**Crime Analysis in Chicago from 2012 to 2017**

**Abstract:**

In today’s world, crimes are an integral part of our lives. Every day the print and Social Media are filled with Crime Incidents to make us aware and cautious of the different techniques used by Criminals. Crime has now become an international concern. However, it is documented and handled in very different ways in distinct countries. No more, being cautious and improve safety is a simple instruction anymore. The use of modern technology and data science techniques wisely can act against this problem. Over the years, so many records and documentation are gathered in the police department, which eventually can be used as a vital source of data for the data analytics tasks. To increase the safety of our society and lower the crime rate we can apply the analytical tasks to these data and bring us valuable information.

In the United States, Federal Bureau of Investigation (FBI) records the violent crimes and property crimes. Also, documentation of crime is done separately by each city. However, some of them release data regarding crime rates. In the United States of America, Chicago is the third most populous city with a population of over 2.7 million people. For this project, we will be conducting an analysis of crime in the city of Chicago. For the analysis work of this project we will utilize the data from 2012 to 2017 in order to avoid using outdated data, and to provide relevant information about the state of crime in Chicago in recent years. To analyze the data, we used Python, R and SQL frameworks.

However, after the analysis of the dataset we concluded that, there is a sharp decrease in the number of crimes reported over the years in Chicago which shows us that the security of Chicago has increased during these years. We infer that the crime rate on an average remains the same throughout the week irrespective of the day. But in numbers number of crimes reported on Tuesday are slightly more than rest of the day. Our reports also say that most crimes were reported in the month of January. The locations of highest Crime rates were found to be Michigan Ave, Halsted St, Pulaski Rd, Street and Western Ave. Additionally, our results show that crimes are usually reported in apartment, residence, sidewalks, and streets which are very common places. The ward number 28 is more potentially dangerous as most number of crimes were reported in it. We observed that only 28% of the accused are getting arrested. It was also concluded that 2012 had the highest crime rate.

**Introduction:**

The main objective for choosing this dataset is that crimes are a part of our present society and every day we listen and see these crimes. Data plays an important role to know the crime rate and to observe many hidden patterns. Pairing technology to bring a solution to this problem helps in decreasing the crime rate. The primary interest of this project is to identify many unobserved patterns which helps in concluding weather the crimes in the city are increasing or decreasing for each year in the state Chicago. The dataset depicts about the crime incidents occurred in the state of Chicago from 2012 to 2017. Data is collected from Chicago’s Police Department. It also gives an indication about the time period at which most crimes are committed also severity of these crimes. Also how socioeconomic indexes are affecting the occurrence of these crimes.

By exploring the dataset, the following research questions were observed and they are:

* How has crime evolved over time in the city of Chicago?
* Which day of the week and month do most crime occur?
* In which locations of the city is crime more likely to happen?
* Which wards are more potentially dangerous?
* In how many incidents the accused are getting arrested?
* In which year most of the incidents took place ranging from 2012-2017?
* Find the highest crime type from the year 2012 to 2017?

After the analysis of the dataset we concluded that, the crime rate is falling over the year in Chicago. We infer that the crime rate on an average remains the same throughout the week irrespective of the day. But in numbers number of crimes reported on Tuesday are slightly more than rest of the day. Our reports also say that most crimes were reported in the month of January. The locations of highest Crime rates were found to be Michigan Ave, Halsted St, Pulaski Rd, Street and Western Ave. The ward number 28 is more potentially dangerous as most number of crimes were reported in it. We observed that only 28% of the accused are getting arrested. It was also concluded that 2012 had the highest crime rate.

It is important to know the different types of crime and also their frequency of the crimes committed it helps in deciding whether the crime rate went up or down over the years and to find a pattern for these crimes so that many major crimes can be stopped. It is also to find how the crime rate has evolved across many parts of the city.

**Literature Review:**

The following three research reports are relevant to the current study of Chicago Crime Analysis

1. **San Francisco Crime Classification:**

In this paper the factors considered for crimes happening in San Francisco are considered. It also mainly focuses on the predicting future crimes happening in the city. The rapid population has led to increase in more crimes in many red light districts. These crimes are being classified based on many factors like district the crime happened. There are factors like longitude and latitude which tells about where exactly the crime has happened. Category describes the type of crime that had happened. Types of crime is of many categories like theft, attempt to kill someone and many more. In conclusion all these factors are analyzed for the better understanding of the crime and also in predicting the crime.

Relevance to the Project:

Here a lot of factors are discussed and analyzed that will influence the crimes happening in the city. There are many important and common factors. I am analyzing the peak time for crime happening and also at what time the crime could happen at which location. I am also using longitude and latitude factors because of which the crime location can be traced.

1. **Crime analysis in Chicago city**

Security is always one of the most important concern of government for ensuring safety of people. It is a challenge for government to ensure that no crimes are happening. A crime information is a system which can process huge data and predict from the given data what are the crime hotspots and at what time there is a high chance for the occurrence of these crimes. This paper tells about the how the system to identify crime is made and what are the factors which are involved in the crime are discussed. Chicago state has made sure that the past data of the crimes is used in providing meaningful insights by predicting future crime. It discusses many factors like location with latitude and longitude, time of the crime, Date and primary type of crime.

Relevance to the project:

The dataset selected for the project also deals with the crimes in Chicago and the main motive is to predict future crimes by analyzing the data. There are very much similar factors to the selected dataset which makes it very much identical.

1. **Analysis of LAPD Crime Data**

The Los Angeles Police Department(LAPD) always thrives to catch a crime before it’s happening. LAPD recorded many crimes over the years and these crimes are used to find patterns and trends of different crimes. Together all these crimes data are analyzed and insights are drawn from their data. Also there was a decline in crime data once the patterns were found. Because of finding the patterns from unidentified data more patrols can be done on the place where the crime is about to happen which will enable the police to eradicate crimes. This report mainly focuses on characteristics of victim. Clustering and regression techniques are used to achieve the result.

Relevance to the project:

All the three data reports which are mentioned above are using big data as a tool and finding patterns and trends from the data. From all the three data reports many of the factors are very much similar to the selected dataset through some of the factors are excluded majority of the common factors like time of the crime, area, city, district and id are all common and they are very important in drawing patterns

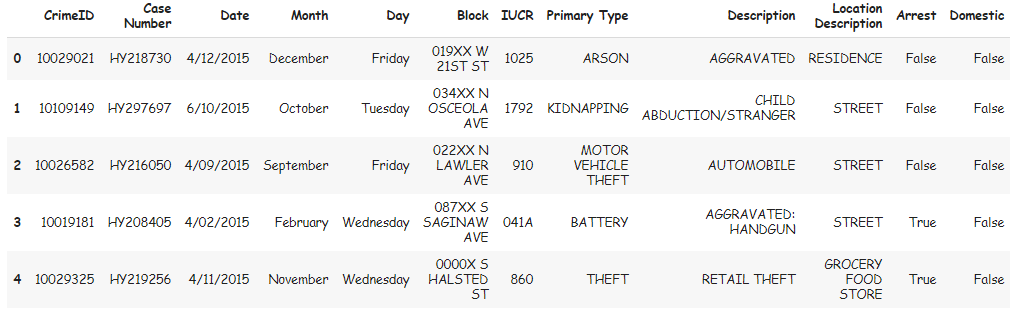
**Methodology:**

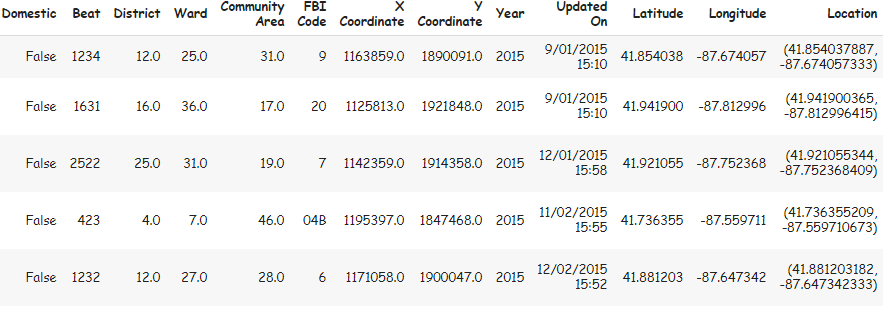
**Data Extraction, Exploration and Pre Processing:**

The data was gathered from the Data World website. The link for the same is given in the Reference Section. According to the information provided along with this data set. The analysis is done with a sample of the Crime dataset from the Chicago Police Department. This dataset contains records of crime that occurred in the City of Chicago from 2012 to 2017. The Dataset used for this study was extracted from the Chicago Police Department’s CLEAR (Citizen Law Enforcement Analysis and Reporting) system. Here, we used R, SQL and Python’s framework to help us with the analysis of the data.

**Overview of the Data:**

The glimpse of the dataset used for the analysis of this research is given below





The dataset selected for the analysis of this project had 361741 records with 24 features. The description of these features are as follows:

* **ID**: A unique identifier for each Crime Record
* **Case Number:** The Chicago Police Department Records Division Number, which is unique for each Crime.
* **Date**: The date on which the crime occurred
* **Block:** Partially redacted address where the Crime occurred.
* **IUCR**: Illinois Uniform Crime Reporting code
* **Primary Type:** The primary description of the IUCR code.
* **Description:** The secondary description of the IUCR code,
* **Location Description**: The location at which the crime occurred
* **Arrest**: whether or not an arrest was made for the crime
* **Domestic**: whether or not the crime was a domestic crime, meaning that it was committed against a family member
* **Beat**: The beat in which the crime occurred. The Chicago police department defines it as the smallest regional division.
* **District**: The district in which the crime occurred.
* **Ward:** The ward (City Council district) where the Crime occurred.
* **Community Area:** The community area in which the crime occurred.
* **FBI Code:** Indicates the National Incident-Based Reporting System (NIBRS) crime classification.
* **X coordinate:** X coordinate of the location where Crime was occurred
* **Y coordinate:** Y coordinate of the location where Crime was occurred
* **Year**: The year in which the crime occurred.
* **Updated On:** Date and time the Crime record was last updated.
* **Latitude**: The latitude of the Crime location.
* **Longitude**: The longitude of the Crime location.
* **Location:** Concatenation of latitude and longitude.

The following steps were performed to extract a Cleaned Data from the selected Dataset:

* Removing Null / NA Values
* Removing Duplicate Rows
* Getting rid of Outliers from the Dataset
* ‘Arrest’ Column was encoded to numeric so in order to be used for modelling

Then, we performed the statistical analysis through Data visualization tools like line graphs, bar graphs, etc. which will answer each of our research questions. For fitting the model, we used Arrest Column as the target Variable and divided that dataset into Train, Validation and Test Sets. The study also intends to use Chi-Square Test of Independence to prove the following Hypothesis:

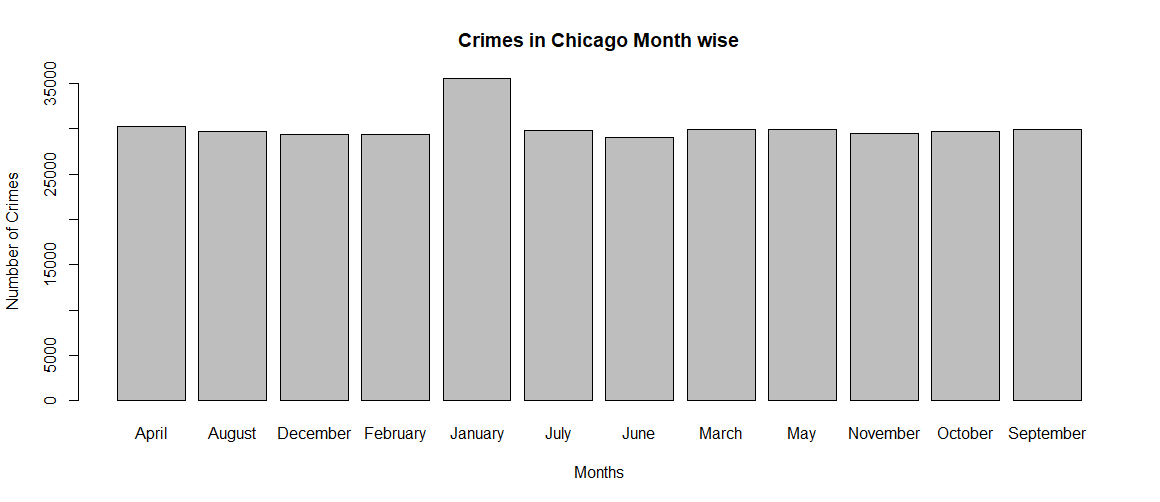
H0 : The arrest made in the city of Chicago is independent of the crime being domestic

H1 : The arrest made in the city of Chicago is dependent of the crime being domestic

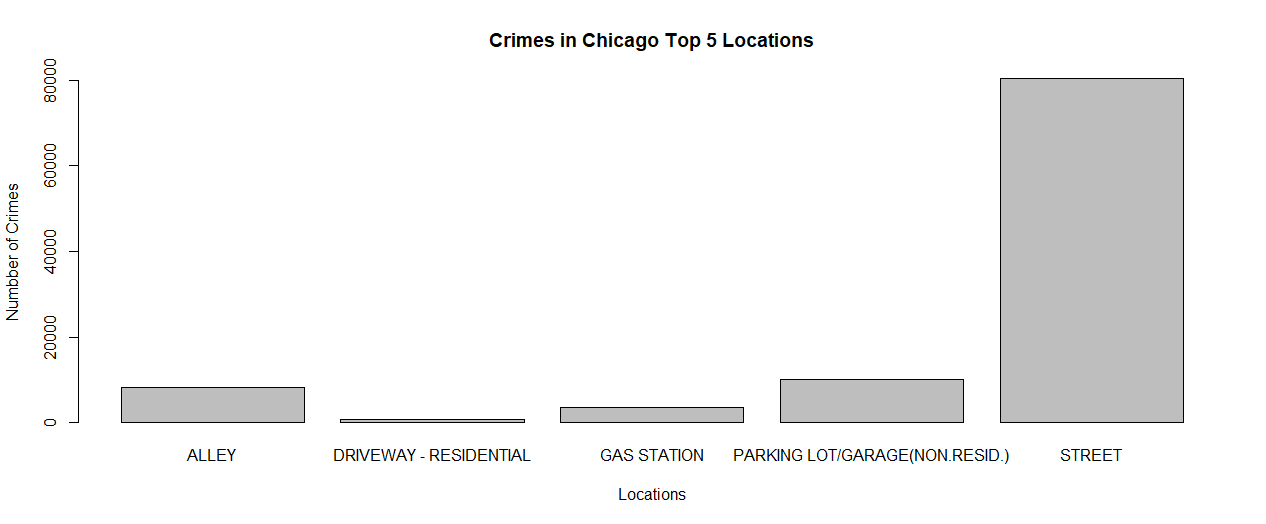
We hope to use the dataset to prove or disprove these hypotheses in order to better understand criminal behavior and whether they can be reduced. Overall, we hope to draw conclusions that can assist the Chicago Police Department to make the city of Chicago a safer and secure place for its inhabitants.

**Results:**

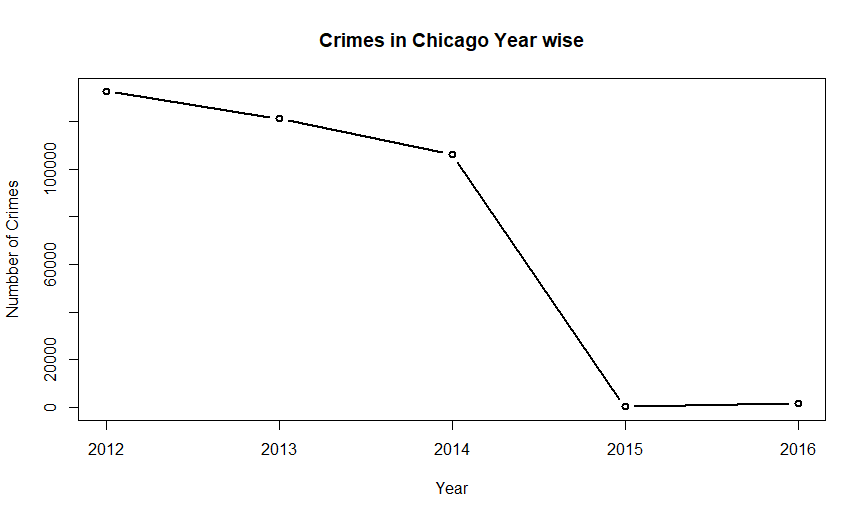
* From the below graph, we can infer the maximum number of Crimes are reported in the month of January in Chicago



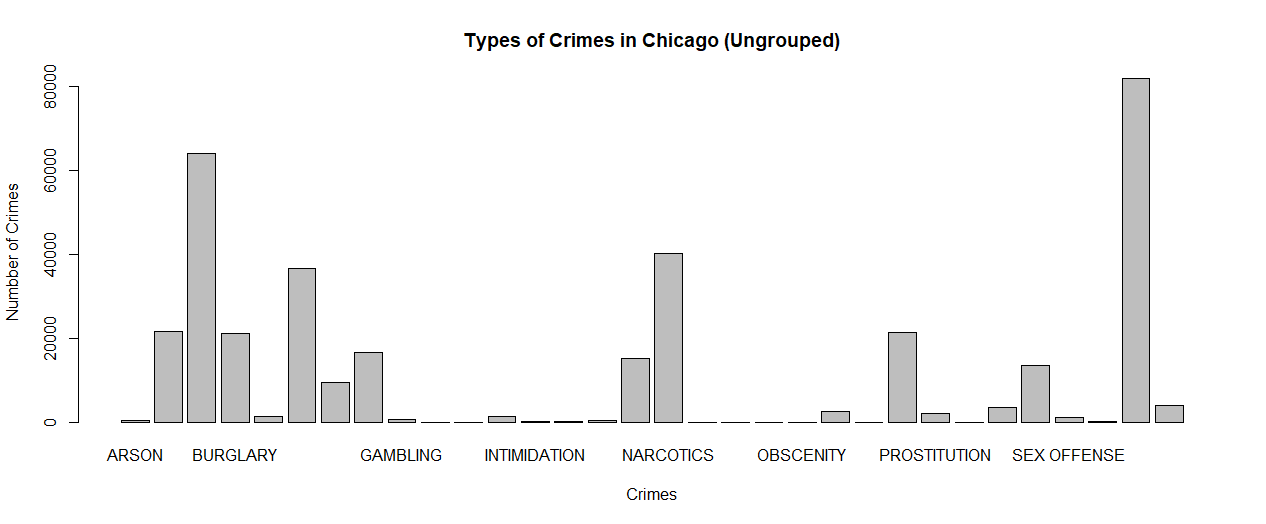
* It was evident from the following graph that Street is the prime location of the majority of Crimes in the City.



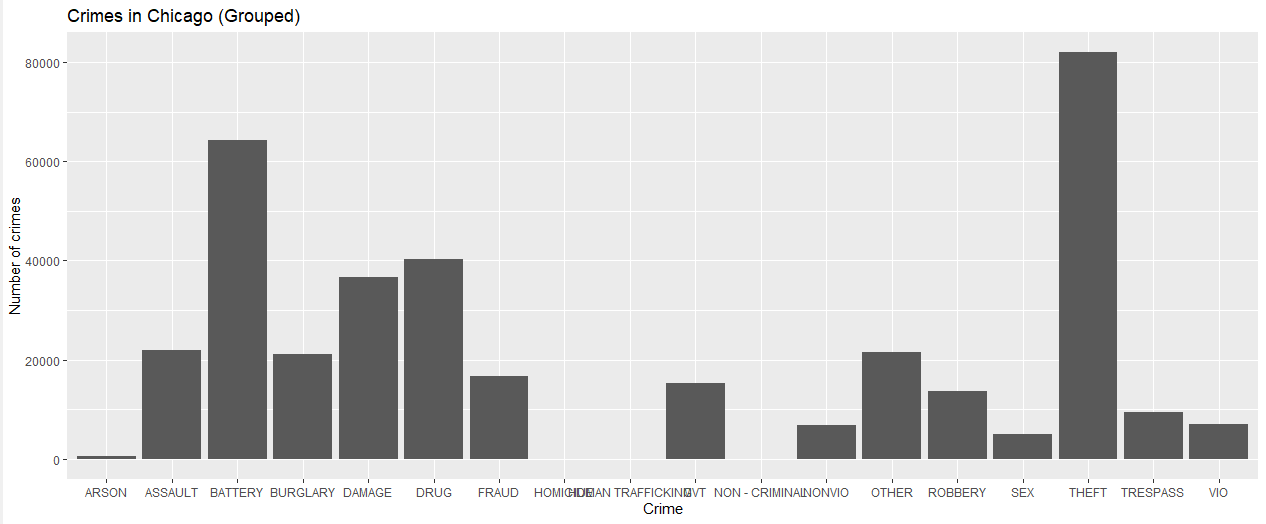
* The below line graphs depicts the sharp decrease in the number of Crimes over the span of 5 years i.e. 2012 to 2017 in Chicago.



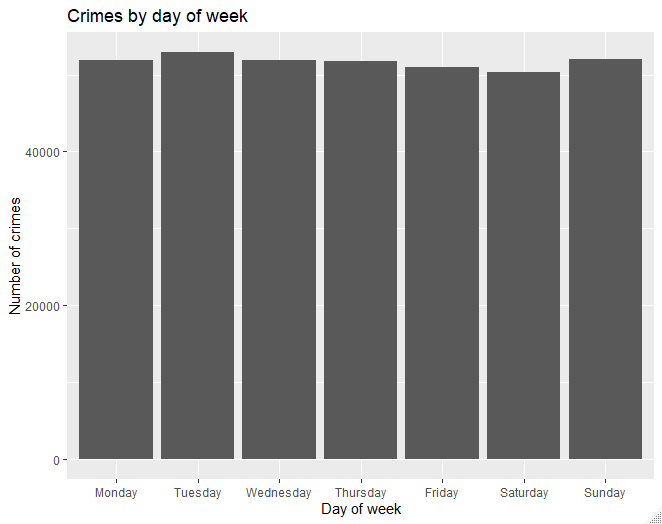
* There were 32 different types of Crime reported in the City of Chicago. Since it was incompatible to plot all the type, we categorized the types under general categories.



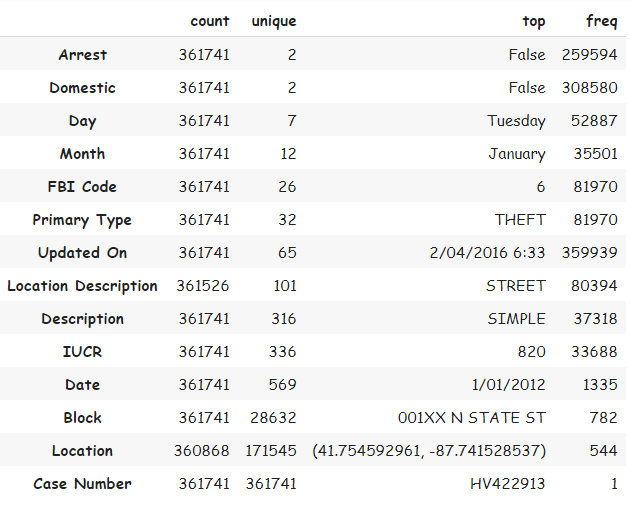
* After grouping the types of Crimes we could infer that the most reported crime was of Theft followed by Battery.



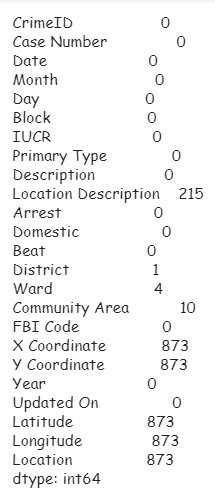
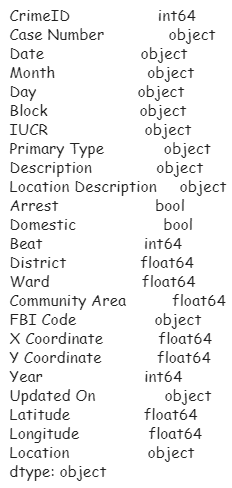
* We can see the crime rate on an average remains the same throughout the week irrespective of the day. But in numbers number of crimes reported on Tuesday are slightly more than rest of the day.



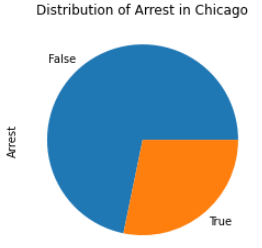
* The below Table shows the number of records in each categorical feature followed by the most frequently occurring category and its respective frequency.



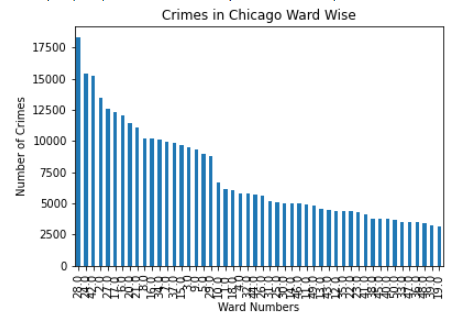
* The below tables show the number of missing values in each column (Left Table) and the type of data in each column (Right Table)

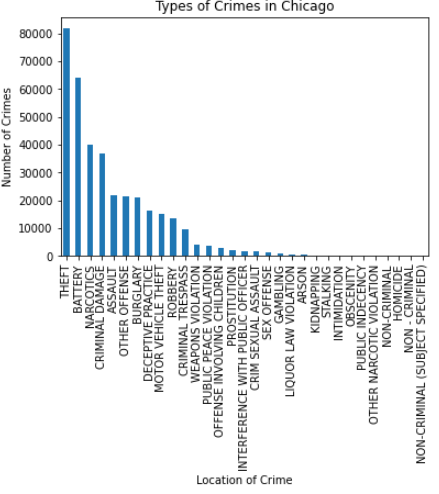
* From the subsequent pie chart, we have observed that only one fourth of the criminals of the reported crimes were arrested.



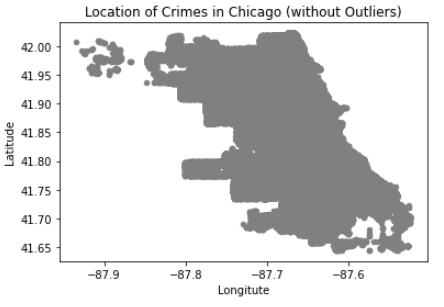
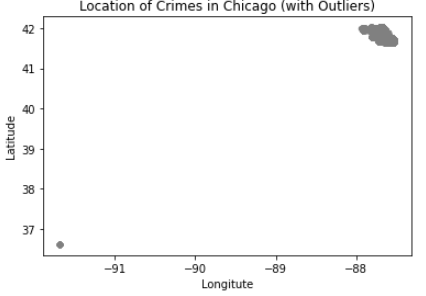
* We can infer that the Ward Number 28 has the most reported crimes in the time period 2012 to 2017 followed by ward numbers 24 and 42



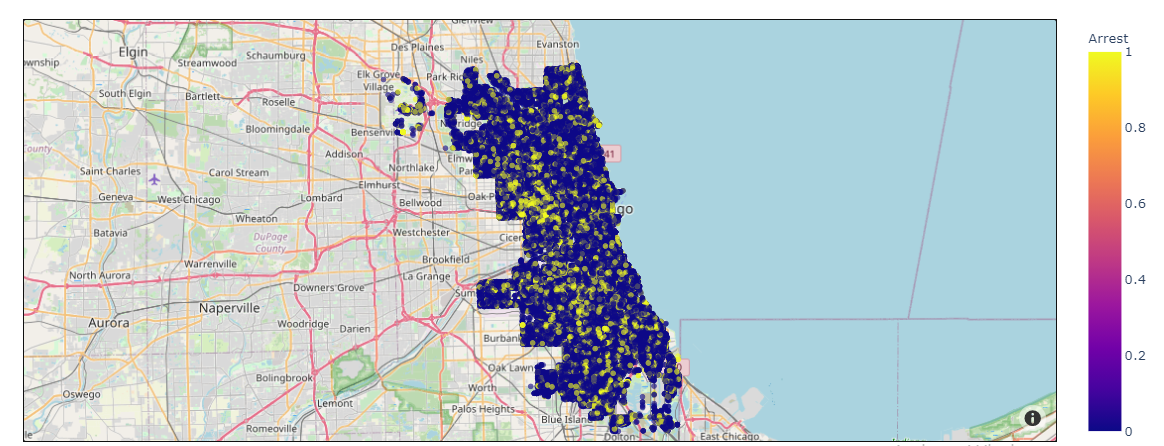
* Most Number of Crimes reported in Chicago during 2012 to 2017 were of Theft followed by Battery and Narcotics.



* The dataset had many Outliers which did not belong to the Chicago City. So, we removed those Outliers. Below is a visualization of latitude and longitude before and after removal of Outliers.



* Below Map shows the geographical locations of the records in the dataset. Here the Yellow dots represents the Crimes where an arrest was made whereas the blue points denote the crimes where an arrest was not made. It is evident from the frequency of Blue points that in majority of Crimes, no arrest was made.



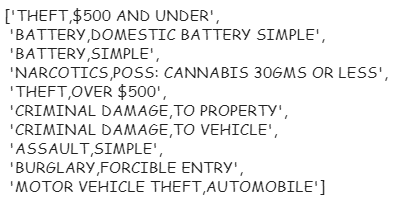
* With our dataset, we can group the data together based on IUCR code, which is a four digit codes that Chicago’s law enforcement agencies use to classify criminal incidents when taking individual reports. The word cloud image tells us that “051A”, “031A”, “041A” and “143A” are the most frequently occurring IUCR codes in Chicago. We can find what the crime codes represent from the Chicago’s data portal link. We can infer that assault, battery and robbery with handguns are the most highly reported crimes in Chicago.



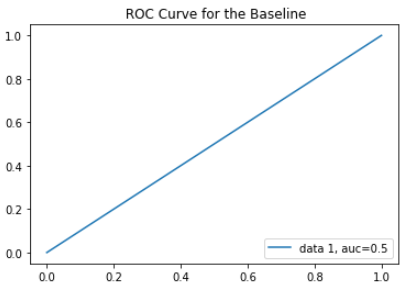
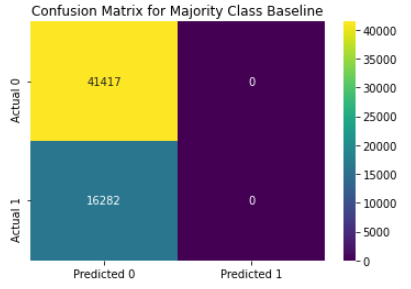
* A word cloud based on the frequency of crimes occurring in different locations was made. As an added visualization tool, the word cloud was mapped over the image of the American flag. This word cloud can be used to quickly tells which locations to avoid in Chicago due to the higher frequency of crime occurrence.



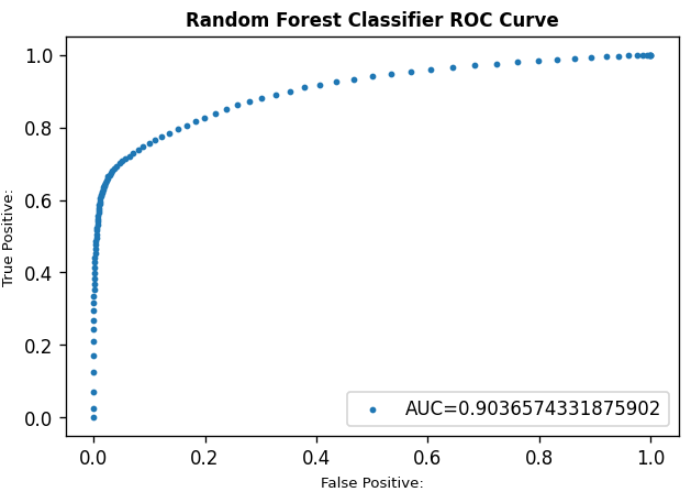
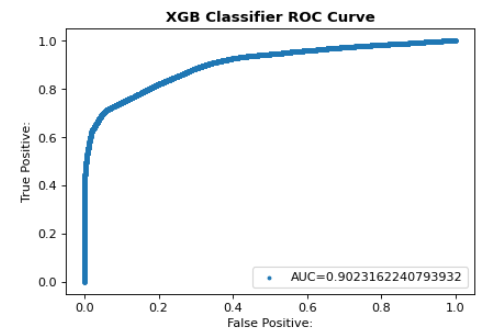
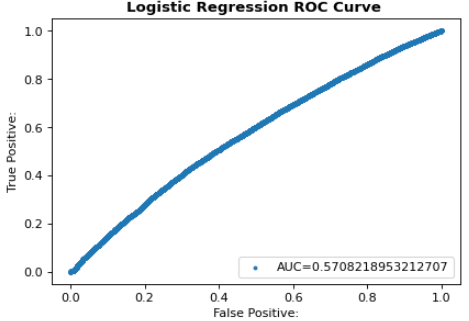
* We extracted the top 10 crimes based on type and description in Chicago. This is a more specific analysis as we focused on the description of the crimes as well.



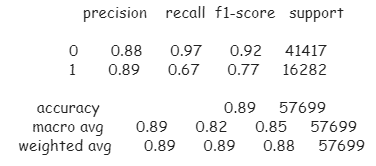
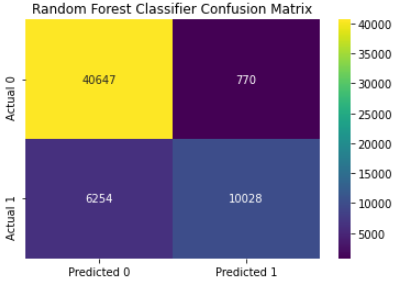
* The baseline model is created for an initial prediction and for calculating an accuracy percentage. This model acts as a benchmark to beat with the future predictive model. For our research, since the target variable was categorical we used the mode for the prediction. The straight line of the ROC curve depicts that there is no discrimination capacity to distinguish between positive class and negative class of the target variable.



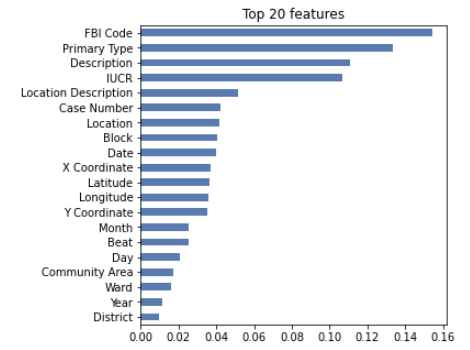
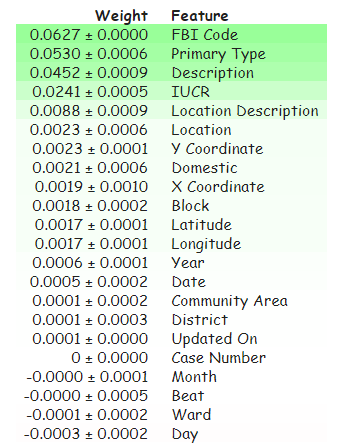
* From the below Receiver Operating Characteristic (ROC) curve of the three models i.e. Logistic Regression, XGB Classifier and the Random Forest Classifier we infer that the Random Forest Classifier is the best to predict whether a Crime Record will result in an arrest.



* The Confusion Matrix of Random Forest Classifier given below clearly indicates low levels of False Positive and False Negative. Also, the Classification Report of Random Forest Classifier indicates good values of Precision and Recall.



* From the below visuals, we can infer that FBI Code is the most important feature to predict whether an arrest was made for a particular crime or not. Some other important features are Primary Type, Description and IUCR Code.

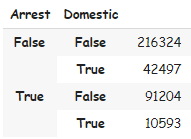
 

* We used the below table to test for independence between the features Arrest and Domestic at 5% level of Significance.

H0 : The arrest made in the city of Chicago is independent of the crime being domestic

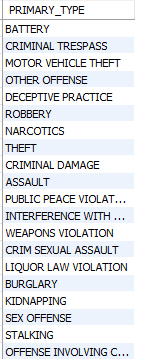
H1 : The arrest made in the city of Chicago is dependent of the crime being domestic

|  |  |  |
| --- | --- | --- |
| **Arrest \ Domestic** | **True** | **False** |
| **True** | 10593 | 91204 |
| **False** | 42497 | 216324 |

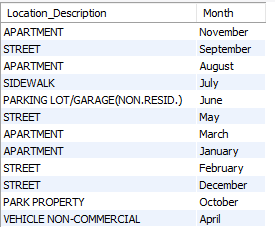


Since , we don not reject H0. Hence, we conclude that The arrest made in the city of Chicago is independent of the crime being domestic or not.

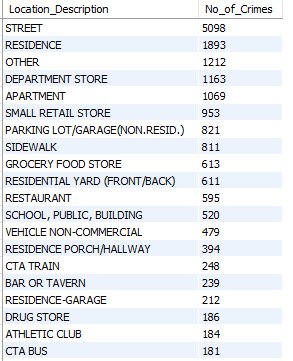
* The following queries were executed in MySQL to extract the following tables. Here the data we used had 90256 records with the 6 features: Year, Month, Day, Primary Type, Location Description and Arrest
  1. SELECT DISTINCT(PRIMARY\_TYPE) from crime\_schema.crime\_chicago\_data WHERE LOCATION\_DESCRIPTION ='STREET';



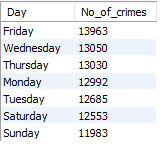
* 1. SELECT DISTINCT(Location\_Description), Month FROM crime\_schema.crime\_chicago\_data GROUP BY Month;



* 1. SELECT Location\_Description, count(Location\_Description) as No\_of\_Crimes FROM crime\_schema.crime\_chicago\_data where PRIMARY\_TYPE = 'THEFT' GROUP BY Location\_Description ORDER BY count(Location\_Description) DESC;



* 1. SELECT Day, count(Day) as No\_of\_crimes FROM crime\_schema.crime\_chicago\_data GROUP BY Day ORDER BY count(Day) DESC;



* 1. SELECT COUNT(\*) as No\_of\_Arrest\_True FROM crime\_schema.crime\_chicago\_data WHERE arrest = 'TRUE';



* 1. SELECT COUNT(\*) as No\_of\_Arrest\_False FROM crime\_schema.crime\_chicago\_data WHERE arrest = 'FALSE';



**Limitations and Further Scope:**

This study focuses on analyzing the crimes only in the city of Chicago. This can be further extended for any area across the globe provided the necessary data is available for the respective area. By adding another dataset like the Census Data the crimes can be forecasted and predicted using various forecasting technique and predictive analytics respectively. A predictive model can be built using the selected dataset to predict the exact location and time of the future crimes.

**References:**

* Chadhar, M. (n.d.). Chicago Crime Dataset [Online] Available at: https://data.world/mchadhar/chicagocrime-dataset
* Abouelnaga, Y. (2016, July 13). San Francisco Crime Classification. [Online] Available at: https://arxiv.org/pdf/1607.03626v1.pdf
* Alqahtani, A. (2019, June). Research Gate [Online] Available at: https://www.researchgate.net/publication/335361962\_Crime\_Analysis\_in\_Chicago\_Cit y
* Cung, B. (2013). ucla. [Online] Available at: https://escholarship.org/uc/item/2v76v571
* Mallula R., Chowdary P. (2018), Predictive Policing in Crime analysis using R, International Research Journal of Engineering and Technology (IRJET), Volume 5, Issue 7 [Online] Available at: https://www.irjet.net/archives/V5/i7/IRJET-V5I781.pdf (Accessed: July 2018)
* Monish N (2019), Chicago Crime Analysis using R Programming, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 5, Issue 2 [Online] Available at: <https://1library.net/document/ydjvk7jy-chicago-crime-analysis-using-r-programming.html?utm_source=related_list>
* Sengupta et al, Crime Analyses Using R, Fractal Analytics [Online] Available at: https://irgn452.files.wordpress.com/2016/03/3-s2-0-b9780124115118000141-main.pdf
* Johnson A., Crime and the Church: An Analysis of Crime in Chicago Based on the Location of Churches [Online] Available at: https://scholar.valpo.edu/cgi/viewcontent.cgi?article=1875&context=cus
* Yerpude P., Gudur V., (2017), Predictive Modelling of Crime dataset using Data Mining, International Journal of Data Mining & Knowledge Management Process (IJDKP) Volume 7, Number [Online] Available at: https://aircconline.com/ijdkp/V7N4/7417ijdkp04.pdf (Accessed: July 2017)
* Alqahtani et al (2019), Crime Analysis in Chicago City [Online] Available at: https://www.researchgate.net/publication/335361962\_Crime\_Analysis\_in\_Chicago\_City (Accessed: June 2019)