Aqsa Mahmood

Abstract:

In this paper, we aim to analyse crime data of san Francisco between 2003 and 2018. The data has been collected by San Francisco Police department and is available on SF open data website. The focus of this research is to use visual analytics approach in order to find distribution of crime over space and time, comparing rates of different type of crimes in different regions of San Francisco. For this purpose, we used line graphs, stacked area charts, stacked bar charts, density maps, space time cube and density-based clustering. By using these visual analytics approaches we have found the most frequent time of crime occurring, regions of high crime rate and crime categories with maximum number of crime records.

Problem Statement:

In this paper, we aim to use visual analytical approaches to analyse patterns and trends in crime. As crime events are increasing over time which are resulting in massive crime datasets. It is worthwhile to analyse crime patterns using historical crime data in order to get useful insights from the data which can be helpful to overcome rising crime rates. Our work is mainly concerned in:

- Time
  - How does crime vary over time? • Over the year • Over the week and month • Over a day
- Space
  - In which districts are there most crimes? Are there any crime hotspots?
- Spatial-temporal
  - Do crime hotspots vary over time?
- Types of crime
  - Are there any striking differences between different types of crimes in terms of: – When and where they occur?
- Temporal correlation
  - How highly correlated are different types of crime temporally?

For this purpose, we choose the historical crime data of San Francisco which has records of crime events occurred in 10 districts of San Francisco between 2003 and 2018. The dataset is contained in csv file. It provides us information about time, date and location of different types of crime occurring. The data is in tabular form and has 2.21 million rows and 13 columns. As the data has information about location and time of different types of crime, it will allow us to examine how crime is distributed over space and time along with making comparison of trends of different categories of crime across various districts.

State of the Art:

(Li et al., 2017) proposed an interactive visualization system for examination the patterns and trends in crime data of Fuzhou, China over 2011.The dataset specifies information about crime events with multidimensional attributes including number of police resources, road networks, population, employment rate for each region of Fuzhou, China in 2011. The analysis is carried out particularly for four visual analytics tasks which are investigating spatial-temporal distribution of crime in order to find anomalies, hotspots and trends in time, comparing the crime rates in different areas during certain time periods, visually analysing the distribution of each attribute and their association with crime events. In order to discover crime trend based on temporal distribution, histograms were manipulated. For spatial visualization of crime hotspots, spatial density on certain time period is used in which different colours are encoded to show the areas of high and low concentrations. The statistics of crime per district is visualized by using a base map with pie charts in which the size of each pie chart and colour of each district presents the number of crimes in different districts. Furthermore, parallel coordinate system is used to investigate the patterns of crime events based upon high dimensional features in which a crime event is represented by line and axis characterise an attribute. Lastly, dependencies of crime events are investigated in order to make coefficient matrix of relevant features in which the intensity of correlation is represented by colours.

(Bayoumi et al., 2018) have worked on crime data of Maryland state, USA and considered interactive visualization to attain useful information about crime fluctuation after July 1st,2016. The dataset provides information about type of crime, location, time and date of each crime and number of victims of each crime event. In their research, they focused on three main categories of crimes: against a person, against property, against society and used visual analytics to find areas of high crime rates, the time and type of the most occurring crimes along with reasons behind it. The samples of each category of crime are plotted on map separately for various subsets of data which provided information that crime against property is the most occurring category of crime. The cause found behind this is high unemployment rate in the region. They considered the most frequent time for each category of crime is investigated. For this purpose, time is divided into six sections early morning, late morning, early afternoon, late afternoon, evening and night and being analysed using stacked bar chart. Similarly, the week day of most occurring category of crime is investigated. Pie Charts are used to examine the percentage of common types of crime against person, society and property. At last, user interaction interface of the data is created which allows user to visually explore data and filter attributes.

These papers use very similar datasets to ours as our data is also multidimensional data and have information about location and time of different types of crime happening. In addition, the problems addressed in the research papers, particularly, spatial temporal distribution of crime and finding the most frequent time and location for different types of crimes are similar to the aforementioned research questions. We will use the proposed approaches in these papers to view and examine temporal distribution of crime, spatial distribution of each type of crime..

Properties of the Data :

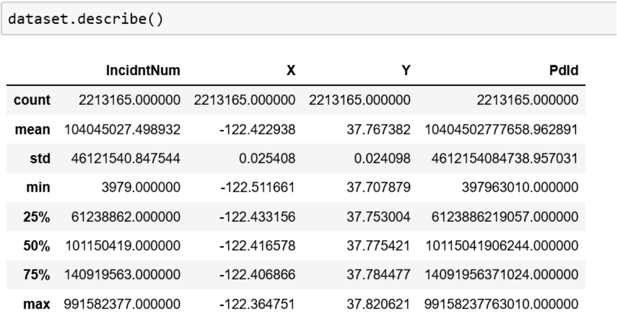
In this paper, the data used to carry analysis on crime has taken from the SF Open Data website. This data has made available by san Francisco Police Department and it has records of crime incidents from the year 2003 to (15 March) 2018. There are 2.21 million rows and 13 columns in the data. The data has memory usage of 270MB. Each attribute in the data is described below:

- IncidentNum: It is of numeric type and has record of incident number of the crimes.
- Category: It is nominal categorical variable and contains information about 39 different types of crime.
- Descript: This column briefly describes the crime and contains information as text.
- DayOfWeek: It has 7 distinct values and gives us information about the day of week when the crime took place. It is categorical variable.
- Date: It has type date-time and contains information about the date when crime occurred.
- Time: it provides us information about the exact time of crime event and has type date-time.
- PdDistrict: It is nominal categorical variable and gives information about 10 distinct police districts where crime occurred.
- Resolution: It specifies the resolution for crime and has object field. This column has 17 distinct values.
- Address: It is object field and contains street address of the crime event.
- X: It is spatial coordinate and gives longitudinal coordinate of crime location.it has float type. The values in column ranges between -122.51364206429 and -120.5.
- Y: It is spatial coordinate and gives latitudinal coordinate of crime location.it has float type. The values in column ranges between 37.7078790224135 and 90.0.
- Location: It is Location field and has the form of pair of geo coordinates (latitude and longitude).
- PdId: it is numeric field and it has records of each complaint registered in Police Department.

The provided dataset contains only 1 missing value in “PdDistrict” column which has been handled by inserting “Unknow” value. Standard deviation-based detection is used to detect and remove outliers from the data. A total of 143 outliers were found in the longitude and latitude columns of the data and were removed. The column "Date" is very important for our analysis, as it is used for temporal analysis; thus, we extract new features Month and Year.

Additionally, we consider using shape files that correspond of polygons that represent neighbourhoods and police-districts of San-Francisco, which are also provide from the same link as the dataset. These shapefiles can be merged or joined with main dataset, so that we analyse crimes by neighbourhood or by district.

At last, we sorted the values in our data by date and time columns. Tableau, Python and Python Modules Plotly, Geo-Pandas, Folium, sklearn are used for the data (pre-)processing and interactive visualizations of the data.



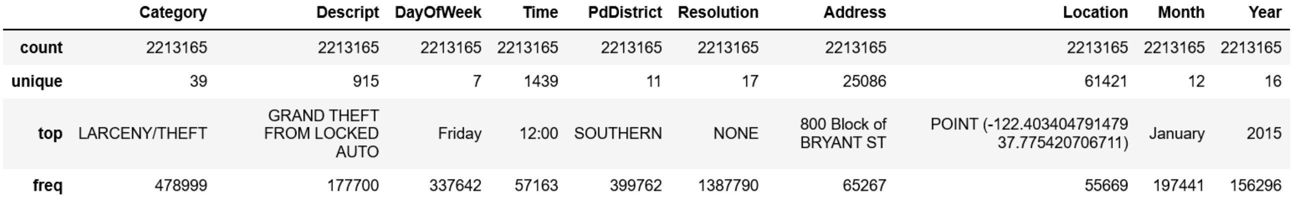
```
dataset.describe()
```

	IncidntNum	X	Y	PdId
count	2213165.000000	2213165.000000	2213165.000000	2213165.000000
mean	104045027.498932	-122.422938	37.767382	10404502777658.962891
std	46121540.847544	0.025408	0.024098	4612154084738.957031
min	3979.000000	-122.511661	37.707879	397963010.000000
25%	61238862.000000	-122.433156	37.753004	6123886219057.000000
50%	101150419.000000	-122.416578	37.775421	10115041906244.000000
75%	140919563.000000	-122.406866	37.784477	14091956371024.000000
max	991582377.000000	-122.364751	37.820621	99158237763010.000000

Figure 1: Summary of Numerical Features of the data

Figure 1 and ***Figure 2*** illustrate the summary statistics, in which we can observe maximum, minimum, standard deviation, mean and quantile values for each numeric and count, unique and frequency for columns of type object.

To answer our research questions, we will limit our work to category, neighbourhood, day of week, district, month, year, time, longitude and latitude columns and descript column will only be used to get more details about categories of crime.



	Category	Descript	DayOfWeek	Time	PdDistrict	Resolution	Address	Location	Month	Year
count	2213165	2213165	2213165	2213165	2213165	2213165	2213165	2213165	2213165	2213165
unique	39	915	7	1439	11	17	25086	61421	12	16
top	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Friday	12:00	SOUTHERN	NONE	800 Block of BRYANT ST	POINT (-122.403404791479 37.775420706711)	January	2015
freq	478999	177700	337642	57163	399762	1387790	65267	55669	197441	156296

Figure 2: summary of categorical features of the data

Analysis

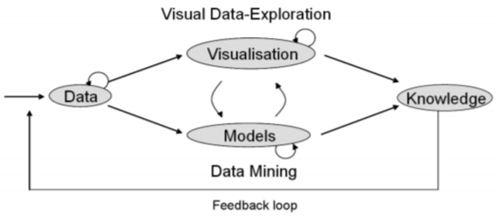


Figure 4: Visual analytics process by Keim et al. [4]

Analysis Approach

To attempt the research questions above, we will start with simple visualizations such as bar charts, line graphs which will help us to understand the general distribution of crime temporally and then progress on to maps & space-time cubes to consider the spatial and spatial-temporal distribution.

To study the effects of crime in the Temporal domain, we will consider the following:

- Line Graph showing number of crimes per year for each crime category.
- Stacked area chart showing number of crimes each hour of the day for each crime category.
- Stacked bar Chart showing number of crimes per month and day of week for each crime category.

We will use **density-based clustering** to find dense groups of each crime category in each district and will be considered for spatial and spatial-temporal analysis for producing maps. In order to achieve this, the “DBSCAN” method form the sklearn is utilised. Before, using this method, that dataset is transformed such that we have a column ‘days since 2013-01-01’, which is essential produced by subtracting every date by the reference date ‘2013-01-01’. Furthermore, we propose the following: variables:

- temporal distance threshold – used primarily for spatial-temporal analysis and is initially set to 20 days
- spatial distance threshold – this variable is used for spatial and for spatial-temporal analysis and is a function of the radius of earth.

In addition, we need to specify the minimal density threshold, i.e., the minimal number of neighbours required for a core object of a cluster. You can start with setting this parameter to 3, which is quite a loose constraint. Additionally, we will need to perform clustering several times with different parameter settings and then decide which result to retain.

For Spatial Analysis, we define a function, that calculates the shortest distance between two latitude and longitude points using the ‘great circle’ formula and is set to the attribute ‘metric’ of the method DBSCAN. One clustering is performed, we can produce several sub-maps to observe the spatial distribution of crime category per district over 15 years. An alternative approach for spatial analysis, to bypass utilising ‘DBSCAN’, could potentially make use of the in-built clustering capabilities of the folium package via the use of ‘MarkerCluster’. Furthermore, we can produce maps to observe the spatial-distribution of crimes in each district and neighbourhoods over the 15 years periods via the use of the geometry attributes within the shapefiles and counting the total number of crimes for each district and neighbourhood.

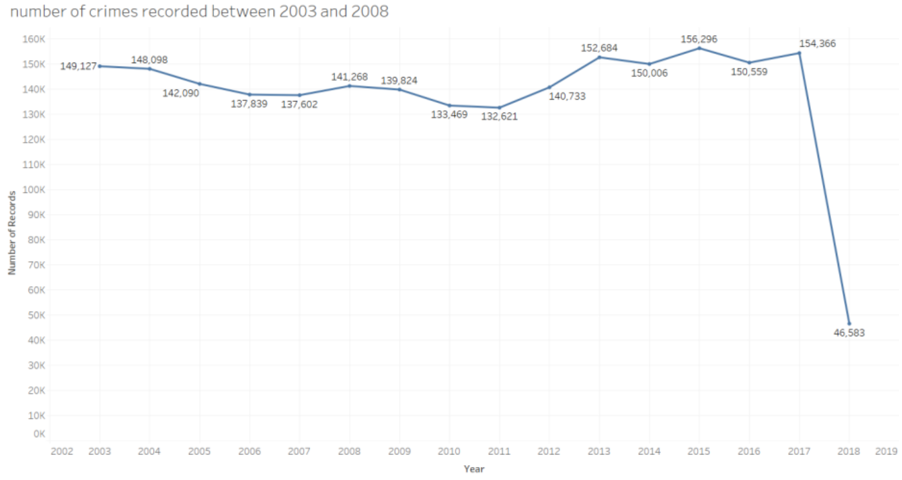
For Spatial-temporal, we define a function, that considers the ‘great circle’ formula for two points and the column ‘days since 2013-01-01’, which is set to the attribute ‘metric’ of the method DBSCAN. One clustering is performed, we will produce

- Space-time cube’s and maps to give a view on how crime rates, either generally or by category, vary over time and space and to try to identify trends or hotspots.

In all cases given that we have 39 categories, it may not be feasible to run clustering on each category due to computer resource requirements, and thus would probably apply it to the highest crime category or maybe several.

Analysis Process

Temporal Analysis



The trend of sum of Number of Records for Year. The marks are labelled by sum of Number of Records.

Figure 3: Overall number of crimes recorded by year

Figure 3 illustrates that there were about 150, 000 crimes recorded in 2003 which has seen a slight fluctuation until 2017, but after this time span number of crimes has dropped sharply to 46,583 in 2018. This can be possibly due to an increase in the number of police officers and advancement in technology.

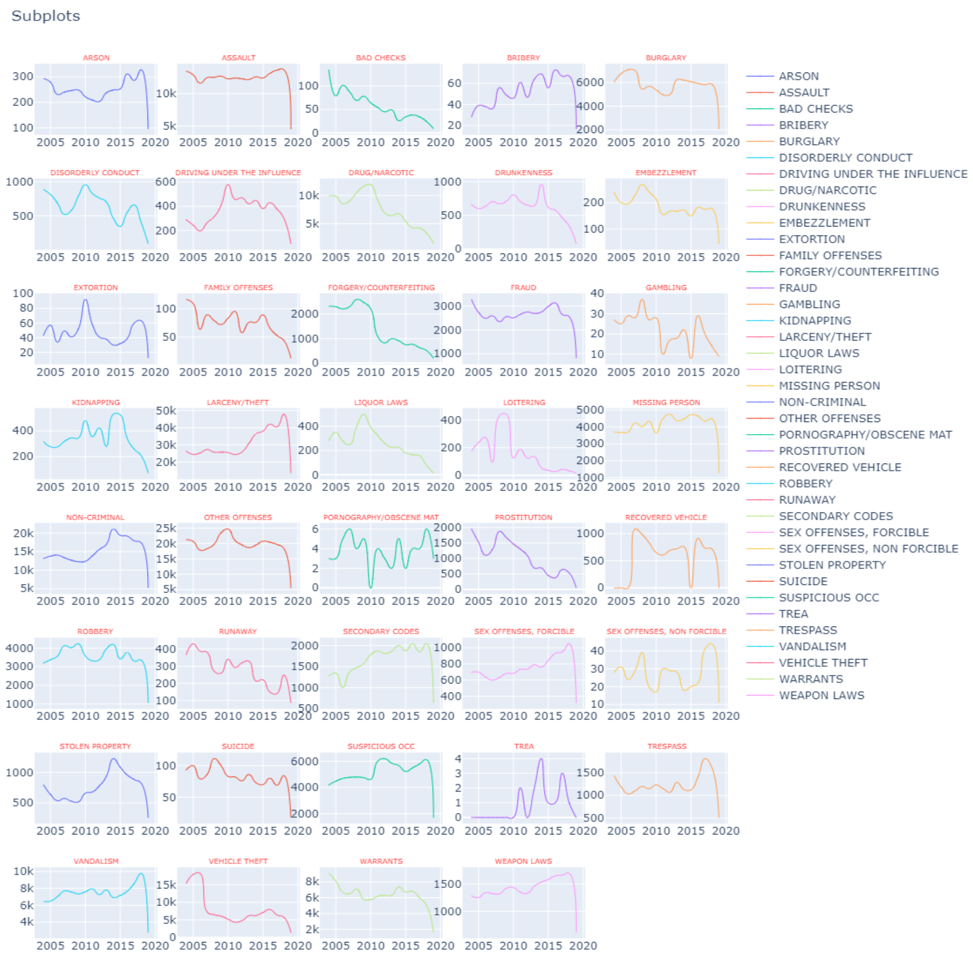


Figure 4: Crime rate for each category over 15 years

From *Figure 4*, we can observe that many of crime categories, namely assault, bribery, burglary, fraud, missing person, Robbery, suicide, vandalism, warrants, larceny/theft, weapon laws have shown slight fluctuation between 2003 and 2016, however, after this period a sharp drop is observed. There are some categories which has a downward trend over the period such as drug/narcotic, family offences, vehicle theft, runaway, prostitution, liquor laws. An interesting point is that there is not any crime category which has an upward trend over the given period.

Stacked area chart showing number of crimes each hour of the day for each crime category.

Figure 5 shows a stacked area chart to observe trends in number of crimes for each category over 24 hours. It can be analysed that, most of the crime events happened between 12:00pm and 10:00pm and 18:00pm is the hour with the highest number of crime records. Also, the highest numbers of crime recorded are related to Theft/Larceny category. Generally, crime rate is lower at morning hours between 2:00am and 9:00am.

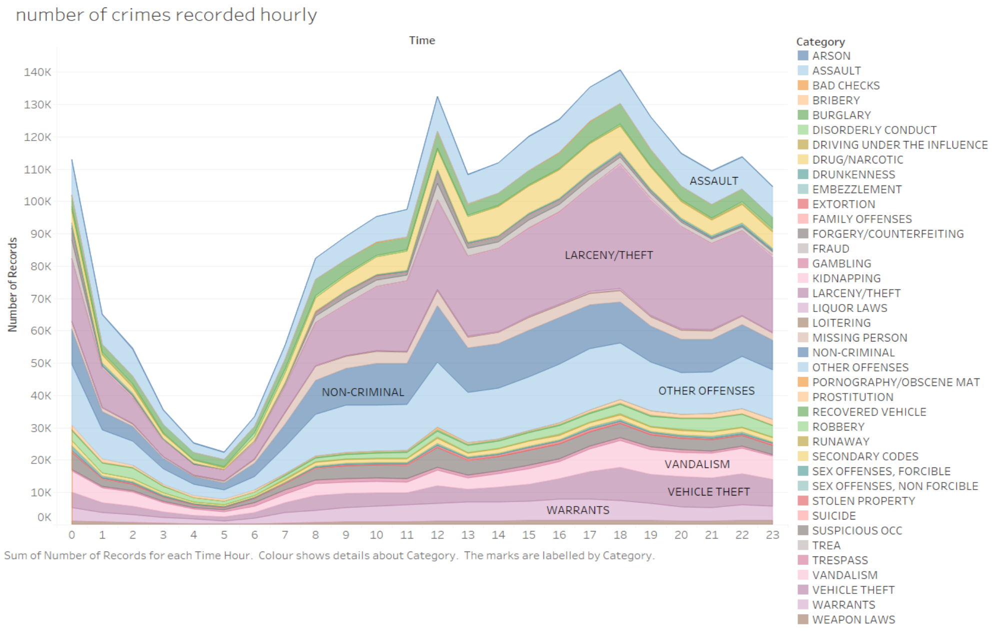


Figure 5: Number of crimes recorded hourly

Stacked Bar Chart showing number of crimes per month for each crime category

Figure 6 illustrates a bar graph representing number of crimes for each category per month. It can be analysed that January and March have the highest number of crimes recorded than all other months. Furthermore, the highest number of crimes recorded for each month are related to larceny/theft.

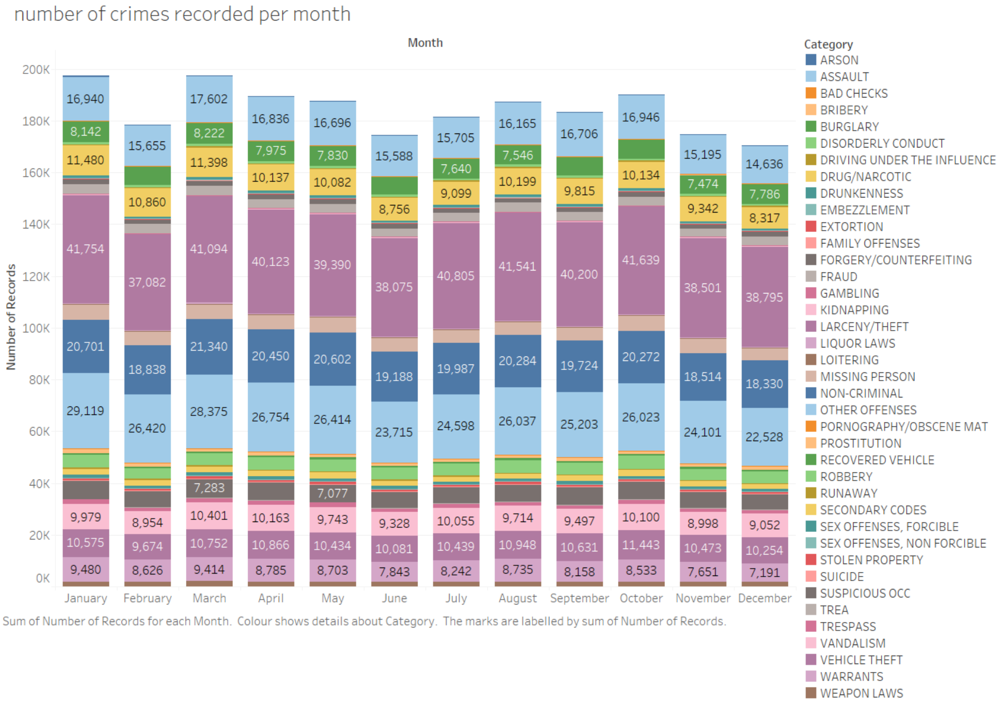


Figure 6: number of crimes per month

Stacked Bar Chart showing number of crimes for each day of the week for each crime category.

Figure 7 demonstrates that Friday has the highest crime rate over other weekdays. The category larceny/theft has the maximum number of crime records and 'other offenses', non-criminal and assault are the next mostly occurring crime categories respectively. The descript column has been used to get information about 'other offenses' and 'non-criminal' crime categories. It has been analysed that 'other offenses' involves events such as traffic violation, lost/stolen license plate, harassing phone calls and non-criminal category specifies incidents like death report natural causes, lost property, found property.

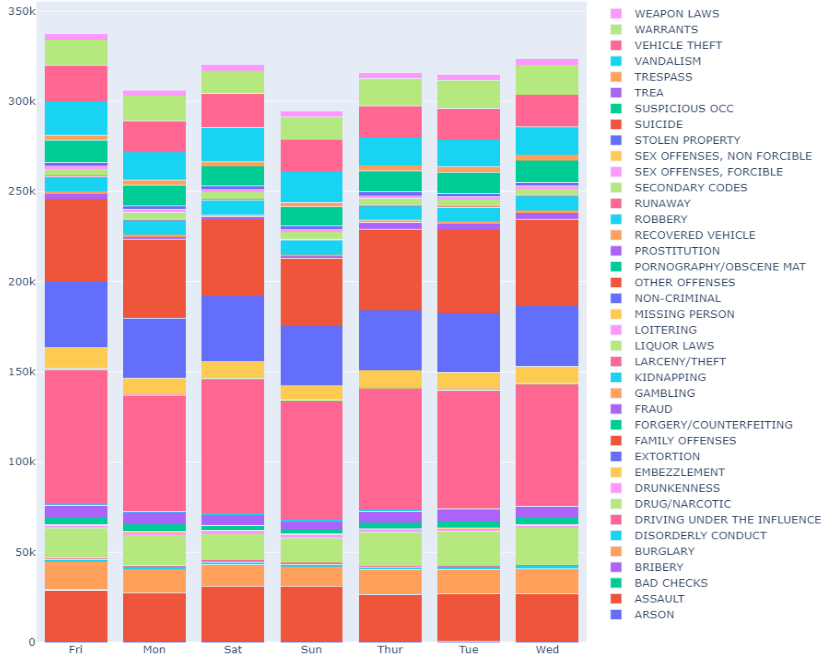


Figure 7: Rate of crime per week

Spatial Analysis:

Using the geo-panda's library, we can read the shapefiles neighbourhoods & police districts to produce a geo-panda dataset, which we can join these using 'sjoin' from the geo-pandas library, with the main dataset based

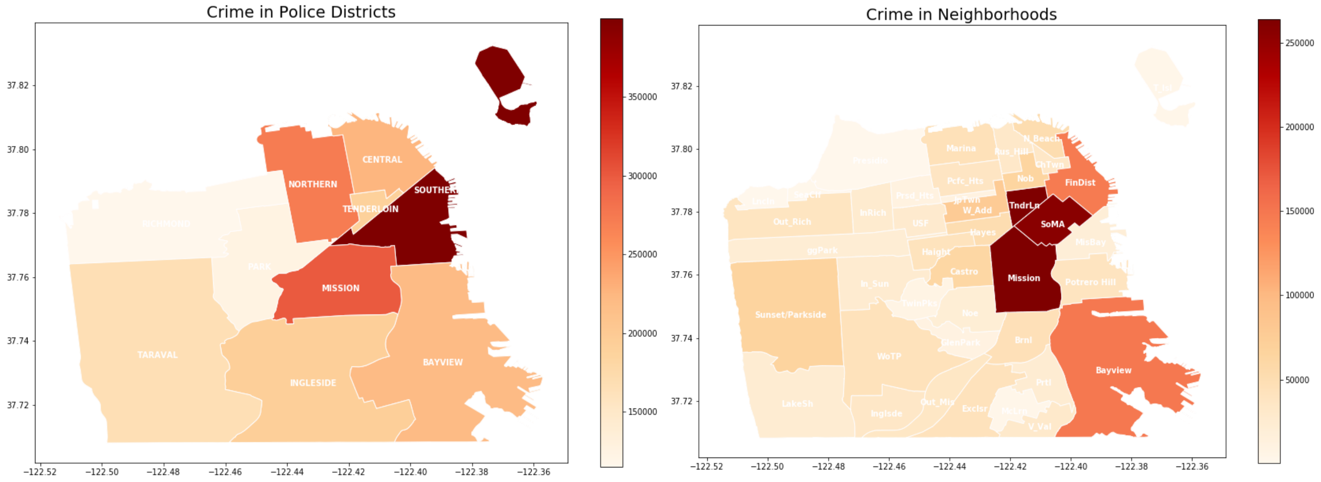
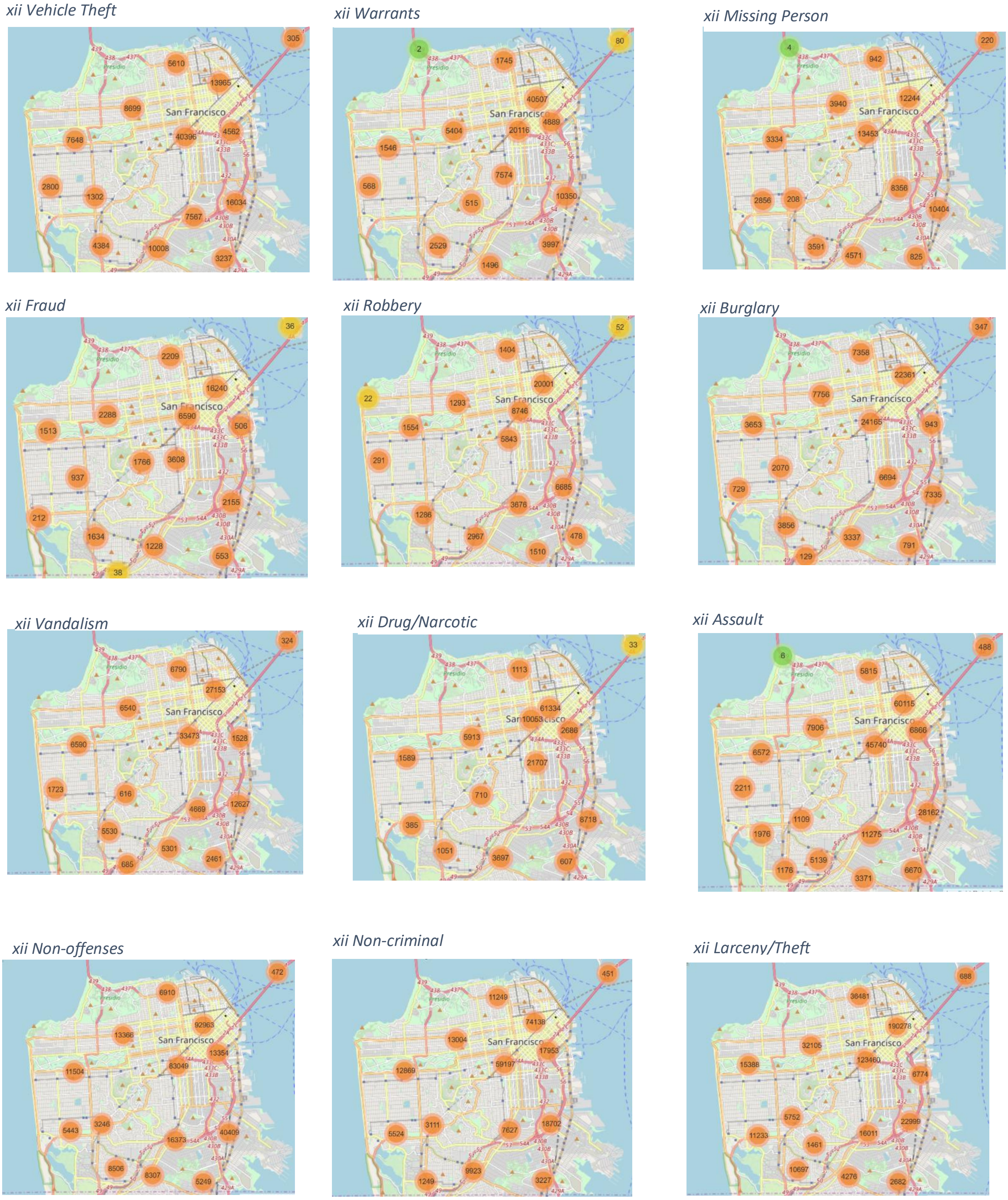


Figure 8: Distribution of crime in police districts and its neighbourhoods

on incident coordinates. Considering calculating the total number of incidents per neighbourhood and police district, we can produce ***Figure 8*** .It is observed that Southern district has the highest crime rate than other districts. After Southern, Mission and Northern have the maximum crime rates respectively.

As we have large number of crime categories, we just filtered 12 crime categories which have been analysed as the most frequently occurring crime categories during temporal analysis of crime. To visualize the spatial distribution of these 12 crime categories, we created folium maps using Marker Cluster for each category which are helpful to visualize geo-spatial data. ***Figure 9*** shows how different crime categories are distributed over all districts. Using ***Figure 8*** and ***Figure 9***, we can see that southern district has the highest crime rate for crimes related to larceny/theft, robbery, fraud, assault, drug/narcotic, warrants, non-offenses, and non-criminals with 190278, 20001, 16240, 60115, 61334, 40507, 92963 and 74138 total number of incidents respectively. Furthermore, crimes related to vandalism, burglary, missing person and vehicle theft has frequently occurred in Mission district with number of incidents 33473, 24165, 13453 and 13965 respectively. Moreover, larceny/theft has the largest number of crime records over all other categories. From above analysis, we can conclude that, southern district has the highest crime rate and mission is the second district with high crime rate.



xii Vehicle Theft

xii Warrants

xii Missing Person

xii Fraud

xii Robbery

xii Burglary

xii Vandalism

xii Drug/Narcotic

xii Assault

xii Non-offenses

xii Non-criminal

xii Larceny/Theft

Figure 9 Spatial distribution of different types of crime

Spatial-temporal Analysis:

We can produce a space-time cube sublots for each crime category via the use of matplotlib to show the variation of each crime in the spatial and temporal domain as shown in Figure 11. The z axis is essentially the number of days since 2003, which has a range of over a five thousand, and has been calculated for each row by subtracting each date-time value by the reference date 01/01/2003. The x axis is the longitude, and the y axis is the latitude.

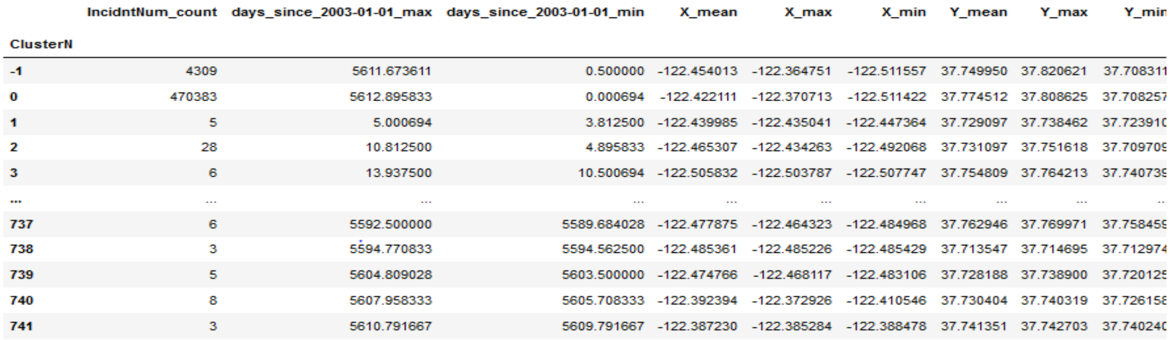
Note density-based clustering has **not** been used, thus why we observe significant clutter. We can observe that in that temporal domain that a significant few of offences including non-criminal, assault, fraud, vandalism, warrants, burglary, robbery, theft shows that number of incidents continue to increase as time passes and are numerous.; whereas other less numerous incidents such prostitution, embezzlement, loitering, extortion, suicides and bribery do not display the same effects. We can also observe that there are hardly any treason or pornographic offences at all.

Density based clustering

Density based clustering was performed using DBSCAN initially for each crime category using the columns 'X', 'Y' and 'days since 2003', with parameters:

- Eps as set to as the ratio of 1.5 and radius of the earth;
- Minimum cluster size as 3,
- a bespoke metric functions the calculates distance in km and the difference in days between two points taken into consideration the temporal and spatial maximum threshold.

As time progressed with the computation, ran into memory errors. Due to the significant number of categories its unfeasible to perform clustering on each category in isolation due to computer resource restrictions, however can consider performing density-based clustering on one of the categories such as theft or robberies, as they are the most prominent offence throughout the spatial and temporal domain, but this would prevent us from partially answering the questions on spatial and temporal correlation on all the crime categories. Result for clustering based on the theft category is shown below in Figure 10 with 741 clusters, and the noise cluster with 4309 observations. We can now visualise the dataset in space-time, but this time taken into consideration the clusters plus noise. Looking at the table we can observe that cluster zero, has by far the greatest number of observations (470383), whereas the clusters are in-significant in comparison. This suggests hyper-tuning maybe required.



	IncidntNum_count	days_since_2003-01-01_max	days_since_2003-01-01_min	X_mean	X_max	X_min	Y_mean	Y_max	Y_min
ClusterN									
-1	4309	5611.673611	0.500000	-122.454013	-122.364751	-122.511557	37.749950	37.820621	37.708311
0	470383	5612.895833	0.000694	-122.422111	-122.370713	-122.511422	37.774512	37.808625	37.708257
1	5	5.000694	3.812500	-122.439985	-122.435041	-122.447364	37.729097	37.738462	37.723910
2	28	10.812500	4.895833	-122.465307	-122.434263	-122.492068	37.731097	37.751618	37.709709
3	6	13.937500	10.500694	-122.505832	-122.503787	-122.507747	37.754809	37.764213	37.740739
...	...	...	...	...	...	...	...	...	..
737	6	5592.500000	5589.684028	-122.477875	-122.464323	-122.484968	37.762946	37.769971	37.758459
738	3	5594.770833	5594.562500	-122.485361	-122.485226	-122.485429	37.713547	37.714695	37.712974
739	5	5604.809028	5603.500000	-122.474766	-122.468117	-122.483106	37.728188	37.738900	37.720125
740	8	5607.958333	5605.708333	-122.392394	-122.372926	-122.410546	37.730404	37.740319	37.726158
741	3	5610.791667	5609.791667	-122.387230	-122.385284	-122.388478	37.741351	37.742703	37.740240

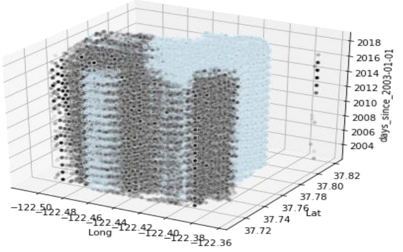


Figure 10: Cluster Table for Theft Category and space-time Cube

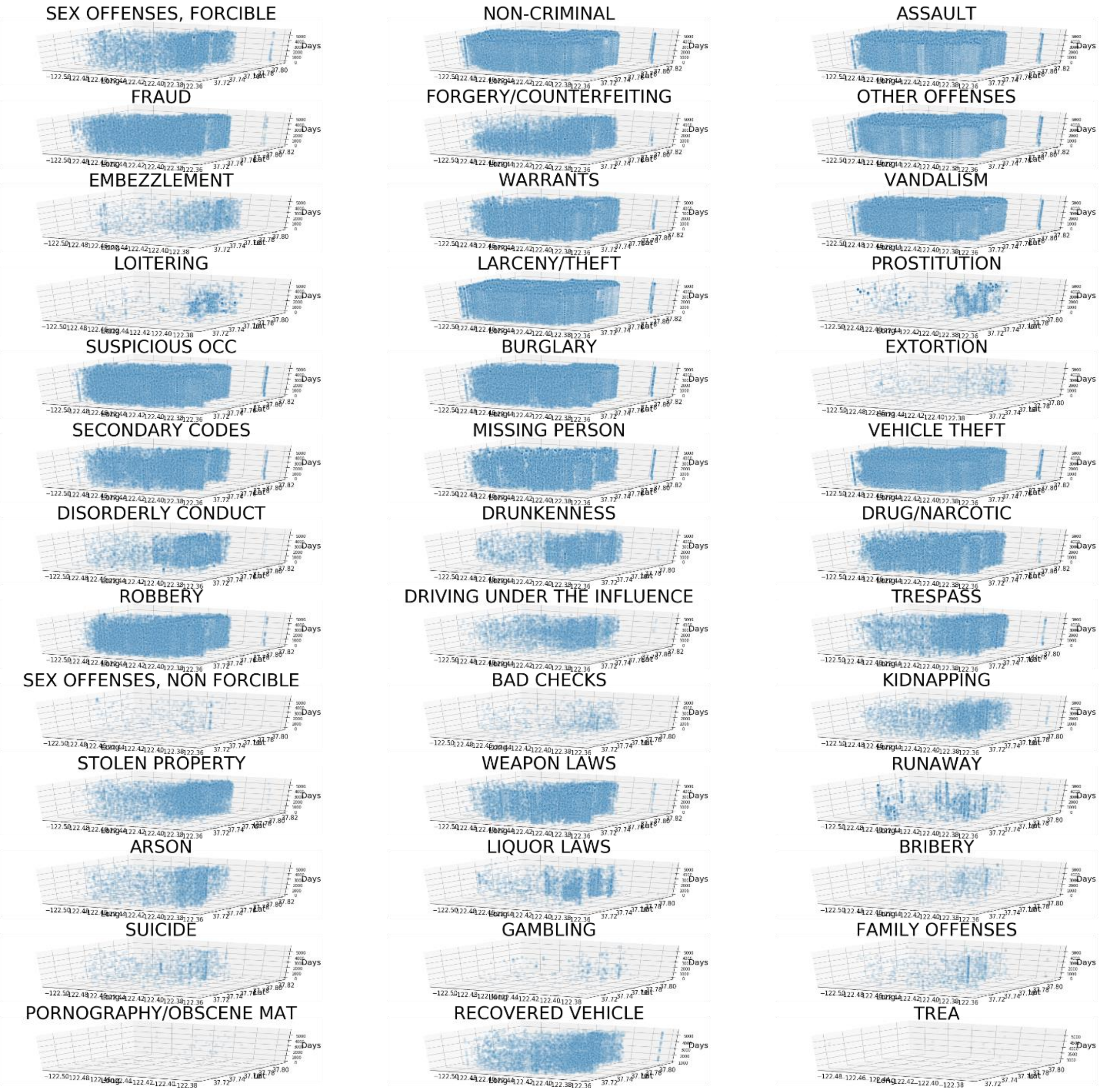


Figure 11: Space-time cube for each crime (un-clustered)

Due to the lack time we were unable to perform hyper-parameter tuning to find the optimum settings.

Analysis Results.

In this section, we will briefly explain the findings for our research questions. During temporal analysis, we analysed trends in crime rate per month, per year, per day and per week using line plots, stacked bar charts, stacked area charts.

-Trends in crime rate by time

-Yearly

Crime rate was higher in 2003 which has been fluctuating slightly until 2017 while between 2017 and 2018, it has dropped sharply.

-Monthly

Crime rate was higher in January and March compare to other months.

-Weekly

The maximum number of crime events were recorded on Friday.

-Hourly

Crime rate was low in morning hours, but it has increased between 12:00pm and 10:00pm.

-Trends in crime rate by space

Southern district has been found with the highest number of crime rates and Mission district is the second district with highest crime rate. The maximum number of crimes recorded were relevant to larceny/theft. Southern district has the highest crime rate for categories larceny/theft, robbery, fraud, assault, drug/narcotic, warrants, non-offenses, and non-criminals whereas vandalism, burglary, missing person and vehicle theft are the most occurring crime categories in Mission district

Critical Reflection

In hindsight, should have considered weeks or even months since 2003, instead of days since 2003, for the z axis in order to produce less cluttered and better visual plots.

Furthermore, as there are more than 2 million rows, it would have been feasible to consider performing density based clustering on the GPU using libraries such as 'cuML' which is a suite of libraries that implement machine learning algorithms and mathematical primitives and can enable engineering to run traditional tabular ML, including density based clustering, tasks on GPUs without going into the details of CUDA programming. The boost in performance whilst perfuming machine learning on the GPU is of O(n) , i.e. the larger the dataset the better the boost in performance in comparison to the CPU. Alternatively, should have considered the analysis on a subset of the data instead of the whole data ranging from 2003 to 2018.

I believe the visual representations were suffice in allowing us to answer most of the questions, however what was lacking, was in addition to producing the space-time cubes subplots, which were difficult to observe, producing space-time maps for each crime category would have been beneficial in releasing the answers to the questions, and providing better spatial clarity.

Table of word counts

Problem statement	248
State of the art	559
Properties of the data	535
Analysis: Approach	516
Analysis: Process	1167
Analysis: Results	183
Critical reflection	206

References

Bayoumi, S. A. (2018). A review of Crime Analysis and Visualization. Case study: Maryland State, USA. 2018 21st Saudi Computer Society National Computer Conference (NCC).

Keim, D. A. (2008). *Visual Analytics.*

Li, D. W. (2017). An visual analytics approach to explore criminal patterns based on multidimensional data.

Data.sfgov.org. (2019). [online] Available at: https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-Historical-2003/tmnf-yvry/data [Accessed 20 Dec. 2019].