**Chicago Crime Data Analysis**

**Data Wrangling:**

In this project, I will be dealing with the data set containing of all the crimes that are reported by the police in their directory from 2001-present.

This dataset is available in the Chicago city data repository. This dataset consisting of 7 million row of data by the columns of 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

On this data first objective was to remove the unwanted data, columns containing the unwanted data or un relevant information like ID, Case Number, Block, IUCR, Beat, Updated On and location were removed. The second objective was to convert some of the object type columns to integers. The questions we are associated to different types of crimes and how they are located.

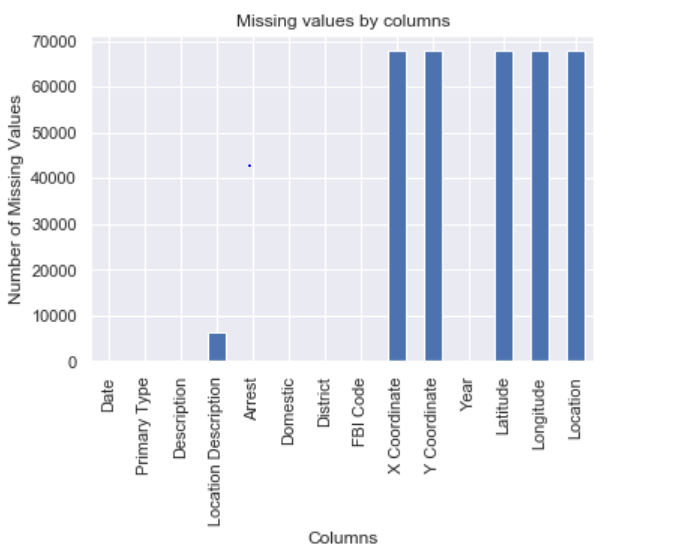
After preprocessing, the data frame consists of following information with our dropping null values

Dropped the following columns from the main Data Frame called crimes



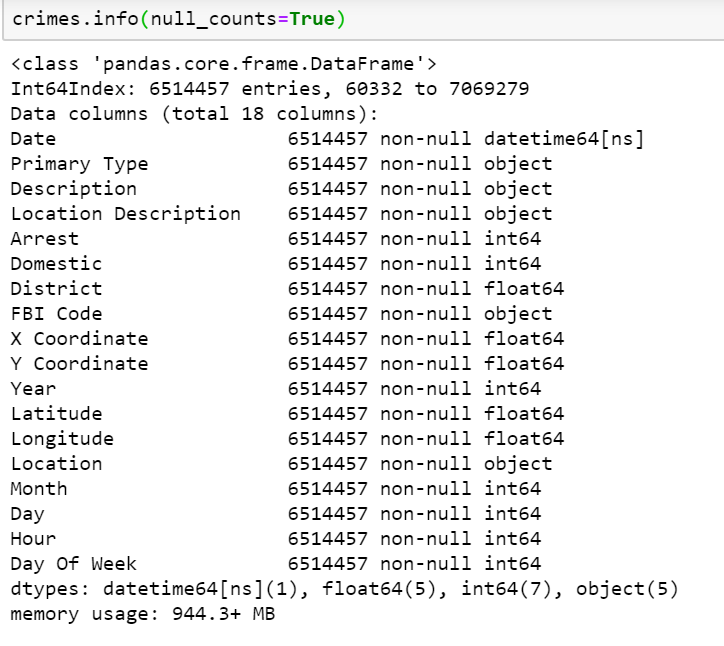
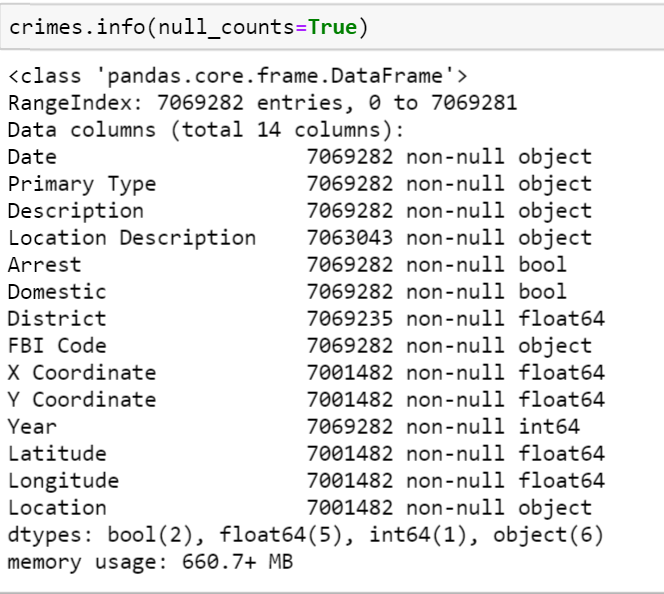
Difference between the columns containing null values and without null values (which is after dropping all null values)

Inspecting the features, we see that all the features that have a large count of missing values are features that relate to the geographical location of the crime scene. This is No Surprise as the Chicago Crime Dataset is based on firsthand accounts of people involved in or around the crime. It is not necessary that such firsthand reports need to contain the specific locations of the crime. We have 3,45,286 missing values in the whole dataset that are present in Location Description, Community, X Co-ordinate, Y Co-ordinate, Latitude, Longitude and Location.

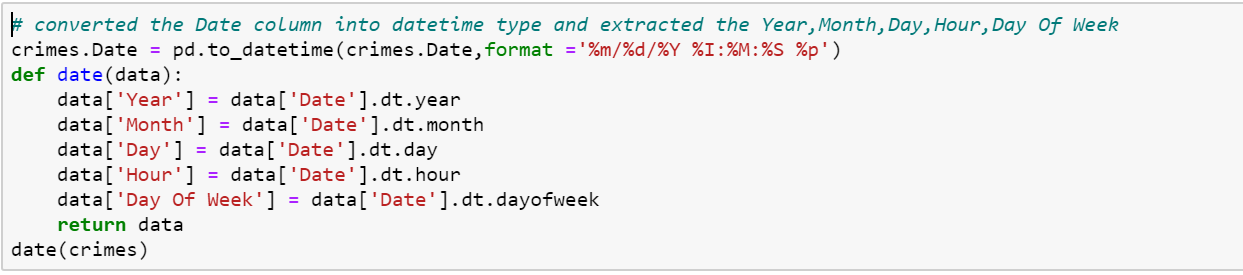


Since, these features are not direct numeric values, we can't use summary statistical functions to fill in the missing values. Hence I thought to remove all values containing null using dropna().92.5 percent of data is retained after dropping null values.

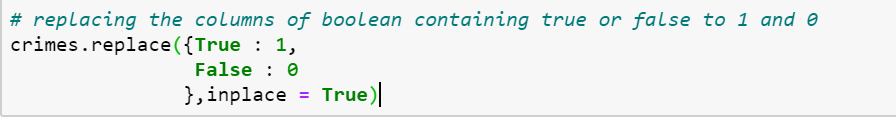
Below figure shows the comparison of no-null values counts before and after dropping null values



I have created 3 additional columns Day, Hour, Day Of Week by converting the Date column to the datetime type and then slicing the required column attributed for the Date. Below figure gives the code that I used for slicing. This will be helpful for me to identify the crime pattern in different sections of day



The columns [‘Arrest’, ‘Domestic’] contains the values in terms Boolean values, I have converted to their respective 1’s and 0’s in the column places using the following code



All the changes done to the Data Frame is saved in the form of CSV file and PICKLE files.

**Chicago crime data analysis EDA**

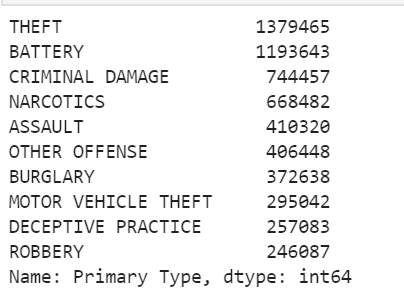
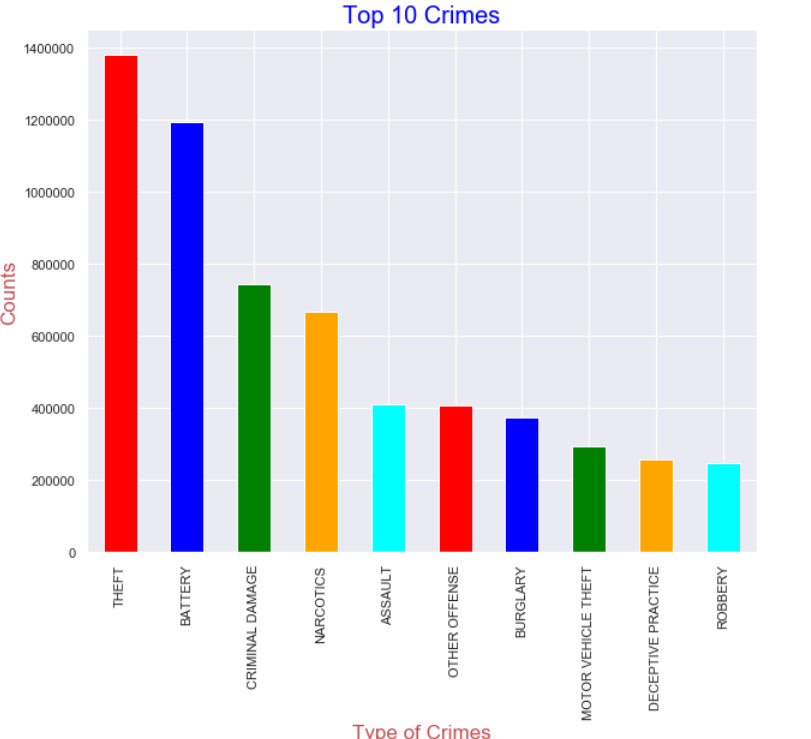
**Imported Packages:**

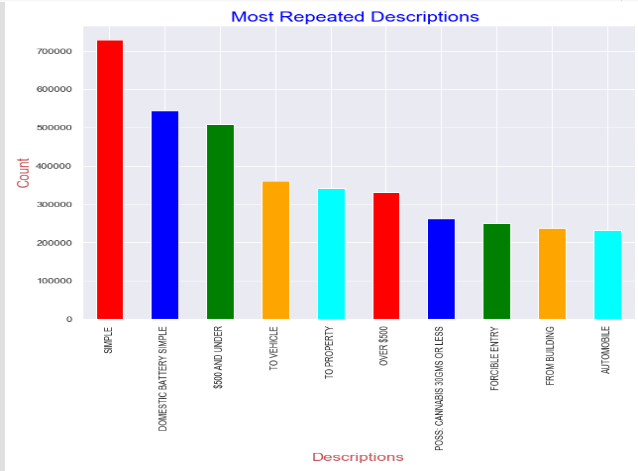
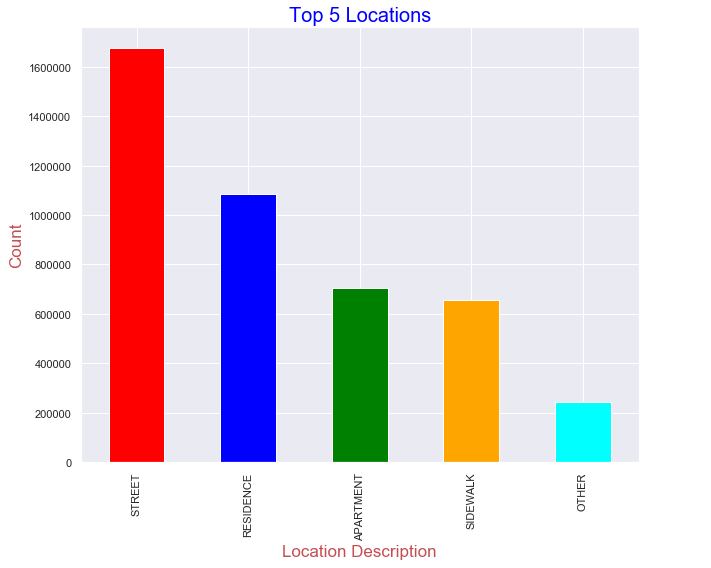
Here is the list of packages that are Imported to do my Exploratory Data Analysis. I have used Matplotlib, Seaborn for my visualization of data and Folium for the interactive maps



**1.All About Crimes in Chicago**:

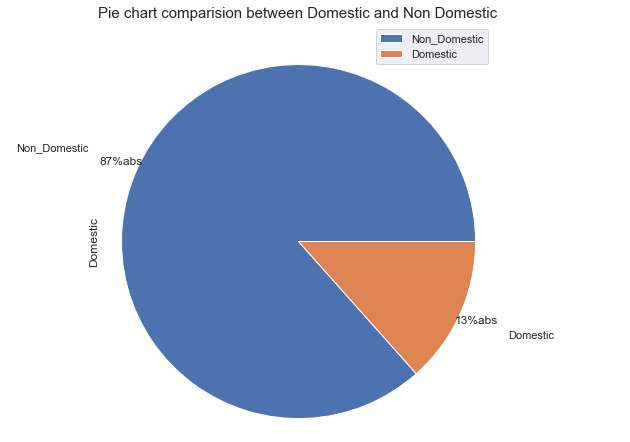
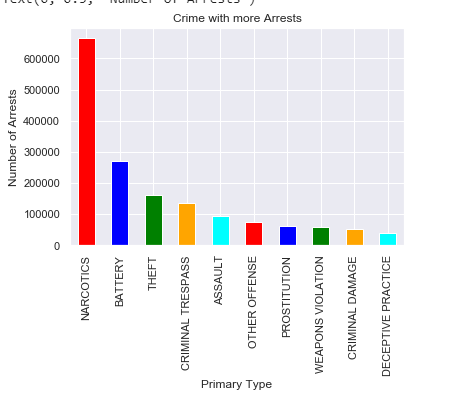
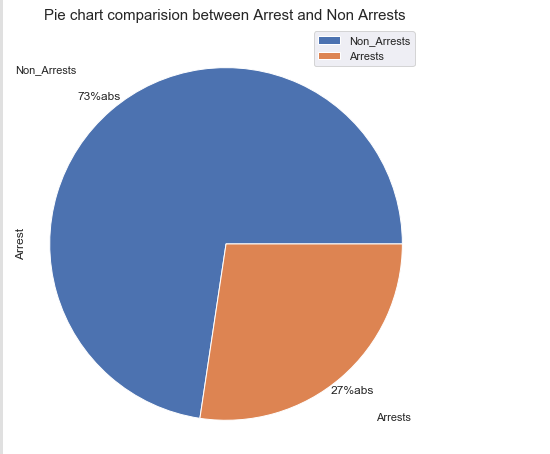
By seeing the data, I was very curious to find the topmost crimes in the Chicago. Below figure represents the top 10 crimes in the Chicago. Out of the top 10 crimes, Theft was the most occurring crimes with an count of 1379465.Higher counts of Battery and criminal Damage indicates the presence of physically violent community.



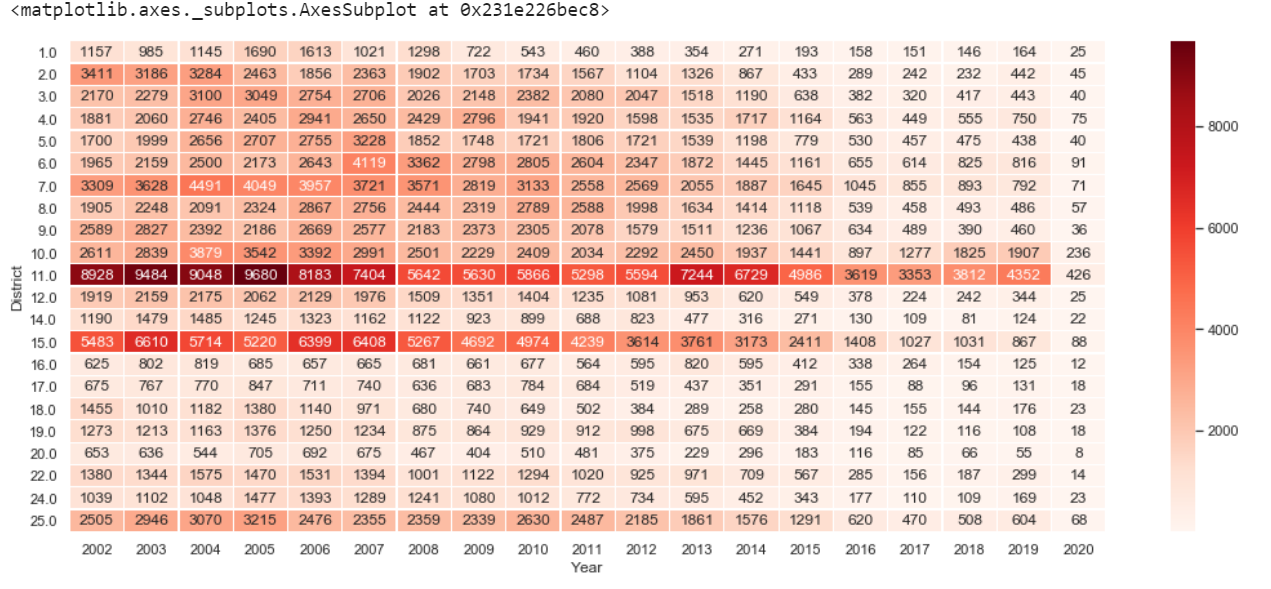
The figure represents the Descriptions that are reported along with the crime. Majority of the crimes are reported as Simple and the battery crimes in the domestic reported as simple, some were under 500$ which could be Theft or Robbery or Burglary. Most of the crimes are at street level and at house level  


**2.Arrests in city of Chicago:**

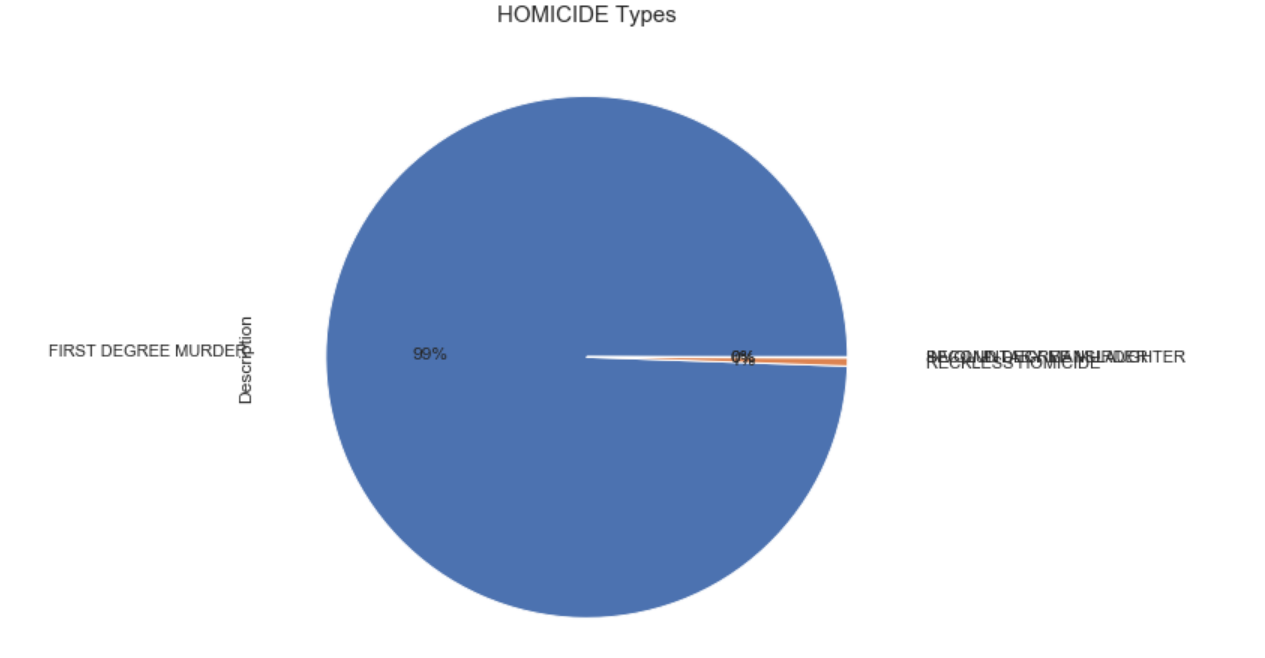
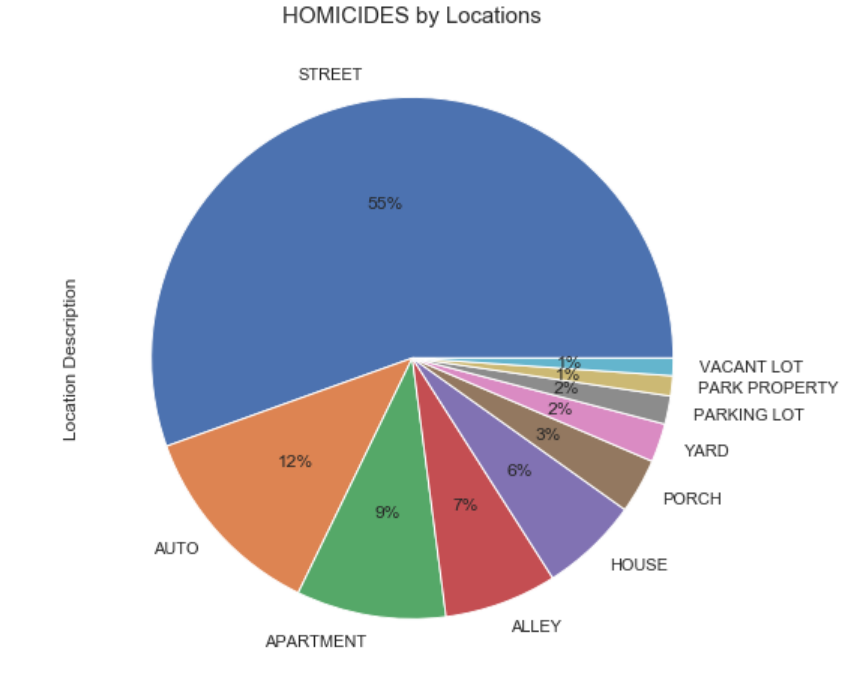
73% of the crimes see No Arrests, means only 23 percent of the crimes has been solved from 2001 to present. As we see in the bar chart which represents the crimes with a greater number of arrests, Arrests related to Narcotics are more. 99%of the Narcotics cases we arrested (total narcotics cases were 67000(approximately). As we overall, they are 1379465 theft cases but only 18000 cases are solved, May be police were not taking serious to cases which are involved with the theft and Battery. 87 percent of the crimes are related to Non-Domestic areas.



Let’s how the Narcotics arrests distributed across the districts.District 11 is more involved with the crimes.There is down trend of narcotics arrests over the years except some hikes in the year 2013-2014

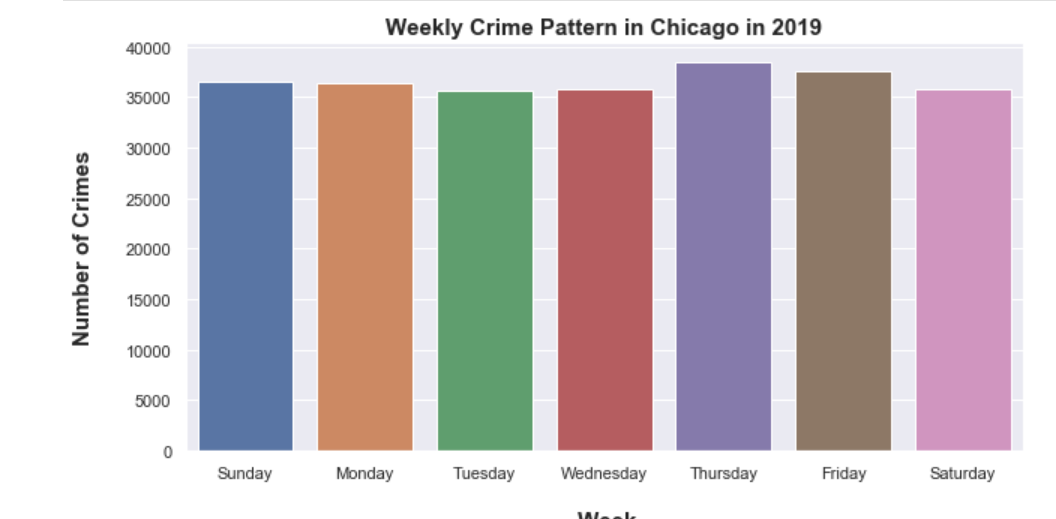
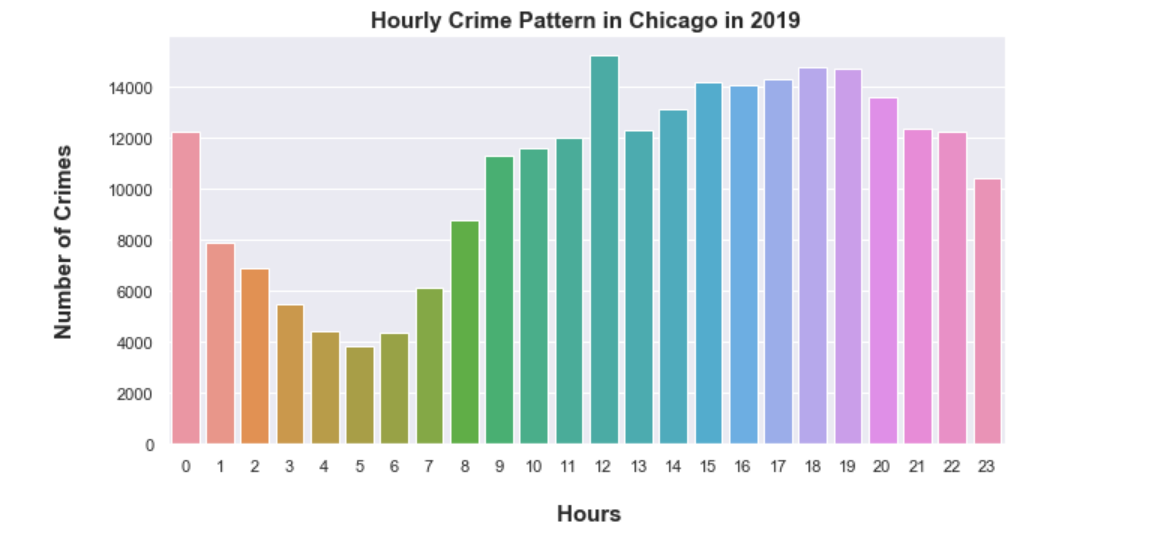


Out of 9421 cases of Homicides,99 percent of the cases are involved with First Degree Murder.56% of homicides are occurring in the streets and followed by Auto, Apartment which contributes 10 percent cases each.



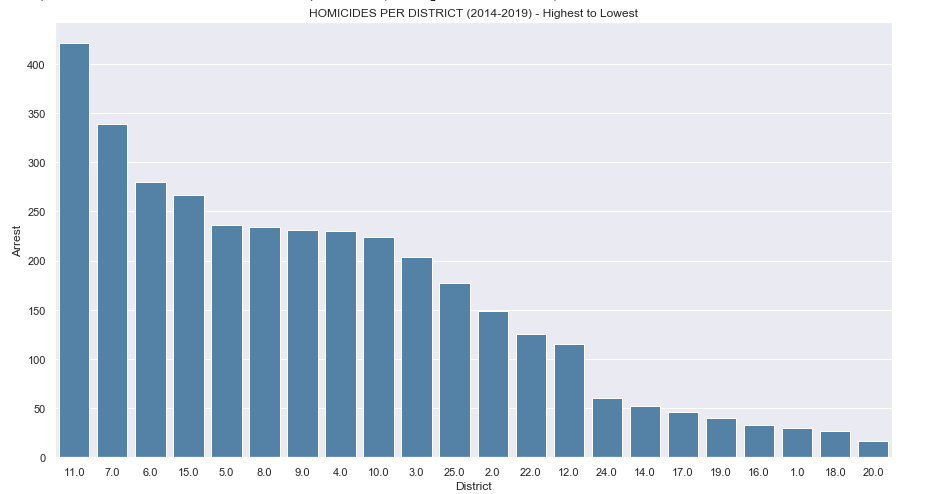
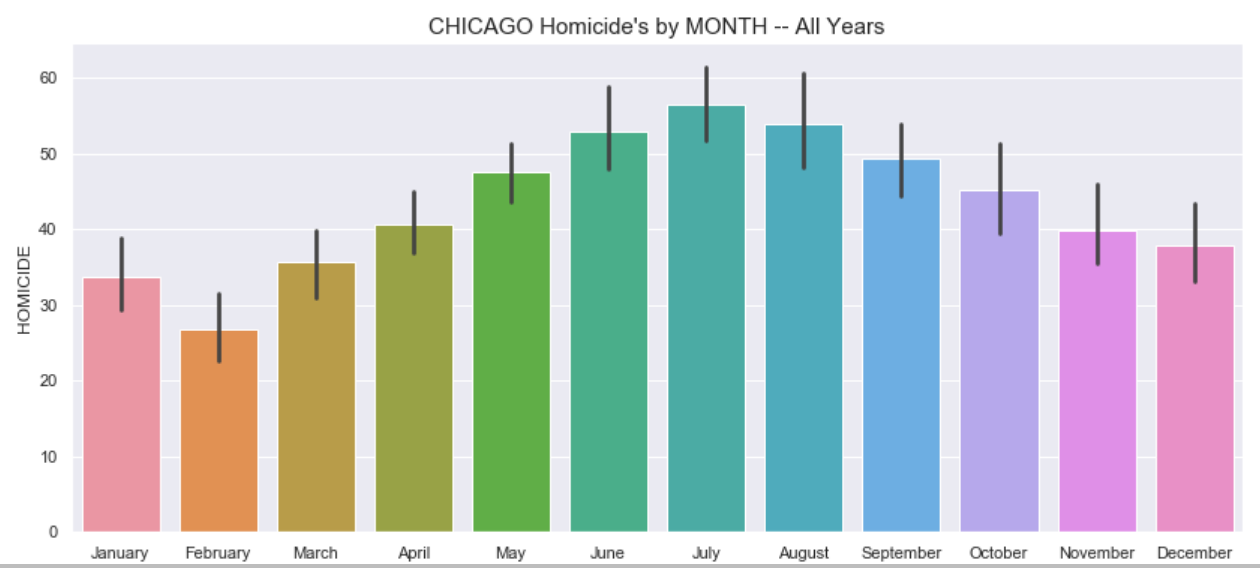
**3. Crime vs Time**:

Here is the Monthly, Hourly, Weekly crime patterns of the crimes in 2019.We see more crime in the summer months, may more people coming out for vacations that might lead to increase in the crimes count. Crime rate increases with the day, needs some rest to crime, less crimes rates between 2 am to 7 pm. coming to weekly part all the days reports the same number of crimes but Thursday and Friday reports slightly greater than remaining days



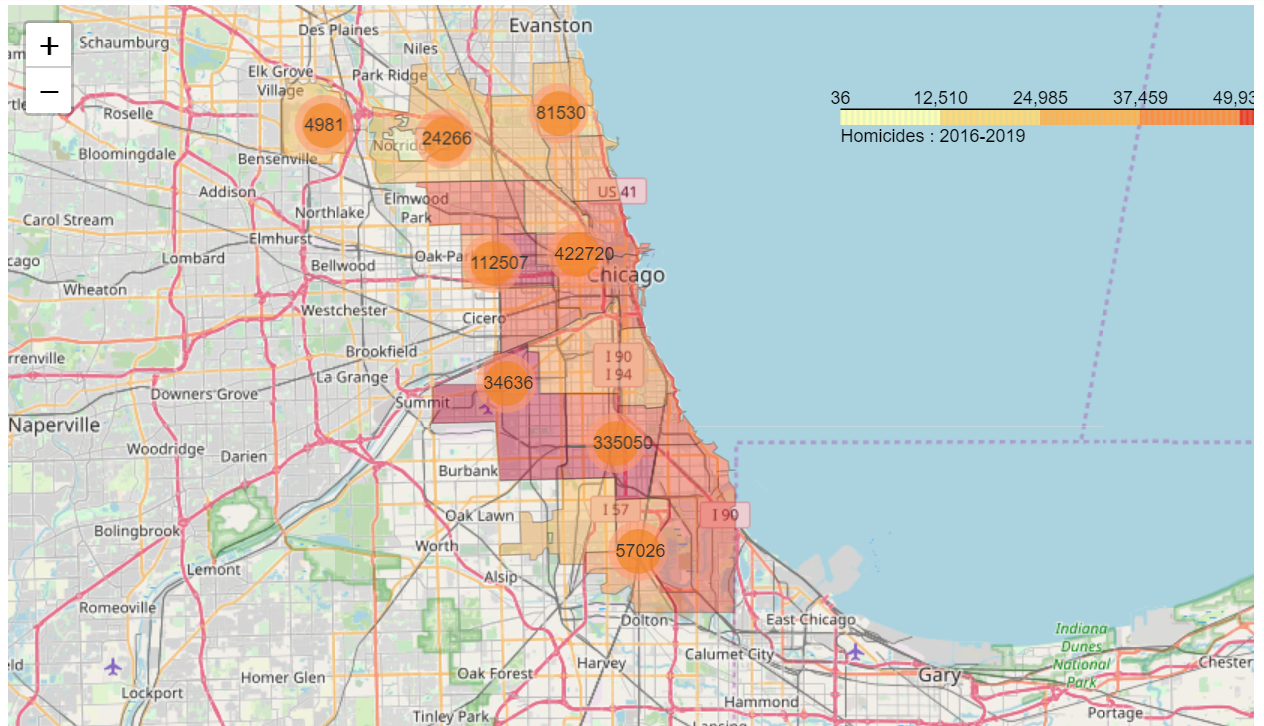
Homicides crime rates vs years

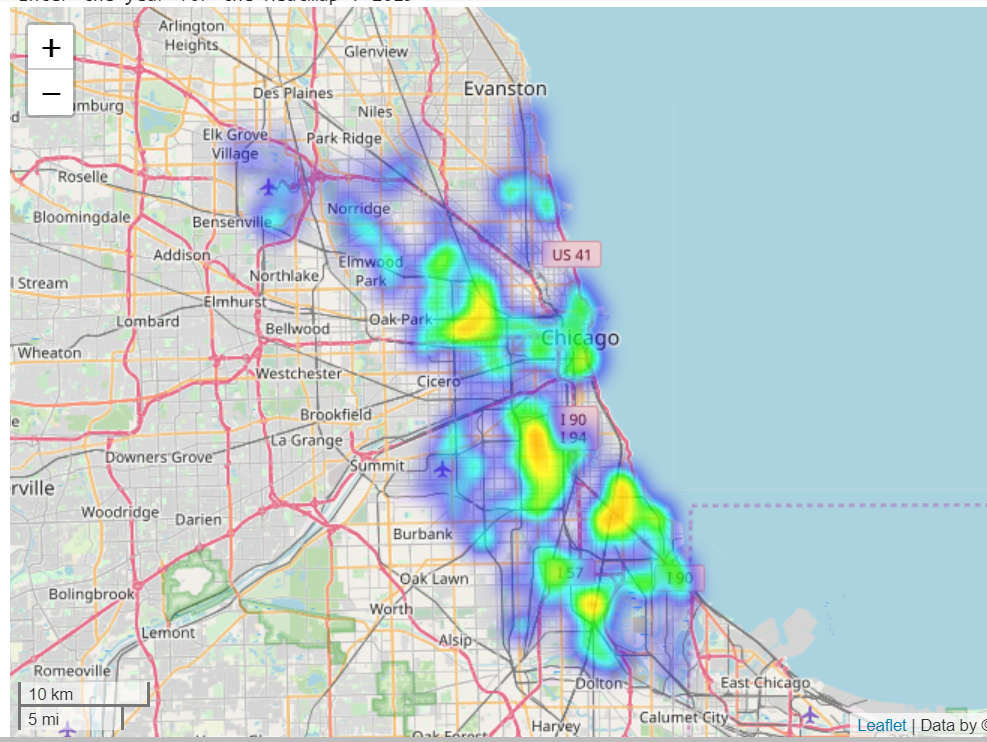
Most of the homicides were happening in the night between 5pm to morning 5am, records very low homicide rate between 7 to 11 am. Districts 11,7,6,15 has more arrests towards homicides compared to remaining districts.



**4.Crimes with Maps**

Below chart shows the arrest count per police district



Heat map of all crimes for the year 2019

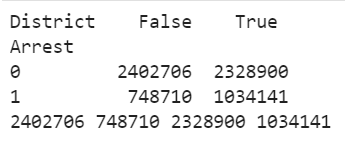
**Inferential Statistics**

**Is there any association between the arrest and Districts :** Inorder to find the relation between the districts, divide the districts which have more blacks population and the districts with other populations which includes white,asian and etc.The districts 15.0,11.0,10.0,21.0,2.0,7.0,9.0,3.0,6.0,4.0,5.0 majority blacks populated districts,where as rest of the districts are other populations.Applied hypothesis testing inorder find the whether there is an association between the Arrest in black majority populated districts and other.

**Hypothesis Testing :**

 H0 : There is no Association with arrests between the Majority blacks districts and others.

 H1 : There is an association with arrest between the districts.



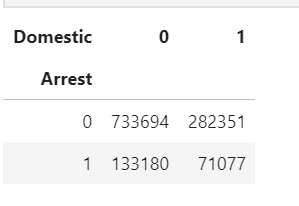
True represents the districts with majority black’s districts

1 indicates the arrests Counts and 0 represents the Non arrests Counts

By using the chi-squared test, p value = 0.000. Since the  P value < 0.05, hence we reject the null hypothesis. Arrest has an association between the districts with more black’s and without.

* These districts 2,3,4,5,6,7,9,10,11,15,21 have more arrests compared to other districts.
* There might be  racial disparities among the blacks than others.
* Police may be more interested in arresting the blacks.

**Is there any association between the Arrest and Domestic majority populated Black’s:**



Count of arrests are more in the more in the Non Domestic areas

**Hypothesis Testing:**

H0 : There is no Association with Domestic and Arrest.

H1 : There is an association with Domestic and Arrest.

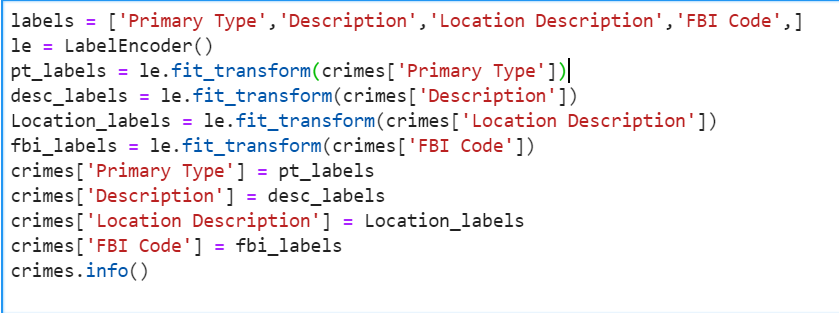
By applying the chi square test, p values was found as 0.00 which is less than 0.05.Therefore Reject the Null hypothesis.There is a strong association between the arrest and Domestic

Most of the Arrests are related to Non domestic due to the crimes that are reported in chicago where Theft ,battery and murders are more probable to happen in the Non Domestic Regions.

* People in the districts are more related to Non domestic type crimes.

**Predictive Modeling:**

The dataset consists most of the features as the categorical values, all these columns are converted into labels using the Label Encoder, in the which all the values are converted into unique number types.

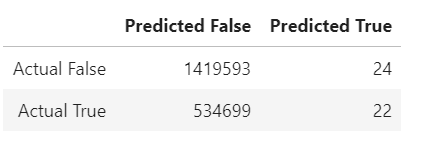


The target variable that I selected was the ‘Arrest’ Column which indicates that whether the person is subjected to arrest or not based on the crime, district and location. These features will contribute to models are the other columns of the data frame. The dataset is randomly split with 70 percent of training data and 30 percent test dataset at a random state of 42.

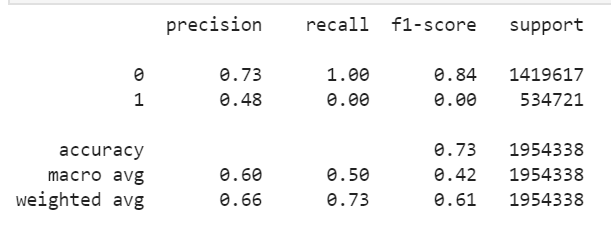
The models that I used are Logistic regression, Decision Tree, random forest classifier, Ada boost Classifier, Gradient Boosting Classifier.

Logistic Regression: Logistic regression is applied on the training dataset; the performance of the model was the evaluated on the testing dataset. The accuracy of the model was reported as 72 percent.

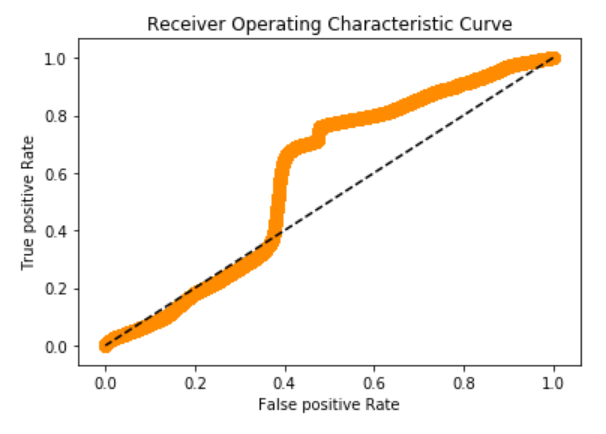
The confusion matrix of the predicted vs actual value are as follows



Classification Report:

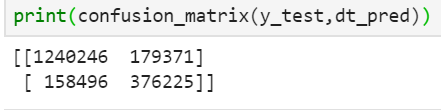


ROC CURVE:

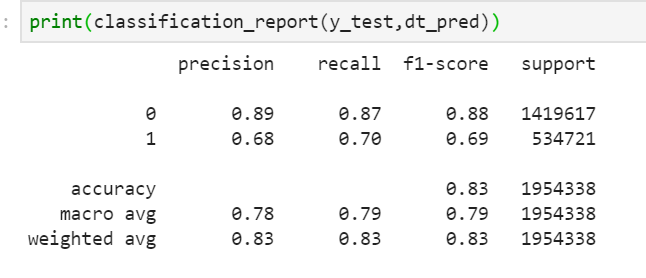


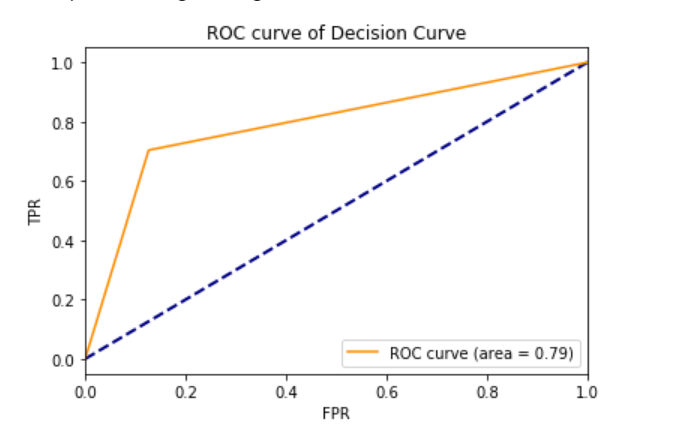
The area under graphs is 0.50,mean weighted F1 score as 0.61

Decision Tree: All the values in the confusion matrix, which contains more false positive and false negatives. Despite of accuracy , high number of false positive and false negative indicates



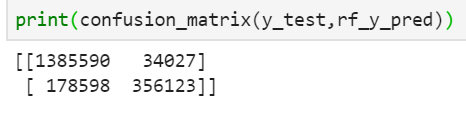
F1 score, precision, recall values are improved from the logistic Regression



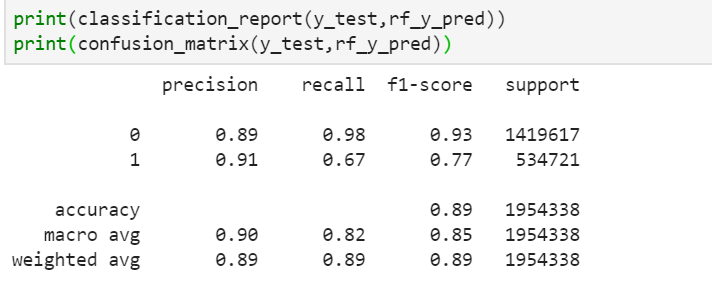


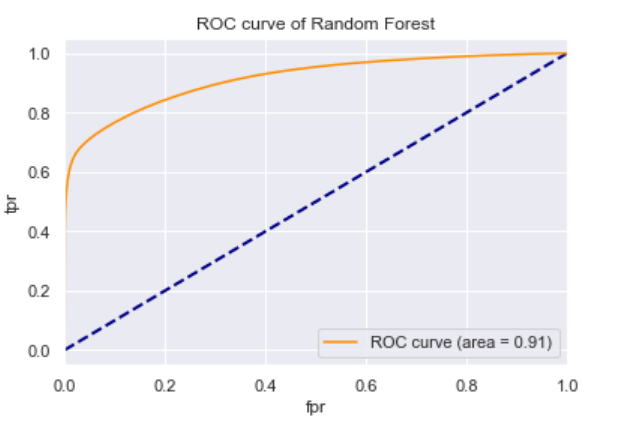
Accuracy 82 percent

**Random Forests:** The confusion matrix shows that there is large true positives and true negatives are high.



F1 score of the random forests was 0.89, with an accuracy of 89 pecent,area under the graph was





Accuracy 89 percent.

Conclusion:

By comparing all these models in terms of accuracy random forests keeps good in predictions, area under the curves are also pretty high which indicates high true positive rate. Accuracy is not the main thing that I have to calculate, based on the F1 scores, precision and recall values. I recommend the use of Random forests model for the prediction of Arrests.