Crime in Chicago

By Snehal, Cesar, and Ubemio

Problem Statement

- Chicago has one of the highest crime rates in United States.
- The city's crime rate is higher than the US average .
- Our project aims to explore crime patterns in Chicago and build model to predict
  1)whether an arrest is possible
  2)the type of crime

Data Source

- The dataset that we used to solve this problem was:

  Chicago crime data of 2019

  https://data.cityofchicago.org/Public-Safety/Crimes-2019/w98m-zvie

  Chicago crime data of 2018

  https://data.cityofchicago.org/Public-Safety/Crimes-2018/3i3m-jwuy

- In both our crime datasets from 2018 and 2019 we had approximately 500,000 observed crimes and in these data sets we have the following variables: ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordinate, Year, Updated on, Latitude, Longitude, Location.

Data Sources

- Chicago police stations data location- https://www.chicago.gov/city/en/depts/cpd/dataset/police_stations.html

- The Chicago police station data location set has 23 police stations with the following variables:

  District, District Name, Address, X coordinate, Y coordinate, Latitude, Longitude, Location.

- We used the distance between the police station and crime location as one of the attribute for our prediction

Data preprocessing

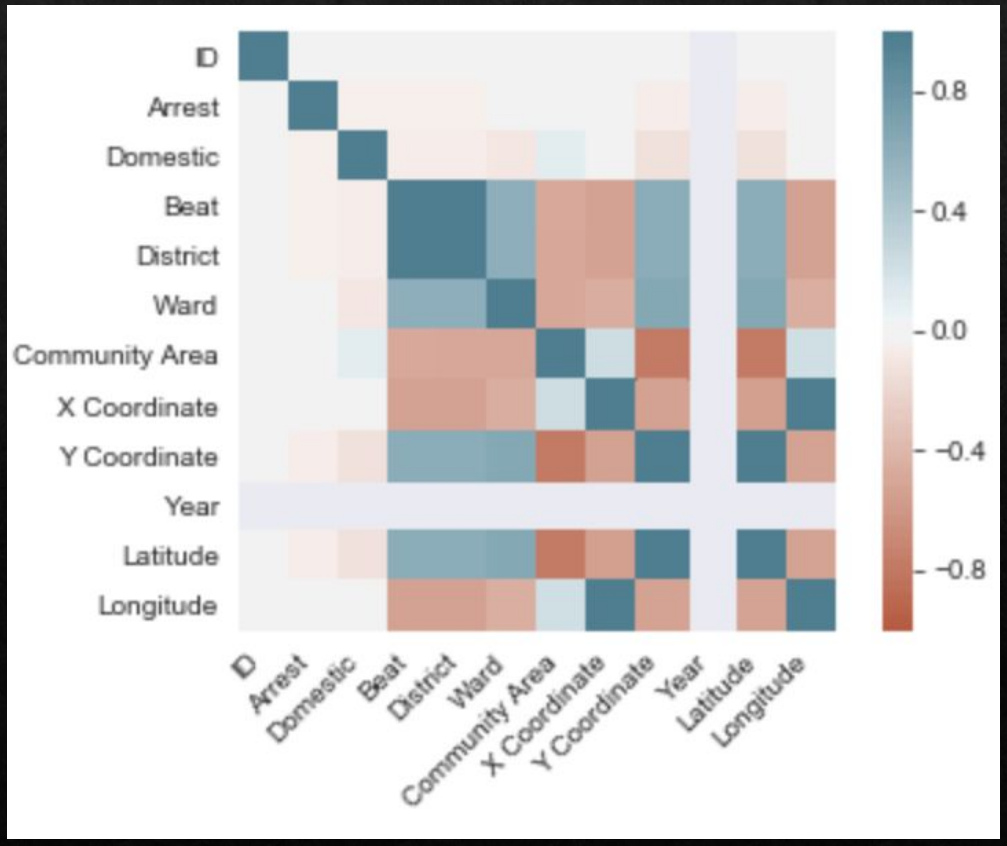
- Columns with missing values: Latitude, Longitude,X Coordinate, Y Coordinate and Location Descriptions.
- We dropped the rows containing missing values
- Other irrelevant attributes like case number, FBI code were dropped.
- Mapped the string labels to numerical values

  Arrests (0-False and 1- True)

  Domestic (0-False and 1- True)
- Dummy variables were created for categorical variables
- Type of crime were grouped in 5 categories

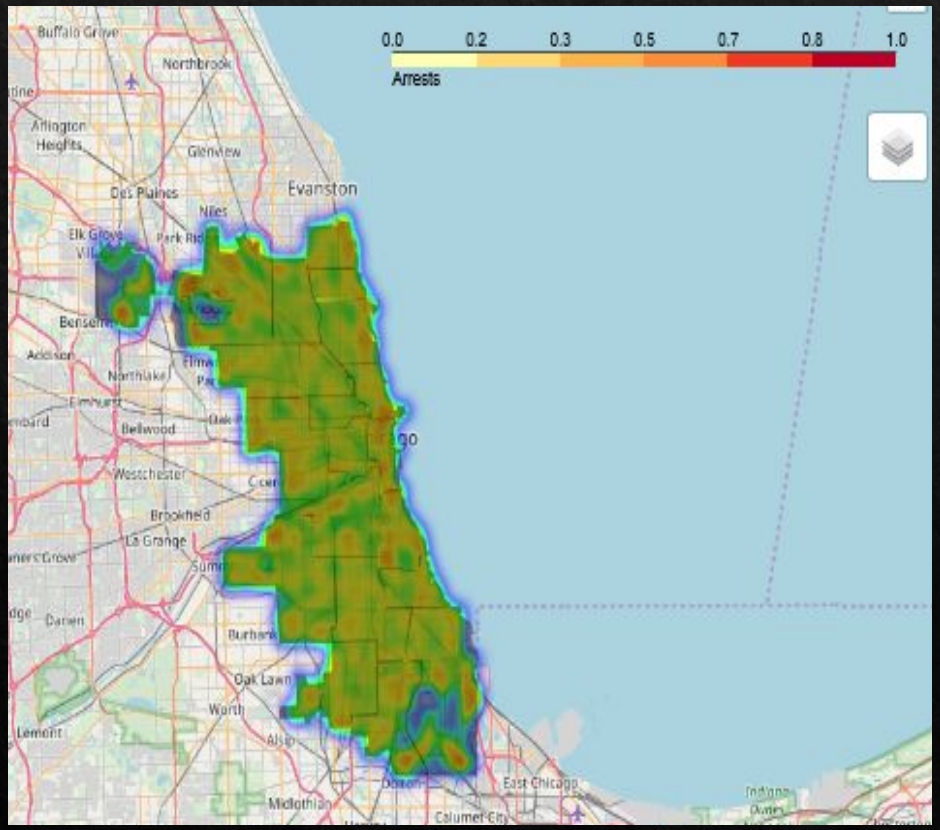
Exploratory Data Analysis

Correlation Matrix : We used the below matrix to find linearly correlated attributes

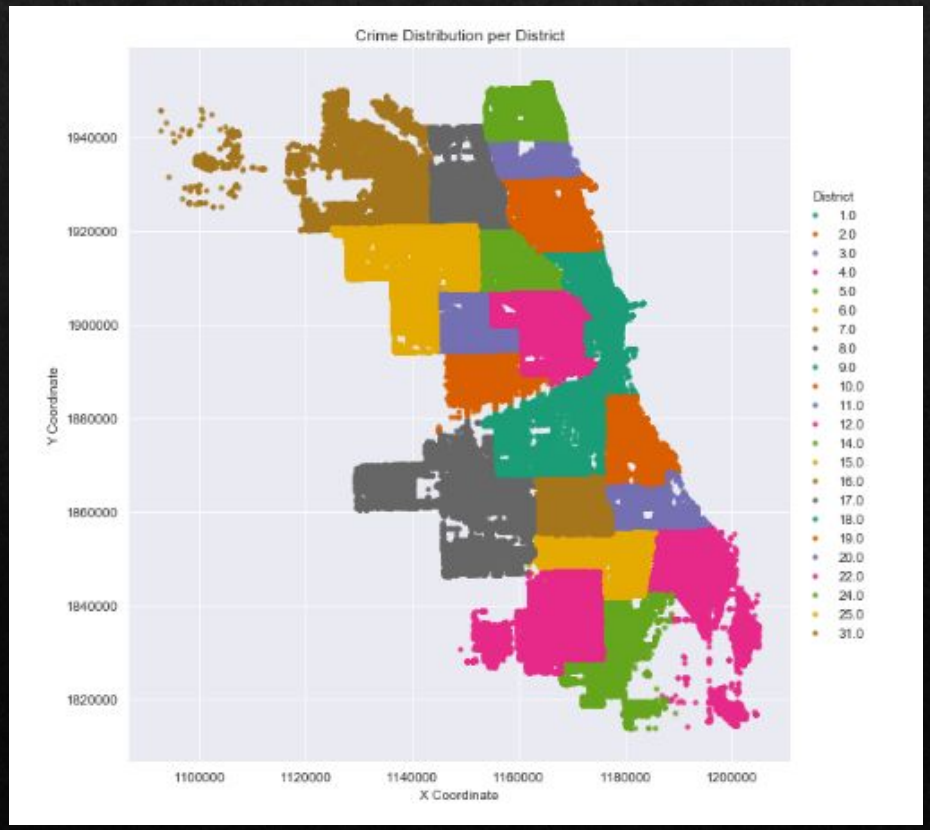


Distribution Map

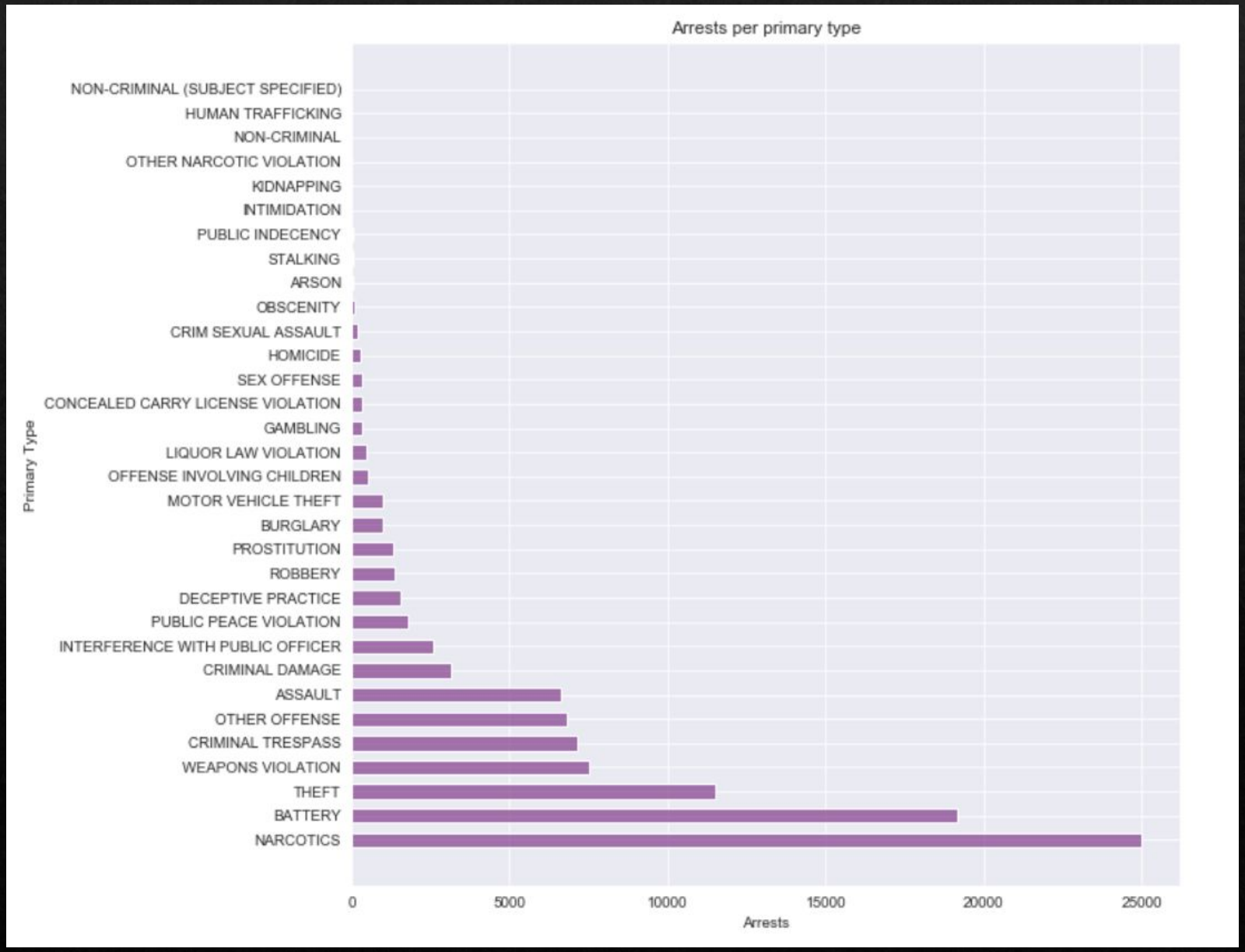
Heat Map of arrests across districts of Chicago



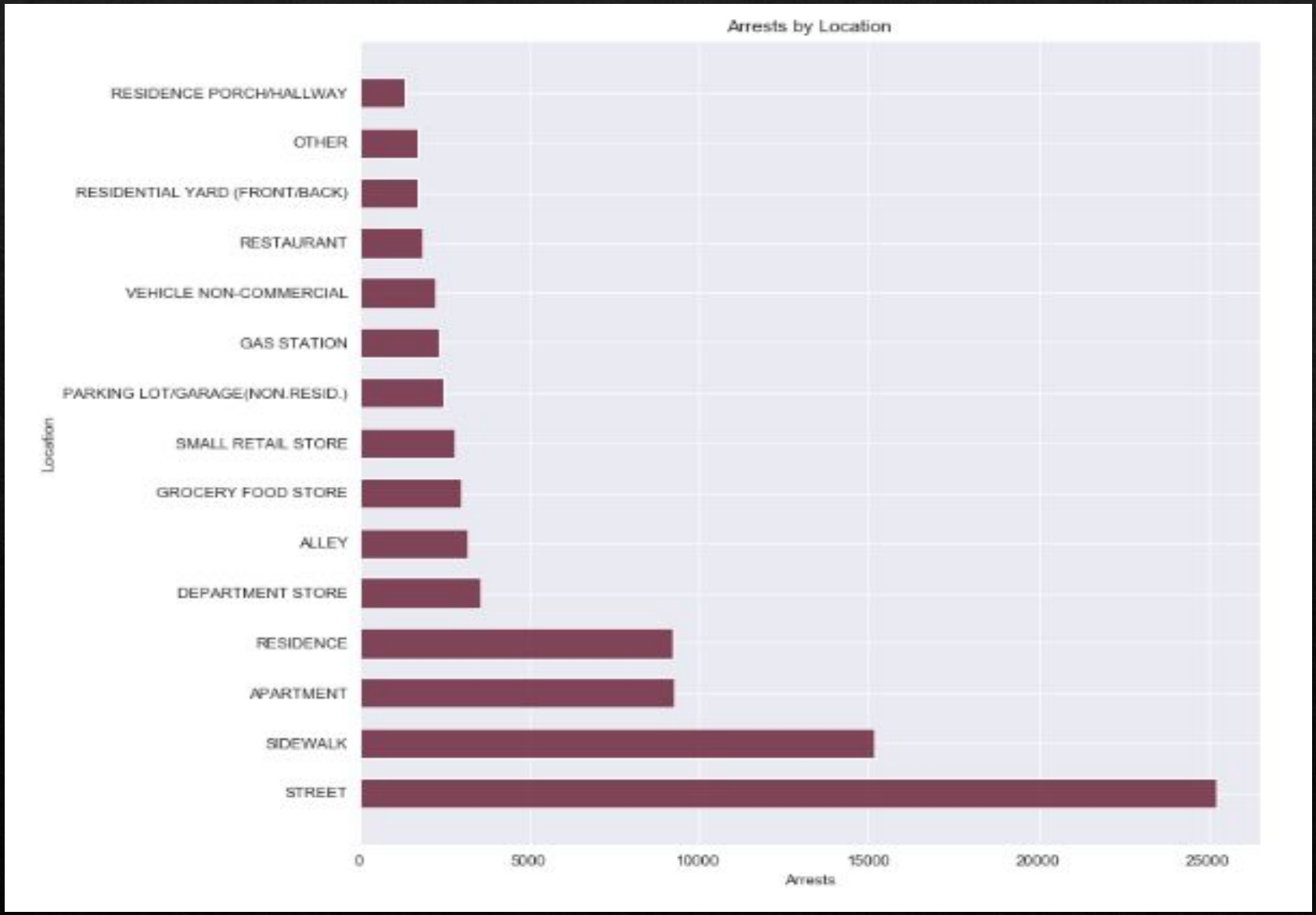
Scatter plot of crimes across districts



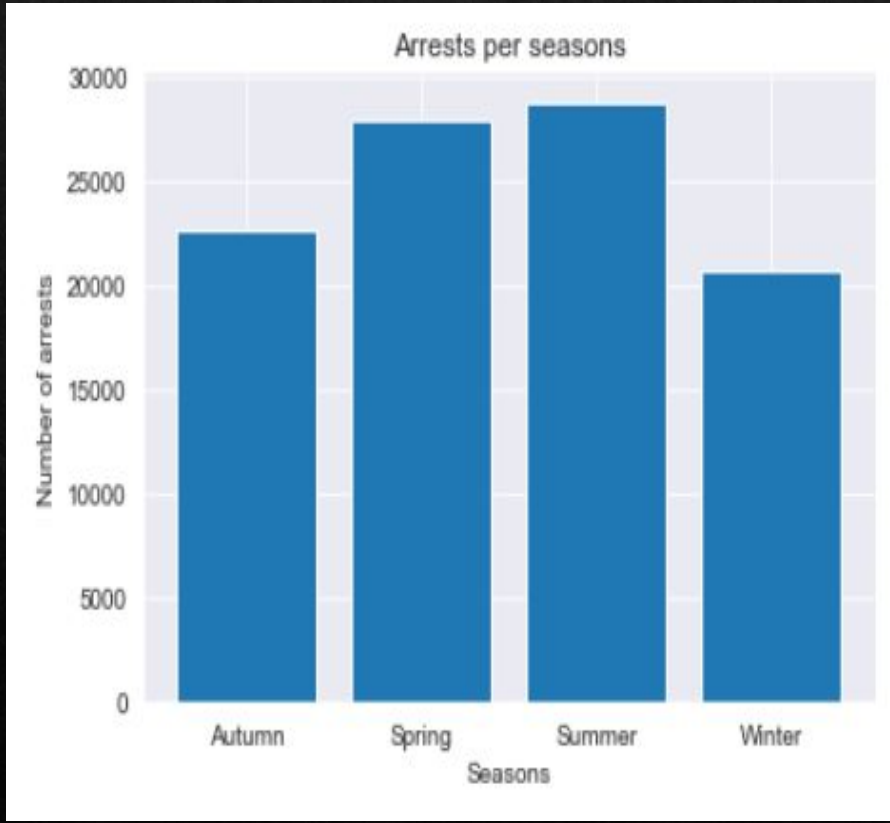
Primary Crime Vs Arrests

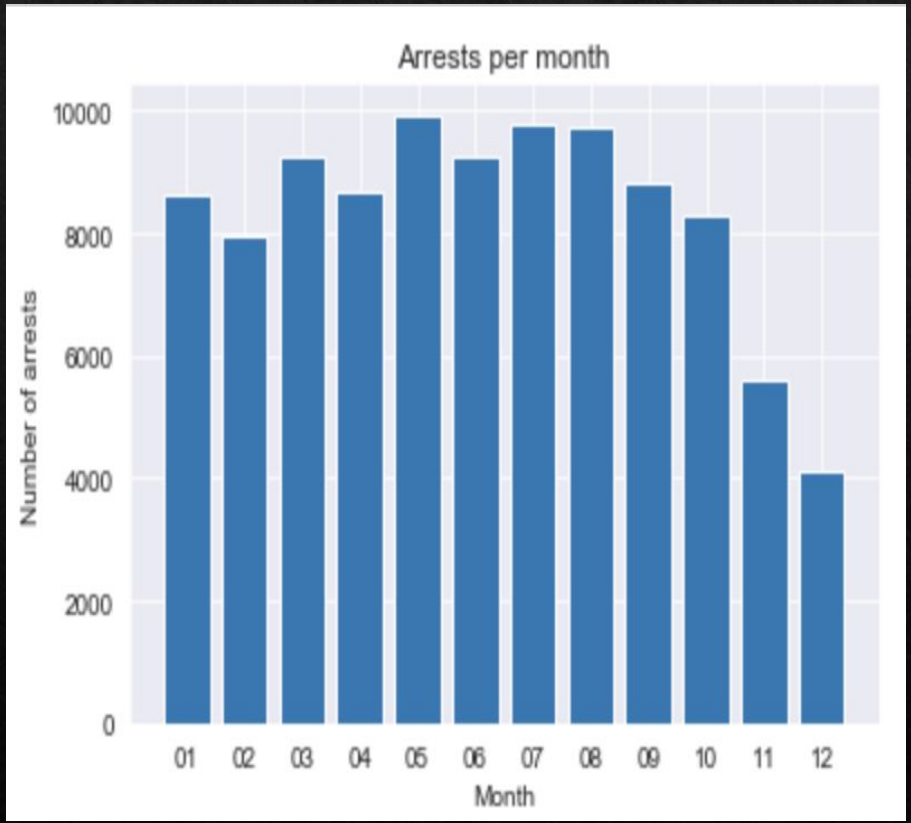


Arrests by location

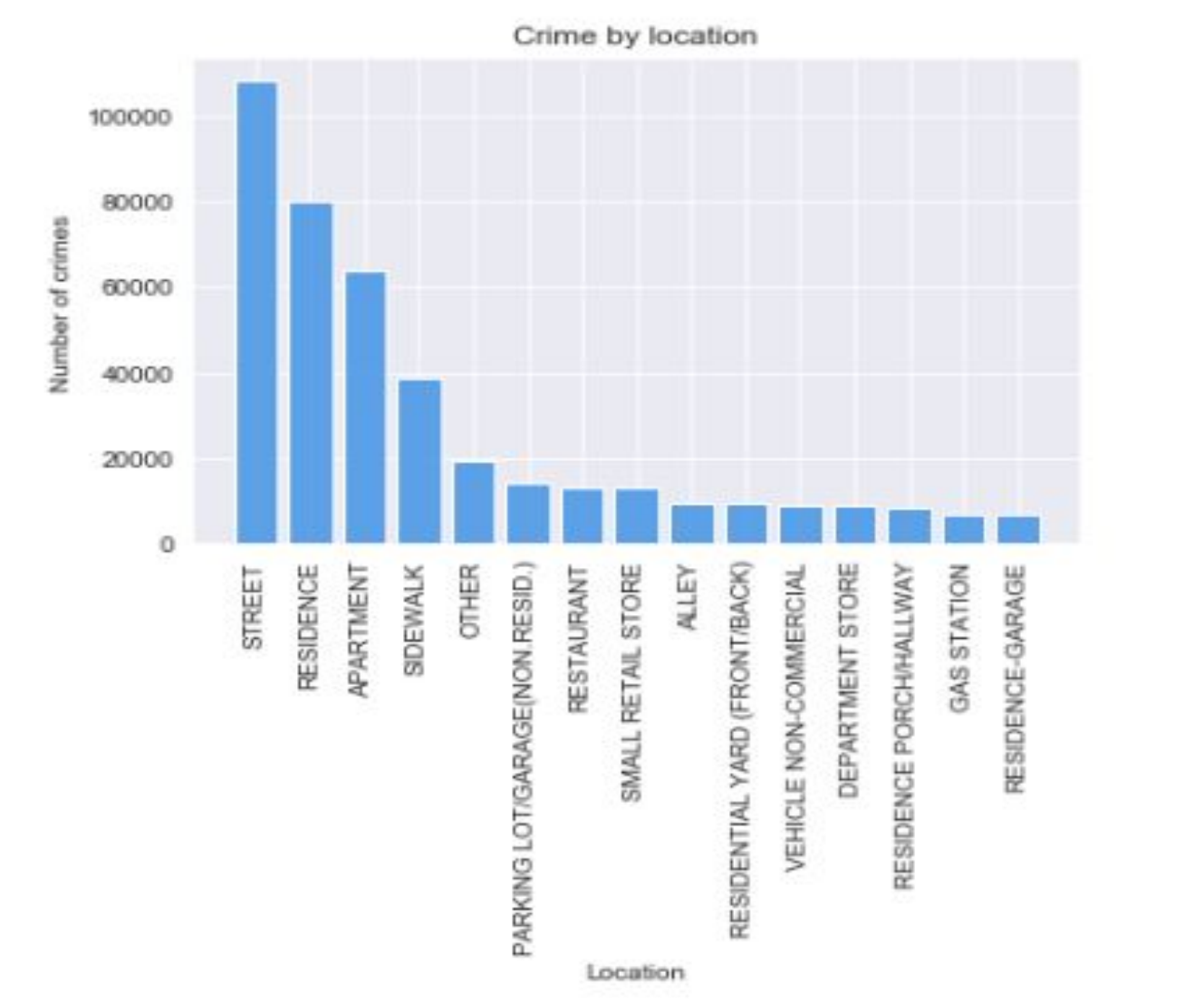


Arrests Vs Seasons and Months





Top 15 Crimes per location



Data science solution

- A classification model to predict whether an arrest will be made.
- And a classification and clustering model to predict the type of crime

How was the data prepared and techniques used to build models?

- 20% of data was used as Test set
- 20% as validation set
- 60% as train set
- Based on the plots and correlation matrix, the parameters and attributes were chosen for building the model.

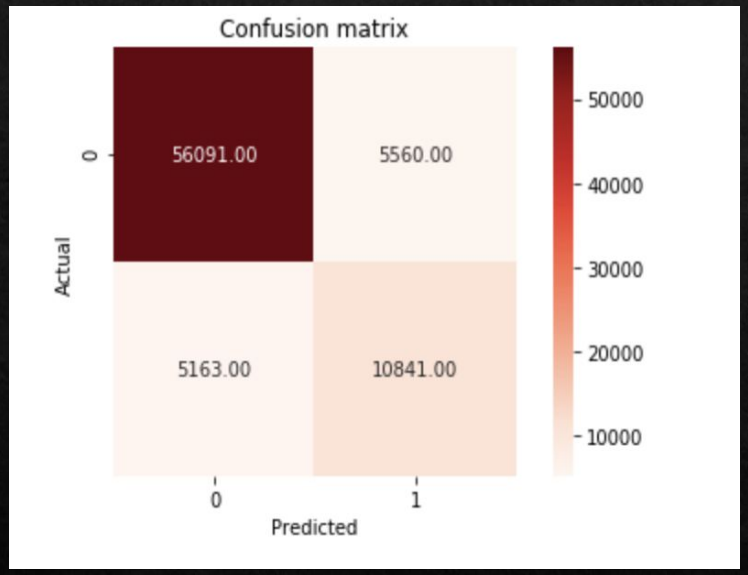
- The classification techniques used were:

  Logistic Regression, Random Forests, Naïve Bayes,

- The clustering techniques we used were:
- K means , K-means++

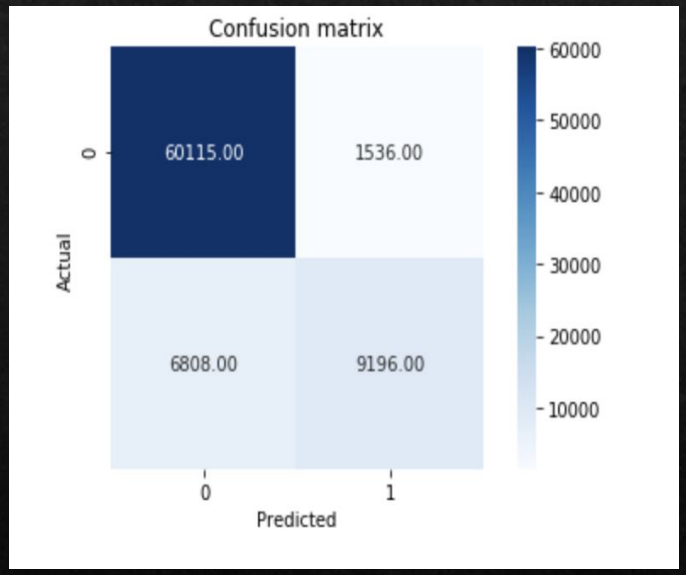
Classification

- Logistic Regression After SMOTE



Accuracy - 86.19%
F1 Score - 86.25%

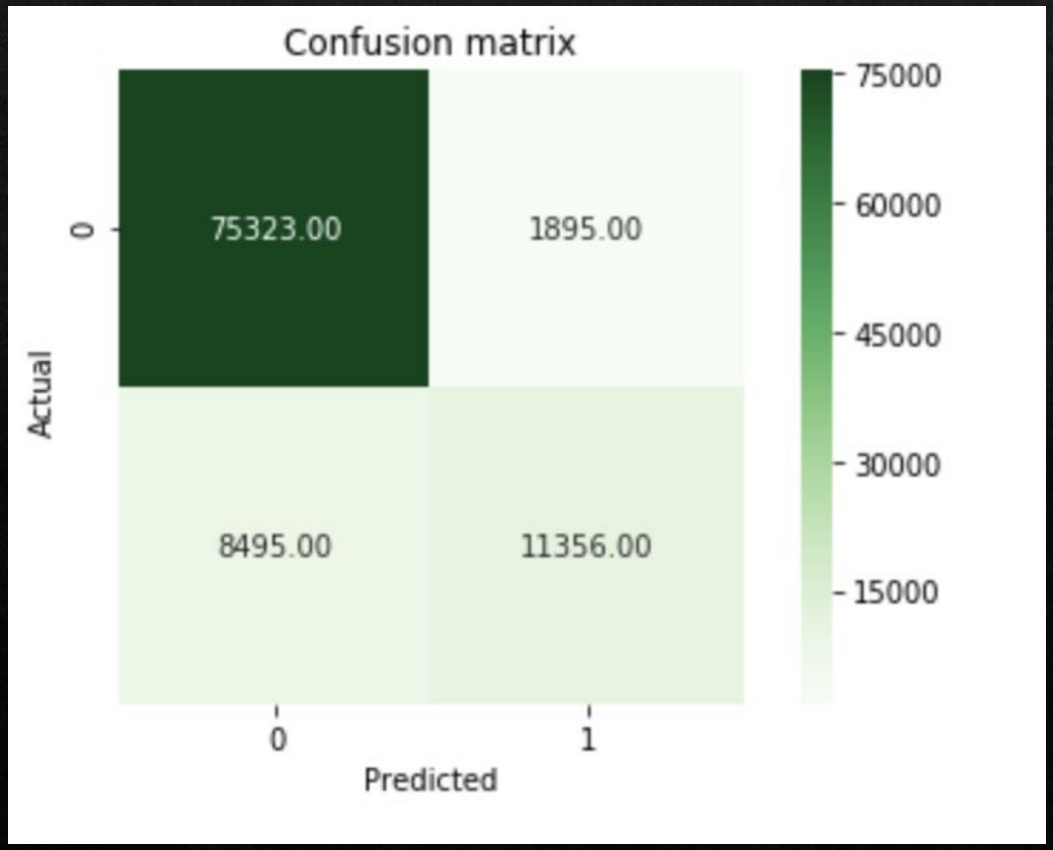
- Logistic Regression after Near miss



Accuracy - 89.25%
F1 Score- 88.41%

Result of Classification

Logistic Regression Final Model

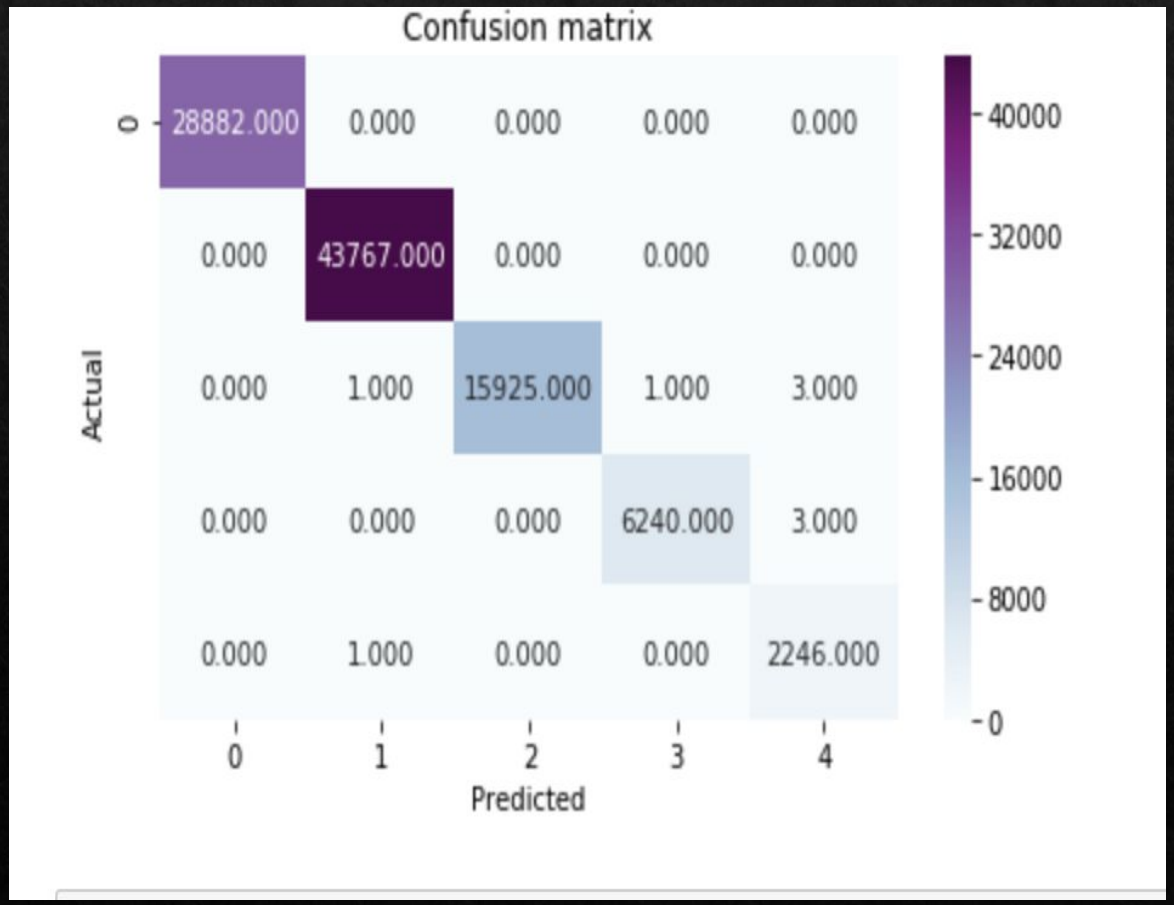


Accuracy- 89.29%

F1 Score- 88.44%

Result of Classification Crime Groups

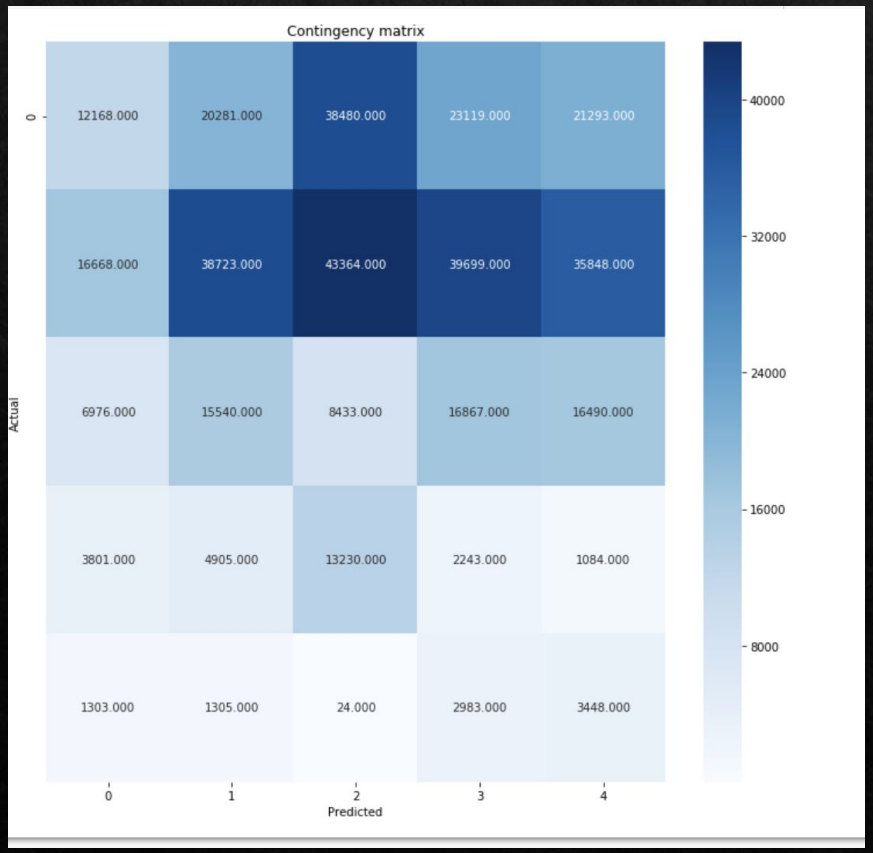
Naive Bayes



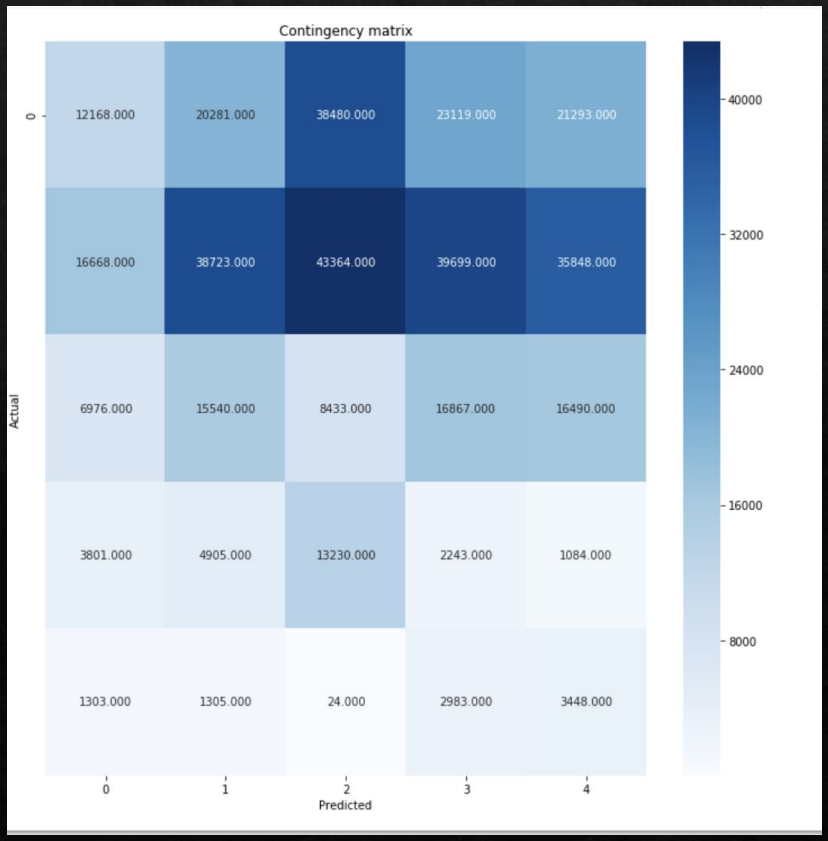
Accuracy- 99.9%
F1 Score- 99.99%

Result of Clustering Crime Groups

K-Means++ [0.0109, 0.2247]



K-means [0.0112, 0.2253]



Conclusions

1. Logistic regression did a good job predicting arrests even though the class was imbalanced.
2. Naives Bayes had a high accuracy in predicting crime groups.
3. Clustering didn't perform good because the data was not well separable and high correlation between crime types.

Thankyou!!