Exploratory Data Analysis

- StudentID: 21900086
- Name: Kim Rubiga
- 1st Major: Life Science
- 2nd Major: AI

Crime Data Analysis

Introduction

Project Summary The data file consists of data on incidents of crimes in Dallas, Texas in the United States. The Police Department records these records daily in a crime log. With analysis these events can be better visualized geographically to observe trends on the location, type of incidents and even the likelihood of events based on the timings. The analysis of crimes are recorded by the police in different locations and on different time and dates. Even patterns of racial distribution of the offenders is seen.

Effective police action is crucial for the safety of people and to prevent crimes overall. So it is important to understand the trends of past crimes for effective future prevention. As time passes, the crime rates are increasing around the world so one of the ways to reduce this is can be on understanding how the crimes are happening. Certain patterns over the type of crime and time can be observed in the way the offenses that contribute to crime.

For example, understanding the distribution of crime rates of people belonging to different racial groups can be important for criminal profiling. If any racial bias can be observed from the past trends ofcriminal logs, then measures should be taken to tackle it. Therefore, effective police action is not only about reducing crime but also including various factors with the main purpose of **safety** for the people regardless of their ethnicity.

Objective of Analysis : to find patterns of policing from previous data for inferences.

1. Data overview

Raw Data
Number of Incidents: 663249
Number of Columns: 107

Selection of variables *Reason for choice*:

- Too much missing values
- Not relevant for analysis
- Details of the code not known well
- Not enough details in the codebook to use in the code to make inferences

For example, 'personinvolvementtype' variable cannot be known from the current codebook detail as I cannot be sure on whether the involvement includes on crimes or person involved directly or indirectly in the crime case.

After dropping n/a and missing values
Number of Incidents: 81010
Number of Columns: 20

Note: If you want to observe the overall trend of just crime incidents in an area regardless of other variables, we could infer with missing values included

Main variables

Here is the above list of column names formatted as a Markdown table:

Column Name	Description
incidentnum	Unique identifier for each incident
ctype	Crime type
year	Year of occurrence
typeofincident	Type of incident
typelocation	Type of location where the incident occurred
division	Police division where the incident was reported
year1ofoccurrence	Year of occurrence (alternate field)
month1ofoccurence	Month of occurrence
day1oftheweek	Day of the week of occurrence
time1ofoccurrence	Time of occurrence
victimtype	Type of victim
victimrace	Race of victim
victimethnicity	Ethnicity of victim
drugrelatedistevencident	Whether the incident was drug-related or not
nibrscrime	National Incident-Based Reporting System (NIBRS) crime code
nibrscrimecategory	NIBRS crime category
xcoordinate	X coordinate of incident location
ycordinate	Y coordinate of incident location
geo_lat	Latitude of incident location
geo_long	Longitude of incident location

Note: nibrscrime may give more specific crime category details

- Year is the only numerical variable

Region	Count
NORTHEAST	14518
SOUTHEAST	12457
SOUTHWEST	12114
NORTHWEST	11220
CENTRAL	11181
SOUTH CENTRAL	10708
NORTH CENTRAL	8811

- The Dallas Police Deparment currently has 8 divisions as explained on their website. http://www.dallaspolice.net/division
- It appears that the Jack Evans headquarters has no incidents associated with it, so our analysis will involved the 7 main divisions in Dallas, Texas. **year**: 2014~2018

Note: Before removing the missing values, the data existed from 2014~2020 which possibly seems that the newer data doesn't have other variables updated

2. Univariate analysis

2.1 (ex) Variable 1

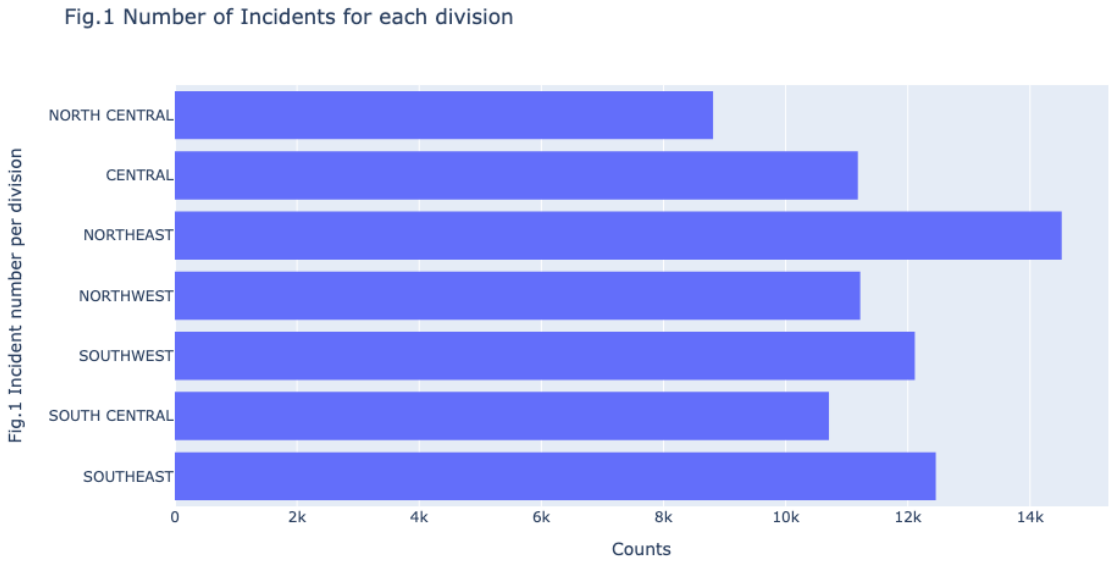

Figure 1. Number of Incidents in the 7 Divisions

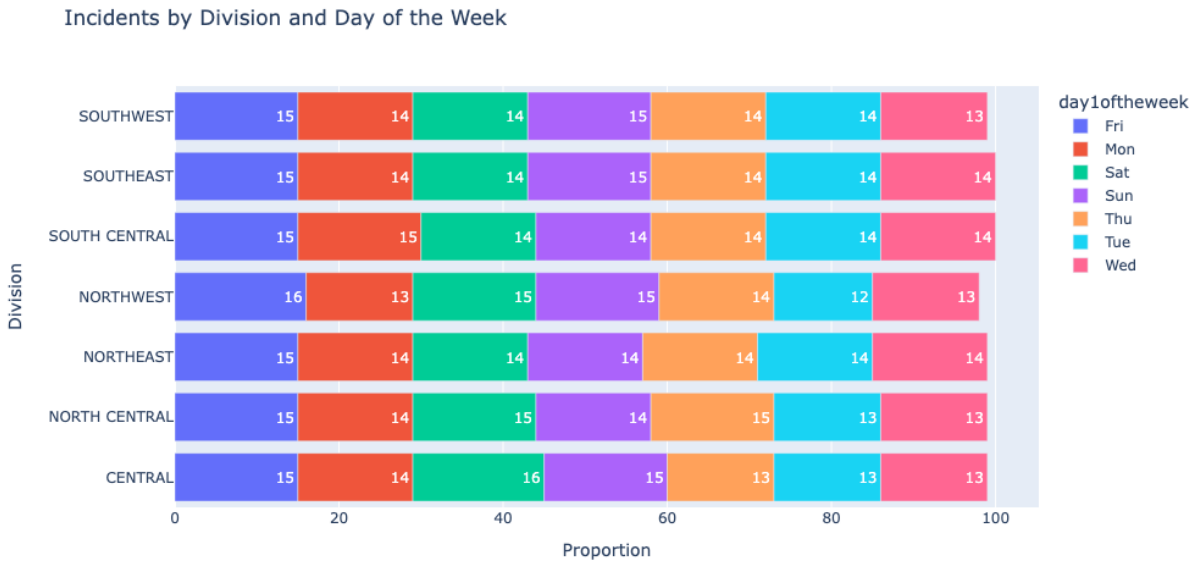

Figure 1. Number of Incidents in the 7 Divisions

- From Fig 1. We can see that the number of incidents recorded is significantly high in the `NORTH EAST` with 14.518K incidents with the least being in `NORTH CENTRAL` with 8811 incidents.
- This pattern can cause more investigation in these areas with other factors.
- From Fig 1.2, we can further analyze of the days of the week in which a particular division has the mose incidents. From this visualization there is no evident patterns found as the pattern of the incidents for the different divisions are almost same on the different days of the week. We can see that `NORTH WEST` has a higher percentage of crime rates on a Friday.

Note: Investivation with otber factors can be further done in the Multivariate analysis

2.2 (ex) Variable 2

It is important to analyze on the descriptions of the subjects that is recorded by the police to analyze the patterns. Understanding the patterns of types of offenses for which the subject was charged for may help on controling certain types of crimes.

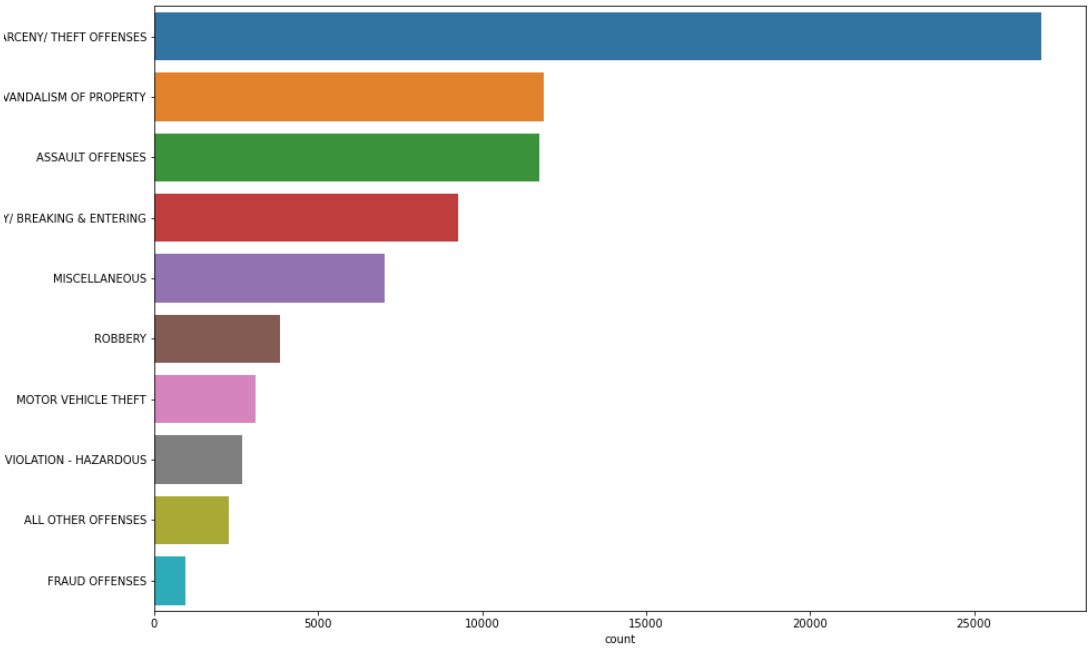

Figure 2. Crime Categories

- From Fig 2. we could notice that a significant proportion was accounted to `Theft` and a small proportion to `fruad offences` .

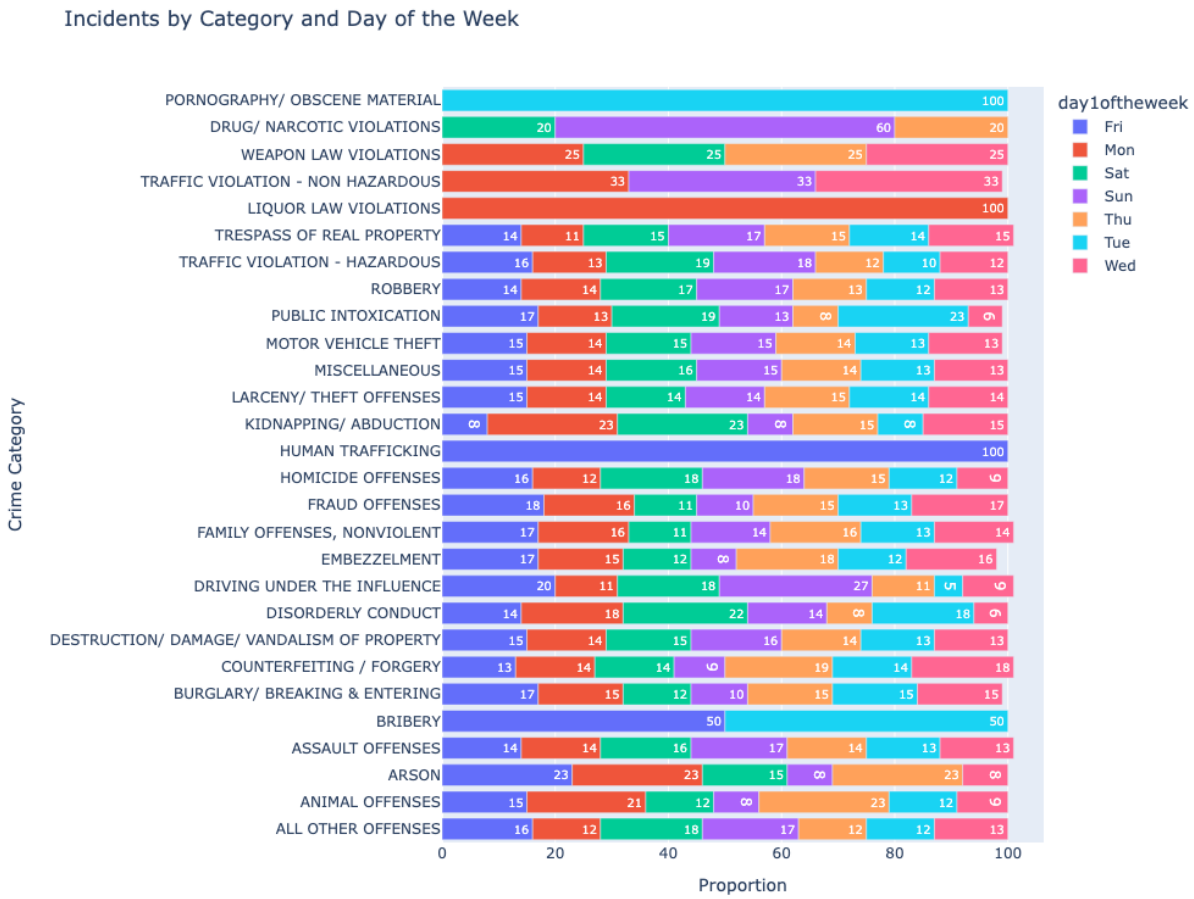

Figure 3. Crime Categories that vary on different days

- In Fig 3, the analysis on different categories of crimes description are done on different days of the week.
- It is interesting to note that 100% of the crime category of `Pornograph/Obcene material` was recorded on a Tuesday and `Human trafficking` on a Friday. This can give account to two reasons: (1) Due to the difficulty of tracking the two types of crime (2) Due to police shifts for the following

3. Multivariate analysis

The dataset contains various informations with various factors on the subjects, crimes and the police officers who dealth with the incident. With the location recorded for the incidents, we can put together the information to coordinate the points on a geographical location to **analyse trends** in the city of Dallas.

3.1 Clustering

Through the geographical locations we can identify the certain areas with certain crimes that are higher, therefore the police can take more measures on controlling it. As there are too many factors, it is almost impossible to analyze all at once so through visualization we can narrow down the options.

My Map

Here is an interactive map:

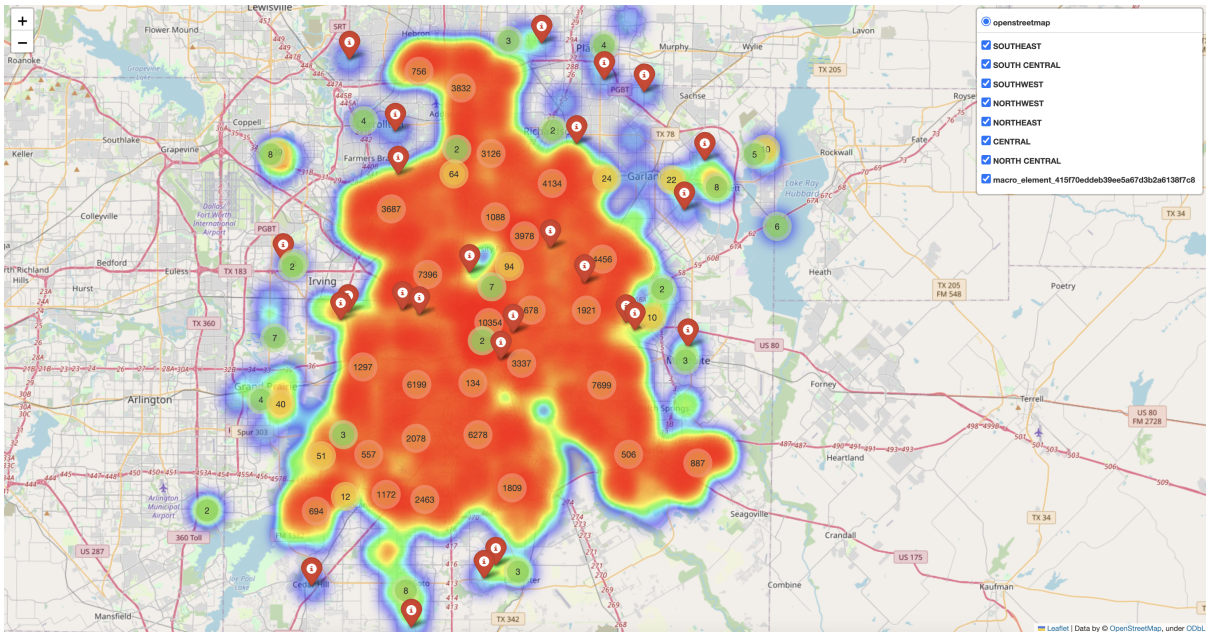

Figure 5. Interactive Map that shows crimes by divisions

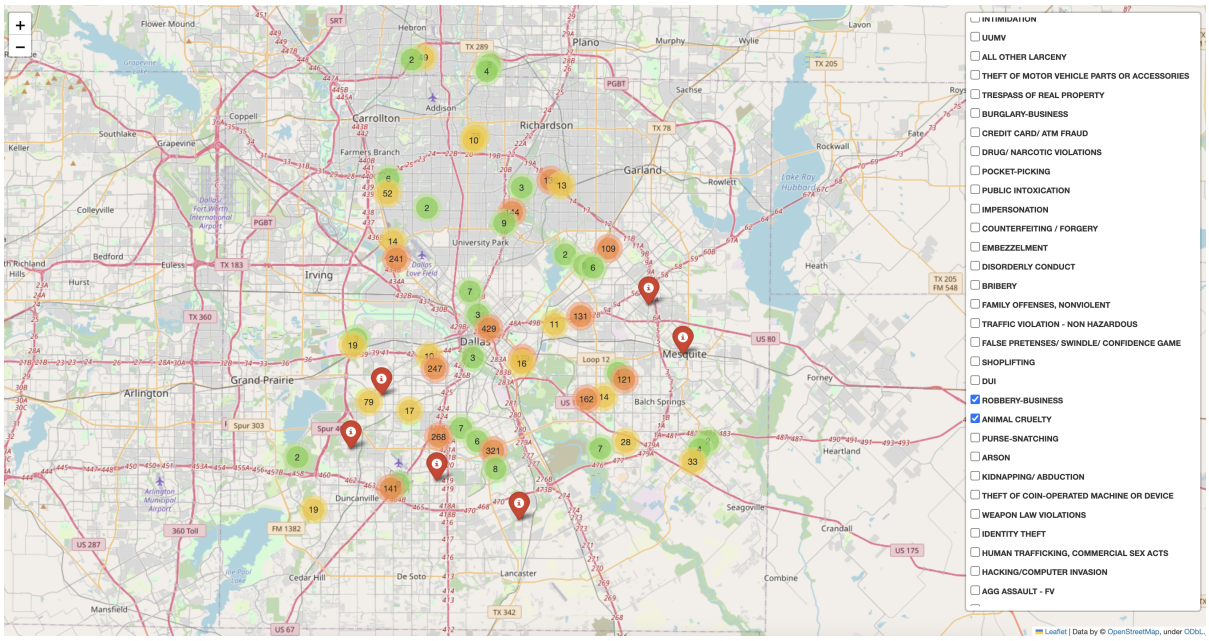

Figure 6. Interactive Map that shows specific crimes in specific regions

4. Suggestion

Tackling the increase in crime rate is hard. But through these analysis we can understand the patterns and from this we could see highlights about various information about the incidents that are recorded in the day. From this analysis we can mainly understand on the divisions that have the highest crime rates. In addition, on the different categories of crime. Futher criminal profiling according to the races may help control the number of offences.