FINAL PROJECT
Crimes in Chicago

Snehal Tikare: stikar2
Cesar Hernandez: cherna83
Ubemio Romero : uromer2

418 Data Science

Problem Statement

Chicago has one of the highest crime rates in the 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

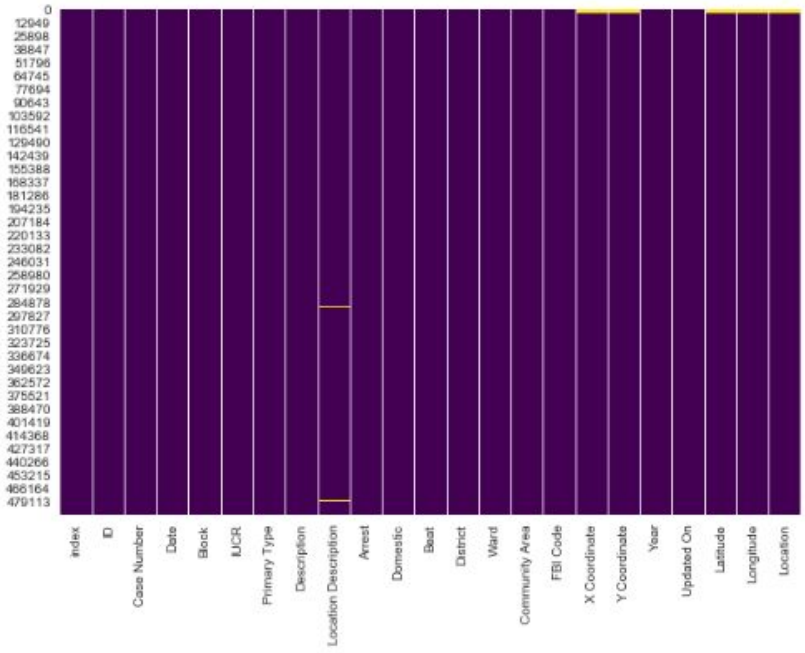
DATA

The primary dataset used in the project is Chicago crime dataset documenting crime records from 2018 to 2019 integrated with data records of Police Station in each district.

- Police_Stations size of (23 rows x 15 col)
- Crimes_-_2018 size of (267,639 rows x 22 col)
- Crimes_-_2019 size of (224,412 rows x 22 col)

PRE-PROCESSING STEPS

1. Preprocessing steps involved taking care of missing values, dropping irrelevant attributes and duplicates
2. The heat map below helps us identify the missing values in the dataset.The yellow area marks the null values in the dataset.



3. Dropped unnecessary attributes that might be redundant

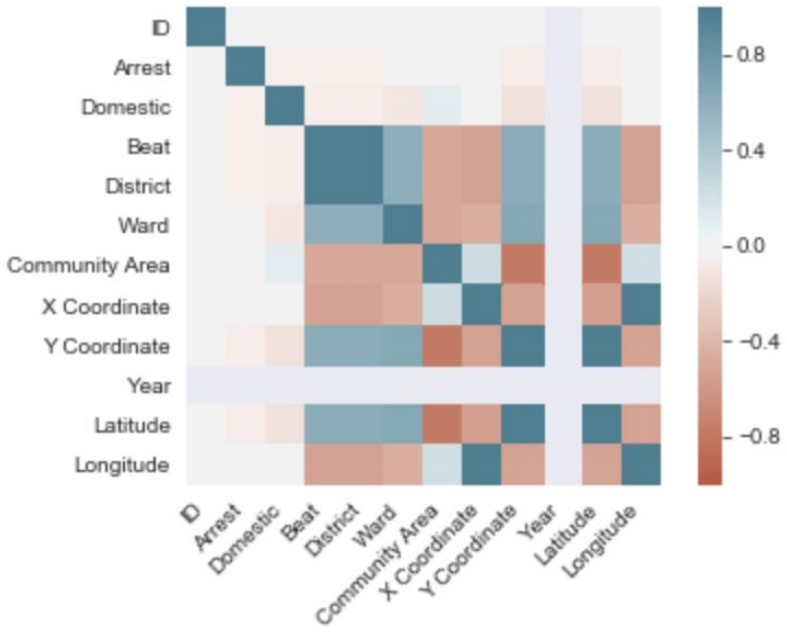
   ex: 'ADDRESS','CITY','STATE','ZIP','WEBSITE','PHONE','FAX','TTY','LOCATION'

4. The main focus is on arrest and the below figure indicates the number of arrests made in the year 2018-2019.

```
False    391229
True     100822
Name: Arrest, dtype: int64
```

-

5. Given the arrest information it is also good to know what type of crimes were committed.
6. Dropping null values as a preprocessing step before building the classifier
7. The correlation matrix shows the attributes that are correlated to each other. The next step was to get rid of one of the highly correlated attributes to avoid redundancy



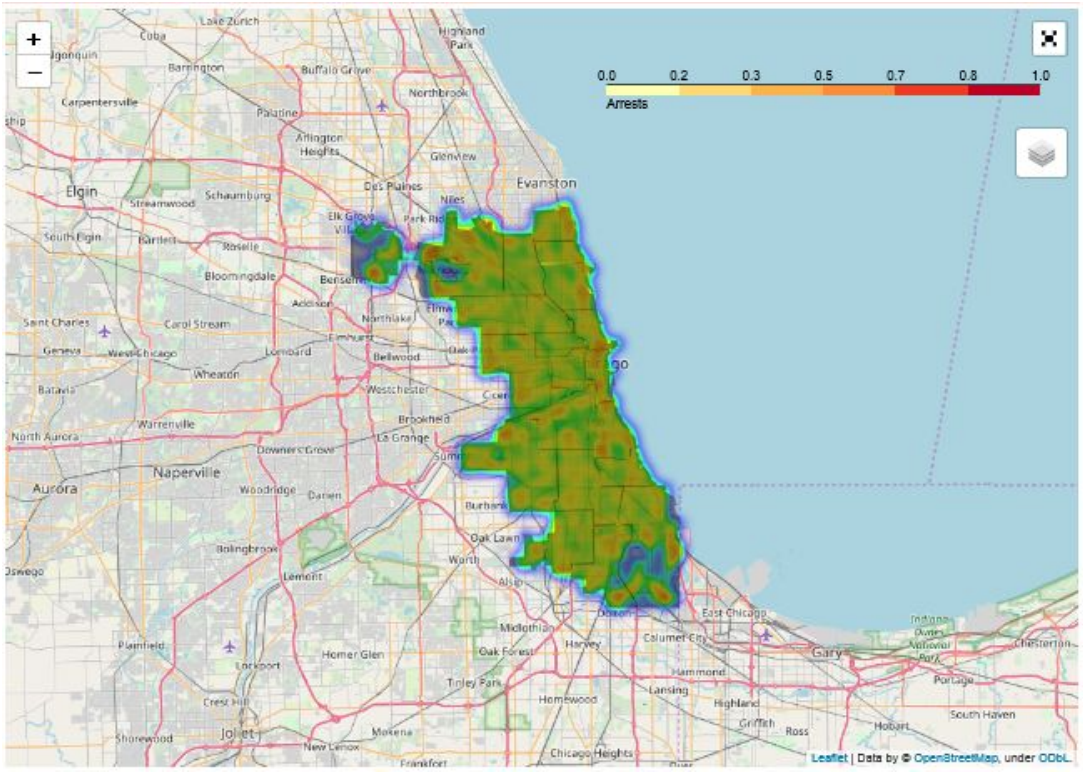
8. Since the goal was to have one merged data set by District, attribute District has to have no duplicates.
9. Merged data set is complete and has a size of 485411.
10. Mapped the string labels to numerical values

    Arrests (0-False and 1- True)
    Domestic (0-False and 1- True)

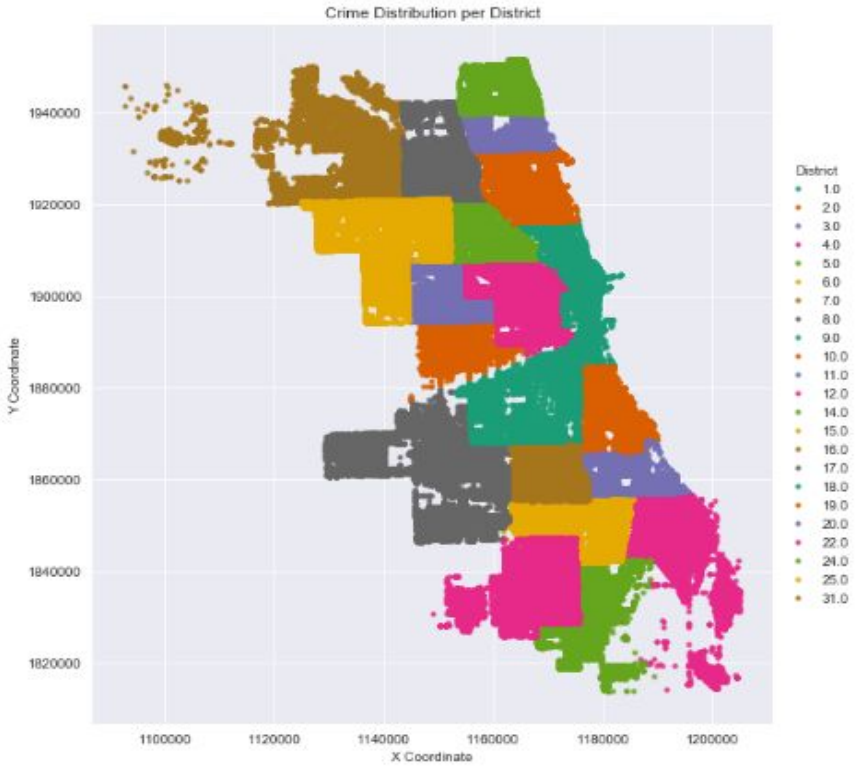
11. Dummy variables were created for categorical variables
12. Type of crime were grouped in 5 categories

Data Visualization

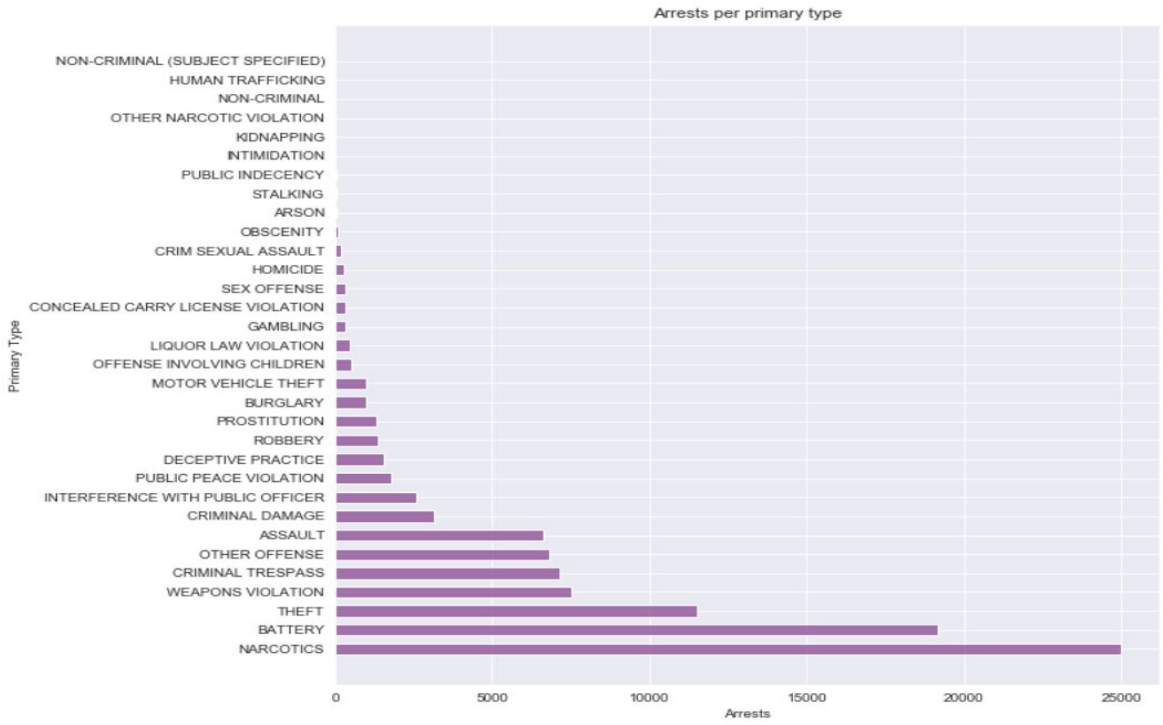
1. To Visualize the data, a performed exploratory data analysis is required to better understand how the crimes are spread across different attributes.And how Arrests are affected.
2. The below map shows a scatter plot of crime distributions across various districts.
3. The Heat Map indicates that arrests that have happened across district.These map help us visualize Arrests against crimes across district.



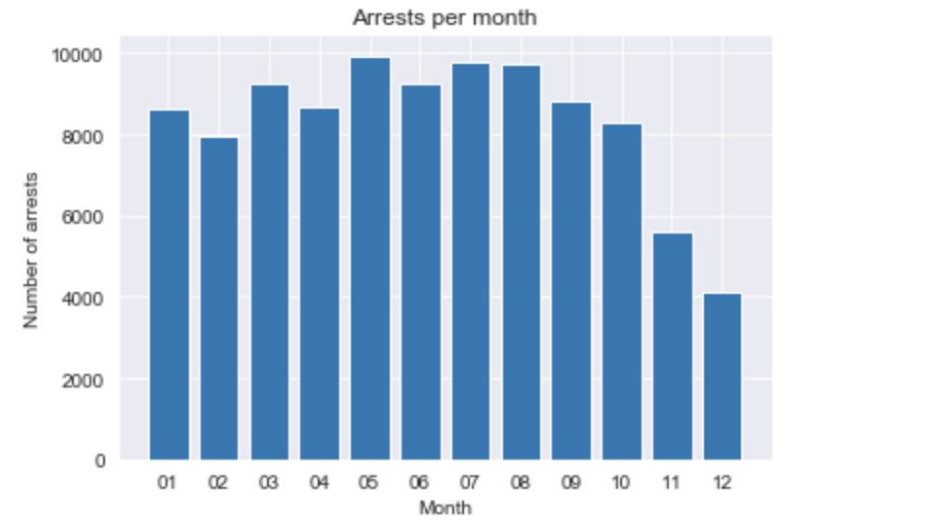
4. The map shows the scattering of crimes at different districts



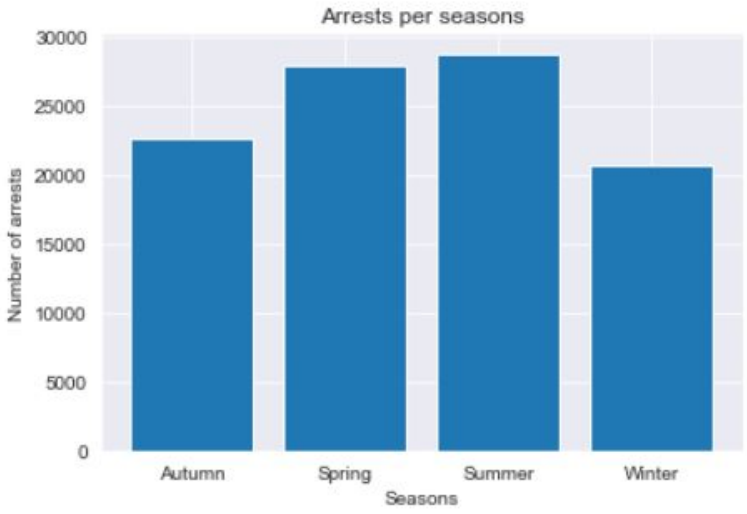
5. The Bar plot shows the number of arrests made for each crime type.Maximum number of arrests were made in criminal activity related Narcotics.



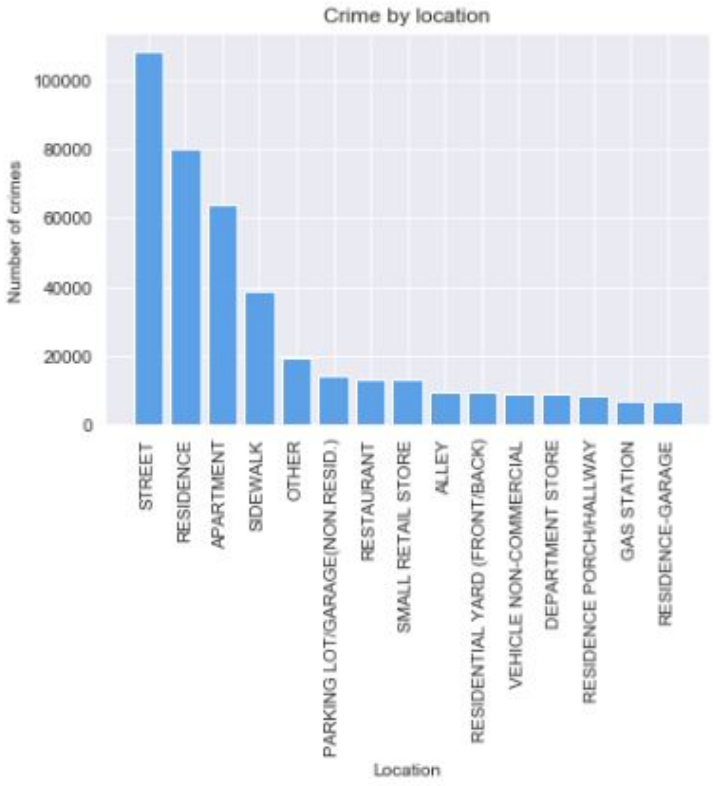
6. The graph plots the arrest per month. Noticing that high amount of arrest were made during the months April-September(the summer)



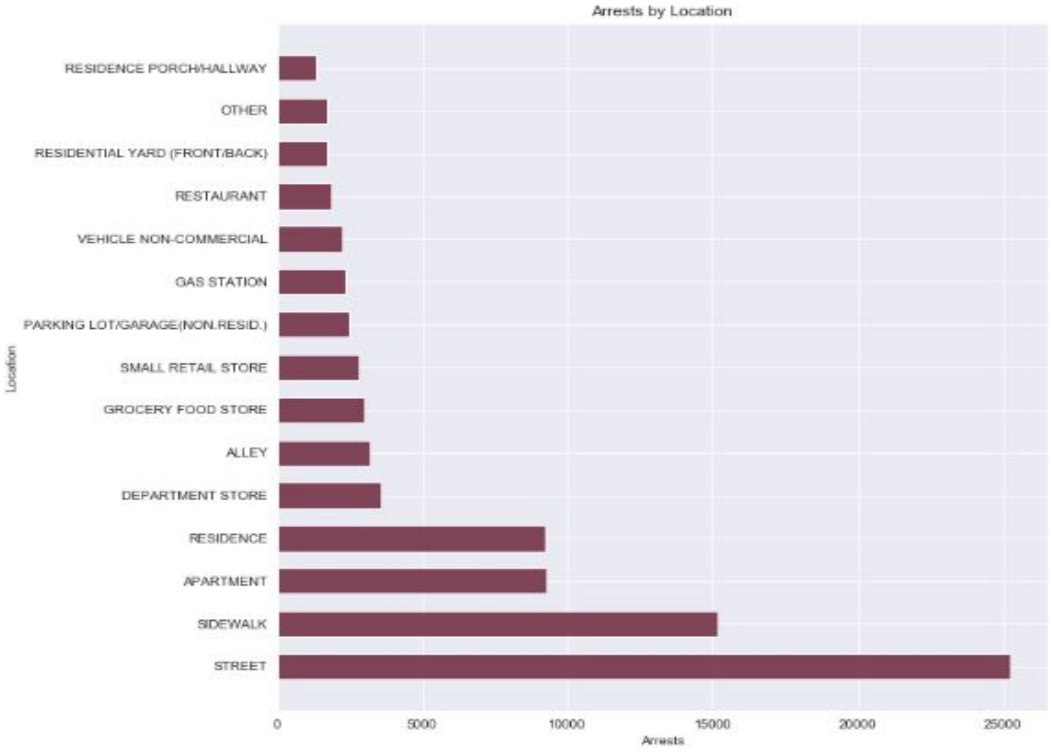
7. Also, the data was grouped into Seasons- Autumn, Spring, Winter and Summer. The below graph shows the arrests made in each season.



8. As previously from months, Summer is the season where a lot of crime takes place in chicago given from the data that consist of years 18' and 19'
9. The top 15 locations where highest numbers of crimes takes place. Noticing a high number of crimes taking place in the streets attribute.



10. Number of arrests that were made at each locations



CLASSIFICATION

The data was prepared for classification by dropping irrelevant columns in like IUCR, Beat, Ward and X coordinate etc. Because they do not assist in helping predict whether an arrest is made or not. The next step is to make a train(60%), test(20%) and validation(20%) sets to run on the classifiers after scaling the data and then the data is ready to be ran with all the different types of classification techniques and see their performance scores.

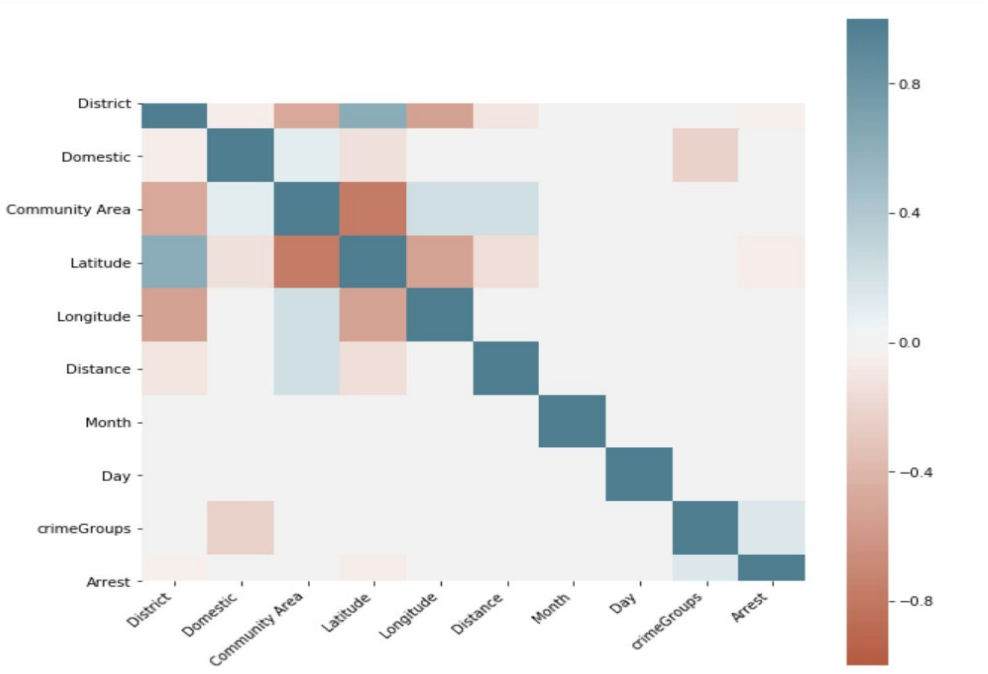
Main goal then is to predict the arrest and crimeGroups on the committed crimes from the reported crimes.

- First approach was to use a couple of methods to validate predictions and that was by Naive Bayes, Random Forest, Logistic Regression for Arrest.
- For crimeGroups, the following methods used: Random Forest and Naive Bayes

Classification Steps

Dropping irrelevant columns

1. 'IUCR','Beat','Ward','XCoordinate','YCoordinate','p_x','p_y','p_latitude','p_longitude','CaseNumber','Block','FBIcode','Date','p_dname','Time','Season'
2. Verified by correlating the attributes that will be used for classification making sure they are not redundant.



3. Next for categorical attributes, they're in object type therefore want to convert them to dummy variables
4. Now the process of splitting data into training and test can begin.
5. For predicting Arrest using predictors
    - District, Primary Type, Description, Location Description, Domestic, Community Area, Latitude, Longitude, and Distance

    For predicting crimeGroups the following attributes are used

    District, Description, Location Description, Domestic, Community Area, Latitude, Longitude, Distance, Seasons, and IUCR
6. Finally scale/standardize data after this step and proceed with classification/Clustering methods.
7. After using Logistic Regressions and Random Forest, there are a couple of discrepancies. Data is unbalanced. Therefore further steps like SMOTE and Near Miss are used to balance the data
8. A final check is done by performing a test with test set data and check for optimal performance.

Predicting Arrest

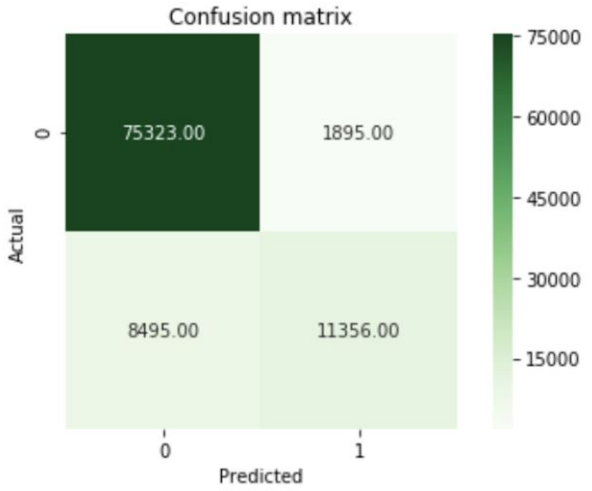
Techniques used for predicting arrest were

1. Logistic Regression
2. Naive Bayes
3. Random Forest

Each of these techniques were built with the original data, Over sampled data and Under sampled data

1. Over sampling technique used : SMOTE
2. Under sampling technique used : Near Miss

Out of the classifier Logistic regression performed the best with accuracy 89.29% and F1 Score of 88.44% even though the data was imbalanced

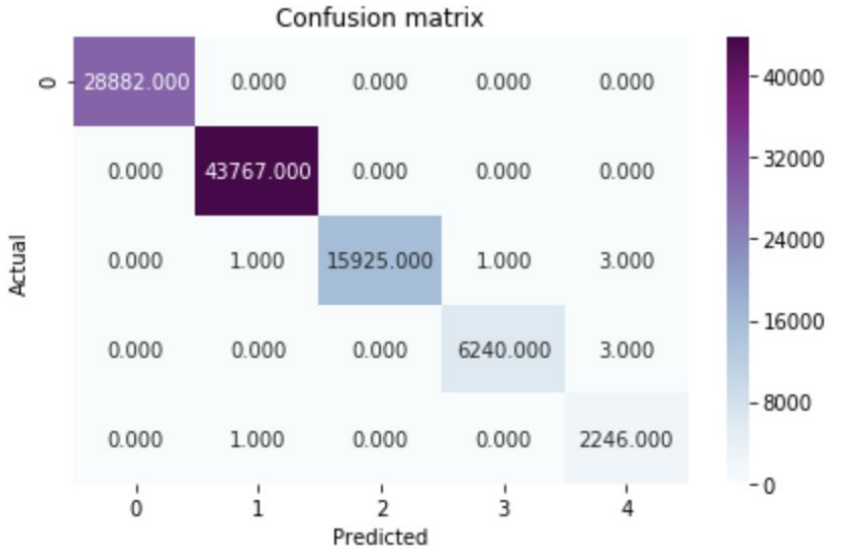


Predicting for crimeGroups

Techniques used

1. Naive Bayes
2. Random Forest

Predicting crimeGroups using classification method indicated that Naive Bayes test case has a better performance than Random Forest with an accuracy score of 99.99% and an F1 score of 99.99%

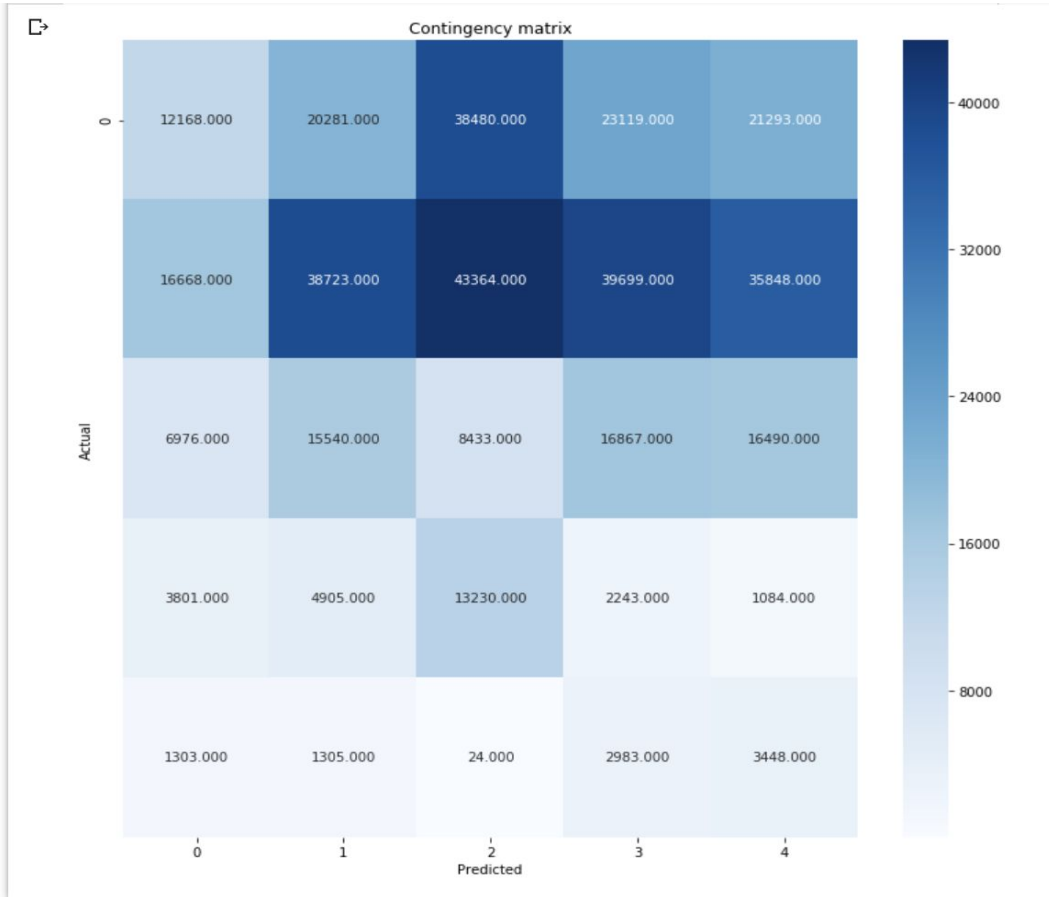


CLUSTERING

K-Means

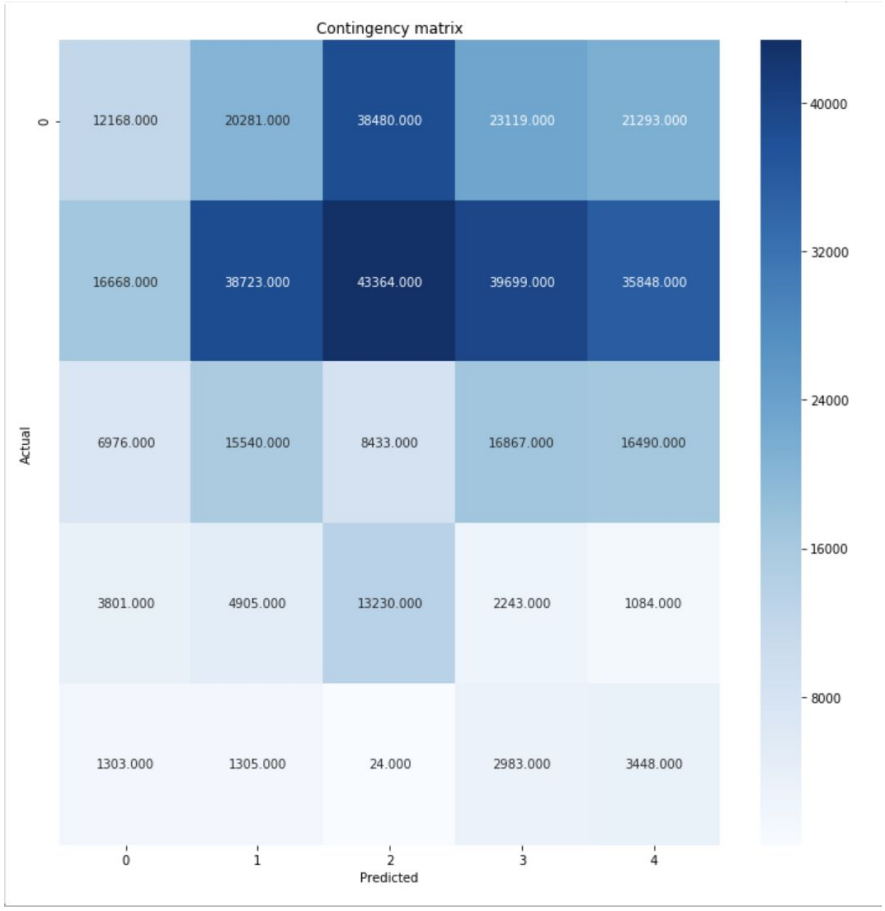
- For this step, K-Means is used to cluster crimeGroups by the amount of Arrest.
- Using K-means++

```
clustering = KMeans(n_clusters = 5, init = 'k-means++', n_init = 20,random_state=0).fit(X_scaled)
clusters = clustering.labels_
```

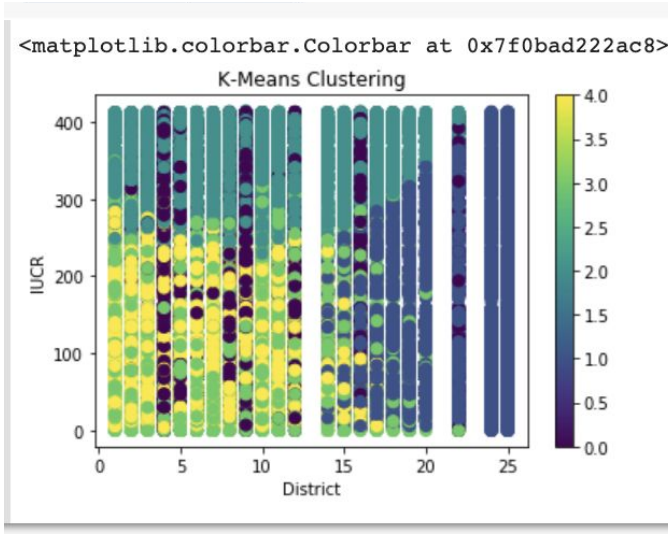


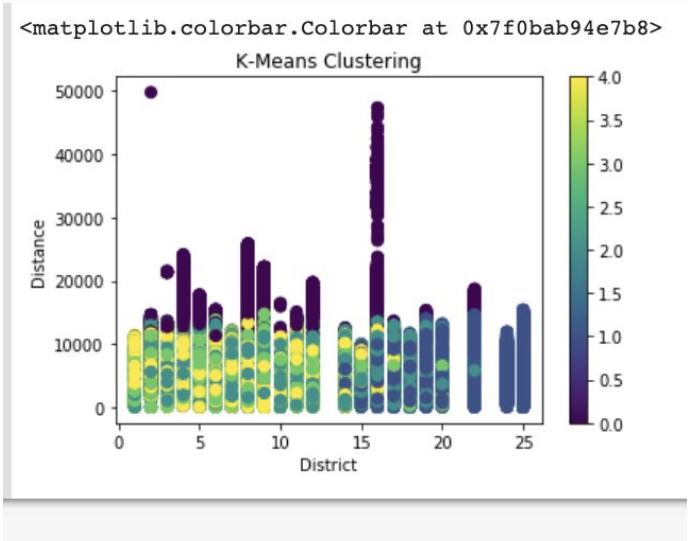
- Using K means

```
clustering = KMeans(n_clusters = 5, init = 'random', n_init = 10,random_state=0).fit(X_scaled)
clusters = clustering.labels_
```



Scatter plot of the clusters





- The step for clustering was a little difficult. After trying to cluster for crimeGroups, based on the contingency matrix the predicted and actual parameters were indicating that there are other higher values on different clusters.
- To further observed this case, a plot was created to visually see whether certain attributes were clustered correctly or not.
- For example, IUCR and District were plotted and based on the plot, district 25 illustrates that a certain primary type only exists in that cluster. That is inadequate since it's safe to say that there is no utopia or perfect district where either no crime exist or only one certain type of crime exist.
- Another example used to plot was clustering District and Distance. Based on the observations, District 0 and 5 had some outliers. Therefore it's uncertain whether this information is reliable. Again, it's noticeable that district 25 is only showing one type of crime. Surely there is some sort of mistake. However it was difficult to detect this error.
- From our earlier classification model, Naive Bayes used to predict crimeGroups demonstrates that there should be a reliable model based on accuracy and f1 score. However when applying this into a cluster method, this was not the case.

CONCLUSION

- Regarding Classification, Logistic Regression gave the best performance predicting Arrest using all respected attributes. As a result, after applying validation set and test set with “Logistic Regression”, an accuracy score of 89.25% and an F1 score of 88.446%.
- Regarding Clustering, the models created did not show a promising result. In fact the results were very low and are not fit for model. The clustering method that was used to predict a crimeGroups was K-means. It’s not really clear why the data was not accurate. Based on the classification method used to predict crimeGroups, did not show the same results. In fact it had an accurate score of .9998. The only problem as of now that can be detected was the way we grouped/categorize the primary type of crimes into groups. Perhaps some sort of biased went in when grouping the type of crimes.

REFERENCES

1. https://data.cityofchicago.org/Public-Safety/Crimes-2019/w98m-zvie
2. https://data.cityofchicago.org/Public-Safety/Crimes-2018/3i3m-jwuy
3. https://www.chicago.gov/city/en/depts/cpd/dataset/police_stations.html