Practical Assignment: Visualizing Data with Tableau

1. Introduction

This document describes the practical assignment coming together with the Practical Information Visualization course. The aim of this assignment is to put yourself in the position of actual designers and users of interactive information visualizations. During this step-by-step assignment, you will use Tableau to analyze a real-world dataset to answer several questions concerning the data at hand. For this, you will design and use several interactive visualizations with Tableau. The techniques to use during this assignment combine the theoretical and practical information provided earlier 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.

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.
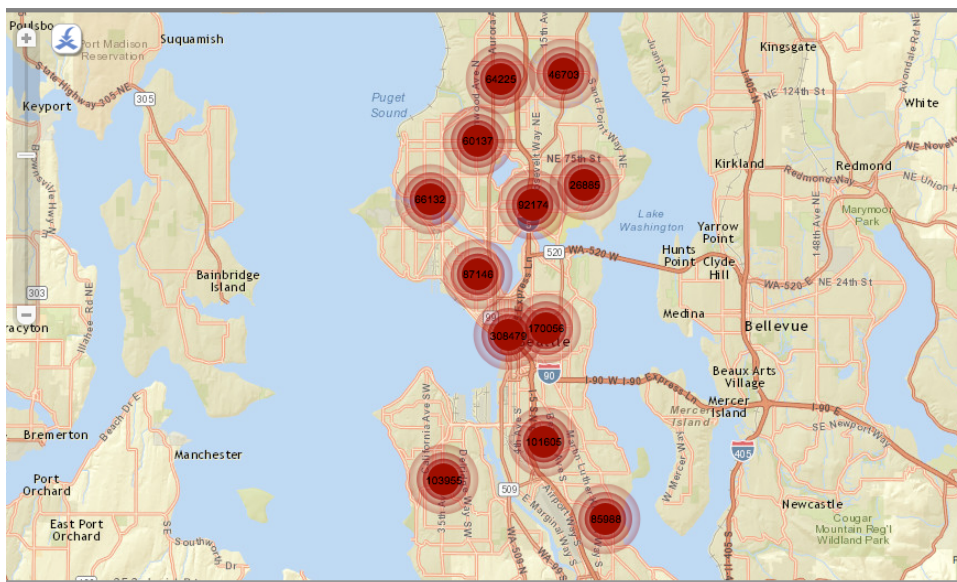

Figure 2: Visualization of incidents density

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.

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. For this, we have already saved a copy of the dataset under the name

Seattle_Police_Department_911_Incident_Response.csv

in the assignment folder. The dataset is roughly 315 MB. It is a CSV (comma separated value) text file containing the raw data table already shown in Figure 1.

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.

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.

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. Each task can be approached at two different levels, as follows.

- *Advanced level:* Read the tasks (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 the visualizations (without looking at the solution), and try to answer the questions.

  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. This can be best done by creating one PowerPoint slide that contains one or a few snapshots of the visualization that answers the question and a few notes that describe what the actual answer is.

- *Easy level:* Read the tasks (as stated on the next pages). The visualizations that cover each task are available online in the lecturer's public Tableau account at

  https://public.tableau.com/profile/alex.telea#!/vizhome/SeattlePoliceDeptIncidents

  Next, use these pre-constructed visualizations to answer the questions for each task. You can use the visualizations directly online via the above link. However, a better solution is to download the entire workbook on your local computer and use Tableau to open it. This is much faster than interacting with the online visualizations. Also, this lets you actually see how the visualizations have been constructed in Tableau.

  A separate document (**Practical Assignment – Construction**) contains snapshots showing how these visualizations have been built in Tableau. Read more below.

  To get the most of this course, you should try to solve the tasks at the advanced level. Only when you do not find your way, use the description of visualization constructions the

assignment-construction document and the online material in Tableau's public server to construct the visualizations and answer the task questions. Repeat the procedure for all tasks.

In detail, proceed as follows:

1.   For each task
2.   Try to construct one or several visualizations that solve it
3.   If you succeed
4.   Use these visualizations to answer the task's questions
5.   Take snapshots of the results and document your findings in PowerPoint
6.   Else, if you did not yet try steps 7 and 8
7.   Read the visualization construction guidelines in the assignment construction document
8.   Continue from step 2
9.   Else (you tried steps 7 and 8 without success)
10.   Download the Tableau worksheet provided online
11.   Proceed from step 4

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?