Capstone Project

Chicago Crime Analysis

Data Inspection

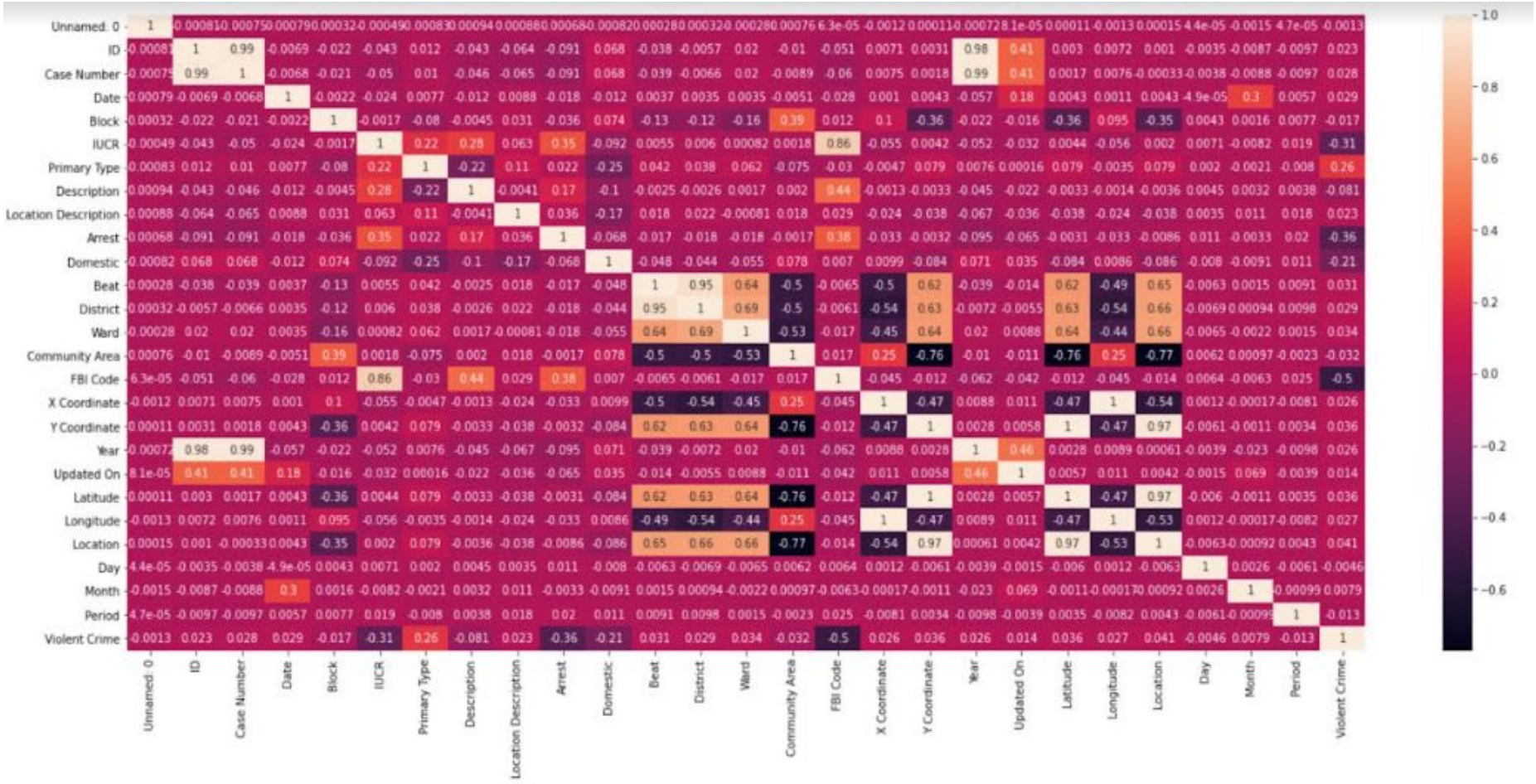
- ➔ The chicago crime data was observed for insights before preparing it for modeling and the following were noted;
  - The has 23 columns ('Unnamed: 0', 'ID', 'Case Number', 'Date', 'Block', 'IUCR', 'Primary Type', 'Description', 'Location Description', 'Arrest', 'Domestic', 'Beat', 'District', 'Ward', 'Community Area', 'FBI Code', 'X Coordinate', 'Y Coordinate', 'Year', 'Updated On', 'Latitude', 'Longitude', 'Location')
  - Missing values were observed
  - Duplicate Case numbers were observed in the data
  - Skewness in the data was observed

Data Treatment;

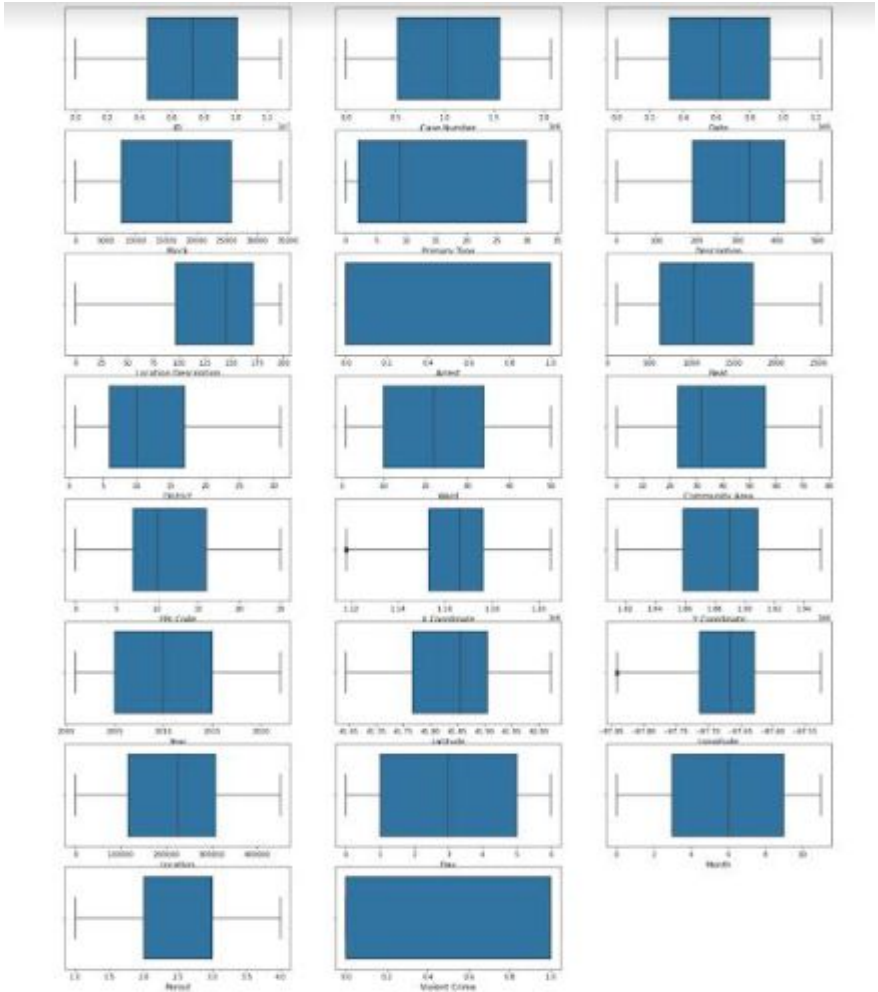
- All columns with missing values were dropped given that they only constituted 9% of the entire sample size and to ensure that the model only focuses on correct data
- Duplicate Case Numbers were still present in the data after dropping rows with missing values as a result the entire data was sorted based on the 'Updated On' column then duplicates case numbers were dropped taking the last(latest) duplicate of each case number as the instance of case number
- The extra columns needed(Period, Day, Month and Violent crime/Non-violent crime) for modelling were included in the data before treating for outliers
- All non-numeric columns were transformed using labelencoder to allow for treatment of outliers across all columns and to enable further processing of data
- A correlation plot was also done to observed the relation between the columns of the dataset and the following were observed;
  - We can deduce the following from the correlation matrix;
  - Case number and ID have a very strong correlation of 0.99
  - Beat and District have a very strong correlation of 0.94
  - ID and Year have a strong correlation of 0.98
  - Case Number and Year have a strong correlation of 0.99
  - IUCR and FBI Code have a strong correlation of 0.86
  - X coordinate and Longitude are have a strong correlation of 1
  - Y coordinate and Latitude are have a stronq correlation of 1

Data Treatment;

- Outliers were also plotted to observed the columns with the most outliers. The following actions were taken after observing both the outliers and correlation plot;
  - The 'Unnamed: 0' column was dropped because it does not have any strong correlation with both the features and output variables, appears to be a serial number for the dataset
  - The 'IUCR' column was also dropped because it does not possess any strong correlation with any of the output variables. It possesses a lot of outliers and the data stored in the IUCR is further explain in the 'Primary Type' and 'Description' columns
  - The 'Updated On' column was dropped as well because it does not appear to be useful anymore after using it to sort the data, it also does not possess any strong correlation with any of the output variables and it also possesses a lot of outliers
  - The 'Domestic' column was dropped because it does not possess any strong correlation with any of the features or output variables. The classified data is also encoded in other columns and it also possesses a lot of outliers which when removed result in a offset in the data
- The outliers were dropped using a range of ±1.5 of the Inter Quartile Range. The appear to be much more presentable after removing the outliers



Correlation Plot for all features



Outliers plot after removing outliers

Histogram plot for all pictures

Data Preprocessing;

- Dataframes were created for each prediction. For models with three predictions, all three output labels were removed from the dataframe and 3 new dataframes were created using the remaining columns in the data and one of three output labels i.e. each of the three dataframes will only have one output label
- After creating the dataframes. The following preprocessing techniques were used;
  - Rescaling
  - Standardizing
  - Normalizing

Feature selection;

- The Extra tree classifier was used to select the 5 most important features for the the prediction of each label

Prediction Models;

- The first two models required three predictions as a result the predictions were done for each of the newly created dataframes and their results were stacked using for loops
- The last prediction had only one output label as a result no stacking was required
- Several algorithms were used in an aim to achieve accuracy and the algorithms with the most accurate values were chosen
- Find below the details for each prediction model and the Mean Square Error (MSE);
  - *Model to predict crime type by day type and district*
    - *Regression models(Decision Tree Regression and Ridge Regression were used given that they gave the highest accuracy values while testing other modelling techniques)*
    - *PrimaryType label MSE: -0.21750304093820927*
    - *Day label MSE: -4.042402571580112*
    - *District label MSE: -5.502889792033616*
  - *Model to predict violent and non-violent crimes by day period and ward*
    - *Models (CART Classification, Decision Tree Regression and Ridge Regression were used given that they gave the highest accuracy values while testing other modelling techniques)*
    - *ViolentCrime label MSE: -0.024470654982165347*
    - *Period label MSE: -0.4896754569554302*
    - *Ward label MSE: -17.632330223928697*

Prediction Models;

- *Model to predict where the next crime would happen by district*
  - *Regression model(Decision Tree Regression was used given that it gave the highest accuracy value while testing other modelling techniques)*
  - *District label MSE: -0.04307613723451299*

Thank you