Visual Analytics for Big Data
Practical Assignment: Visualizing Data with Tableau

Eser Comak
s.e.comak@student.rug.nl
S3432548

Hichem Bouakaz
h.bouakaz@student.rug.nl
S2525763

February 7, 2019

1 Introduction

In this report we will try to answer 5 main tasks and the sub questions that belongs to these tasks. In the first task we will focus on the high vs low crime density zones. In the second task, we will explore the temporal and geographical distributions of crimes. In the third task, we will observe how fast or slow crimes are resolved. In the fourth task, we will inspect to correlation between the count of event clearance group which had the same entry with the initial type group. In the fifth task, we will look at what kind of temporal trends do these different type of crimes display. To access the reported visualizations please visit the team member's Tableau Public profiles.

Eser Comak: `https://public.tableau.com/profile/eser8458#!/`
Hichem Bouakaz: `https://public.tableau.com/profile/hichem5756#!/`

2 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.

2.1 Solution

For a matter of simplicity, all the questions can be answered using this visualization. To access the visualization please visit the link embedded in the screenshot below. On the top left, the map shows the regions in different colours. The barplot at the bottom also shows the same regions. Clicking on any bar(zone) filters out rest of the map

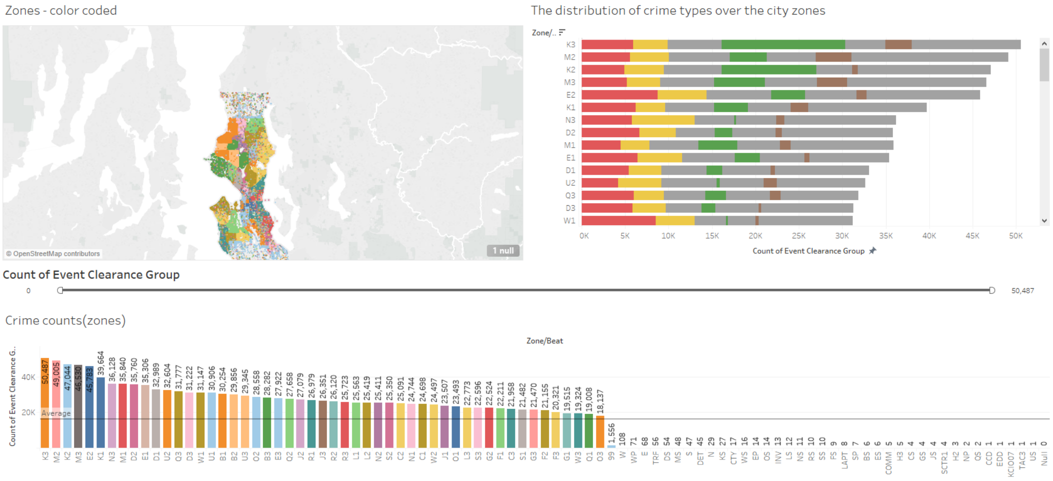


Figure 1: Dashboard for observing high/low crime density zones and incident types.

only the selected zone will be shown on the map when the filter for the crime count is triggered. In order to see high vs low crime rate zones, one can use the sliding filter on top of the bar plots to select the range of the number of crimes. This filtered range will also be applied to the bar plots at the bottom. Finally, on the top right, we have horizontally placed bars showing the different types of crimes for each zone. Using crime count filter will also modify these barplots and only the zones that fall inside the range will be shown.

For the top right bars, red group symbolizes the traffic-related calls. Yellow is for suspicious circumstances. Brown is for shoplifting. Green is for liquor violations. Green seems to have the highest variation among zones.

High Density Zones (top 10 percentile)	Low Density Zones (bottom 40 percentile)	Incident types with most variation across zones(descending)
K3,M2,K2,M3,E2,K1	99,WP,TRF,MS,DET...CCD,US	Liquor violations Traffic related calls Shoplifting Suspicious circumstances

Table 1: High/low density zones and incident types with highest variance between zones.

3 Task 2: Exploring the incidents' geographical and temporal distribution

3.1 Solution 2.1

After exploring the data We could notice that the years 2009 and 2004 has the least amount of records (which we predict it has to do with data missingness). In our analysis we kept data that has null values; however are aware that the years 2009 and 2004 has missing data.

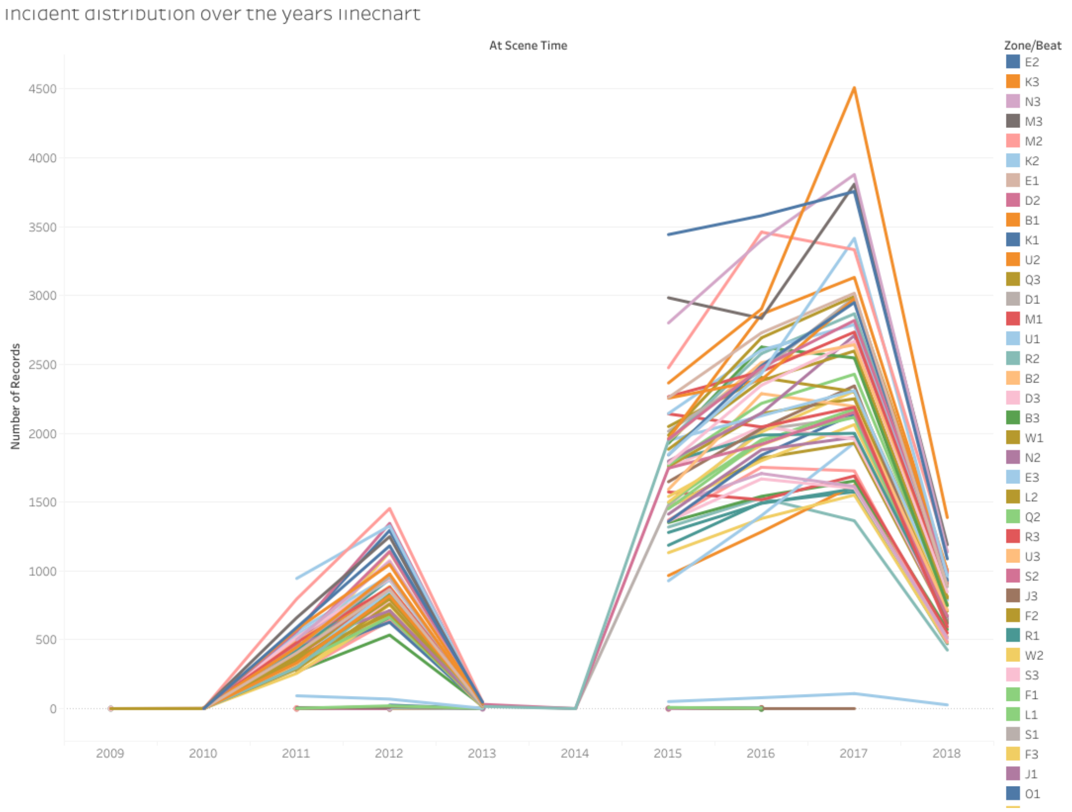


Figure 2: Incidents distribution over the years line chart.

To find out if there is different spatial distribution of incidents over the different years we plotted the years on the x axis and the number of records on the y axis, we also encoded the zones in to categorical colors so that each line in the chart is presenting the number of incidents over the years for each zone separately. From Figure 2 we can see that there is not a significant difference in spatial distribution over the different years , all the zones follow the same curve when it comes to the number of records.

3.2 Solution 2.2

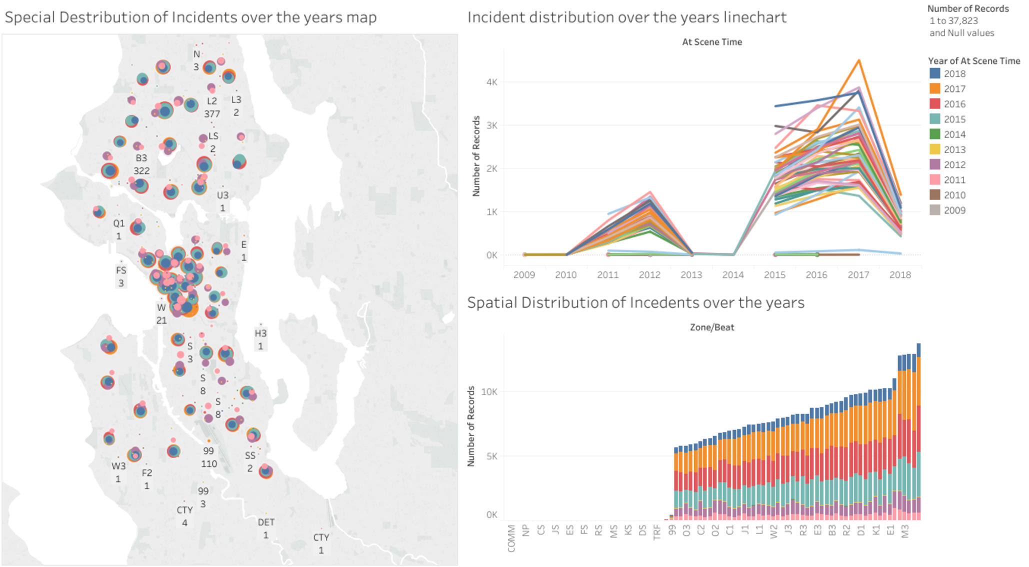


Figure 3: Spatial Distribution of incidents over the years.

To find out which zones have a consistent low incident density over all the years and which have high incidents density, we have plotted the data into bar plot chart in which the we put the zones on the x axis and the number of incidents on the y axis. We encoded the years into colors, we also wanted to see the where the zones geographically are located, for this purpose we plotted the same data using Seattle map, finally the line chart helps us to visually see which years have the highest number of incidents.
On the map we plotted the years as discs , to make sure the discs are visible we sorted the discs by the number of incidents so that the the smaller discs will be displayed on top of the bigger ones, as for the color map choice we chose a color map that represents categorical data.

The following zones have the lowest incident density over the years:
CCD, COMM, EDD, KCI007, NP, QS, US, CS, H3, JS, SCTR1, EP ES, GS, BS, FS,LS, NS
The following zones have the highest incident density over the years:
E2, K3, N3, M3, K2, E1, D2, Q3, D1, M1

We notice that the zones with low crimes are distributed over the Seattle map, so we cannot conclude that north Seattle has low crimes or south Seattle, same thing goes for zones with high incidents density, this observation excludes the data rows with null value of at scene time column,this observation includes the years with the least amount of incidents recorded (2009, 2013,2014), although we have kept those years on the plot but we do not take them into consideration when we are answering this question.

4 Task 3: Exploring the resolution speed

4.1 Solution 3.1

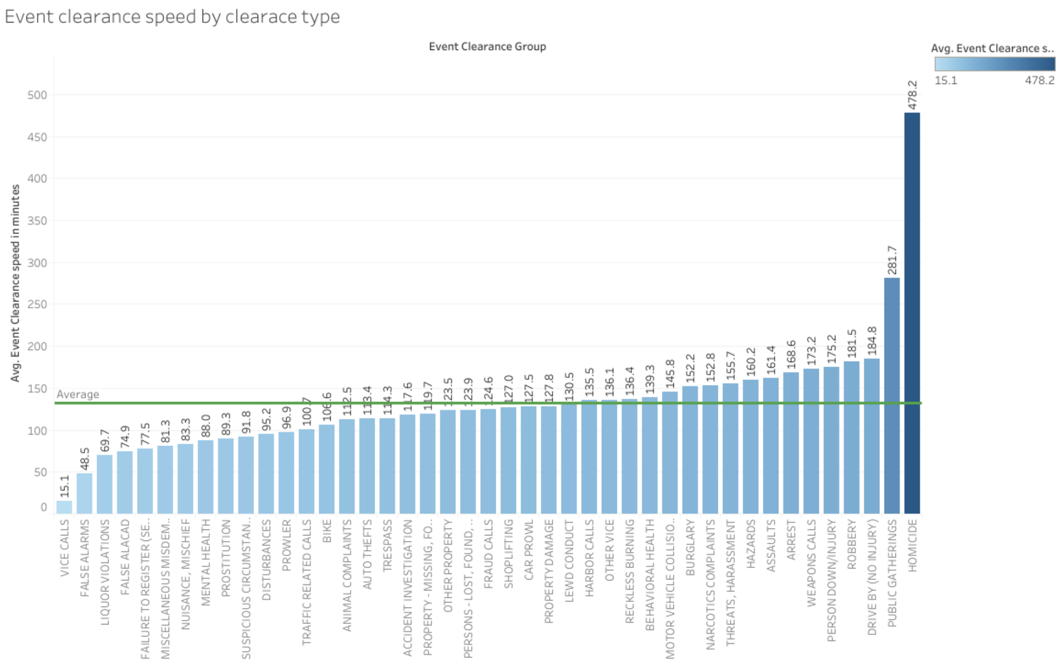


Figure 4: Event clearance speed by clearance type.

To find out the average resolution speed of an incident we had to create a calculated field where we calculate the difference between At Scene Time Value and Event Clearance Date value and plot them in minutes on the y- axis and put the incident types on the x- axis since the incident type is the independent variable, and add a an average line that represents the total average of the event clearance speed. We can see that the total average is (132 minutes).

4.2 Solution 3.2

To answer this question we used the same plot used for 3.1 where we have each incident type is presented by a bar, and we have the incidents sorted on the x axis by the resolution speed; from the plot we can also see that Homicide has the longest resolution speed (477 minutes) followed by Public Gatherings (281 minutes) and Drive By (184 minutes). The incidents that have the shortest resolutions speed are Vice Calls (15 minutes) and False Alarms (48 minutes).

4.3 Solution 3.3

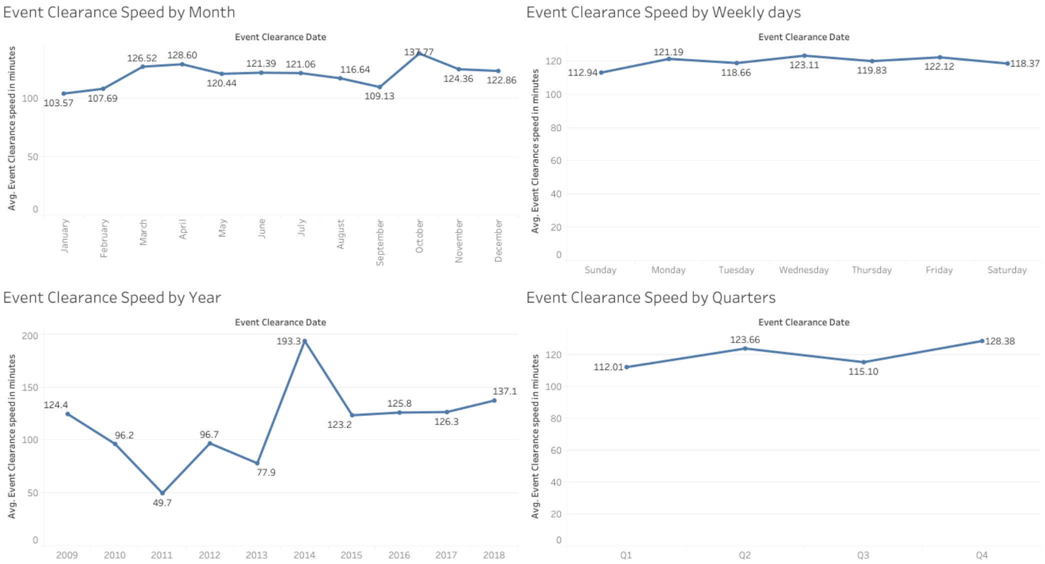


Figure 5: Resolution speed of incidents using different measures.

Figure 06 Resolution speed of incidents for each event clearance group To find out if the resolution speed depend on the time period we made a dashboard to investigate the following time periods: yearly , quarter of year , days of the week, and monthly. For the design choice we used linechar instead of bar plots, because it is easier to observe how each line representing an incident type group fluctuates. We plotted the time variable on the x axis, and the resolution speed on the y axis see ??.

From the dashboard we can notice that on the yearly basis the resolution speed varies the most compared to other temporal comparisons that we chose, the year 2014 has the slowest resolution speed (193 minutes) followed by the year 2018 (137 minutes) while the year 2011 has the fastest resolution speed (49 minutes) followed by 2013 (77 minutes).
On the quarter of the year basis we can notice there is not so much variance in the speed of resolution between each quarter of the year but having quarter 1 with the fastest average resolution speed (112 minutes), and quarter with the slowest average of resolution speed (128) minutes.
On the monthly basis the results are similar to the ones we got from the quarter basis since the monthly basis is the same visualization but having more details (every quarter is actually aggregated 3 months). We can notice that October has the slowest resolution speed average (137 minutes) followed by April (128 minutes), and the month January have the fastest resolution speed average(103 minutes) followed by February (107 minutes).
On the days of the week basis same as on the quarter basis and the monthly basis there is not much variance of the resolution speed average over the days of the week, we notice that Wednesday has the slowest resolution speed average (123 minutes) followed by Friday (122 minutes), Sunday has the fastest resolution speed average (112 minutes) followed by Saturday (118 minutes), we can conclude that on the days of the week basis even though the variance of the the resolution speed average is low the weekends have faster resolution speed average than the rest of the days.
We also made another dashboard in case we want to investigate the resolution speed average of each type of incident instead of the resolution speed of incidents in total.

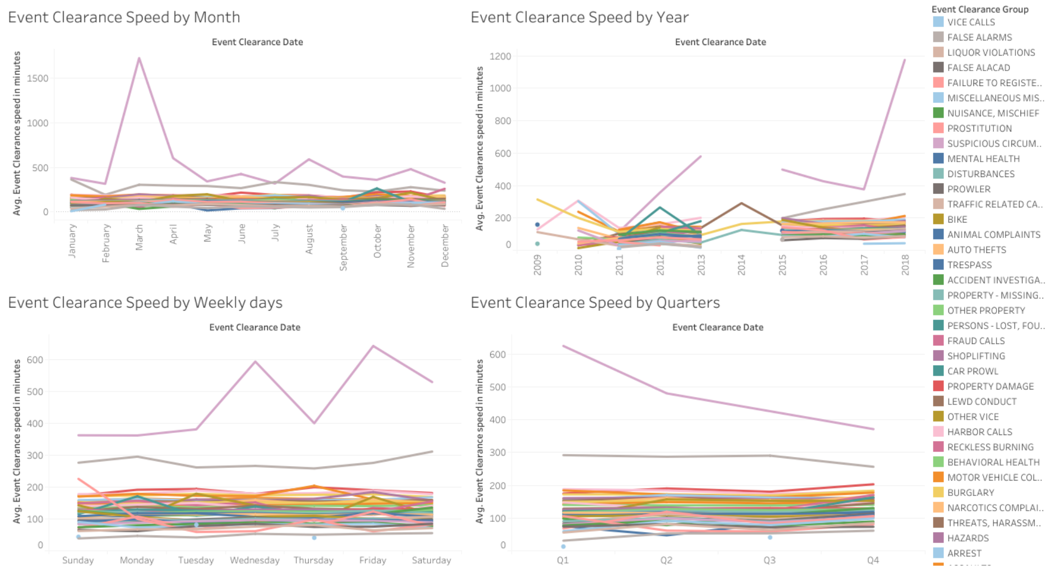


Figure 6: Dashboard showing the correlation and the distribution of crime types.

5 Task 4: Exploring the incidents' classification

For a matter of simplicity, all the questions can be answered using this visualization. To access the visualization please visit the link embedded in the screenshot above. We calculated the correlation between the count of event clearance group which had the same entry with the initial type group. Thus we didn't include the null entries which were later on labeled with an event clearance incident type. The correlation returned R^2 value of 0.917, $p < 0.0001$. Using the above dashboard, one can select a group of crimes on the left top pane. This will return trend line that best fits the selected incident group. After selecting some incidents on the left top pane, the bottom stacked bar chart will be filtered out showing only the selected incidents. These stacked bars are segmented based on the count of event clearance type that each initial type has. Clicking any segmented corresponding to the event clearance type will then reveal the event clearance subtypes on the right top pane. The dashboard enables for a fast access selection covered in one screen.

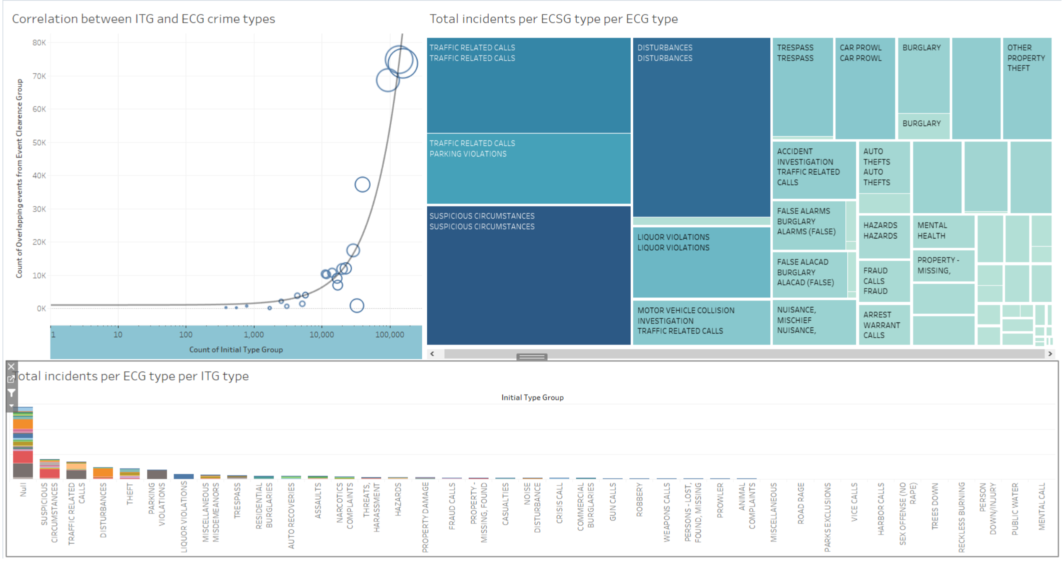
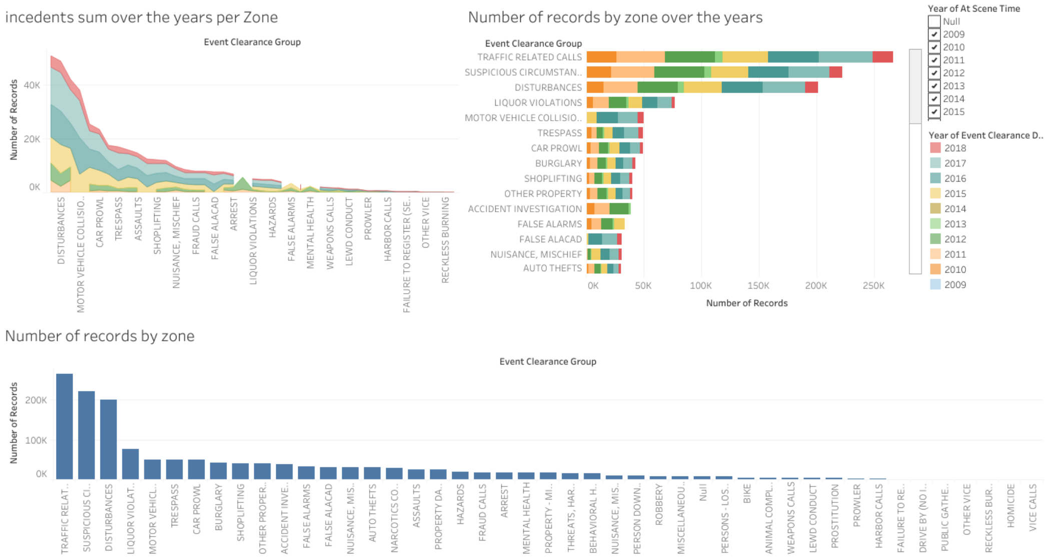


Figure 7: Resolution speed of incidents for each event clearance group

5.1 Solution 5.1:



incedents sum over the years per Zone

Number of records by zone over the years

Number of records by zone

Figure 8: Different types of incidents as represented by event clearance group variance over the years.

To find out how the different types of incidents vary in number over the different years we made a dashboard containing three visualizations first we used that to plot the variance of events over the years where we plotted the events on the x axis and the and the amount of incidents on the y axis, and we gave each your different categorical color, in the second visualization we plotted the same data using barplot where the y axis represents the incident types and the x axis represents the sum of number of records this comes handy when we use want to select only one type of incident because we end up with one bar having each year with different color.
The third bar plot used in the visualization is used as a filter of incident types over the first and second visualization. We sorted the years in an ascending way where we start by the year 2009 and finish by the year 2018 because it is more natural to read this way see Figure 8.
After investigating the data using the dashboard in figure 08 we noticed that most of the incidents have more or less the same distribution of incident types, however we noticed that certain incidents were not recorded in some years that observation can be seen using the area chart graph and using the years filter where we can highlight each year separately for example we have incident investigation, other property and false alarms absent in the year 2018, we speculate that this is could be due to changing the name of the incident type in the event clearance group in the year 2018. So your conclusion that generally different types of incidents have the same temporal variation pattern over the years with few exceptions.

5.2 Solution 5.2

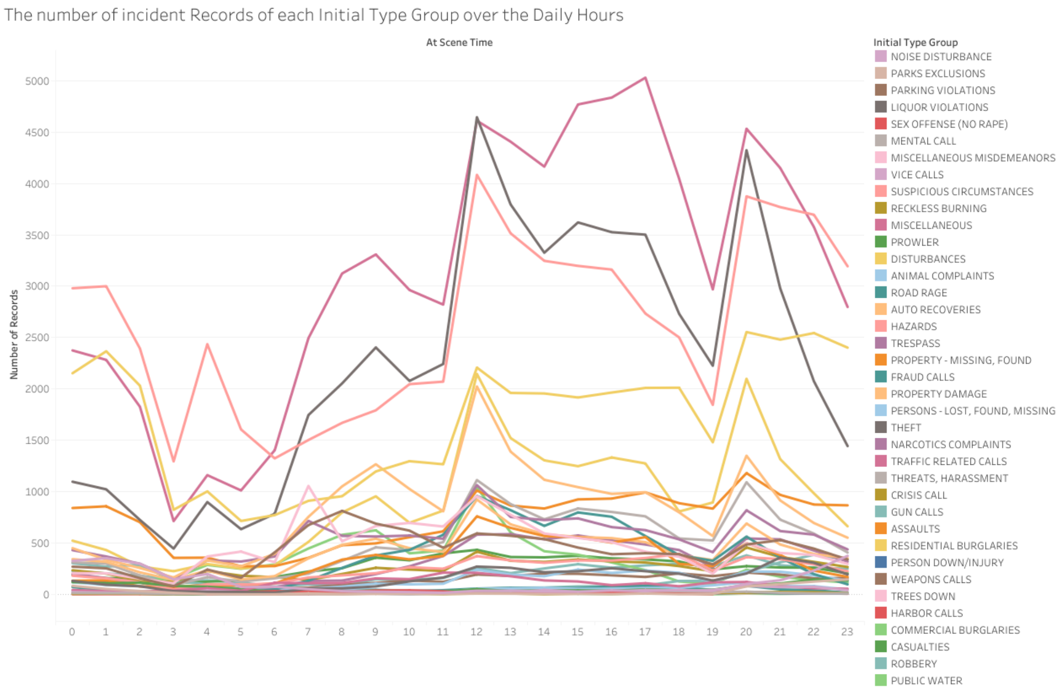


Figure 9: The number of incident records of each initial type group over the daily hours.

To find out how do the different types of the reported incidents vary over the 24 hours of the day we have plotted a line chart with the hours of the day on the x axis and the sum of reported incidents on the y-axis, then we gave each type different categorical color to be able to differentiate between each type of incident.
From the plot in Figure 10 we can see that the number of records fluctuates through the daily hours, but we can split the daily hours into two parts first part where we have high number of incident records (between 12:00 and 22:00) and hours that have low number of incidents record (between 22:00 and 12:00), we can also notice that most of the of the initial group types follow that division with some exceptions like gun calls and noise disturbance .

5.3 Solution 5.3:

To find out how long does it take to clear incidents as a function of the hour when they were reported (At scene time) we had to plot the average speed of the event clearance on the y axis and the at scene time on the x axis, we also plotted the total average as a line to be able to see what hours of the day result in resolution speed above the average and what hours will result on in resolution speed below the average, we also enhanced the visualization by showing the average speed as a number along side the line.

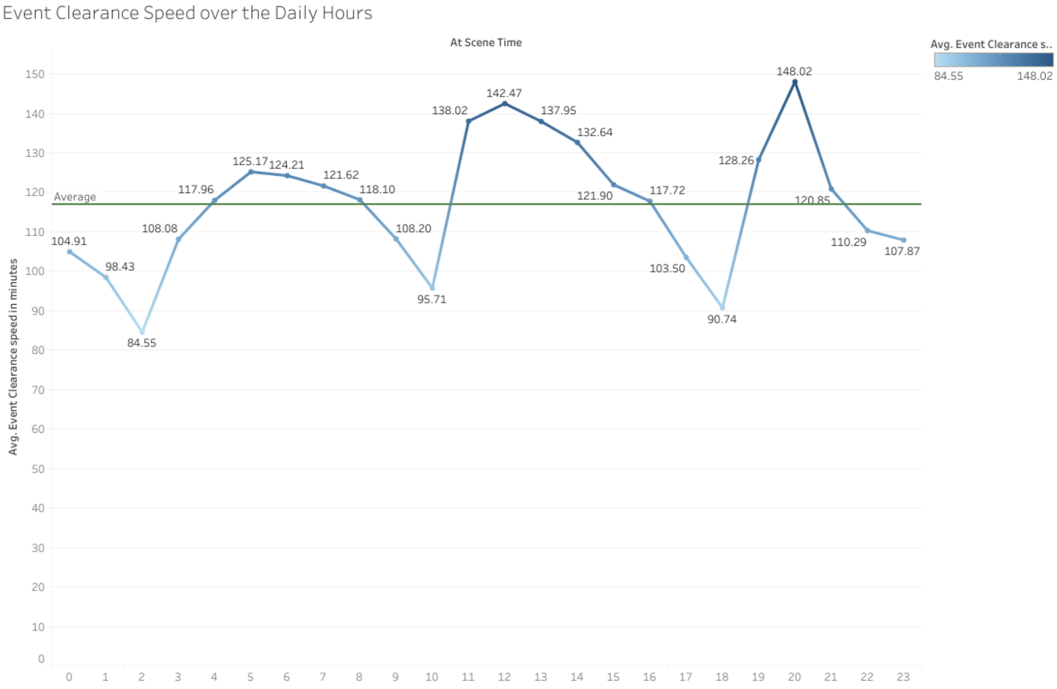


Figure 10: Event clearance speed over the daily hours

From the plot we can notice that the hours that have the slowest resolution speed are 20:00 with 148 minutes and 12:00 with 142 minutes the hours with the the fastest resolution speed are 02:00 with 84 minutes 18:00 with 90 minutes and 10:00 with 95 minutes.
Where the average resolution speed is 166 minutes. As a conclusion we can say that the hours between 22:00 and 3 am have fast resolution speed compared to the hours between 11:00 and 16:00.