

Practical Assignment: Visualizing Data with Tableau

1. Introduction

This document describes the practical assignment coming together with the Visual Analytics for Big Data course. The aim of this assignment is to put you in the position of actual *designers* and *users* of interactive visual analytics applications. During this step-by-step assignment, you will use Tableau – a state-of-the-art visual analytics framework – to analyze a real-world dataset to answer several questions concerning the real-world data at hand. For this, you will first design and next use several interactive visualizations with Tableau. The techniques to use during this assignment combine the theoretical and practical information provided during the course.

2. Dataset

The provided dataset comes from a real-world context. Recently, the city of Seattle (USA) has committed itself to an open data and transparency policy: The information collected by the various departments of the city during their daily functioning is made publicly available for free to all interested parties. Organizations and citizens can now use this data to query it to find answers to specific questions. In particular, they can create *visualizations* of the data to better understand and communicate about how the city operates.

The entire data collection of the city of Seattle is available at <http://data.seattle.gov>. Data is organized under various categories (e.g. city business, education, permitting, public safety and so on). In each category, several datasets are available. These can be used in three different ways:

Online browsing

Users can select a dataset and browse through the actual data values. In most cases, datasets consist of (large) tables where rows indicate records (observations) and columns indicate the various attributes (dimensions) measured per observation. Below is shown an example of such a dataset, called **Seattle Police Department 911 Incident Response**. The raw data is stored as a table, where every row is an incident reported to the police alarm number 911 and the columns indicate various attributes stored per incident (e.g. type of incident, location in the city, response time of the police, and so on).

The dataset is a very good example of real-world information whose analysis requires interactive visualization techniques. Indeed, the data is

- *large* – the dataset contains 1213480 records (incidents having taken place in the city over a period of several years);
- *multidimensional* – there are 20 dimensions or columns;
- *hybrid* – columns are of various types such as date/time, geolocation, categorical, text, and quantitative;
- *time-dependent* – data records are gathered over a period of several years.

Seattle Police Department 911 Incident Response
This dataset is all the Police responses to 9-1-1 calls within the city. Police response data shows all officers

Find in this Dataset

Manage More Views Filter Visualize Export Discuss Embed

	CAD CDW ID	CAD Event Numbe	General Offense N	Event Clearance C	Event Clearance C	Event Clearance S	Event Clearance C	Hundred Block Loc	District/Sector	Zone/Beat
1	1729117	16000154448	2016154448	245	DISTURBANCE, O	DISTURBANCES	DISTURBANCES	05/04/2016 08:53:0	21XX BLOCK OF 2 D	D1
2	1729125	16000154350	2016154350	170	RECKLESS ENDA	MISCELLANEOUS	MISCELLANEOUS	05/04/2016 08:49:2	90XX BLOCK OF 2 J	J1
3	1729120	16000154403	2016154403	470	PARKING VIOLAT	PARKING VIOLAT	TRAFFIC RELATE	05/04/2016 08:46:2	12XX BLOCK OF H W	W1
4	1729122	16000154384	2016154384	064	SHOPLIFT	THEFT	SHOPLIFTING	05/04/2016 08:46:2	26XX BLOCK OF 3 D	D1
5	1729123	16000154374	2016154374	041	HARASSMENT, TH	THREATS, HARAS	THREATS, HARAS	05/04/2016 08:46:0	1XX BLOCK OF ME Q	Q3
6	1729116	16000154347	2016154347	161	TRESPASS	TRESPASS	TRESPASS	05/04/2016 08:44:1	15XX BLOCK OF N B	B1
7	1729118	16000154444	2016154444	280	SUSPICIOUS PER	SUSPICIOUS CIRC	SUSPICIOUS CIRC	05/04/2016 08:43:5	23XX BLOCK OF S F	F1
8	1729124	16000154369	2016154369	130	PROPERTY DEST	PROPERTY DAMA	PROPERTY DAMA	05/04/2016 08:42:5	80XX BLOCK OF 1 J	J2
9	1729119	16000154408	2016154408	470	PARKING VIOLAT	PARKING VIOLAT	TRAFFIC RELATE	05/04/2016 08:40:5	14XX BLOCK OF 1 E	E2
10	1729121	16000154398	2016154398	245	DISTURBANCE, O	DISTURBANCES	DISTURBANCES	05/04/2016 08:40:3	4 AV / TERRACE S K	K3
11	1729114	16000154396	2016154396	280	SUSPICIOUS PER	SUSPICIOUS CIRC	SUSPICIOUS CIRC	05/04/2016 08:39:5	44XX BLOCK OF L B	B2
12	1729115	16000154391	2016154391	161	TRESPASS	TRESPASS	TRESPASS	05/04/2016 08:32:0	15XX BLOCK OF 9 M	M2
13	1729108	16000154419	2016154419	281	SUSPICIOUS VEH	SUSPICIOUS CIRC	SUSPICIOUS CIRC	05/04/2016 08:30:3	5XX BLOCK OF ME Q	L2
14	1729109	16000154394	2016154394	072	AUTO THEFT AND	AUTO RECOVERIE	AUTO THEFTS	05/04/2016 08:23:2	80XX BLOCK OF 1 J	J2
15	1729112	16000154357	2016154357	220	CRISIS COMPLAIN	CRISIS CALL	BEHAVIORAL HEA	05/04/2016 08:22:3	133XX BLOCK OF L	L1
16	1729101	16000154401	2016154401	280	SUSPICIOUS PER	SUSPICIOUS CIRC	SUSPICIOUS CIRC	05/04/2016 08:21:2	90XX BLOCK OF 1 F	F3
17	1729111	16000154372	2016154372	200	ALACAD - COMME	BURGLARY ALACAD	FALSE ALACAD	05/04/2016 08:20:3	24XX BLOCK OF N B	B1
18	1729113	16000154320	2016154320	372	PROPERTY - FOU	PROPERTY - MISS	PROPERTY - MISS	05/04/2016 08:19:3	44XX BLOCK OF V B	B3
19	1729100	16000154412	2016154412	460	TRAFFIC (MOVING	TRAFFIC RELATE	TRAFFIC RELATE	05/04/2016 08:13:5	6 AV S / S HOLTGA O	O1
20	1729110	16000154382	2016154382	280	SUSPICIOUS PER	SUSPICIOUS CIRC	SUSPICIOUS CIRC	05/04/2016 08:12:2	DENSMORE AV N, N	N3
21	1729095	16000154395	2016154395	280	SUSPICIOUS PER	SUSPICIOUS CIRC	SUSPICIOUS CIRC	05/04/2016 08:08:4	2XX BLOCK OF JA K	K1
22	1729096	16000154390	2016154390	280	SUSPICIOUS PER	SUSPICIOUS CIRC	SUSPICIOUS CIRC	05/04/2016 08:07:2	25XX BLOCK OF 3 D	D1
23	1729102	16000154386	2016154386	470	PARKING VIOLAT	PARKING VIOLAT	TRAFFIC RELATE	05/04/2016 08:05:3	1XX BLOCK OF LA G	G2
24	1729103	16000154376	2016154376	280	SUSPICIOUS PER	SUSPICIOUS CIRC	SUSPICIOUS CIRC	05/04/2016 08:05:2	8XX BLOCK OF NE U	U1
Totals	1213480									

Figure 1: Browsing the raw data of Seattle police incident response

Online visualization

The site data.seattle.gov offers several options to create visualizations of this data. These can be constructed directly in the browser, without having to use any external third-party tool or programming skills.

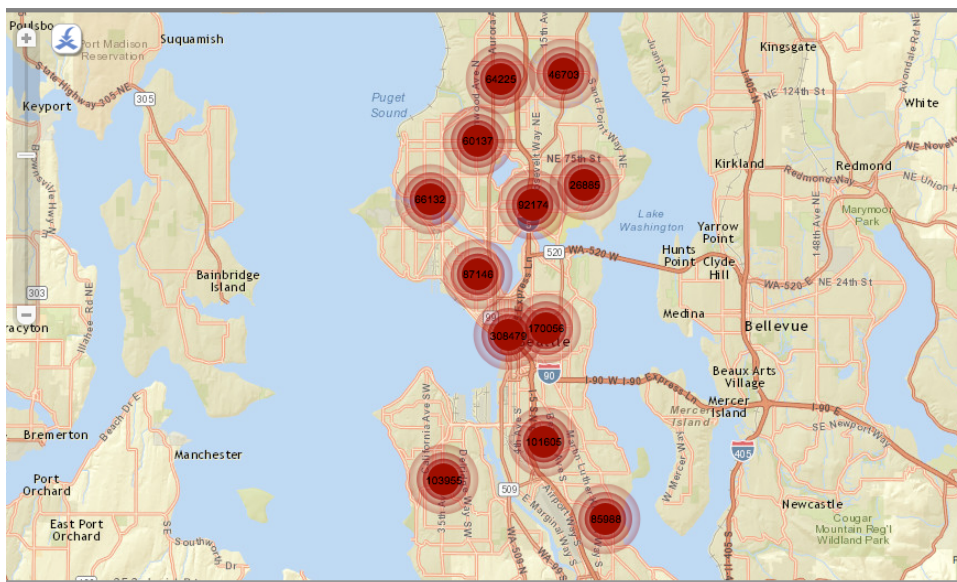


Figure 2: Visualization of incidents density

Figure 2 shows an example of such a visualization. Here, we display the density of incidents on a map of Seattle. Dark red circles with labels indicate the number of incidents per city district. A nice (though quite crude) example of a heat map, or density map!

Offline analysis

While the online options offered by data.seattle.gov help in getting a quick impression of the data at hand, they are limited. Browsing very large raw data tables is not very insightful. The visualization options offered by the site are also limited in terms of flexibility. For large datasets (over roughly 100K records), creating and manipulating visualizations in the browser is also quite slow.

For such situations, it is better to download the data to analyze it *offline*. In this assignment, we will do this using Tableau.

First, download and install Tableau from www.tableau.com. Obtaining a free copy of the tool is very easy – just search for Tableau Public, make an account in the cloud (you will use this to publish your visualizations also), download the tool, and install it. Binaries are provided for the most common platforms, like Windows and Mac OSX.

Next, get the data. Download the aforementioned Seattle dataset and save it as

Seattle_Police_Department_911_Incident_Response.csv

The dataset is roughly 315 MB (can be larger, the file is dynamically updated). It is a CSV (comma separated value) text file containing the raw data table already shown in Figure 1.

Note: An offline copy of this dataset is also made available on the course's website. This offline copy does not offer the most up-to-date version of the dataset. However, do use the offline copy only if you cannot get hold of the live data, and only after having discussed this issue with the course teacher!

3. Loading the data

Use Tableau's 'open text file' option to load the CSV dataset. When doing this, the Data Source view appears showing an image similar to the one in Figure 3. We see here that each record (incident) has several columns (attributes). A short explanation of the most important attributes, including their types, is given next.

Note: As the dataset is quite large, expect some delays (several seconds) when manipulating the data within Tableau, depending on your computer.

#	#	#	#	Abc	Abc	Abc	Seattle_Police_Depart...	Abc	Abc
Seattle_Police_Dep...	Seattle_Police_Depart...	Seattle_Police_Depart...	Seattle_Police_Depart...	Seattle_Police_Depart...	Seattle_Police_Depart...	Seattle_Police_Depart...	Seattle_Police_Depart...	Seattle_Police_Depart...	Seattle_Police_D...
Cad Cdw Id	CAD Event Number	General Offense ...	Event Clearance C...	Event Clearance D...	Event Clearance S...	Event Clearance G...	Event Clearance D...	Hundred Block Location	District/Sector
579,831	12,000,249,934	2,012,249,934	64	null	null	null	null	null	null
2,026,080	14,000,163,328	2,014,163,328	245	null	null	null	null	null	null
1,281,062	13,000,009,701	20,139,701	null	null	null	null	null	11XX BLOCK OF MADISON ST	null
1,267,684	13,000,005,758	20,135,758	null	null	null	null	null	6 AV / UNION ST	null
1,314,495	13,000,051,135	201,351,135	null	null	null	null	null	24XX BLOCK OF 3RD AVE S	null
1,306,309	13,000,040,967	201,340,967	null	null	null	null	null	64XX BLOCK OF CALIFORNIA AVE SW	null
1,662,188	13,000,418,498	2,013,418,498	null	null	null	null	null	96XX BLOCK OF RAINIER AV S	null
1,876,810	14,000,014,151	201,414,151	null	null	null	null	null	101XX BLOCK OF W MARGINAL PL S	99
1,878,225	14,000,016,195	201,416,195	null	null	null	null	null	161XX BLOCK OF 32 AV S	99
1,313,936	13,000,050,450	201,350,450	null	null	null	null	null	103XX BLOCK OF 8TH AVE NW	B
1,845,904	13,000,434,340	2,013,434,340	null	null	null	null	null	105XX BLOCK OF DAYTON AV N	B
1,314,478	13,000,051,031	201,351,031	null	null	null	null	null	107XX BLOCK OF EVANSTON AVE N	B
1,867,095	14,000,000,264	2,014,264	null	null	null	null	null	110XX BLOCK OF GREENWOOD AV N	B
1,291,428	13,000,022,861	201,322,861	null	null	null	null	null	110XX BLOCK OF GREENWOOD AVE N	B
1,859,478	13,000,455,347	2,013,455,347	null	null	null	null	null	115XX BLOCK OF PALATINE AV N	B
1,874,729	14,000,011,460	201,411,460	null	null	null	null	null	117XX BLOCK OF GREENWOOD AV N	B

Figure 3: View of the dataset in Tableau's Data Source

At Scene Time (date & time):

Records the actual moment when the police arrived at the scene of an incident.

District/Sector (text/categorical):

Records the location of the incident in a specific sector of Seattle.

Event Clearance Date (date & time):

Records the actual moment when the incident was declared closed (fully analyzed) by the police.

Event Clearance Description (text/categorical):

Records a short textual description of how the event was classified after closure.

Event Clearance Group (text/categorical):

Records a high-level classification of the type of event clearance (how the event was categorized by the police after it was analyzed). 43 different categories exist here.

Event Clearance SubGroup (text/categorical):

Records a finer-grained classification of the type of event clearance. That is, per type of event clearance group, we have several subgroups of events. For example, disturbances (event clearance group) can be sub-classified in generic and noise-related disturbances (two event clearance subgroups).

Hundred Block Location (text/location):

Records the location of the incident, stored in terms of a block-numbering scheme for the city of Seattle. An example value is: *1XX BLOCK of 22ND AVE NE*.

Incident Location (location):

Records the (latitude,longitude) location of the incident as a pair of quantitative values.

Initial Type Description (text/categorical):

Records a short textual description of how the event was recorded when initially reported. Similar to Event Clearance Description, but referring to the initial moment when the event was reported.

Initial Type Group (text/categorical):

Records a high-level classification of the type of the event when it was initially reported. 41 different categories exist here. Note that these are not the same types of categories as present in the Event Clearance Group (though, the two attributes share some common categories).

Initial Type Subgroup (text/categorical):

Records a finer-grained classification of the type of event as it was reported. That is, per type of Initial Type Group, we have several subgroups of events. This is similar to how we have several Event Clearance Subgroups in the Event Clearance Group categories.

Zone/Beat (text/categorical):

Records the location of the event in terms of the Seattle police zones.

Cad CDW ID (measure):

Records the ID of the reported incident.

CAD Event Number (measure):

Records another type of ID of the reported incident. As far as we can see, this is an unique ID per incident (row in the data table).

Latitude (location):

Records the latitude of the location of the event. Same data as the 1st component of the Incident Location pair.

Longitude (location):

Records the longitude of the location of the event. Same data as the 2nd component of the Incident Location pair.

Number of records (measure):

This data field is automatically synthesized by Tableau and gives the number of records (table rows) having the same values for a given set of columns (attributes). We will see next how to use this field in our visualizations.

4. Creating visualizations

To create visualizations, click the 'New sheet' icon in the bottom toolbar of Tableau. Typically, you will create *one new visualization per sheet*. This is the easiest way.

Important Note: When browsing the data (either in Tableau's Data Source view or the online raw table view of seattle.data.gov), you will notice that several cells are empty. These indicate *missing data*. It is important for several of the assignment steps to filter out missing data values. This can be done by dragging various fields, in a sheet, into the Filters panel. Then, you can configure filters to exclude data records having *Null* values for specific fields (*Null* is the way Tableau indicates that there is missing data for a certain field for certain records). Sometimes, however, it is important to keep the *Null* values, depending on the type of question to answer.

Way of working

You will get next several tasks to complete on the Seattle police incidents dataset. For each task,

- Read the task description (as stated on the next pages). Each task has a number of questions that it should answer. Try to construct yourself one or several visualizations with Tableau that cover this task – that is, answer the task's questions. Next, use these visualizations, and try to answer the questions. Be critical in finding the limitations of your visualizations (e.g. they don't answer the questions fully, or they do so but they are hard to use or explore). If so, refine the design. Save your iterative designs, and explain in which way latter designs improve on the former ones.

Save the visualizations as a Tableau workbook in your Tableau Public online account. Alternatively, if you have Tableau Desktop, you can save the visualizations as a workbook file (*.twbx) on your local computer.

- For each question, describe the answer you have obtained. Document your answer with the description of the visualization construction; the reasons why you used this visualization; a description of the parameter settings and why you preferred these; a description of the results (based on snapshots); and a critique of the strong and weak points of your solution, including how you could improve it, e.g. using different visualization tools discussed during the course.

Task 1: Exploring the incidents' geographical distribution

Aim: Incidents happen at different locations throughout the Seattle area. The aim of this task is to get an idea of how the incidents are distributed over Seattle.

Questions

Q1.1: Are there high variations of incident densities over Seattle? Which are the low-density zones? Which are the high-density zones?

Q1.2: Are there high variations of incident densities over Seattle for specific types of incidents? Which are these types and which are the variations you found?

Task 2: Exploring the incidents' geographical and temporal distribution

Q2.1: Do you see a different spatial distribution of incidents over the different years? If so, which are the differences you found?

Q2.2: Are there zones with a consistent low incident density over all years? Are there zones with a consistent high incident density over all years?

Task 3: Exploring the resolution speed

Q3.1: What is the average resolution speed to an incident? Resolution speed is defined as the difference between the *Event Clearance Date* value (moment when the police closed the file of an incident) and the *At Scene Time* value (moment when the police arrived at the scene of the incident).

Q3.2: Are there certain types of incidents having a much lower resolution speed than others? If so, which are these?

Q3.3: Does the resolution speed depend on the time period (e.g., year, season of the year)?

Task 4: Exploring the incidents' classification

Q4.1: Incidents have two types: the type of the incident at the moment it was reported (Initial Type Group) and the type of the incident after it has been studied and classified (Event Clearance Group). Obviously, there should be some kind of correlation of the latter with the former (for instance, incidents reported as robberies are very likely to be also classified finally as robberies). How strong is this correlation?

Q4.2: How do the total number of incidents break down per incident reported type (Initial Type Group) and incident resolution type (Event Clearance Group)?

Q4.3: How do the total number of incidents break down per incident resolution type (Event Clearance Group) and incident resolution subtype (Event Clearance Subgroup)?

Task 5: Exploring the incidents' temporal distribution

Q5.1: How do the different types of incidents (as represented by Event Clearance Group) vary in number over the different years? Do you see different types of incidents having the same temporal variation pattern over the same year(s)?

Q5.2: How do the different types of reported incidents (as represented by Initial Type Group) vary over the 24 hours of a day, for the entire data collection? Are there certain hours having a higher rate of reported incidents?

Q5.3: How long does it take to clear incidents, as a function of the hour when they were reported? For example, do incidents reported at noon get cleared faster than incidents reported in the middle of the night?