutter than benefit (8.7%).

DIFFERENTIAL ANALYSIS AND DIMENSION EXCLUSION Participants rated the usefulness of the differential analysis tool with a mean score of 5.74 (SD = 1.03) and the dimension exclusion with a score of 5.74 (SD = 1.42).

COMMENTS Comments on the system elements mentioned several labels being hard to read due to their assigned colours or font size, suggested a fixed sorting mode, confusion between variance ranking and the standard deviation whiskers and being unsure how large to make the selection circle.

6.4 SUMMARY

Overall, participants from a wide range of experience levels participated in the study. After a short tutorial introducing the various system elements, participants collectively responded with what we consider the correct answers to *all* control questions. In the live exploration questions, their insights and conclusions generally coincided with our conclusions reached through thorough analysis of the projections.

Participants rated the various system elements highly in terms of usefulness, with a preference of the value ranking over the variance ranking, and provided minor feedback on the display and practical details of several elements.

7 DISCUSSION AND CONCLUSIONS

In this thesis we looked at explaining multidimensional data projections. We identified two limitations in past research that led us to form the following research question:

How can local point patterns in projection embeddings of highly multidimensional data be explained in terms of their original dimensions?

In answering this research question, we proposed a linked view visualisation system consisting of a projection view augmented with global explanations and a local analysis view that provides local explanations based on user-guided exploration of regions in the projection. The explanations are supported by a differential analysis tool for explicitly finding differences between two regions of the projection, and the ability to exclude dimensions from the explanations.

In this chapter we return to the requirements stated in 3.2 that we believe are the pillars of a visualisation solution that answers the research question. We now evaluate to what degree our proposed solution satisfies these requirements. To recap, a satisfactory explanatory mechanism must:

1. Explain local point patterns in the projection in terms of their original dimensions
2. Be easy and intuitive in use
3. Scale well in the number of observations and dimensions
4. Scale well in terms of computational cost
5. Be applicable to projections in a black-box manner

We will first consider requirement four and five which can already be evaluated. The other requirements will be evaluated after discussing the results of the evaluation study.

COMPUTATIONAL SCALING In evaluation of requirement four, we can already say the following: The local explanations scale very well due to their linear computation complexity in both the number of observations and the number of dimensions. They compute in real-time on a projection with 300,000 observations without parallelisation. The global explanations have a worse complexity in the number of observations however only need to be computed once. The precomputation takes roughly one second for a dataset with 10,000 points and roughly 25 seconds for a dataset with 100,000 points using parallelisation. That

being said, the algorithm is embarrassingly parallelisable and therefore could be computed much more efficiently on the graphics card. Considering that most of the analysis happens in local exploration mode, and the global explanations are just an initial guide of where to explore, we consider the system to scale well in terms of computational cost. All timings were measured on a machine running Windows 10 equipped with an Intel i7-4790K CPU running at 4.00GHz clock speed.

BLACK-BOX APPLICABILITY In terms of requirement five, the explanations have no intrinsic reliance on any specific projection technique, and will function for any given projection, no matter how it was generated. Of course, if the projection quality is very low, the explanations will be meaningless, however a projection with such low quality should not be analysed for insights in any case. Overall, we therefore consider the fifth requirement to be met by the system.

Next, we will evaluate the results of the evaluation study and examine them in light of the first three requirements.

7.1 EVALUATION STUDY

We created a set of new explanatory tools that allow for interactive exploration and explanation of projections of high-dimensional data on both a global and local level. As stated in the requirements we put on a solution, we would like the tools to explain local point patterns in terms of their original dimensions, be easy and intuitive to use and scale well in the number of observations and dimensions. In order to get an idea of to what degree these requirements are satisfied we performed an initial evaluation study. This study had participants performing a series of analyses on datasets of increasing dimensionality, from both static snapshots and by live exploration using our system.

As far as we are aware this is the first study of its kind, evaluating how these kinds of explanations are understood by users. We believe this study is incredibly important to substantiate our contributions. As we introduce a lot of new explanatory elements, without knowledge of, whether these tools make sense, whether they produce consistent and accurate insights into the data, and whether they are understood and valued by people, the contributions have no justification.

From the results of the study we saw an overwhelming number of participants picked the right answer on the control questions, as well as answers that were congruent with our analysis of the data on the live exploration questions. Not only did the participants in their majority give correct and plausible answers, but they consistently did so over all control questions in the study. Surprisingly, never in the study did an answer we considered unlikely to be correct receive a majority of votes.

In terms of difficulty, the control questions were designed to be fairly easy. This likely plays a role in the high accuracy of the answers. However, as the questions were designed to test whether participants could read

the various visual elements of the system correctly, this consistency and accuracy suggest strongly that they indeed did understand how to read the explanations. This was important to establish, considering the live exploration questions rely heavily upon being able to correctly read the explanations the system provides. Moreover, considering the participants had only followed a short ($\sim$ 15 minute) tutorial on the features of the system, this is a crucial first piece of evidence these elements are easy and intuitive to learn and understand.

The live exploration questions presented participants with more complex and abstracted analysis questions that required them to use the system as they saw fit to come to an answer. We saw that answers given by participants were less consistent than the control questions, which is not strange considering the free nature as well as the increased complexity of the questions. After all, participants were free to select any region in the projection, at any scale, and even integrate results from multiple regions to come to a conclusion. Therefore, answers can differ based on what exactly the participant was looking at, the amount of effort put into the analysis, different understandings of the question, and a host of other factors.

Nevertheless, despite this, we were surprised to see that the vast majority of answers given by participants lined up exactly with the answers we considered likely based on our own extensive analysis. Moreover, we saw that only a few answers got a majority of votes from participants, while the others barely got any. This conveys participants came to very similar and consistent insights, suggesting that the system is indeed capable of helping users solve analysis problems.

Not only did participants give consistent and plausible answers to questions on datasets having just a few dimensions (like the wine dataset), neither of these qualities noticeably degraded as the datasets became higher dimensional (breast cancer and spam datasets). This is evidence that explanations generated by the system scale beyond just a few dimensions, to datasets of much higher dimensionality.

Finally, we asked participants to rate the various tools in terms of their usefulness on a scale from 1 (not useful) to 7 (very useful). Participants responded very positively to the contributions, giving high ratings to all elements of the system. The lowest rating (4.83 SD=1.63) was given to the variance ranking mode, which can indicate a number of things. Firstly, it is possible that the variance ranking requires more time to fully understand, as the concept itself is not complicated, but the implications of a dimension having low or high variance can be hard to grasp. Secondly, it is possible that the variance ranking is just not that useful in answering the type of questions formulated in the study. Apart from the control questions, only one live exploration question really dealt with using the variance mode, while the rest could be solved in value mode. This bias in the questions is accidental and could contribute to participants not perceiving it as useful. The rest of the elements were perceived as very useful to participants. This corroborates the idea that the new tools are actually helpful to users in their analysis of the data, otherwise, if participants found the answers to the questions in different

ways without help of the tools, then they likely wouldn't rate them as very useful.

In conclusion, we ran an evaluation study to find out if our visualisation system is capable of explaining local point patterns in projections in terms of their original dimensions, is easy and intuitive to use and scales well in the number of observations and dimensions. Participants in the study gave consistent and accurate answers that strongly suggest the system is easy to learn and use with minimal training and results in consistent insights, even when the analysis process is not controlled and analysed datasets become more and more high-dimensional. This satisfies the remaining requirements we believe necessary for an explanatory mechanism capable of adequately explaining local point patterns in high-dimensional projections. Therefore, we consider our visualisation solution to be an answer to the research question we set out to investigate.

In the following section, we will look at several limitations of our visualisation solution and evaluation study and discuss potential future work.

7.2 LIMITATIONS AND FUTURE WORK

We now discuss several of the limitations of our work, and suggest future improvements and directions for further research. We divide the discussion of these items up by topic.

GLOBAL EXPLANATIONS We start our discussion by looking at the limitations that apply to the global explanations. Firstly, the global explanations inherit their mechanism from the original paper by da Silva et al. We discussed before that as the intrinsic dimensionality of the dataset becomes higher, assigning visually distinct colours to observations based on top-ranked dimensions becomes a limitation. This limitation remains for both the original global variance explanation proposed by da Silva et al. as well as for the global value explanation proposed in this thesis. Nonetheless, section 5.6 shows that there are datasets with a significant number of dimensions and high intrinsic dimensionality, for which the explanations do still work very well.

Secondly, the global explanations are designed to work on a projection without knowledge of the specific projection technique that generated it, nor its parameters. It could however be interesting to provide additional explanation metrics that can be used when such information is known. For example, if we know the projection technique used a particular distance function in order to generate the projection, then it may make more sense to analyse it with explanations based on that same distance function, rather than a more generic metric like variance. As users could be given the choice which metric to evaluate their projection with this would not detract from the generality of the system. However, it may make the system less accessible to users who don't have the experience or knowledge to pick the right metric.

Lastly, the precomputation required for the global explanations becomes increasingly computationally costly as the number of observations in the data increases. However, as it only needs to be computed once this cost is not prohibitive. That being said, if a projection does contain millions of points then the computation will likely take a significant amount of time. This problem may be mitigated by the fact that the computation is embarrassingly parallelisable and could easily be modified to run on a graphics card. In practice however, a projection containing that many points at once is likely suffering from more issues such as a high number of projection errors and misrepresentation of relative distances between clusters due to a lack of visual space. In this case, a hierarchical projection or a projection showing just a subset of the data points is a better approach. Therefore, it would be interesting to extend the application of these global explanations to such projections.

LOCAL EXPLANATIONS The local explanations also have a number of limitations that we will discuss here. Firstly, while using the local explanations allows for the analysis of many dimensions at once, exactly how many dimensions is limited by the size of the local analysis widget and therefore by the size of the screen. While it is possible that there exists a visual layout that could display more dimensions at once at the same screen size, it soon becomes a trade-off with the readability. For many datasets however, the current number of dimensions shown on an average screen size is more than sufficient for proper analysis of the data.

Secondly, in some datasets we are equally interested in dimensions that have unusually high values as well as unusually low values. If these dimensions are shown on either sides of the sorted dimension list, this makes the analysis cumbersome. Therefore, users might be given the option to choose between either sorting from unusually high values to unusually low, vice versa, or a combination of the two. This could easily be implemented by slightly modifying the value ranking computation. Furthermore, keeping track of a particular dimension while moving the lens can sometimes be tricky as it can appear higher or lower in the sorting based on the brushed points. Therefore, an additional option to have a fixed sorting of dimensions would be valuable.

Thirdly, some datasets have abstract dimensions rather than ones with a semantic meaning. This is commonly the case in image datasets where the pixels are taken as the dimensions. In these cases, the dimensions are often simply labelled by a numerical value indicating which pixel they refer to. Analysis of such datasets is still possible using the described explanatory mechanisms, but much less intuitive. An interesting direction for future research would be to visualise the ranking or statistics of the dimensions as pixel colours in a representative image widget.

Lastly, in variance mode it can occur that there is an incongruence between the ranking of the dimensions in the local explanation widget and the displayed standard deviation whiskers. That is, as dimensions are sorted from least variance to most variance, one would expect the

width of the standard deviation whiskers to uniformly increase in size for dimensions going down the list. However, this is not always the case. The reason for this is that we choose to base the sorting of the dimensions in the local analysis widget on the rankings computed for the global explanations, which are based on neighbourhoods around individual points, while the standard deviation whiskers are computed based on the variance of solely the brushed points. There is therefore a trade-off between congruence of the top-ranked dimension in the global explanation and in the local explanation versus congruence of the ranking in the local analysis sorting and the displayed standard deviation whiskers. We think the former is more important as we believe that exploring a group of points coloured in purple in the global explanation and finding that the top-ranked dimension in the local explanations is the orange dimension is very jarring and confusing. However, it may be possible to find some middle ground where both are mostly congruent, barring exceptional cases.

EVALUATION The main limitation to the evaluation study is that it doesn't conclusively prove anything, however this was not the intention. The study was purely intended to get an initial sense of how quickly people can learn to understand our tools, how much experience is necessary to use them and whether our tools results in consistent and plausible insights when users are given complete freedom in their use. Given that, as far as we are aware, this is the first study of its kind in the field of explaining projections, it serves as a first foray into substantiating the contributions of new explanatory mechanisms.

VISUAL DESIGN Through comments from participants of the study as well as colleagues we became aware of several minor issues in the visual design of the tools. Firstly, the bars indicating the difference between the local and global mean in the local analysis widget are coloured in green and red, however these colours are hard to distinguish from another to people with red-green colour deficiency. Secondly, our chosen colourmap [14] still contains colours that are hard to distinguish from the background of the local analysis widget. Leaving these colours out would leave around 16 colours left in the palette, which we expect to only have a minor impact on the quality of the explanations on datasets with higher intrinsic dimensionality.

7.3 CONCLUSION

In this thesis we discussed how local point patterns in projections of highly multidimensional data can be adequately explained in terms of their original dimensions. This has applications in an incredible amount of fields, such as healthcare, artificial intelligence, business and research.

We looked at the previous research that has been done in this area and concluded that there is a gap in the ability to explain projections

in enough detail, especially when the intrinsic dimensionality of the dataset increases.

Therefore, we presented a visualisation solution that allows for in depth explanation of projections at multiple levels of detail, whilst simultaneously scaling better in the number of dimensions of the data. The solution is implemented in the HDPS [7] software application for exploring high-dimensional data. The source code and binaries of our implementation are available on GitHub [13].

Importantly, we demonstrated the functionality and usefulness of our system by running an evaluation study. Results of the study show that, with minimal training, users arrive at consistent insights using the explanations generated by the system, and that those insights coincide with the insights gathered over a much more extensive analysis.

Therefore, we believe the visualisation solution we present can provide a lot of benefits for all fields dealing with high-dimensional data in trying to obtain novel insights into their data.

APPENDIX A

Evaluation

Thank you for taking the time to participate in this evaluation.

In this evaluation we will ask you to download a software tool for Windows 10 or Linux (Ubuntu or Debian-based) machines. Then we will ask you to get familiar with the tool and use it to answer several questions. Finally, some additional opinion questions will be asked. [In total ~50 mins]

Next Page 1 of 46 Clear form

The Research

A lot of fields of research and business deal with high-dimensional data. Analysis of such data is problematic and hard to understand. The domain of Visual Analytics attempts to understand the data by visual exploration. Visualizing such data can be done by using dimensionality reduction and producing a (typically two-dimensional) projection of the data.

These projections can be inspected in a scatter plot to find groups of observations that have similar attribute (dimension) values or outliers with very uncharacteristic dimension values. However, from just a scatter plot visualization it is unclear what dimensions cause these observations to be similar or different.

Several efforts have previously been made to explain the structure of these projections, however they tend to fall short in adequately explaining patterns in the projection, or the techniques don't scale well to a large number of dimensions.

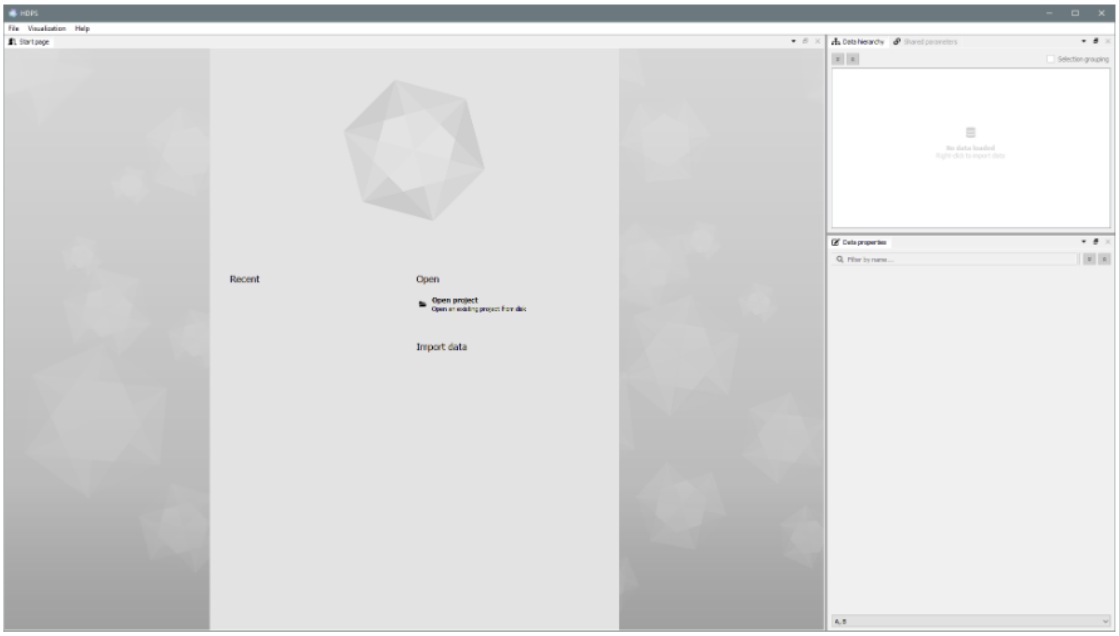
We aim to improve these techniques with additional explanations and tools that can potentially unlock the ability to adequately explain datasets with a larger number of dimensions.

For this purpose, we have created a software tool that we will ask you to download and get basic familiarity with. We will then ask you to answer several questions about projections of three datasets. Finally, some additional opinion questions will be asked. [In total ~50 mins]

Installing the Tool

In order to use our tool, please be aware that a Windows 10 or Linux (Ubuntu or Debian-based) machine is required to run it.

If you have access to such a machine, please visit the following link to download the tool for one of these platforms and follow the few steps there to get it running: https://docs.google.com/document/d/1klpvBm4S_CCF3ATXuSzl7V0wFYWTzr7l9xuTZl8V-bM/edit?usp=sharing



If the application has started successfully and you are presented with this screen. Continue to the next section.

Getting Familiar (1 of 8)

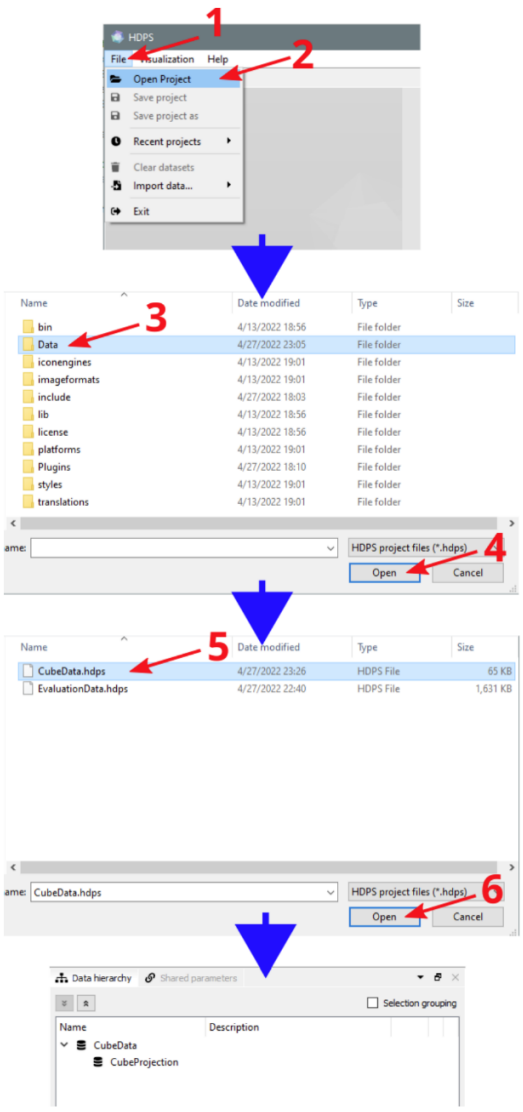
Before starting the evaluation, we would like you to get familiar with using the tool. In the following sections we will walk through a basic dataset and analyze it using the tools created for this research.

Back Next Page 4 of 46 Clear form

Getting Familiar (2 of 8): Loading Basic Data

In order to get familiar the tool, we will load some basic data to explore.

In the menu bar on the top-left of the application, click on File -> Open Project, and navigate to the folder where you extracted the application. In this folder next to HDPS.exe is a folder called Data. Navigate into this folder and select and open the CubeData.hdps project. The process is explained in images below:

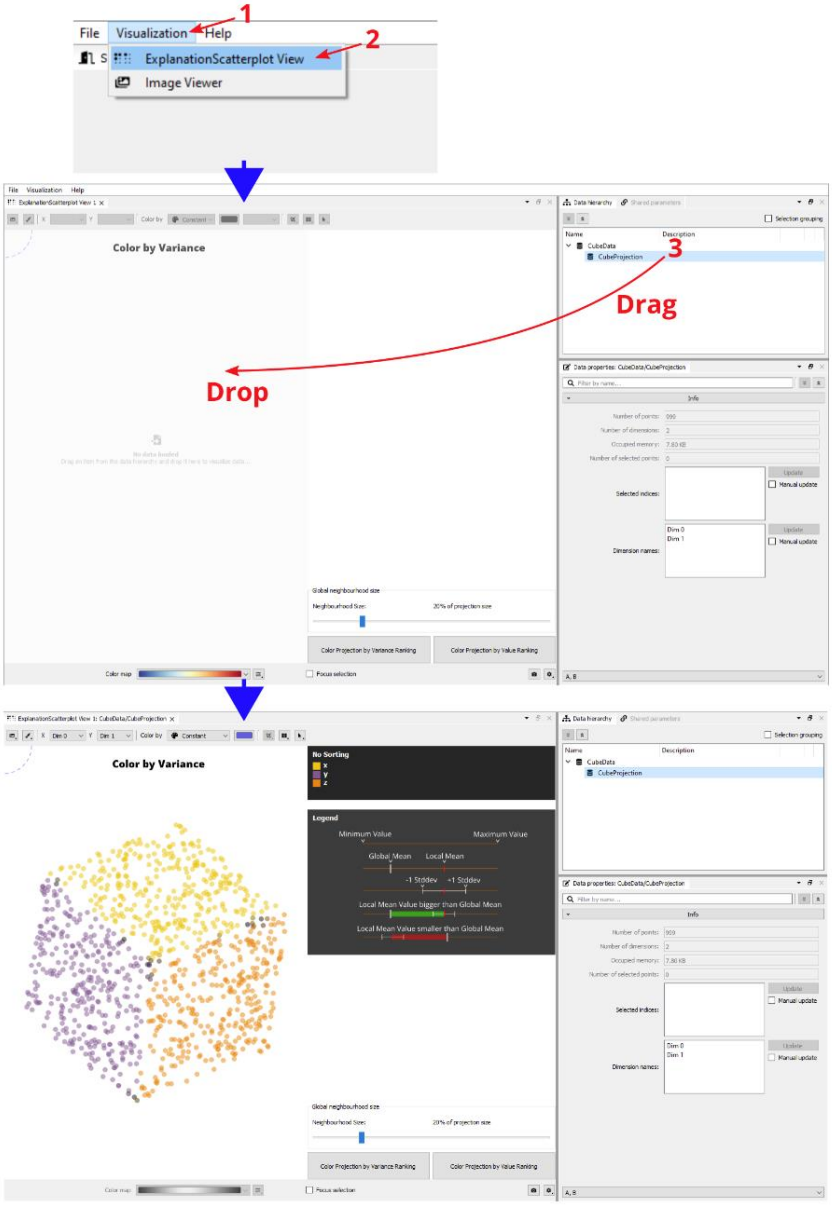


If all went well, the Data hierarchy window on the top right should now list a CubeData and CubeProjection dataset. If that is the case, please continue to the next section.

Back Next Page 5 of 46 Clear form

Getting Familiar (3 of 8): Visualizing the Data

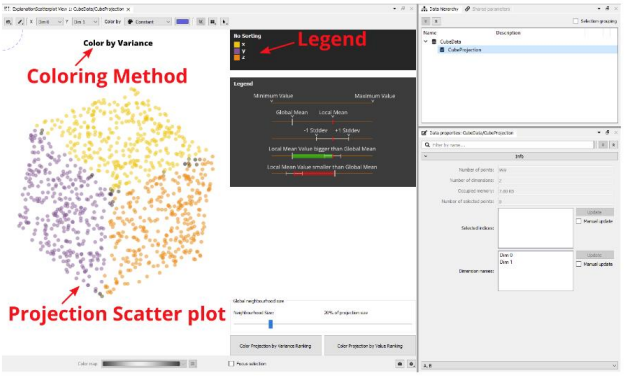
In order to visualize the loaded data, we will open a Scatter plot view and show the data in it. In the menu bar on the top-left of the application, click on Visualization -> ExplanationScatterplot View. A view will open up in the central area of the application. Now click the dataset called 'CubeProjection' in the Data Hierarchy window on the right, and drag it into the central visualization view. The process is explained in images below:



Getting Familiar (4 of 8): Understanding the Data

The data visualized in the scatterplot is a 2D projection of 999 points scattered randomly on 3 faces of a 3-dimensional cube. These points have three spatial dimensions (coordinates) named x, y and z.

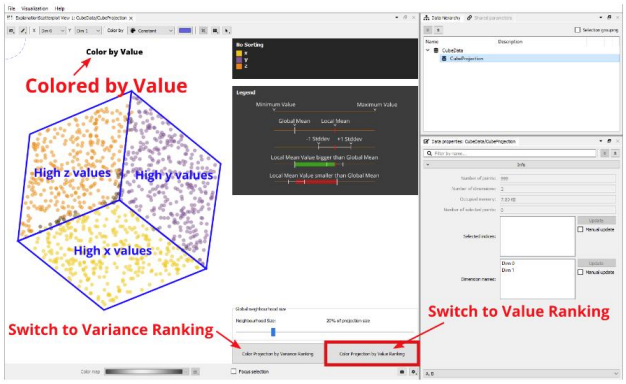
Each of the points of the data is represented as a dot in the projection scatterplot. These dots are currently colored according to which dimension has the lowest variance in that region of the projection (Indicated by the Color by Variance label at the top of the scatter plot). In the top-right of the visualization, a legend shows which color corresponds to which dimension (in this case, yellow for x, purple for y and orange for z). As an example, points colored yellow in the scatterplot will have the lowest variance in the x-dimension. Points colored purple will have the lowest variance in the y-dimension, etc..



Similarly to coloring the points by lowest variance, we can color them by which dimension has the highest values in that region of the projection. On the bottom-right of the visualization view there are two buttons, "Color Projection by Variance Ranking" and "Color Projection by Value Ranking".
Click the latter (outlined in red below) to switch the projection coloring to Value mode. Again the color yellow is assigned to x, purple for y and orange for z.

Take some time to switch between the different coloring modes and observe the label called Coloring Method in the previous image change to indicate the current mode.

As an example, in Color by Value mode, the points colored in yellow have the highest values for the x-dimension, relative to the average value of the x-dimension over all points. Points colored in orange have the highest values relative to their global average for the z-dimension.

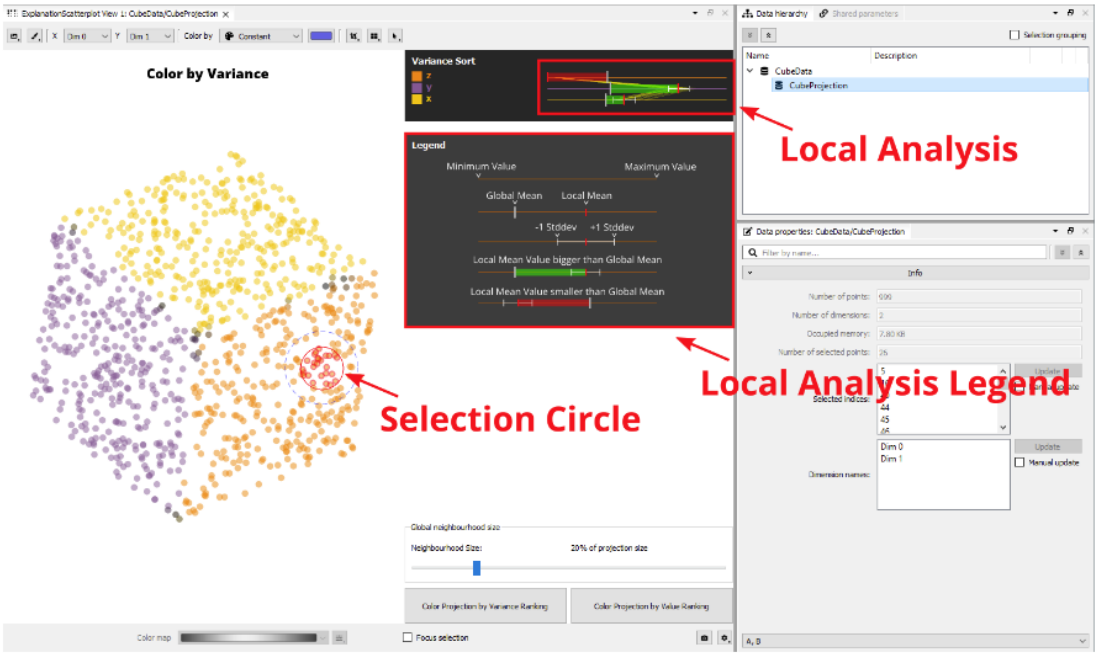


Getting Familiar (5 of 8): Exploring the Data

While the coloring of the points already gives some insights into the data, it is at a very global level. In order to understand the local differences between projected points we can explore the data on a more local level.

Clicking and dragging anywhere in the scatterplot will move a selection circle (red circle) around the projection. Points within this circle are selected and can be analyzed. Scrolling with the mouse wheel will make the selection circle bigger or smaller.

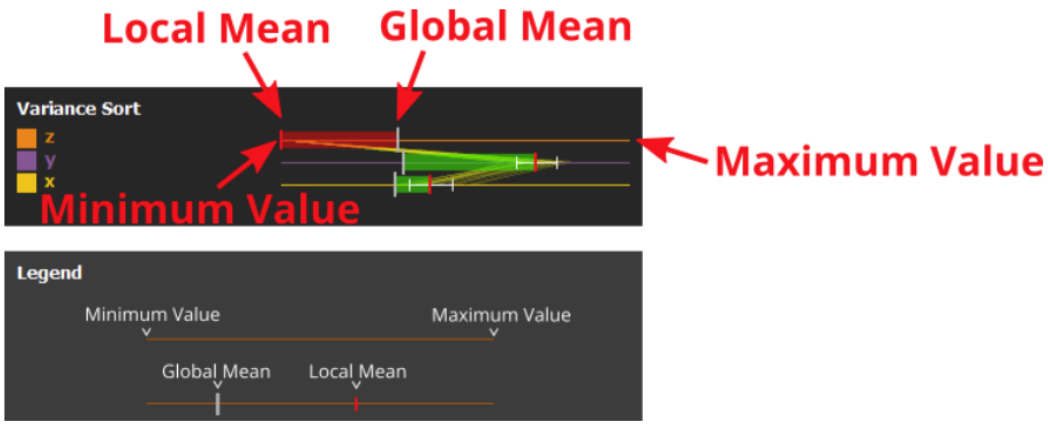
If at least some points are selected, the dark top-right widget now shows additional information. We will explain the different elements of it in more detail in the next section.



If you feel comfortable with clicking and dragging the selection circle around, and you have tried to scroll the mouse wheel to make it bigger or smaller, please continue to the next section.

Getting Familiar (6 of 8): Local Analysis

Let's have a closer look at the basic elements of the local analysis widget (the widget with a dark-gray background shown in the top-right of the central view, see image below).



Range, Local and Global Mean

To the right of each listed dimension, a horizontal line is drawn. This line represents the full range of values for that particular dimension in the whole dataset. The left endpoint represents the minimum value the dimension takes over the whole dataset, and similarly, the right endpoint represents the maximum value.

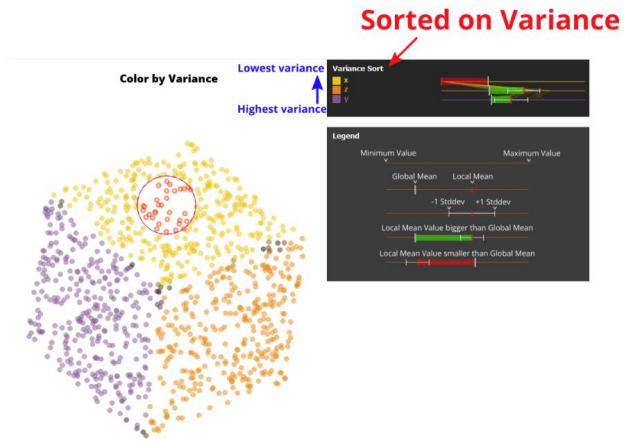
Intersecting this line is a little vertical red line which represents the average value of this dimension over the selected points. We call this the Local Mean. As an example, in the previous image the local mean (little vertical red line) of the z-dimension is all the way on the left. This means that the average value of the points we selected is equal to the minimum value of the z-dimension over the whole dataset (The minimum value is 0 in this cube dataset).

Another vertical grey line intersects the range line, which represents the average value of this dimension over the whole dataset. We call this the Global Mean. As an example, in the previous image the global mean of the z-dimension is around 1/3 of the way along the range line. The minimum value of the z-dimension is in dataset is 0, while the maximum value is 1. Therefore, the global average value of the z-dimension over all points is around ~0.33.

Getting Familiar (7 of 8): Variance vs. Value Ranking

As you move the selection circle around the projection of the cube faces, you will observe that the order of the dimensions changes. Depending on whether the points are colored by variance or by value, the dimensions in the local analysis are ranked and sorted according to variance or value.

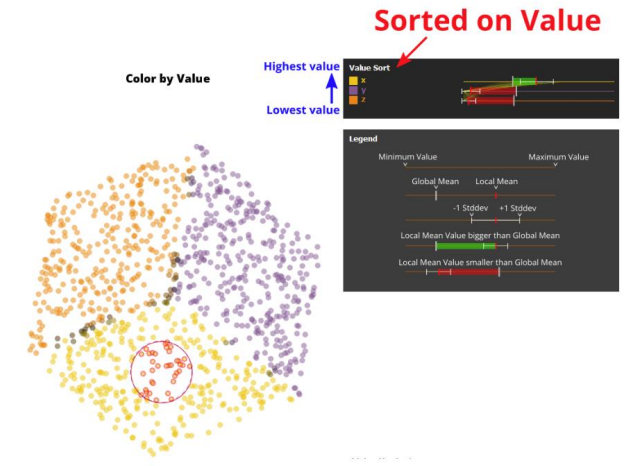
In the following image the points are colored by variance, and the dimensions of the selected points are ranked from lowest variance (at the top) to highest variance (at the bottom).



In the following image the points are colored by value, and the dimensions of the selected points are ranked from highest value (at the top) to lowest value (at the bottom).

You can see the x-dimension has a higher average value for the selected points than its average value over all points (the red vertical line is further to the right than the gray vertical line). If this is the case, a green bar is drawn to quickly visually indicate the local mean is higher than the global mean.

For the y-dimension the average value of the selected points is lower than its average value over all points (the red vertical line is left of the gray vertical line). If this is the case, a red bar is drawn to quickly visually indicate the local mean is lower than the global mean.



Getting Familiar (8 of 8): Conclusion

If you feel comfortable with the various elements described in the previous sections, please proceed to the next section where we start the evaluation.

If you want to review the Getting Familiar section during the evaluation, you can find all of it here: https://docs.google.com/document/d/1PVCC543H8tYbQwXTq1l-xWTlZ1LROzB6LheSQfBGoDU as well as in the Manual.pdf document next to the software executable.

The Evaluation

In order to evaluate the effectiveness of the tools we have created, we will ask you to use our tool to analyze projections of three different datasets. For each of these datasets you will be asked to analyze specific portions of the projection using the explanatory mechanisms described in the previous section.

In the next section, we will present the first dataset. For each dataset we will ask you to answer 4 questions using screenshots of the tool, and 3 questions where you will use the tool yourself to analyze the dataset and answer the questions.

If you are ready, please proceed to the next section.

Dataset 1 of 3: Wines

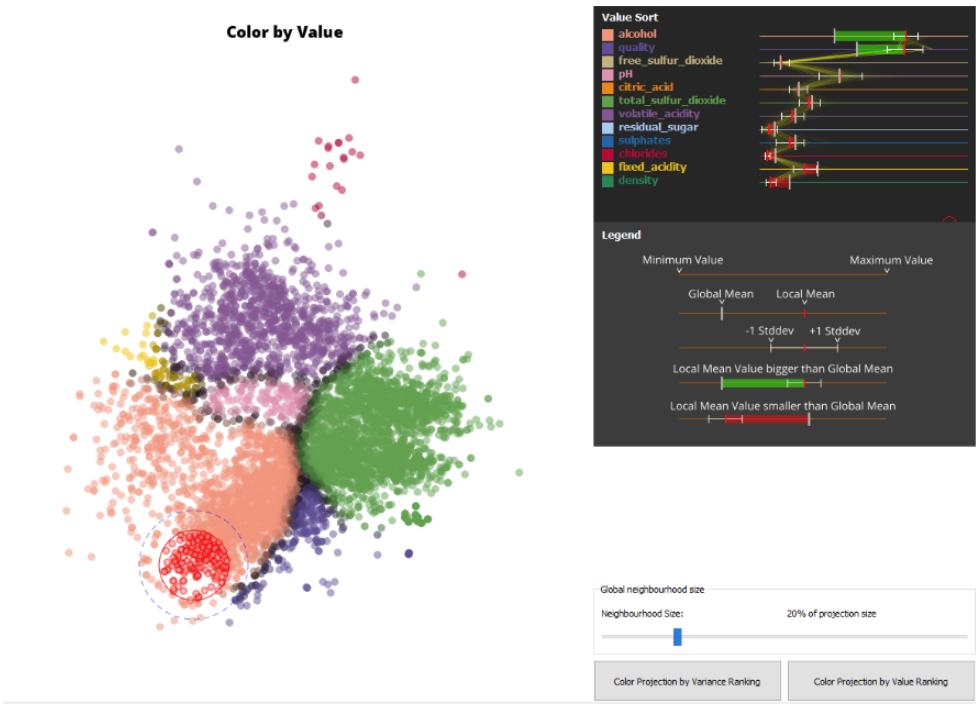
This dataset contains roughly 6500 samples of white and red 'vinho verde' wines from Portugal. For each wine sample 12 attributes are measured (e.g. pH, alcohol %, total sulfur dioxide). One of the attributes (dimensions) is the perceived quality of the wine on a scale from 0 to 10 (0 = worst quality, 10 = best quality). This dimension is called 'quality' and will be the subject of several questions.

In each of the following four questions, we will show you a projection of this dataset and ask you to explain a selected region in the projection using a screenshot of the tool. Each screenshot will show the projection of the data on the left, and the explanations for the selected points on the right.

Wines: Question 1 of 4

For each of the wines in the dataset, the alcohol percentage is measured. This is encoded in the 'alcohol' dimension.

Looking at the following image, several points are selected using the selection circle (points in the red circle). In the ranking (order of dimensions in top-right widget) we see that the average alcohol percentage of the selected wines is much higher than the average alcohol percentage of all wines (see the big green bar to the right of 'alcohol').



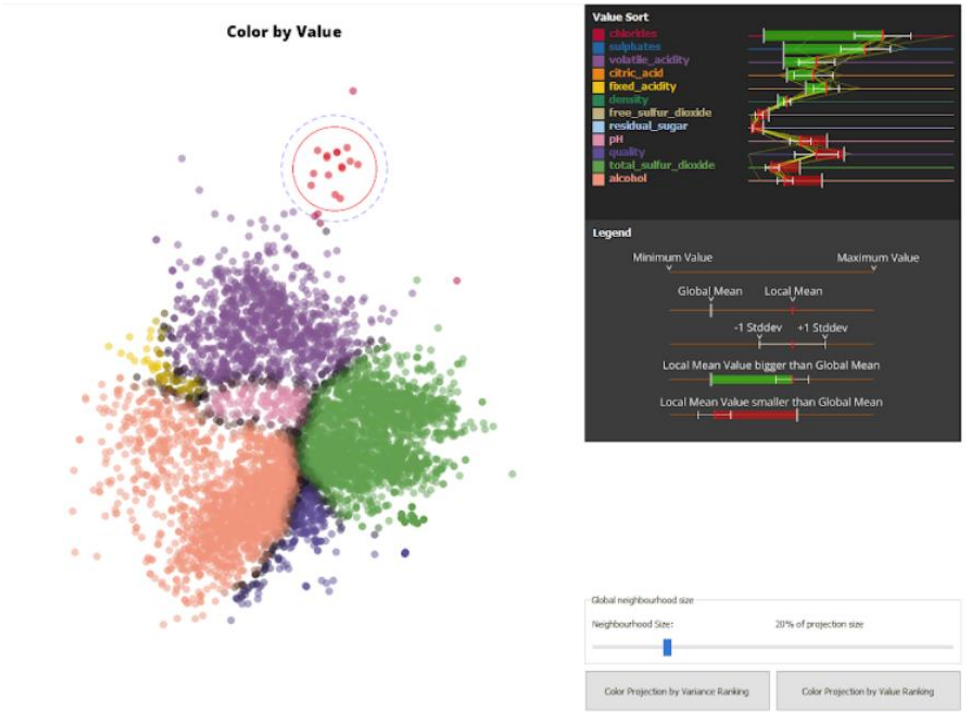
For the selected wines, a high alcohol percentage seems to be associated with a: *

- ◯ Higher than average perceived quality
- ◯ Lower than average perceived quality
- ◯ Nothing can be said about the perceived quality

Wines: Question 2 of 4

Another one of the attributes measured for each wine is their 'chloride' level. This corresponds to the amount of salt in the wine.

Looking at the following image, several points are selected using the selection circle (points in the red circle). In the ranking we see these wines have an unusually high chloride level (see the big green bar, i.e. the local mean is much higher than the global mean).



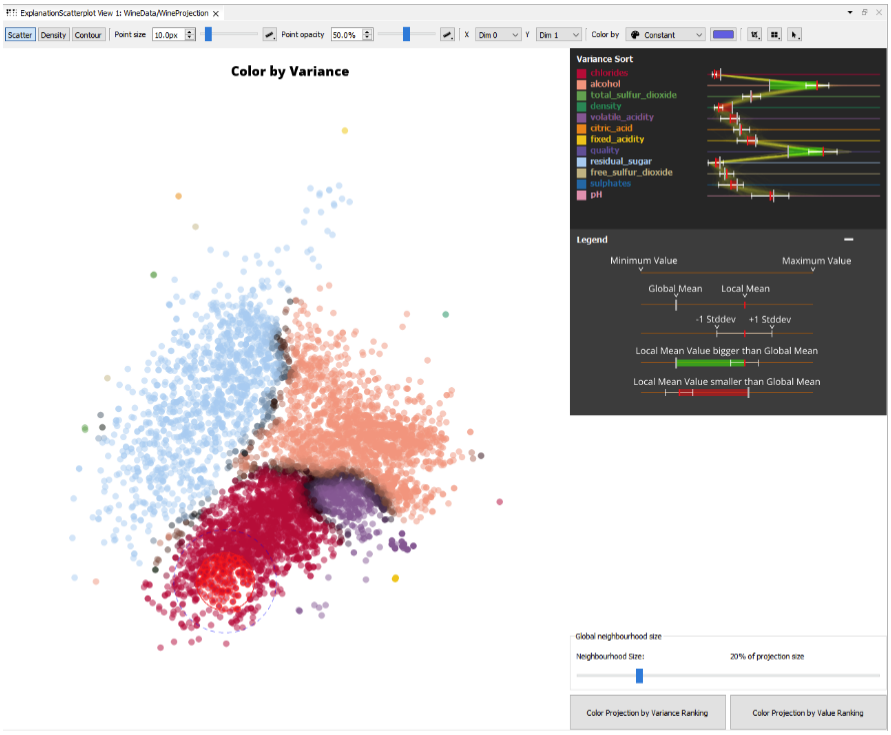
What can be said about the perceived quality of these wines? The perceived quality of these wines is: *

- ◯ Higher than average
- ◯ Lower than average
- ◯ Nothing can be said about the perceived quality

Wines: Question 3 of 4

Below a scatterplot is shown in 'Color by Variance' mode. Meaning the colors correspond to dimensions with the lowest variance in that region of the projection, and the dimensions in the top-right widget are ranked from least variance to most variance (top-to-bottom).

Again several points are selected using the selection circle (points in the red circle). In the ranking we see these wines have a low variance in their chloride levels (chlorides is ranked as the top dimension based on variance).



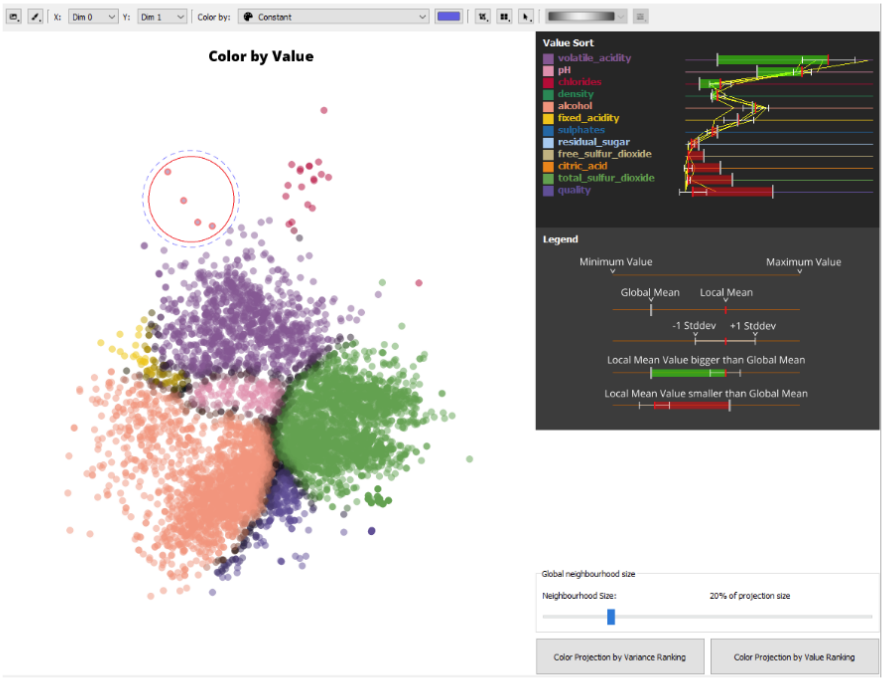
What can be said about the pH value for the selected wines? *

- ◯ The pH value is constant over the selected wines
- ◯ The pH value is higher than usual over the selected wines
- ◯ The pH value changes more than other dimension values over the selected wines
- ◯ None of these statements can be concluded from the image

Wines: Question 4 of 4

Volatile acidity in wines refers to the presence of acids in the wine that easily evaporate (mostly acetic acid, the main component of vinegar). These acids are typically formed by bacterial spoilage and large amounts are considered unfavorable for the taste. Wines with a lower pH value are typically less susceptible to this bacterial spoilage.

Looking at the following image, several points are selected using the selection circle (points in the red circle). In the ranking we see these wines have relatively high pH (the local mean of the pH dimension is higher than the global mean), as well as a high volatile acidity. Therefore it is likely they were affected by bacterial spoilage.



What else can be observed about these wines? *

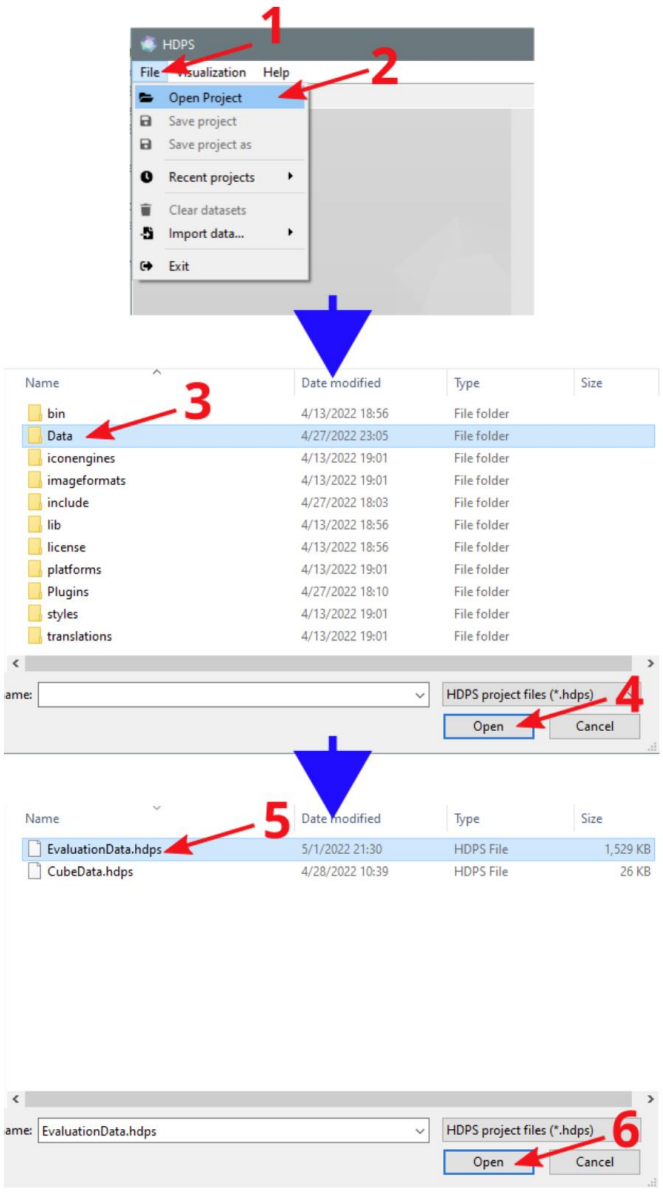
- ◯ They have an unusually low level of chlorides
- ◯ They have an unusually low density
- ◯ They have an unusually low level of citric acid
- ◯ None of the above can be observed

Wines: Live Exploration

Now please restart the tool application by closing any previously opened instances, and double-clicking HDPS.exe in the unzipped folder.

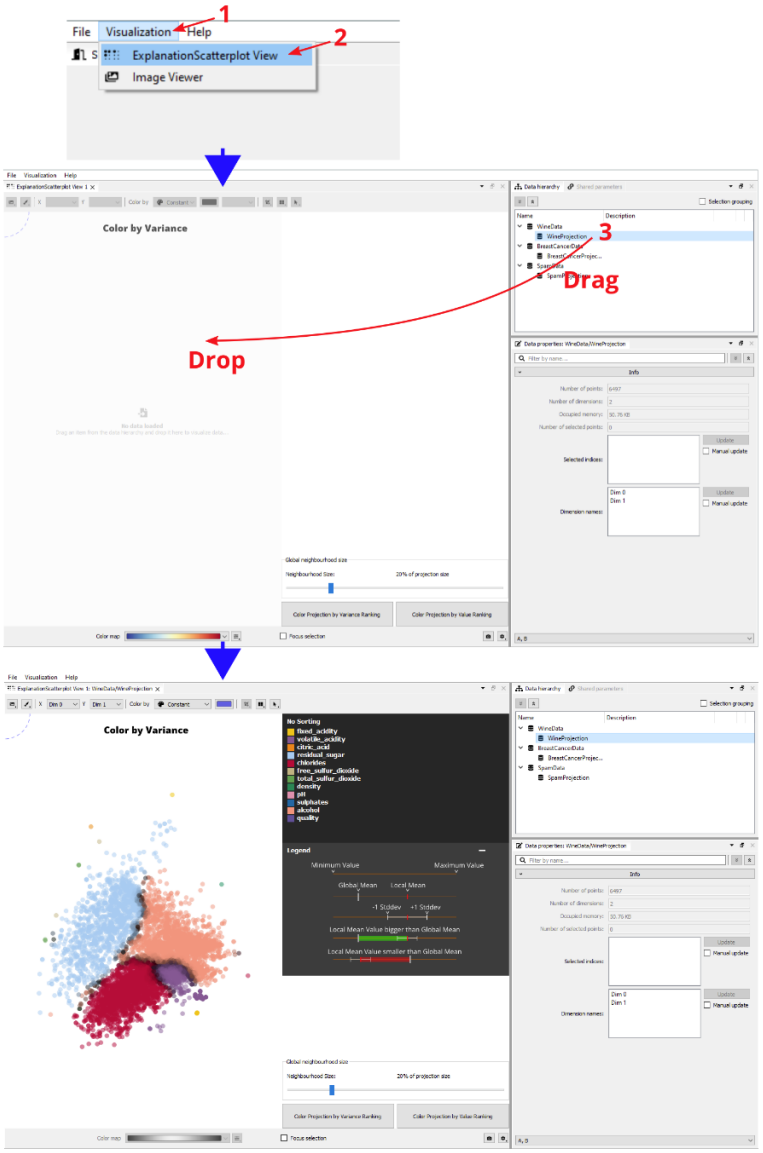
For the next 3 questions, we will ask you to perform your own exploration of the data using the tool.

First we must load the three datasets used for this evaluation in the tool. In the menu bar on the top-left of the application, click on File -> Open Project, and navigate to the folder where the application is stored. In this folder next to HDPS.exe is a folder called Data. Navigate into this folder and select and open the EvaluationData.hdps project. The process is explained in images below:



Wines: Live Exploration

Now we can visualize the wine dataset by clicking Visualization -> ExplanationScatterplot View at the top and dragging and dropping the WineProjection dataset into the central view. See image below:



If your window looks similar to the final step in the image, please continue to the next section in this evaluation form. Note that the tool is in 'Color by Variance' mode in these images, the colors in your projection may be different if you are not in this mode.

Wines: Live Exploration, Question 1 of 3

The density of wine is generally close to that of water, however some of the wines in the plot have a particularly low density.

Find out which wines have the lowest density by moving the selection circle over the projection and looking at the dark widget to see where the density dimension has the lowest local mean. It may help to switch to _Value Ranking_ by clicking the 'Color Projection by Value' button.

For these wines with an extraordinarily low density, tick below which attribute is similarly out of proportion and likely the cause of the low density. *

- ○ Fixed Acidity
- ○ Volatile Acidity
- ○ Citric Acid
- ○ Residual Sugar
- ○ Chlorides
- ○ Free Sulfur Dioxide
- ○ Total Sulfur Dioxide
- ○ pH
- ○ Sulphates
- ○ Alcohol

Wines: Live Exploration, Question 2 of 3

For wines with a high quality, attributes with a low variance can be considered important for this high-quality. After all, if for the high-quality wines a particular attribute has very different values, then this attribute clearly does not affect the quality of the wines very much.

Find the highest quality wines in the plot and tick below which attributes, if any, appear to be the most important to the quality of these wines [Max 3]: *

- [ ] Fixed Acidity
- [ ] Volatile Acidity
- [ ] Citric Acid
- [ ] Residual Sugar
- [ ] Chlorides
- [ ] Free Sulfur Dioxide
- [ ] Total Sulfur Dioxide
- [ ] Density
- [ ] pH
- [ ] Sulphates
- [ ] Alcohol
- [ ] I don't know which dimensions are important

Wines: Live Exploration, Question 2 of 3 Follow-up

In the previous question, you have ticked 0-3 attributes that appeared to be important to the perceived quality of the selected wines.

Please tick below, which methods you used to arrive at your answer *

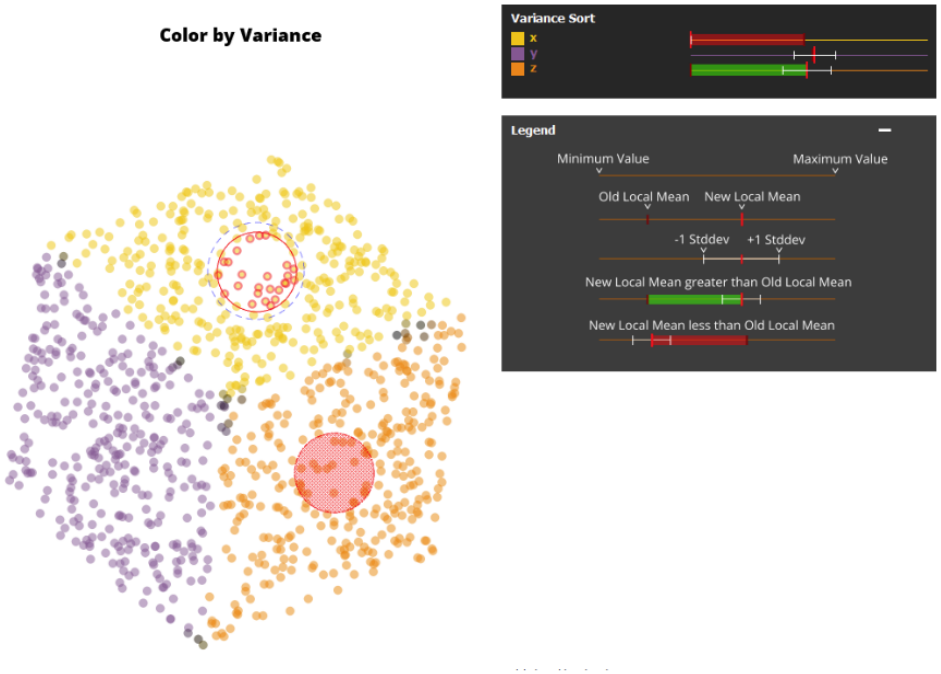
- [ ] I found the highest quality wines by moving the lens over the projection, and observing the local mean of the 'quality' attribute.
- [ ] I took the purple points in the 'Color by Value' mode to be the high-quality wines
- [ ] I took the purple points in the 'Color by Variance' mode to be the high-quality wines
- [ ] I ticked the attributes that also had high values for high-quality wines
- [ ] I ticked the attributes that had both extraordinarily high or low values
- [ ] I ticked the attributes that had extraordinarily low variances
- [ ] Other: _____

Differential Analysis

Before proceeding to the next question, we introduce another tool at your disposal.

Many times in understanding data it is useful to compare two groups of data with each other. Looking at the difference between these groups is called differential analysis. The software tool allows for doing easy differential analysis between different regions of the projection.

In order to do such an analysis place the selection circle on one region of the projection, press and hold the Ctrl button on the keyboard, and drag the selection circle somewhere else in the projection. Doing this on the Cube dataset would look something like:



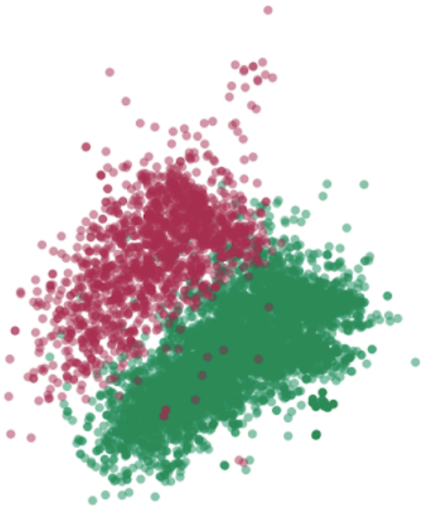
The filled-in red circle shows the region originally selected, the open red circle is the current selection. The dark widget on the right shows the difference between the original and newly selected region. For example, here the new selection has a much lower x-value, a similar y-value and a much higher z-value.

Try this differential analysis on the wine dataset, and then continue to the next question.

Wines: Live Exploration, Question 3 of 3

The wine dataset we are examining consists of both red and white wines.

Below is an image of the same projection as the one used in this live exploration phase, but the points are colored differently. Points colored in red correspond to red wines, whereas points colored in green correspond to white wines. Keeping this image of the separation of which wines are red and which are white in mind, please answer the following question (No need to reproduce this image in the tool, it is just for reference).



Using the differential analysis tool, find out which attributes are most different between the red and white wines. [Max 4] *

- [ ] Fixed Acidity
- [ ] Volatile Acidity
- [ ] Citric Acid
- [ ] Residual Sugar
- [ ] Chlorides
- [ ] Free Sulfur Dioxide
- [ ] Total Sulfur Dioxide
- [ ] Density
- [ ] pH
- [ ] Sulphates
- [ ] Alcohol

Dataset 2 of 3: Breast Cancer

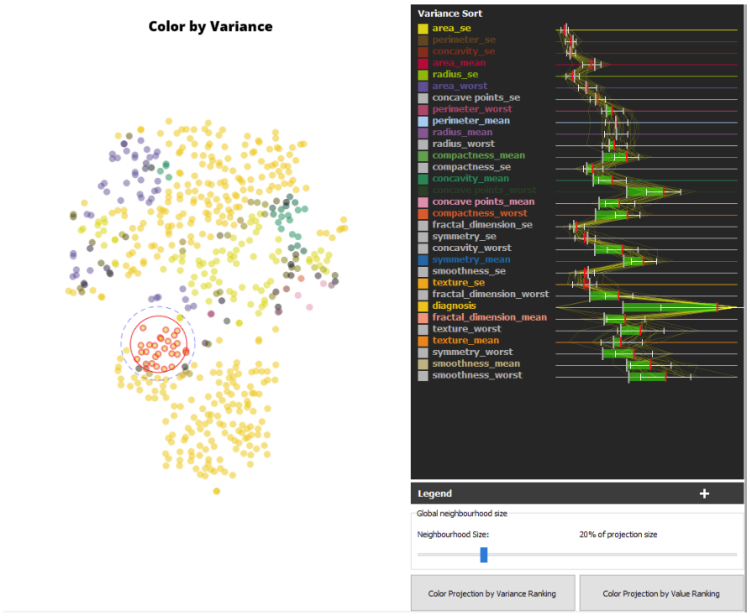
In order to check for breast cancer, one diagnostic procedure takes a small amount of breast tissue and checks it for cancer cells. This dataset examines roughly 570 slices taken from such a tissue sample. Each slice (each data point) contains a number of cells and 10 attributes are computed that describe the size, shape and texture of their cell nuclei. For each of these attributes the mean value (mean), the largest value (worst) and the standard error (se) are found, resulting in a total of 30 attributes per slice.
As an example, one slice may have a number of cells whose nuclei have a mean perimeter value ('perimeter_mean') of 1.3, a largest or worst perimeter value ('perimeter_worst') of 1.7 and a standard error value ('perimeter_se') of 0.15.

In each of the following 4 questions, we will show you a screenshot of our tool containing a projection of this dataset and ask you to explain a selected region in the projection based on information in the screenshot.

Following this, we will ask you to answer 3 questions by exploring the projection in the actual software tool.

Breast Cancer: Question 1 of 4

In the following image, the projection is in _Color by Variance_ mode, meaning the attribute with the lowest variance is used for the coloring. Several points are selected using the selection circle (points in the red circle). The dimensions of the selected points are listed in the dark widget and ranked top-to-bottom from least variance to most variance.



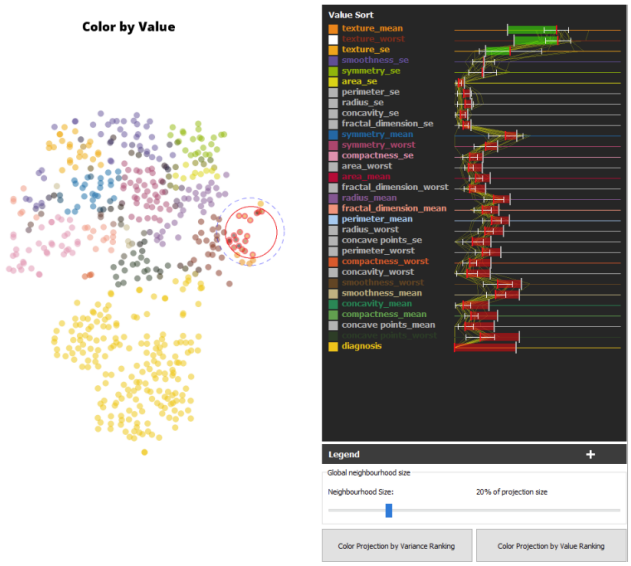
For the selected points which attribute has the lowest variance? *

- ◯ Radius
- ◯ Texture
- ◯ Perimeter
- ◯ Area
- ◯ Smoothness
- ◯ Compactness
- ◯ Concavity
- ◯ Concave Points
- ◯ Symmetry
- ◯ Fractal Dimension
- ◯ No attribute can be said to have the lowest variance

Breast Cancer: Question 2 of 4

The dataset contains one extra dimension not used in the projection process. This dimension is called 'diagnosis' and its value is 0 when the cells were found to be benign, and 1 when they were found to be malignant.

In the following image, the projection is in _Color by Value_ mode, meaning the attribute with the highest value is used for the coloring and appears highest on the list. Several benign (diagnosis of 0) points are selected using the selection circle (points in the red circle). On the right, the values of the attributes are shown. Higher values of attributes indicate a higher likelihood of malignancy.



For the selected points, which of the attributes shows an indication of malignancy? *

- ◯ Radius
- ◯ Texture
- ◯ Perimeter
- ◯ Area
- ◯ Smoothness
- ◯ Compactness
- ◯ Concavity
- ◯ Concave Points
- ◯ Symmetry
- ◯ Fractal Dimension
- ◯ No attribute shows an indication of malignancy

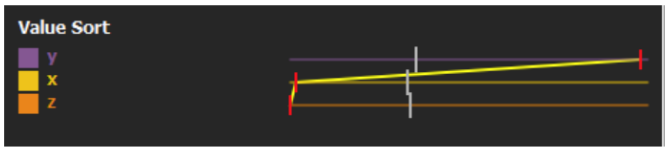
Parallel Coordinate Plot (PCP) Lines

Before proceeding to the next question, a brief explanation of the yellow lines in the widget on the right.

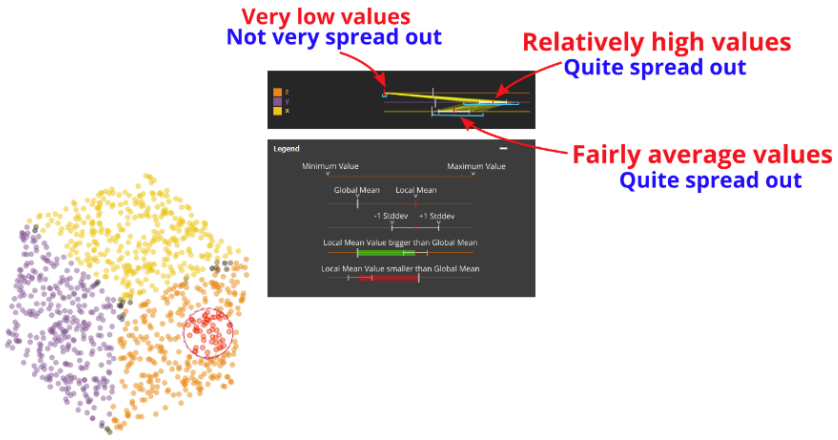
When selecting points in the projection, each selected data point is drawn as a vertical yellow line through the dimension ranking. Where this vertical line intersects the horizontal range lines corresponds to the value of that dimension for the data point.

Remember the horizontal range lines represent the full range of values that particular dimension takes on in the dataset. The left endpoint represents the minimum value over the whole dataset, and the right endpoint the maximum.

As an example in the image below just one data point is selected and we see it has a high value for y (the yellow line intersects the range line of the y-dimension all the way on the right), while having very low values for x and z (it intersects the range lines of the x and z dimensions all the way on the left).



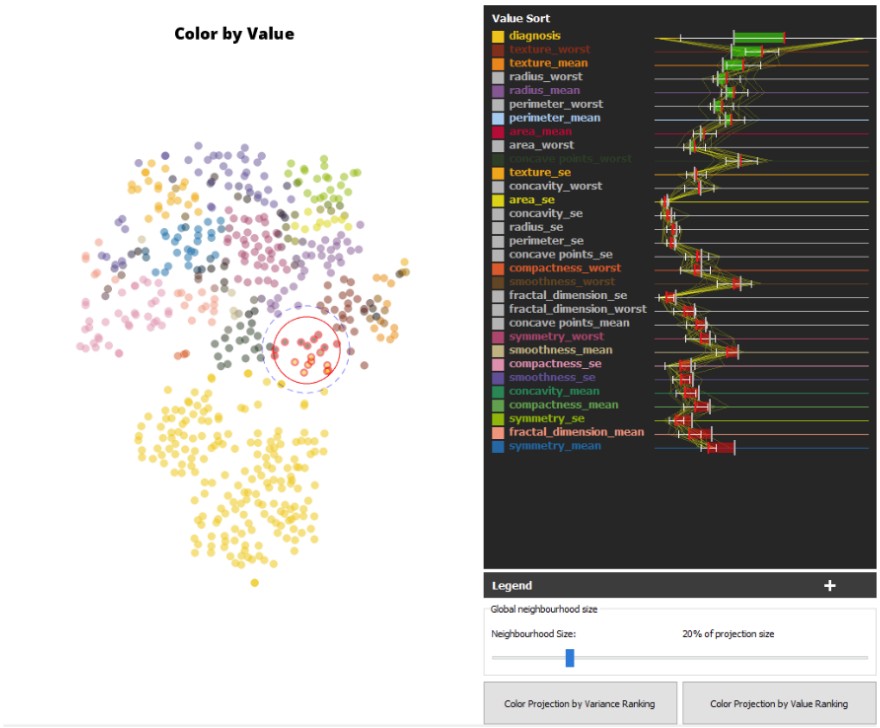
When selecting more data points at the same time, a line is drawn for every data point and it gives you an idea of the distribution of values the data points take on.
See the image below and then please continue to the next section if the concept is clear.



Breast Cancer: Question 3 of 4

In the following image, several points are selected using the selection circle (points in the red circle).

Looking at the PCP lines (yellow lines) in the image. A line corresponding to a benign observation would intersect the 'diagnosis' range line all the way on the left (as diagnosis would be 0), a line corresponding to a malignant observation would intersect the 'diagnosis' range line all the way on the right (as diagnosis would be 1).



What can be said about the diagnoses of these selected points? *

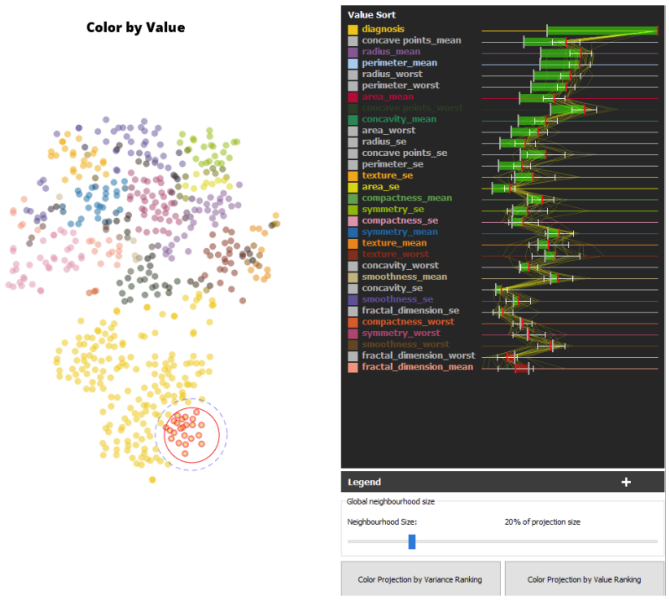
- ◯ All the selected points are diagnosed as benign
- ◯ All the selected points are diagnosed as malignant
- ◯ The selected points contain both benign and malignant diagnoses
- ◯ Nothing can be said about the diagnoses of these points

Breast Cancer: Question 4 of 4

In the following image, several points are selected using the selection circle (points in the red circle). On the right, the values of the attributes are shown.

We know for this dataset that higher values of attributes indicate a higher likelihood of malignancy, while lower values of attributes indicate benign cells.

We see that all the selected points have been diagnosed as malignant (all yellow lines intersect the diagnosis dimension on the far right, high values indicating malignancy).



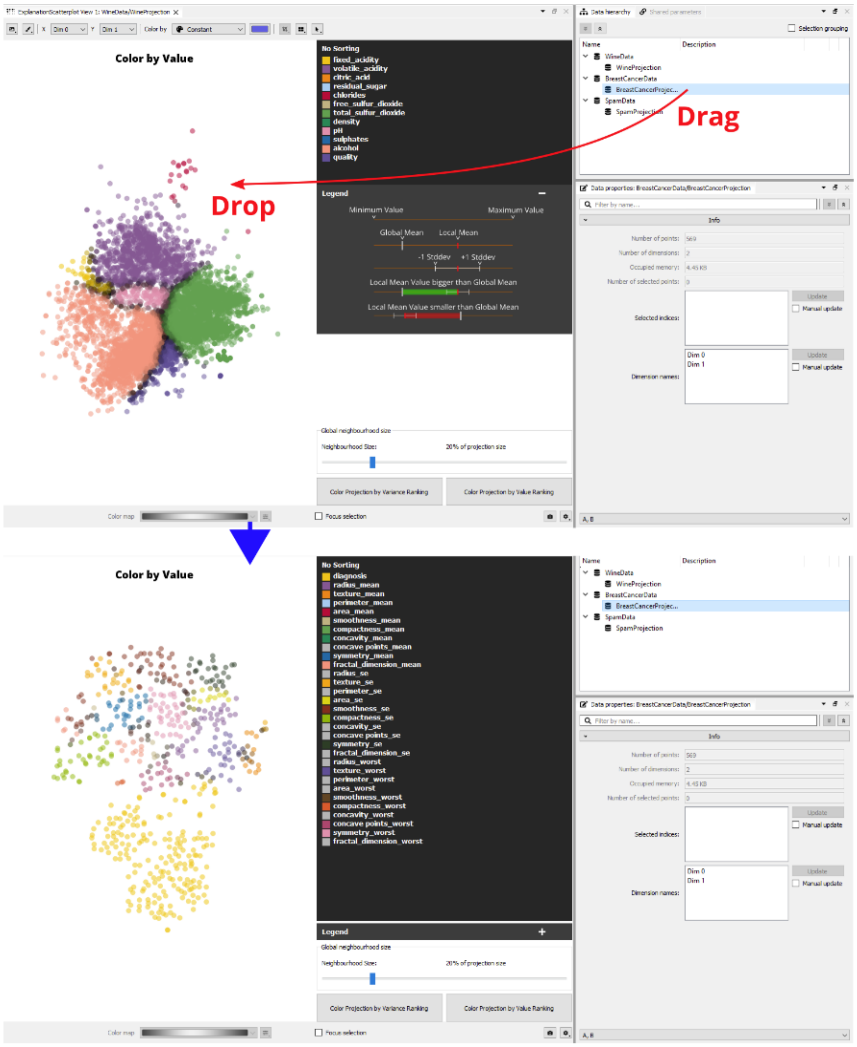
Which of the attributes, however, indicate the selected points are benign? *

- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Concavity
- Concave Points
- Symmetry
- Fractal Dimension

Breast Cancer: Live Exploration

For the next 3 questions, we will ask you to perform your own exploration of the data using the tool.

First, visualize the BreastCancerProjection dataset by dragging it from the right into the central view. This process is shown in the image below:



If the projection looks the same in the software tool running on your PC as in the image, please continue to the next section. Note that the tool is in 'Color by Value' mode in these images, the colors in your projection may be different if you are not in this mode.

Breast Cancer: Live Exploration, Question 1 of 3

Put the tool in __Color by Value__ mode (by clicking the 'Color Projection by Value' button).

There is a big cluster of similarly colored points whose color corresponds to the 'diagnosis' dimension. These are points that have a diagnosis of malign (diagnosis attribute is 1).

Within this malign diagnosis cluster there are multiple hidden subclusters where a particular attribute has higher relative values than other attributes (local mean much higher than global mean).

Find these hidden subclusters, and select which attributes characterize them due to their high relative values: [Max 3] *

- [ ] Radius
- [ ] Texture
- [ ] Perimeter
- [ ] Area
- [ ] Smoothness
- [ ] Compactness
- [ ] Concavity
- [ ] Concave Points
- [ ] Symmetry
- [ ] Fractal Dimension

Disabling and Enabling Dimensions

Before proceeding to the next question, we introduce another tool at your disposal.

When analyzing datasets with multiple dimensions, it sometimes occurs that several dimensions are not interesting for the analysis. In this case, it would be easier to understand the projection without including these dimensions. Therefore, the software tool allows for disabling and re-enabling dimensions from the explanatory tools.

In order to disable a dimension hover your mouse over the colored box in front of the dimension name in the right widget, clicking this box will exclude the dimension, turn it white, and move it to the bottom of the list (if there is a sorting). Clicking the box of the white excluded dimension again will re-enable it. This process is shown in the image below:

Try this disabling and enabling of dimensions on the breast cancer dataset. Make sure all dimensions are enabled again (no dimension box is shown in white) and then continue to the next question.

Breast Cancer: Live Exploration, Question 2 of 3

Disable the diagnosis dimension by clicking the square to the left of it, so that it turns white. In the projection, instead of seeing a big group of points colored with the color of the diagnosis dimension, you should now see different colors there.

Now try once more to identify the different subclusters in the malignant observations and attempt to answer the same question below.

For the malignant subclusters, select which attributes characterize them due to their high relative values: [Max 3] *

- [ ] Radius
- [ ] Texture
- [ ] Perimeter
- [ ] Area
- [ ] Smoothness
- [ ] Compactness
- [ ] Concavity
- [ ] Concave Points
- [ ] Symmetry
- [ ] Fractal Dimension

Breast Cancer: Live Exploration, Question 3 of 3

Radius, Perimeter and Area are all attributes that relate to the size of the cell nuclei in the observation. Higher values indicate bigger cells, and lower values indicate smaller cells.

Concavity and Concave Points are attributes that relate to the number and severity of concavities (indentations) in the cells nuclei.

Compactness is a measure of how compact the cell nuclei are. A circular disk shape would have a low compactness, while a shape with many irregularities would have a high compactness (remember, high attribute values indicate malignity).

Looking at the points with a malignant observation, which of the following statements are likely true: *

- [ ] Bigger malignant cells tend to be more concave than smaller malignant cells
- [ ] Bigger malignant cells tend to have lower compactness than smaller malignant cells
- [ ] Malignant cells with a bigger perimeter tend to have a bigger area
- [ ] None of the above options are likely true

Dataset 3 of 3: Spam E-mails

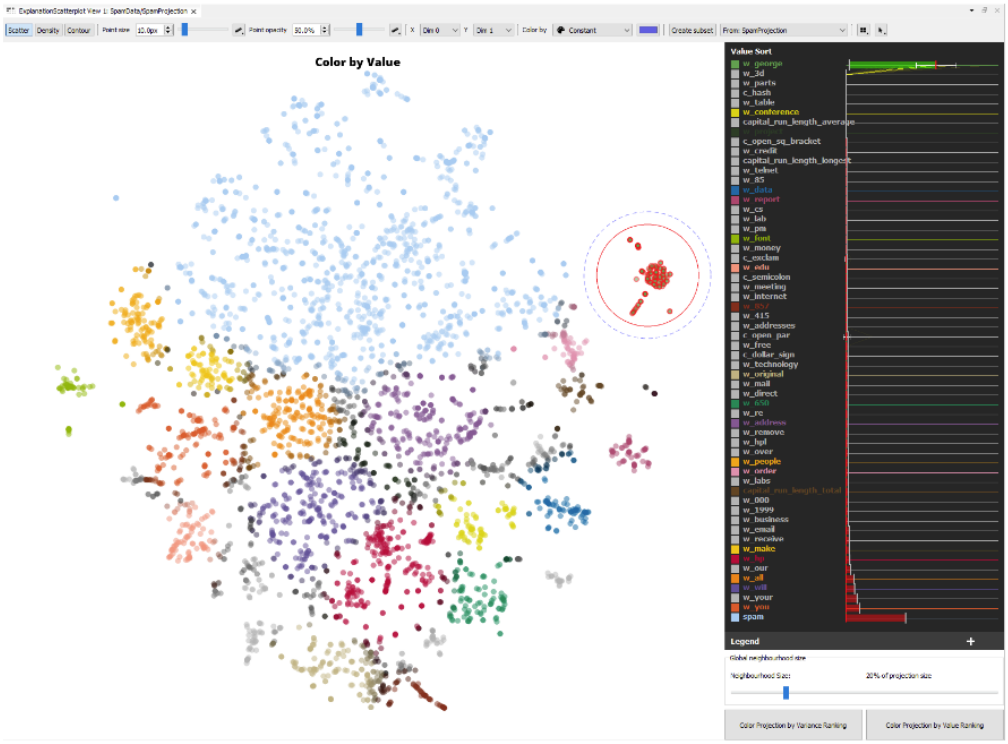
This dataset examines roughly 4600 e-mails, of which a number have been classified as spam (an unsolicited commercial e-mail). For each e-mail, the frequencies of 48 words and 6 characters have been counted (higher numbers mean the word or character occurred more often in the mail), as well as the run length of capital letters in the mail (how many sequences of capital letters there are on average, the longest run, and the total length of all sequences). Therefore, in total for each e-mail 57 attributes are computed.

In each of the following 4 questions, we will show you a screenshot of our tool containing a projection of this dataset and ask you to explain a selected region in the projection based on information in the screenshot.

Following this, we will ask you to answer 3 questions by exploring the projection in the actual software tool.

Spam: Question 1 of 4

In the following image, the projection is in _Color by Value_ mode, meaning the attribute with the highest relative value is used for the coloring. Several points are selected using the selection circle (points in the red circle). The values of the attributes are shown on the right.



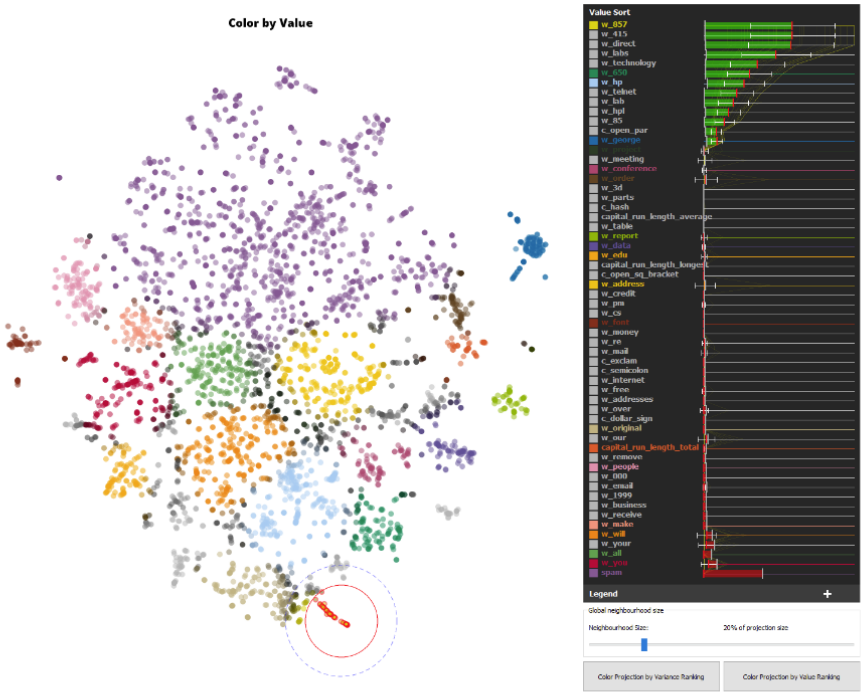
Which of the following statements is likely true: *

- ◯ These mails are primarily about money
- ◯ These mails primarily mention a george
- ◯ These mails are written with a lot of capital letters
- ◯ Nothing can be said about these mails

Spam: Question 2 of 4

The mails in this dataset have been taken from work and personal e-mails at HP Labs. Several of these mails have some combination of the company name, address and phone number in them. Their phone number is (415) 857-4144. Both the parts 415 and 857 have been used as attributes in this dataset (w_415 and w_857 respectively).

In the following image, several points are selected using the selection circle (points in the red circle). The values of the attributes are shown on the right.



For the selected points, what can be said about the variances of the attributes called 415 and 857? *

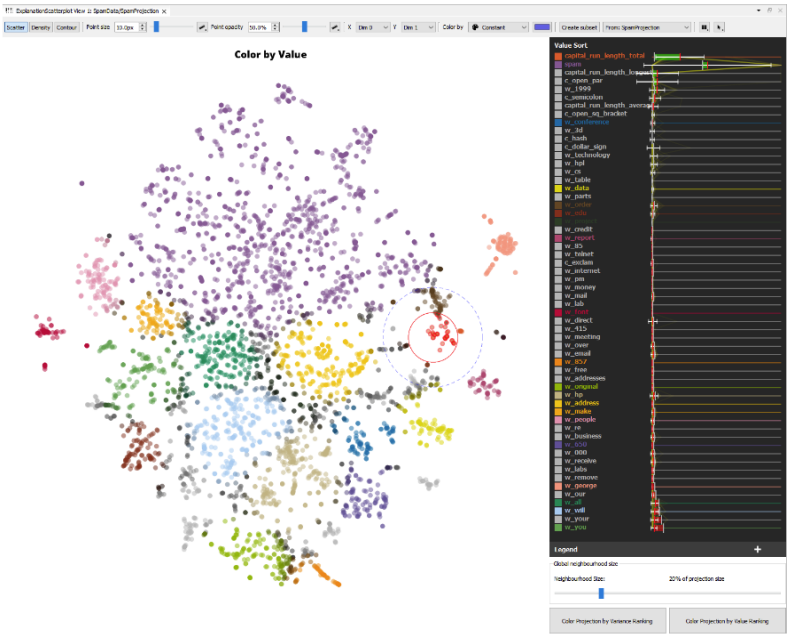
- ◯ Their variances appear to be the same
- ◯ The variance in the word frequency of attribute 415 is larger than of 857
- ◯ The variance in the word frequency of attribute 857 is larger than of 415
- ◯ Nothing can be said about their variances

Spam: Question 3 of 4

As a short review of the PCP plot lines (yellow lines in the widget with the dark background on the right):
The horizontal range lines next to each dimension define the full range of values a dimension takes on, where the left endpoint of the line is its minimum value, and the right endpoint its maximum. Each of the yellow lines represents a data point and where it intersects a range line represents the value of that dimension for that data point.

Apart from the attributes in the dataset mentioned already, there is one additional dimension called 'spam'. This is a classification of whether the e-mail is considered spam (value is 1) or not (value is 0).

In the following image, several points are selected using the selection circle (points in the red circle). The values of the attributes are shown on the right. It seems the selected mails have an extraordinarily high amount of capital letters (see big green bar next to capital_run_length_total).

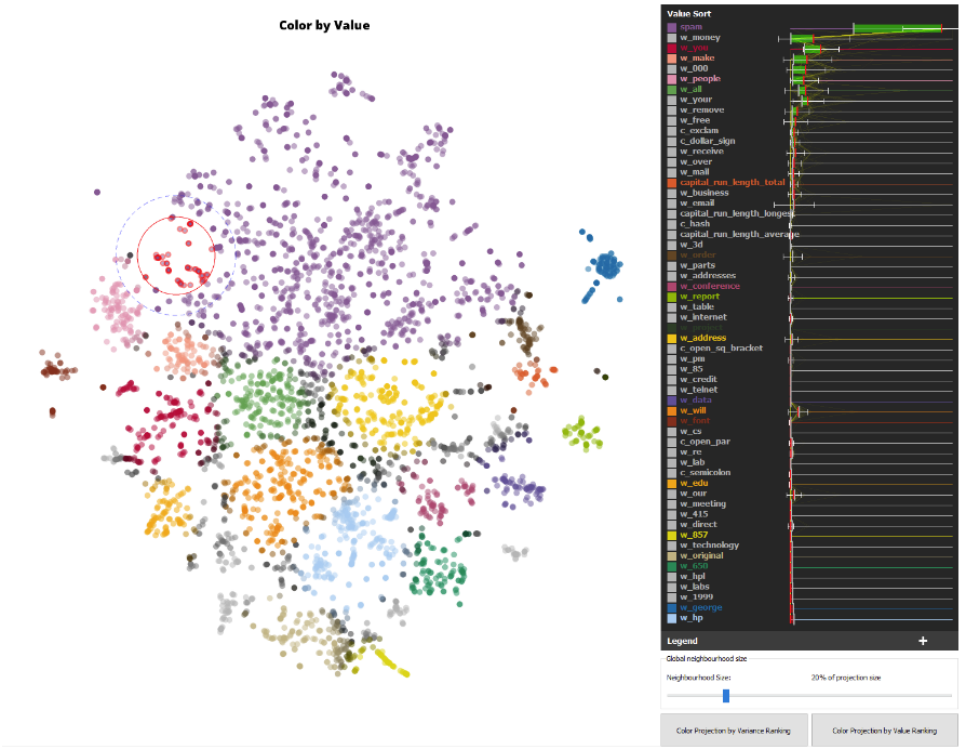


What can be said about whether these mails are spam? *

- ◯ All of these mails are spam
- ◯ None of these mails are spam
- ◯ Some of these mails are spam, and some are not
- ◯ Nothing can be said about this

Spam: Question 4 of 4

In the following image, several points are selected using the selection circle (points in the red circle). The values of the attributes are shown on the right. Most of the selected e-mails appear to be spam (see big green bar next to spam).



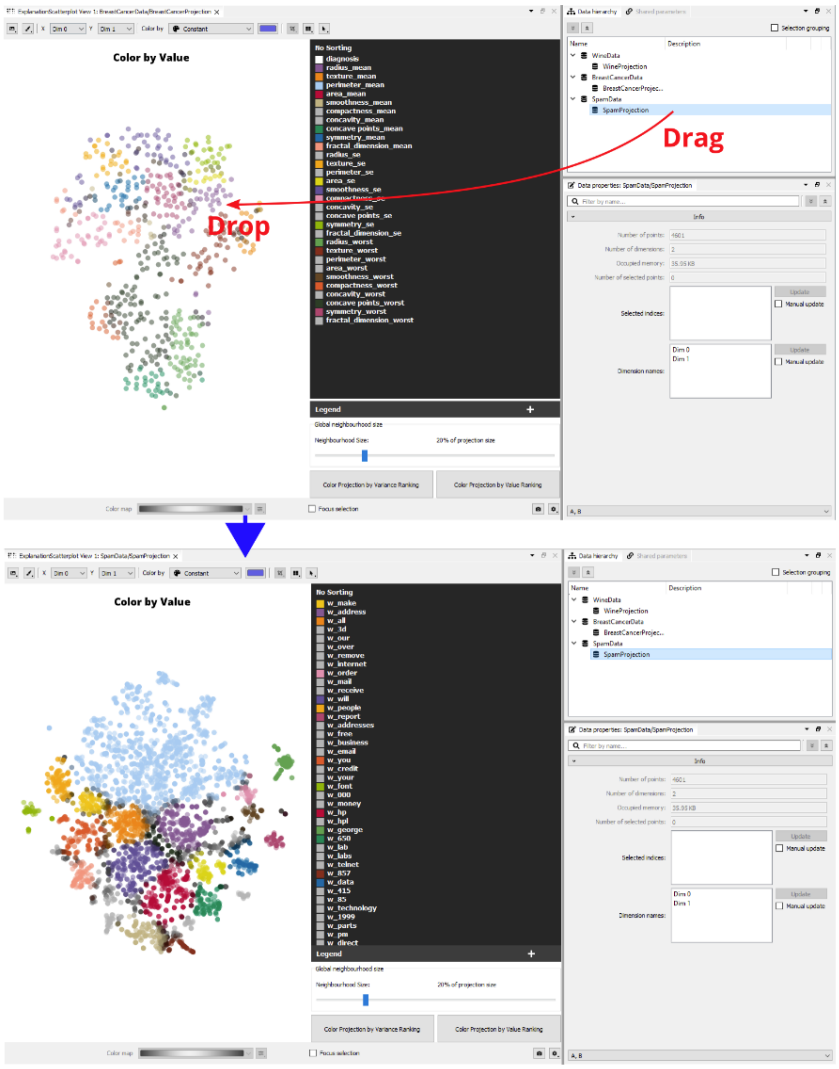
Looking at the other attributes, which of the following topics are likely the content of these spam e-mails? *

- ◯ These e-mails are about making money
- ◯ These e-mails are advertising a product
- ◯ These e-mails are about improving credit scores
- ◯ None of these topics are likely the content of the e-mails

Spam: Live Exploration

For the next 3 questions, we will ask you to perform your own exploration of the data using the tool.

First, visualize the SpamProjection dataset by dragging it from the right into the central view. This process is shown in the image below:



If the projection looks the same in the software tool running on your PC as in the image, please continue to the next section. Note that the tool is in 'Color by Value' mode in these images, the colors in your projection may be different if you are not in this mode.

Spam: Live Exploration, Question 1 of 3

Put the tool in _Color by Value_ mode (by clicking the 'Color Projection by Value' button).

There is a big cluster of similarly colored points whose color corresponds to the 'spam' dimension. These are points that have been classified as spam (the value of the spam dimension is 1).

Moving the selection circle over these points classified as spam, dimensions with a green bar correspond to words that occur more often than usual among spam e-mails.

For e-mails _not_ classified as spam, find out and tick below which words occur more often than usual: [Max 3] *

- [ ] Receive
- [ ] George
- [ ] Money
- [ ] Addresses
- [ ] Conference
- [ ] Data

Spam: Live Exploration, Question 2 of 3

Making optional use of the disabling/enabling of the spam dimension, answer the following question.

Find out and tick below which words occur more often than usual among e-mails that _are_ classified as spam: [Max 3] *

- [ ] Edu
- [ ] Remove
- [ ] Money
- [ ] Receive
- [ ] Original
- [ ] Report

Spam: Live Exploration, Question 3 of 3

Take some time to find the regions in the projection where the word 'will' has the highest frequencies for both spam and non-spam e-mails. Using the differential analysis tool (Holding Ctrl and dragging the selection circle to another place) answer the following question.

For the e-mails with the highest frequency of the word 'will', which attributes make a difference between the mail being classified as spam or not. *

- [ ] Your
- [ ] Data
- [ ] George
- [ ] Receive
- [ ] Parts
- [ ] Business
- [ ] Remove

Questionnaire

In this section we will ask you several questions regarding some of the elements of the tool you have used to answers questions about the dataset, as well as several open questions about the evaluation process.

What is your opinion on the Variance Ranking mode? *

- [ ] It helps me find important dimensions
- [ ] It helps me identify clusters to explore
- [ ] It provides no additional value
- [ ] Other: _____

What is your opinion on the Value Ranking mode? *

- [ ] It helps me find important dimensions
- [ ] It helps me identify clusters to explore
- [ ] It helps me identify which values are extraordinarily high
- [ ] It helps me identify which values are extraordinarily low
- [ ] The green/red bars are confusing to understand
- [ ] The standard deviation bars are confusing to understand
- [ ] It provides no additional value
- [ ] Other: _____

What is your opinion on the parallel coordinates plot (PCP) (yellow lines)? *

- [ ] It helps me understand the distribution of attribute values for the selected observations
- [ ] It provides additional explanatory value
- [ ] It provides no additional value
- [ ] It makes the local analysis widget more confusing
- [ ] Other: _____

How useful did you consider the Variance Ranking mode? *

	1	2	3	4	5	6	7	
Not Very Useful	○	○	○	○	○	○	○	Very Useful

How useful did you consider the Value Ranking mode? *

	1	2	3	4	5	6	7	
Not Very Useful	○	○	○	○	○	○	○	Very Useful

How useful did you consider the differential analysis? *

	1	2	3	4	5	6	7	
Not Very Useful	○	○	○	○	○	○	○	Very Useful

How useful did you consider enabling/disabling of dimensions? *

	1	2	3	4	5	6	7	
Not Very Useful	○	○	○	○	○	○	○	Very Useful

How many years of experience do you have with multi-dimensional data analysis and projections? *

- ○ No experience
- ○ Less than two years
- ○ Two to five years
- ○ More than five years

If you have any comments, questions, suggestions for the tool, please put them here:

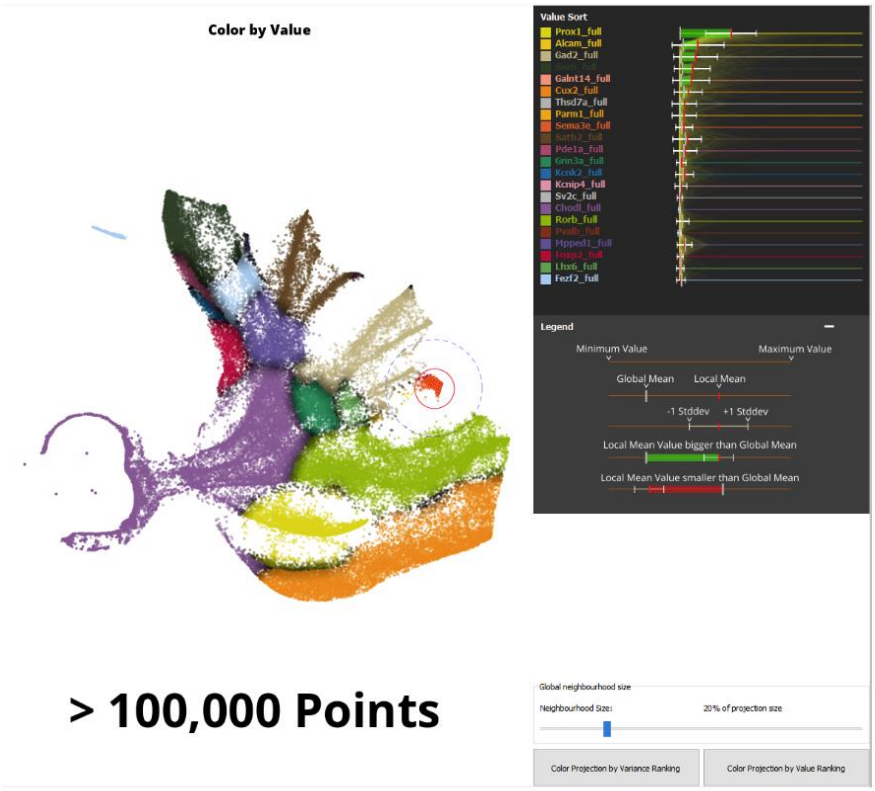
Your answer

If you have any comments about this evaluation, please put them here:

Your answer

Thank you!

Thank you very much for taking the time to participate in this study! We hope that you enjoyed exploring the datasets. As a bonus dataset and a teaser of the potential, below is a picture of a large biological dataset which can be interactively explored in real-time with the tool.



If you would like to receive additional updates on the software and/or explanatory techniques, feel free to leave your e-mail below:

Your answer

Please press the Submit button to submit all your answers and feedback. Thank you again for helping us to improve our developments and feel free to send me a mail at julianthijssen@gmail.com if you have any further questions!

Back Submit Page 46 of 46 Clear form

APPENDIX B

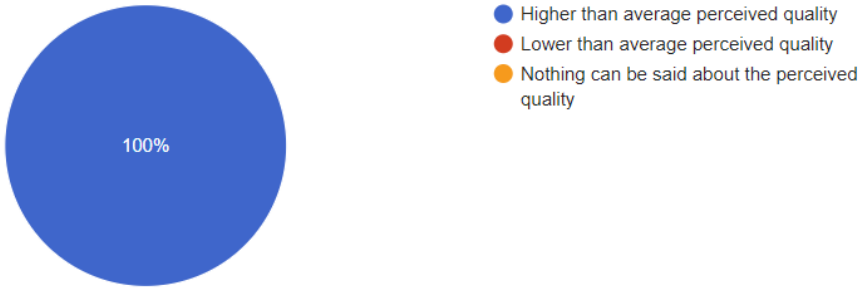
B

Wines: Question 1 of 4

For the selected wines, a high alcohol percentage seems to be associated with a:

Copy

23 responses

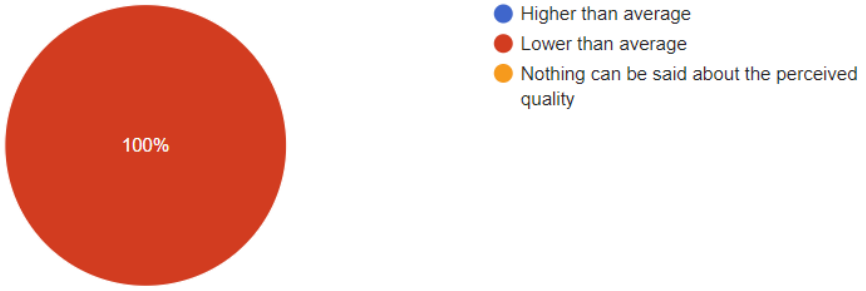


Wines: Question 2 of 4

What can be said about the perceived quality of these wines? The perceived quality of these wines is:

Copy

23 responses

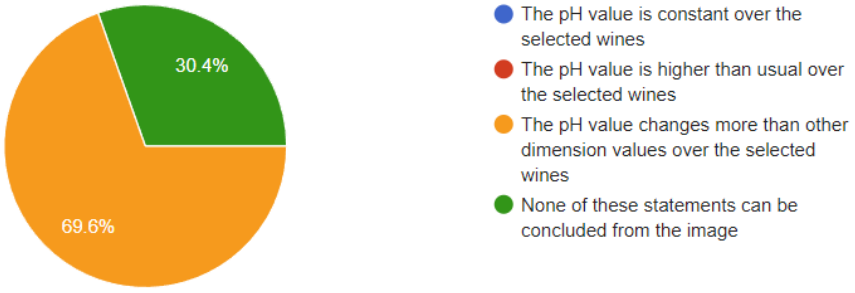


Wines: Question 3 of 4

What can be said about the pH value for the selected wines?

Copy

23 responses

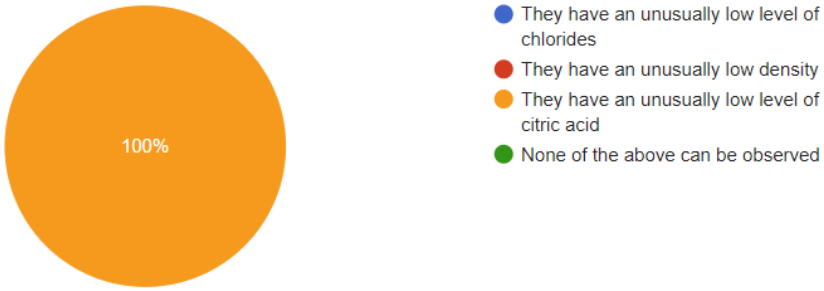


Wines: Question 4 of 4

What else can be observed about these wines?

Copy

23 responses

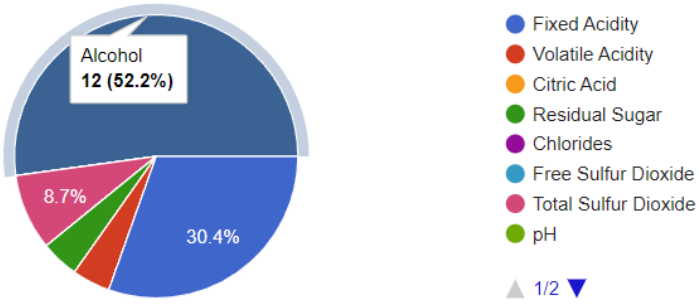


Wines: Live Exploration, Question 1 of 3

For these wines with an extraordinarily low density, tick below which attribute is similarly out of proportion and likely the cause of the low density.

Copy

23 responses

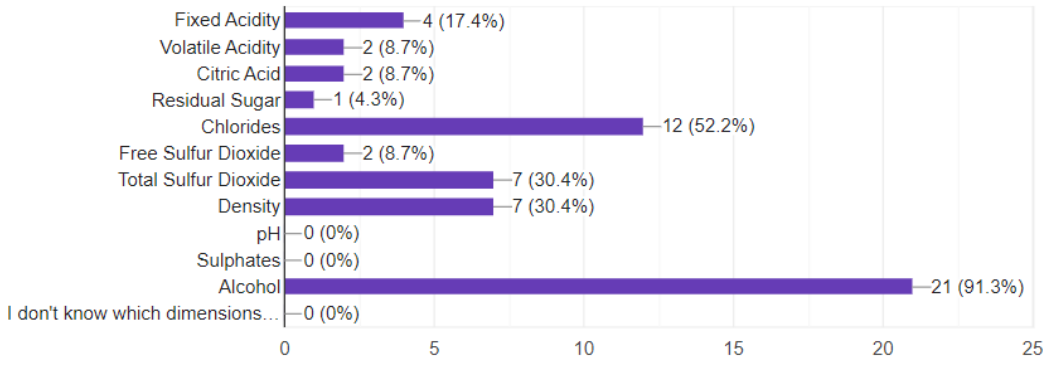


Wines: Live Exploration, Question 2 of 3

Find the highest quality wines in the plot and tick below which attributes, if any, appear to be the most important to the quality of these wines [Max 3]:

Copy

23 responses

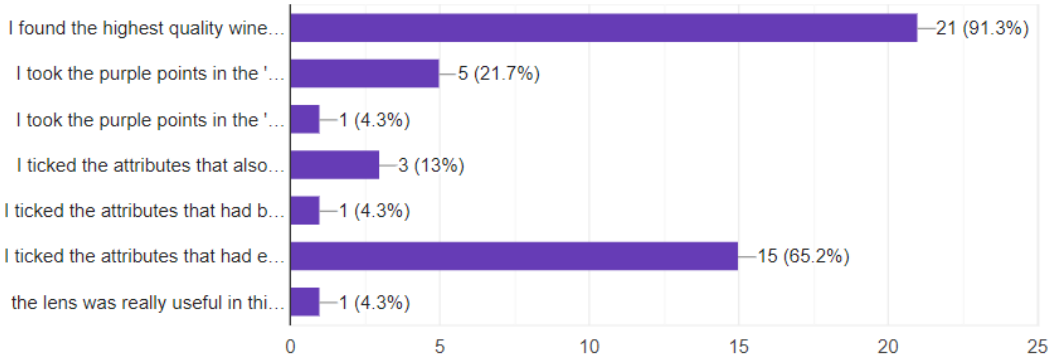


Wines: Live Exploration, Question 2 of 3 Follow-up

Please tick below, which methods you used to arrive at your answer

Copy

23 responses



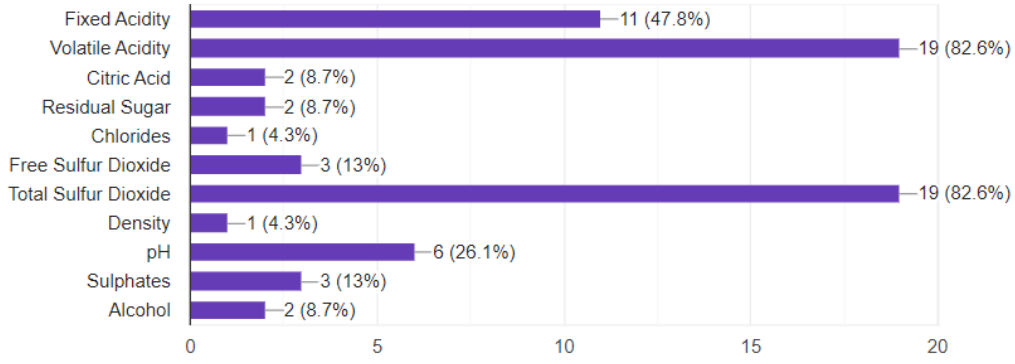
Differential Analysis

Wines: Live Exploration, Question 3 of 3

Using the differential analysis tool, find out which attributes are most different between the red and white wines. [Max 4]

Copy

23 responses



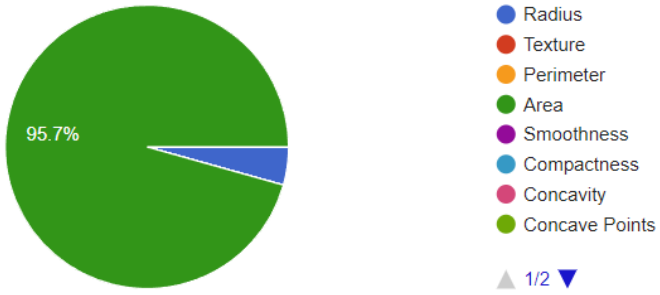
Dataset 2 of 3: Breast Cancer

Breast Cancer: Question 1 of 4

For the selected points which attribute has the lowest variance?

Copy

23 responses

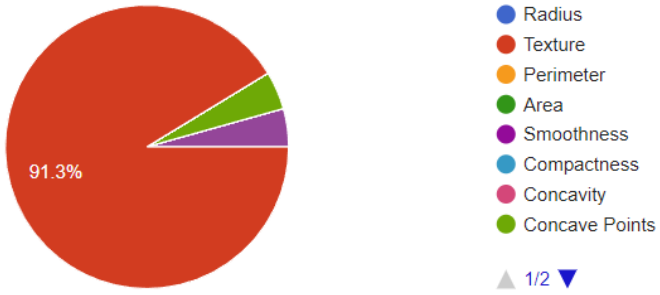


Breast Cancer: Question 2 of 4

For the selected points, which of the attributes shows an indication of malignancy?

Copy

23 responses



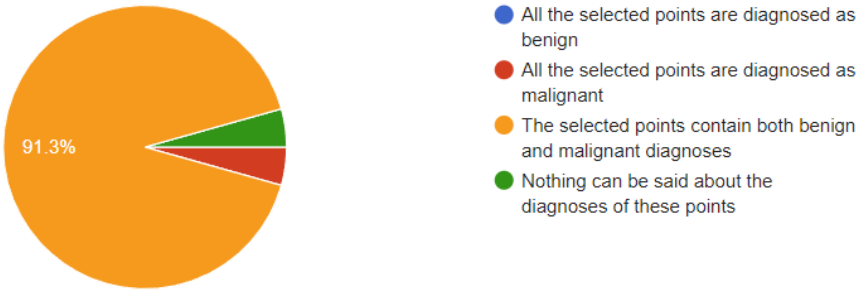
Parallel Coordinate Plot (PCP) Lines

Breast Cancer: Question 3 of 4

What can be said about the diagnoses of these selected points?

Copy

23 responses

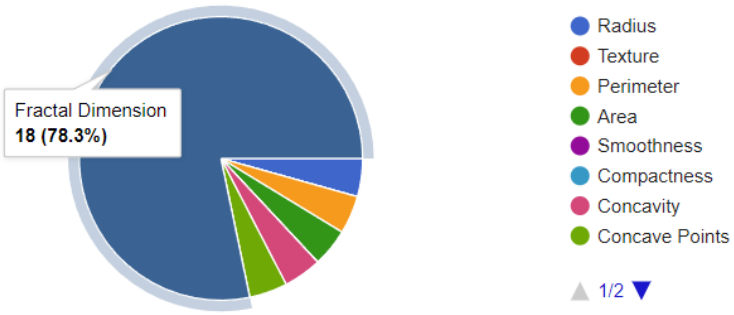


Breast Cancer: Question 4 of 4

Which of the attributes, however, indicate the selected points are benign?

Copy

23 responses

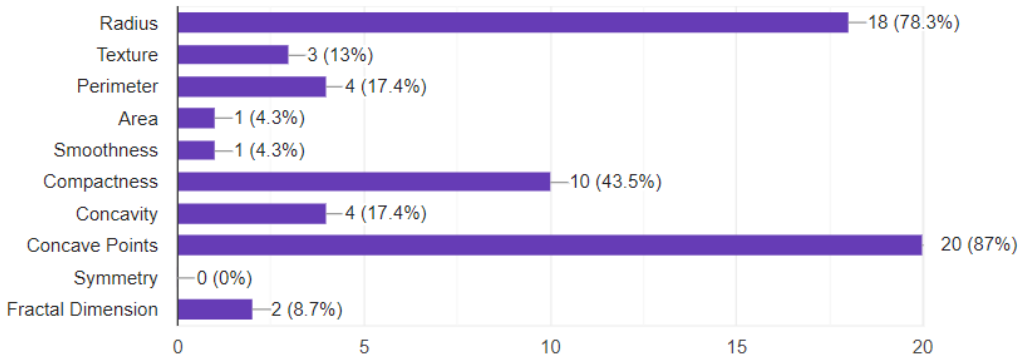


Breast Cancer: Live Exploration, Question 1 of 3

Find these hidden subclusters, and select which attributes characterize them due to their high relative values: [Max 3]

Copy

23 responses



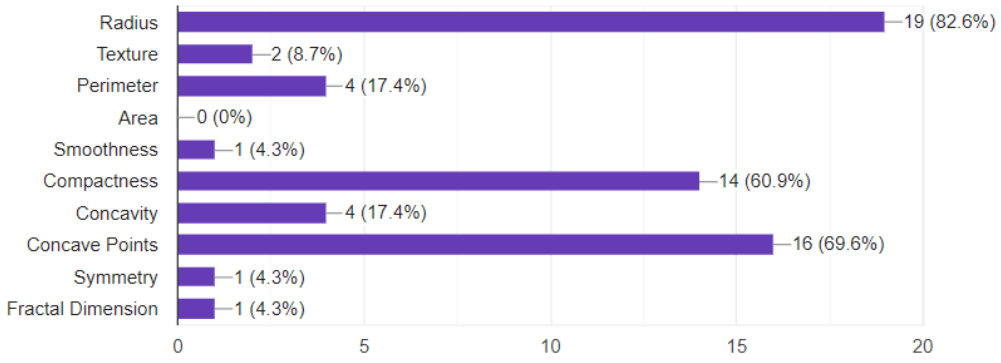
Disabling and Enabling Dimensions

Breast Cancer: Live Exploration, Question 2 of 3

For the malignant subclusters, select which attributes characterize them due to their high relative values: [Max 3]

Copy

23 responses

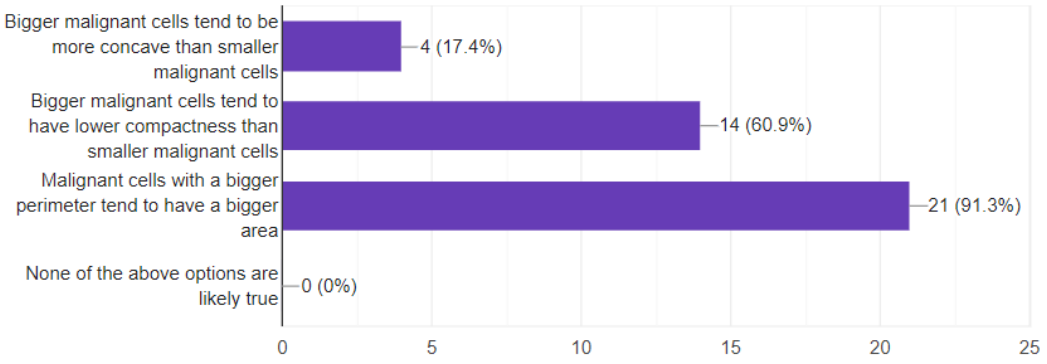


Breast Cancer: Live Exploration, Question 3 of 3

Looking at the points with a malignant observation, which of the following statements are likely true:

Copy

23 responses



Spam: Question 1 of 4

Which of the following statements is likely true:

Copy

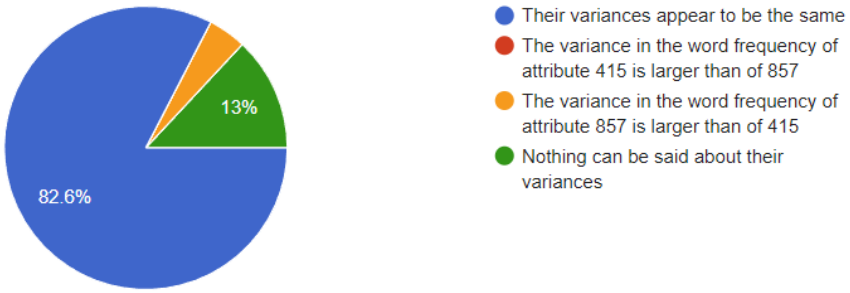
23 responses

Spam: Question 2 of 4

For the selected points, what can be said about the variances of the attributes called 415 and 857?

Copy

23 responses

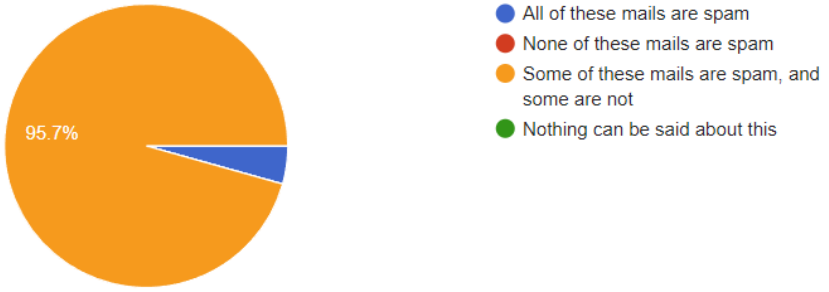


Spam: Question 3 of 4

What can be said about whether these mails are spam?

Copy

23 responses

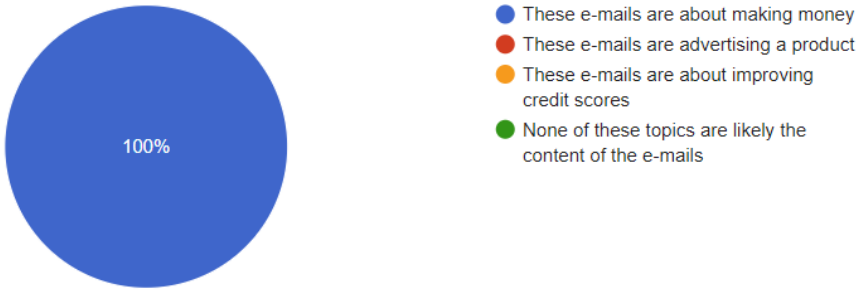


Spam: Question 4 of 4

Looking at the other attributes, which of the following topics are likely the content of these spam e-mails?

Copy

23 responses

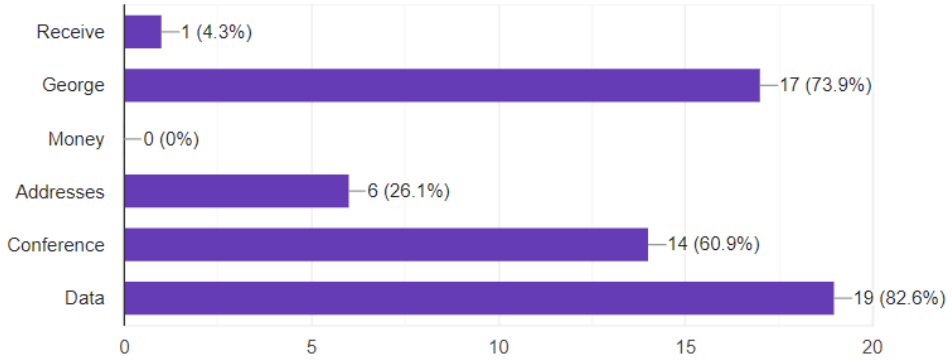


Spam: Live Exploration, Question 1 of 3

For e-mails _not_ classified as spam, find out and tick below which words occur more often than usual: [Max 3]

Copy

23 responses

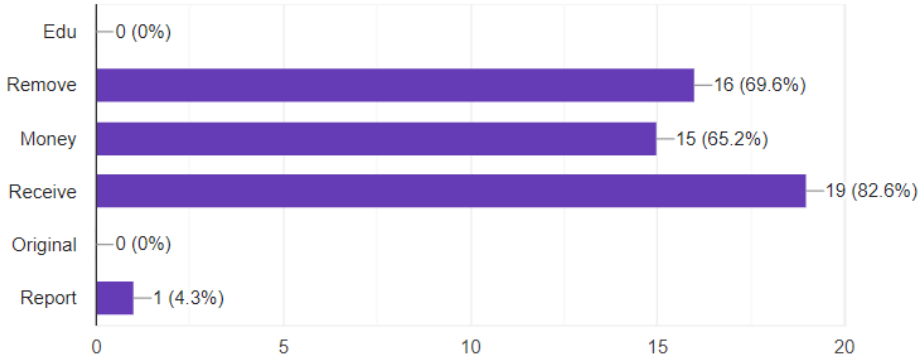


Spam: Live Exploration, Question 2 of 3

Find out and tick below which words occur more often than usual among e-mails that _are_ classified as spam: [Max 3]

Copy

23 responses

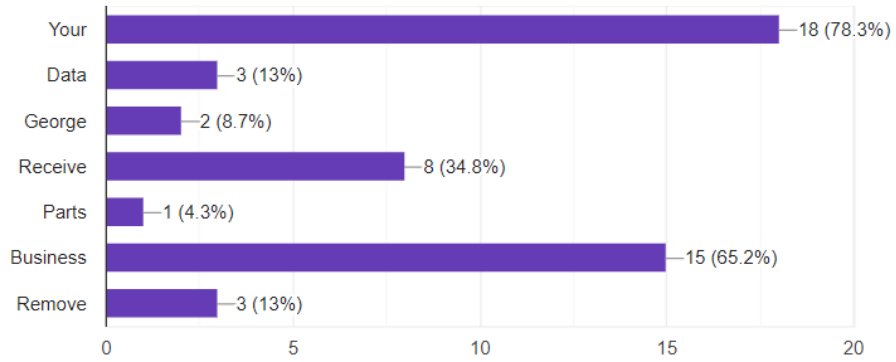


Spam: Live Exploration, Question 3 of 3

For the e-mails with the highest frequency of the word 'will', which attributes make a difference between the mail being classified as spam or not.

Copy

23 responses

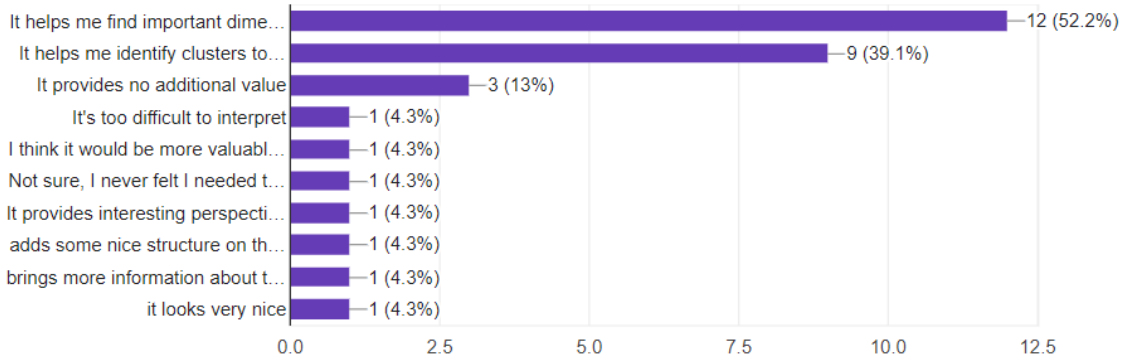


Questionnaire

What is your opinion on the Variance Ranking mode?

Copy

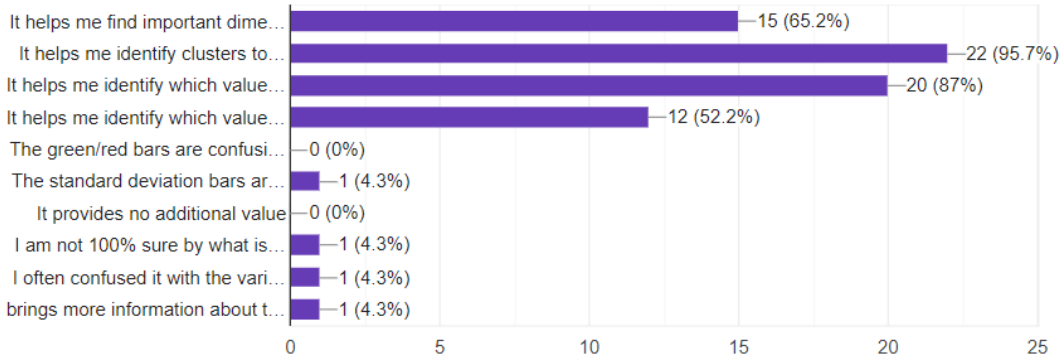
23 responses



What is your opinion on the Value Ranking mode?

Copy

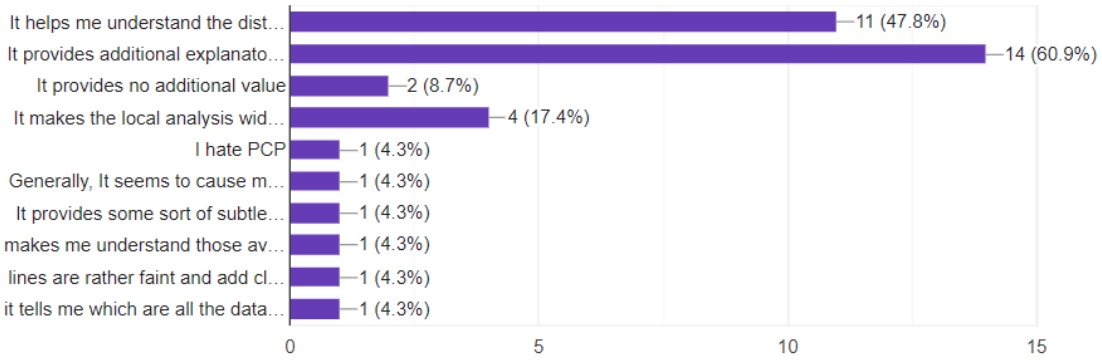
23 responses



What is your opinion on the parallel coordinates plot (PCP) (yellow lines)?

Copy

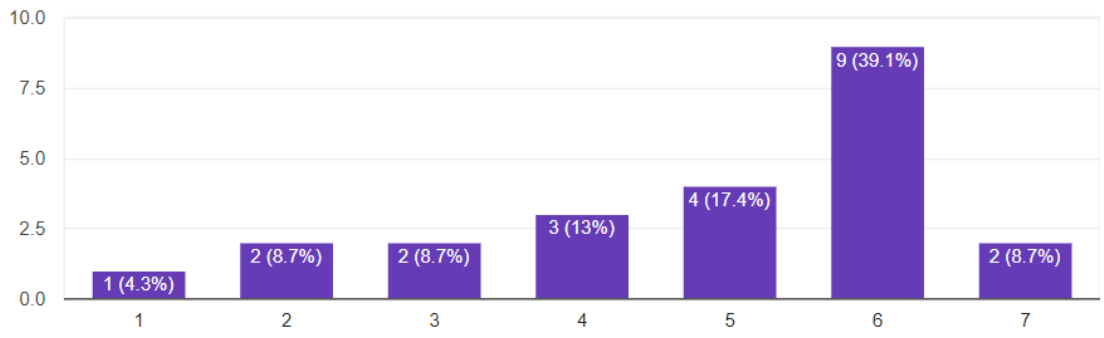
23 responses



How useful did you consider the Variance Ranking mode?

Copy

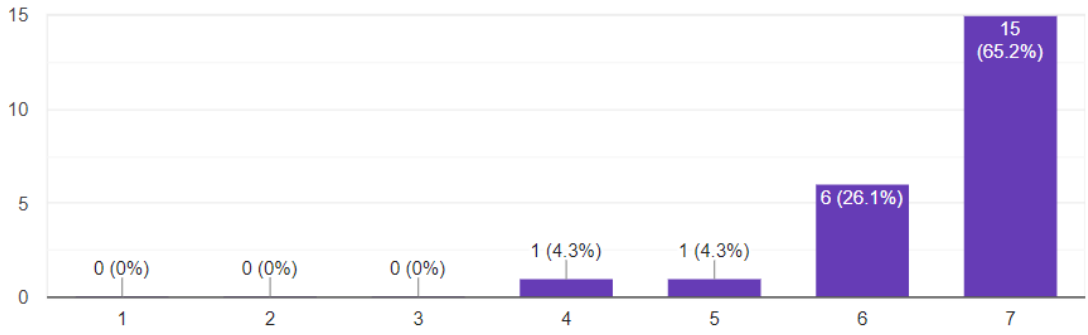
23 responses



How useful did you consider the Value Ranking mode?

Copy

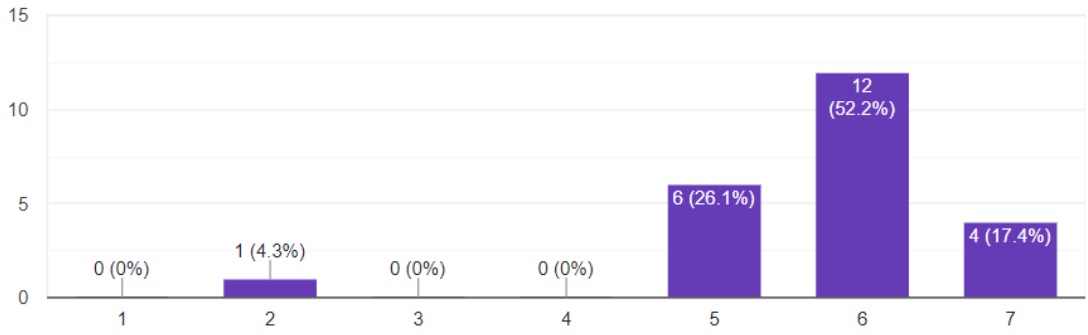
23 responses



How useful did you consider the differential analysis?

Copy

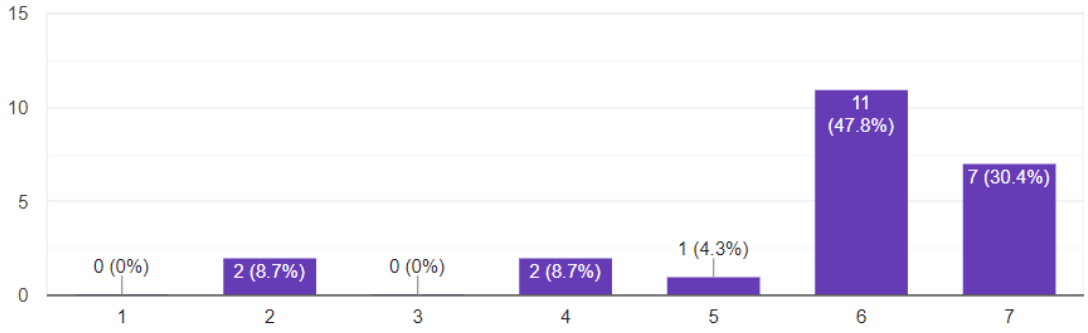
23 responses



How useful did you consider enabling/disabling of dimensions?

Copy

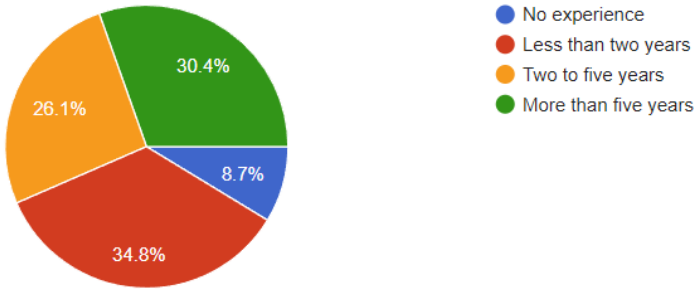
23 responses



How many years of experience do you have with multi-dimensional data analysis and projections?

Copy

23 responses



If you have any comments, questions, suggestions for the tool, please put them here:

7 responses

Making multiple dimensions visualized is super helpful, but it is confusing when choosing the size of the circle (should it be bigger or smaller?). It might be different when more dots are selected in the circle and the important/value rank will also change accordingly.

I the differential mode, the color difference (dark red/light red) is a bit hard to read. Generally, red seems to be a bad choice, since it is also used in the bars.
Some labels are hard to read (e.g., dark greens)
I was a bit confused in the Breast Cancer data set that we generally assigned large values to malign (or benign, I forgot). I would be looking for differentiation in either direction.
I think a mode that has a fixed sorting would be useful to follow easier the change of values in a certain channel. I was extremely confused in the tutorial that the chart always looked the same no matter which side of the cube I sampled, until I realized that the order changed.

无法用鼠标上下滑动右上角的attributes：有时候点一下某个attribute，它跑走了，但不知道怎么弄回来。似乎在最底部，但是当attributes太多的时候，把legend最小化也无法找到这个。

With the number of dimension in the last dataset it was very hard to consistently be able to read the variance sort on the right.

Amazing work

FYI, first of all, I used the tool on VM.

I would say sometimes it was hard to see dark red bars (old local means) on the top-right widget because of its black background. It was also the case for some attributes.

I can understand the constraint on the size of the black background widget. But, the text size could have been a little bit bigger. Or maybe, when I hover on a text, the application might have magnified it.

Finally, I did not understand how to leverage variance ranking mode while I can observe the variance from the white range on the top-right widget.

Very nifty tool!

If you have any comments about this evaluation, please put them here:

6 responses

great job

Good luck with the research and tool development.

In the last question (will and connection to spam) I could not find points with high values for will classified as spam. I was not allowed to check nothing so I checked all.

不太确定我懂了projection的过程和原理，特别是color by variance的。

Nice evaluation!

Some images were hard to read in standard zoom

BIBLIOGRAPHY

[1] El-Ad Amir, Kara Davis, Michelle Tadmor, Erin Simonds, Jacob Levine, Sean Bendall, Daniel Shenfeld, Smita Krishnaswamy, Garry Nolan, and Dana Pe'er. "ViSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia." In: *Nature biotechnology* 31 (May 2013). DOI: 10.1038/nbt.2594.

[2] Charles Anderson. "The end of theory: The data deluge makes the scientific method obsolete." In: 2008.

[3] Richard A. Becker, William S. Cleveland, and Ming-Jen Shyu. "The Visual Design and Control of Trellis Display." In: *Journal of Computational and Graphical Statistics* 5.2 (1996), pp. 123–155. DOI: 10.1080/10618600.1996.10474701. eprint: https://www.tandfonline.com/doi/pdf/10.1080/10618600.1996.10474701. URL: https://www.tandfonline.com/doi/abs/10.1080/10618600.1996.10474701.

[4] Bertjan Broeksema, Alexandru C Telea, and Thomas Baudel. "Visual Analysis of Multi-Dimensional Categorical Data Sets." In: *Computer Graphics Forum*. Vol. 32. 8. Wiley Online Library. 2013, pp. 158–169.

[5] Paulo Cortez, Juliana Teixeira, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. "Using Data Mining for Wine Quality Assessment." In: Oct. 2009, pp. 66–79. ISBN: 978-3-642-04746-6. DOI: 10.1007/978-3-642-04747-3_8.

[6] Renato RO da Silva, Paulo E Rauber, Rafael Messias Martins, Rosane Minghim, and Alexandru C Telea. "Attribute-based Visual Explanation of Multidimensional Projections." In: *EuroVA@EuroVis*. 2015, pp. 31–35.

[7] Division of Image Processing at LUMC & Computer Graphics and Visualisation Group at TU Delft. *HDPS*. Version 0.3.0. Jan. 31, 2022. URL: https://github.com/hdps.

[8] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[9] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[10] Alex Endert, Patrick Fiaux, and Chris North. "Semantic interaction for visual text analytics." In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2012, pp. 473–482.

[11] Karl Ruben Gabriel. "The biplot graphic display of matrices with application to principal component analysis." In: *Biometrika* 58.3 (1971), pp. 453–467.

[12] Paulo Joia, Danilo Coimbra, Jose A Cuminato, Fernando V Paulovich, and Luis G Nonato. "Local affine multidimensional projection." In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2563–2571.

[13] Julian Thijssen. *ProjectionExplorer*. Version 1.0. July 10, 2022. URL: https://github.com/JulianThijssen/ProjectionExplorer.

[14] Kenneth L Kelly. "Twenty-two colors of maximum contrast." In: *Color Engineering* 3.26 (1965), pp. 26–27.

[15] Stephan Kudyba and Stephan Kudyba. *Big data, mining, and analytics*. Auerbach Publications Boca Raton, 2014, pp. 1–3.

[16] John A Lee and Michel Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007, p. 22.

[17] John A. Lee and Michel Verleysen. "Quality assessment of dimensionality reduction: Rank-based criteria." In: *Neurocomputing* 72.7 (2009). Advances in Machine Learning and Computational Intelligence, pp. 1431–1443. ISSN: 0925-2312.

[18] MacroFocus. *High-D*. Version 2019.8.1. Jan. 31, 2022. URL: https://www.high-d.com/.

[19] Rafael Messias Martins, Danilo Barbosa Coimbra, Rosane Minghim, and A.C. Telea. "Visual analysis of dimensionality reduction quality for parameterized projections." In: *Computers and Graphics* 41 (2014), pp. 26–42. ISSN: 0097-8493.

[20] Microsoft Corporation. *Microsoft Excel*. Version 2112 (Build 16.0.14729.20254). Jan. 31, 2022. URL: https://office.microsoft.com/excel.

[21] Bassam Mokbel, Wouter Lueks, Andrej Gisbrecht, and Barbara Hammer. "Visualizing the quality of dimensionality reduction." In: *Neurocomputing* 112 (2013), pp. 109–123.

[22] Luis Gustavo Nonato and Michael Aupetit. "Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment." In: *IEEE Transactions on Visualization and Computer Graphics* 25.8 (2018), pp. 2650–2673.

[23] Charles Nyce. "Predictive Analytics White Paper." In: (Jan. 2007).

[24] Nicola Pezzotti, Julian Thijssen, Alexander Mordvintsev, Thomas Höllt, Baldur van Lew, Boudewijn Lelieveldt, Elmar Eisemann, and Anna Vilanova. "GPGPU Linear Complexity t-SNE Optimization." In: *IEEE Transactions on Visualization and Computer Graphics* PP (Aug. 2019), pp. 1–1. DOI: 10.1109/TVCG.2019.2934307.

[25] Latanya Sweeney. "Information explosion." In: *Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies* (2001), pp. 43–74.

[26] Edward R Tufte, Nora Hillman Goeler, and Richard Benson. *Envisioning information*. Vol. 2. Graphics press Cheshire, CT, 1990.

[27] Laurens Van Der Maaten, Eric Postma, Jaap Van den Herik, et al. "Dimensionality reduction: a comparative." In: *J Mach Learn Res* 10.66-71 (2009), p. 13.

[28] Daan van Driel, Xiaorui Zhai, Zonglin Tian, and Alexandru C Telea. "Enhanced Attribute-Based Explanations of Multidimensional Projections." In: *EuroVA@ Eurographics/EuroVis*. 2020, pp. 37–41.

[29] Yun Zhang, Jeremy Andrew Miller, Jeongbin Park, Boudewijn P Lelieveldt, Brian D Aevermann, Tommaso Biancalani, Charles Comiter, Christoffer Mattsson Langseth, Brian Long, Viktor Petukhov, et al. "Reference-based cell type matching of spatial transcriptomics data." In: *bioRxiv* (2022).