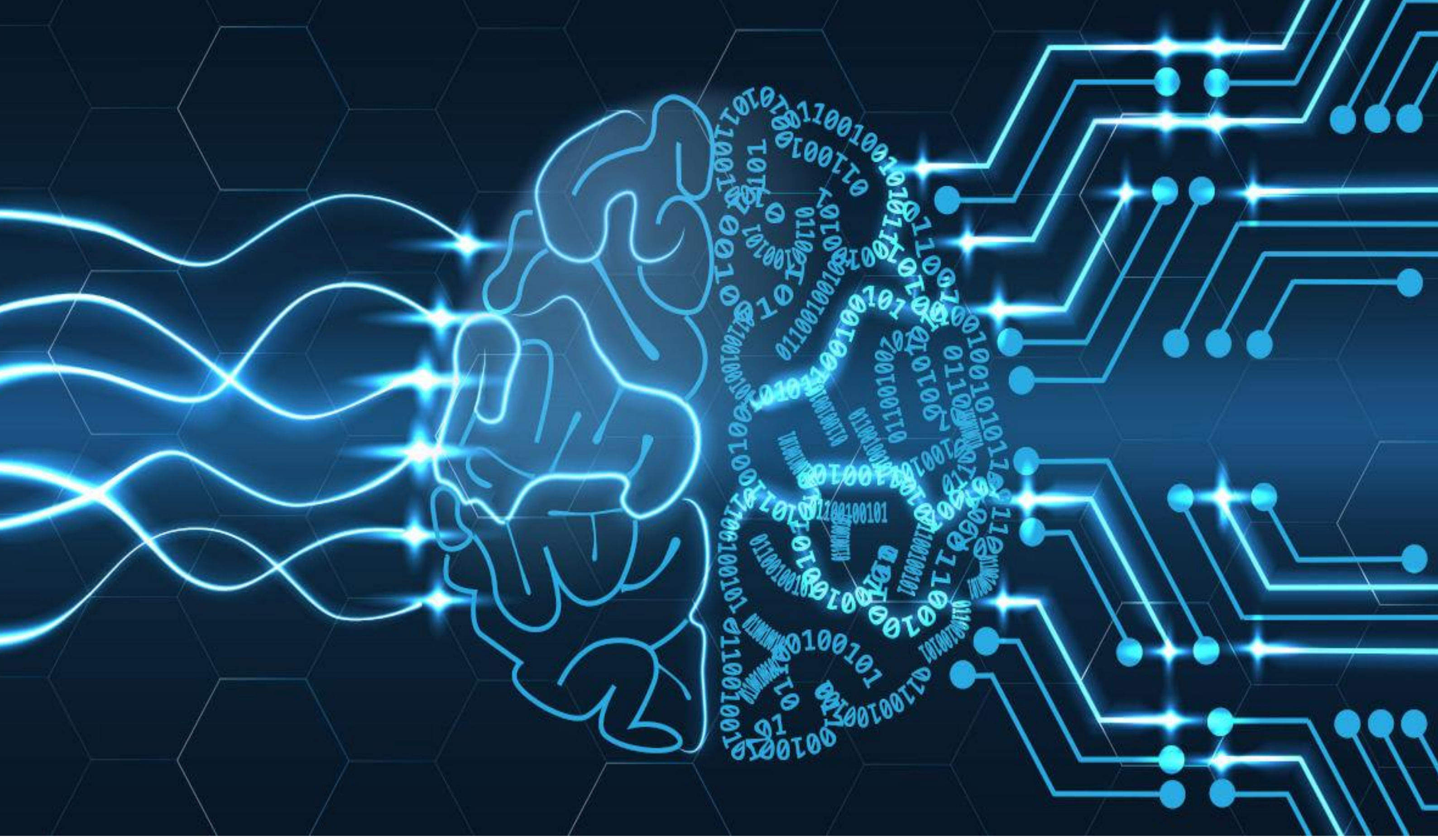


# Philadelphia Crime Analysis

# Group Members

| Name | ID |
| --- | --- |
| Seif Tarek | 92160070 |
| Eslam Hesham | 92160067 |
| Fares Ashraf | 92160078 |
| Abdel-Majid Ihab | 92160071 |

# Table of Contents

1. Introduction & dataset description: .... 4
2. Dataset Cleaning & EDA: .... 4
  2.1 Cleaning .... 4
  2.2 General EDA .... 5
  2.3 Time Series Visualization & EDA .... 20
3. Time Series: .... 25
  3.1 Preprocessing (Making data stationary): .... 25
  3.2 ARIMA Forecasting Model (Autoregressive Integrated Moving Average): .... 32
4. Violent / non-violent Crimes Classification: .... 33
  4.1 Preprocessing .... 33
  4.2 Machine Learning Models Training & Evaluation: .... 34
  4.3 Neural Network Implementation .... 35

# List of Figures

Figure 1. Total number of crimes committed per year. .... 5
Figure 2. Number of crimes committed per Month. .... 6
Figure 3. Number of crimes committed per hour. .... 7
Figure 4. Number of Times a Specific Crime was Committed. .... 8
Figure 5. Number of Times a Specific Crime was Committed. .... 8
Figure 6. A bivariate plot that shows the crime types versus police districts. .... 10
Figure 7. A bivariate figure that shows the types of crimes committed in each year. .... 12
Figure 8. A plot shows the crime counts by police district per year. .... 13
Figure 9. A plot that shows the location of each crime that was determined by latitude and longitude coordinates. .... 14
Figure 10. A plot that shows the location of each crime that was determined by latitude and longitude coordinates, points are coloured with respect to crime type. .... 15
Figure 11. A live heatmap that shows the number of crimes in each zone over the city of Philadelphia along with pins that represents the police districts in city. .... 16
Figure 12. A live heatmap that shows the number of crimes in each zone over the city of Philadelphia. .... 17
Figure 13. A pair plot which shows different relationships with the features, and it is coloured on the classes we made. .... 18
Figure 14. Histograms for the features to further see the distribution of the feature values among the datasets. .... 19
Figure 15. The correlation matrix which gives insights about how strong the features are related to the target value which is in our case the UCR_General. .... 20

Figure 16. Number of Crimes During Period of Time. 21
Figure 17. Number of Non-Violent Crimes During Period of Time. 21
Figure 18. Number of Violent Crimes During Period of Time. 22
Figure 19. The crime prime time per year which means is the average number of crimes per day versus the month. 22
Figure 20. Different categories of crimes based on when this crime was committed during the day. 23
Figure 21. The average number of crimes that are committed in special days. 24
Figure 22. Time series components for violent crimes in annual manner. 25
Figure 23. The first test we had which was on the original data where it fails the augmented Dickey-Fuller Test. 26
Figure 24. The results of the augmented Dickey-Fuller Test on the monthly difference data. 27
Figure 25. The results of using the seasonality data by applying difference in time of 12. 28
Figure 26.The results of using seasonality from our seasonal decomposition and use a difference in time of 1. 29
Figure 27. Autocorrelation (ACF). 31
Figure 28. Partial Autocorrelation (PACF). 32
Figure 29. MA model results. 32
Figure 30. AR model results. 33
Figure 31. AR+MA model results. 33
Figure 32. Feature importance from Random Forest. 35
Figure 33. Learning Curves Comparison between the applied models. 35
Figure 34. Results of the MLP. 36

# 1. Introduction & dataset description

Philadelphia, Pennsylvania's largest city, is notable for its rich history, on display at the Liberty Bell, Independence Hall (where the Declaration of Independence and Constitution were signed), and other American Revolutionary sites. Also iconic are the steps of the Philadelphia Museum of Art, immortalized by Sylvester Stallone's triumphant run in the film "Rocky." In such a large city, it is very important to know what is the most committed crime type and the places that contain the highest crime rates in Philly in order to distribute the police forces in the heavy crime places to prevent such crimes.

In order to being able to determine heavy crime and dangerous places, Machine Learning techniques has been applied on "crime incidents" dataset that collected from "opendataphilly.com" which is a csv file that contains over 2 million records about the crimes that was done in Philadelphia from 2006 until 2017. The dataset contains multiple feature such as the district id, police district number, UCR number (which is the indicator to whither the crime is violent or not), dispatch date (which is the date and time that the police was dispatched to deal with the crime), location block, the crime types (Text_General_Code), the month of the crime, and the location in longitude and latitude coordinates. To start with this dataset some EDA and pre-processing has been performed on the dataset to get some insights from it and to make it ready for the modeling phase. Then, time series forecasting to predict the for new crime rates; machine learning, and neural network techniques was applied on the dataset to classify whither the crime is violent or not.

# 2. Dataset Cleaning & EDA:

## 2.1 Cleaning

The first thing to do is to clean the dataset to make it ready for modeling phase. After some basic data exploration, it was found that there were some missing values, and the time-related features are not in date time data type. So the cleaning starts by removing all the null values which were around 9k so there was no problem of getting rid of all of them, then all the date time related features were set to datetime data type, and there were 4 new columns that were made to express year, month, day, and crime that will be used in the time series forecasting.

## 2.2 General EDA

The EDA starts by printing all the crimes and grouping it by the districts to see which districts has the most crimes and we printed the UCR general against the crime type to see which crime corresponds to which UCR value which gave us the insight that all the crime types that were under the value of 800 were the violent crimes and the crimes above that threshold are the non-violent crimes. We also printed the number of crimes in each location block to see the distribution of the crimes among the block of the city. Then we started plotting some plots to give us more insight about the dataset.

Figure 1. Total number of crimes committed per year. The first figure shows you the total number of crimes committed per year. From this figure we can conclude that almost every year there is a decrease in the number of crimes happening. As shown below, the number of crimes is decreasing as the years increase.

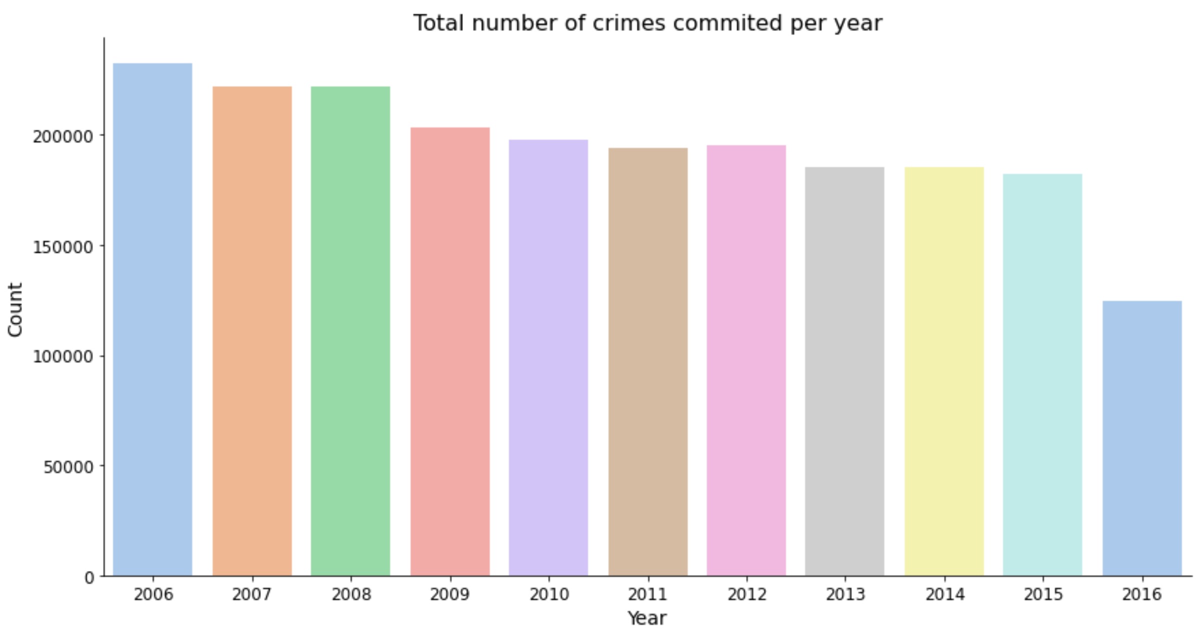



Figure 1. Total number of crimes committed per year.

Figure 2. Number of crimes committed per Month. Figure 2 shows us the number of crimes happening in each month. Given this figure we can see that a decrease happens towards the end of the year throughout the beginning of the year. This might be due to the outside temperature of these months. As shown in the figure, summertime moths have more crime counts than other months.

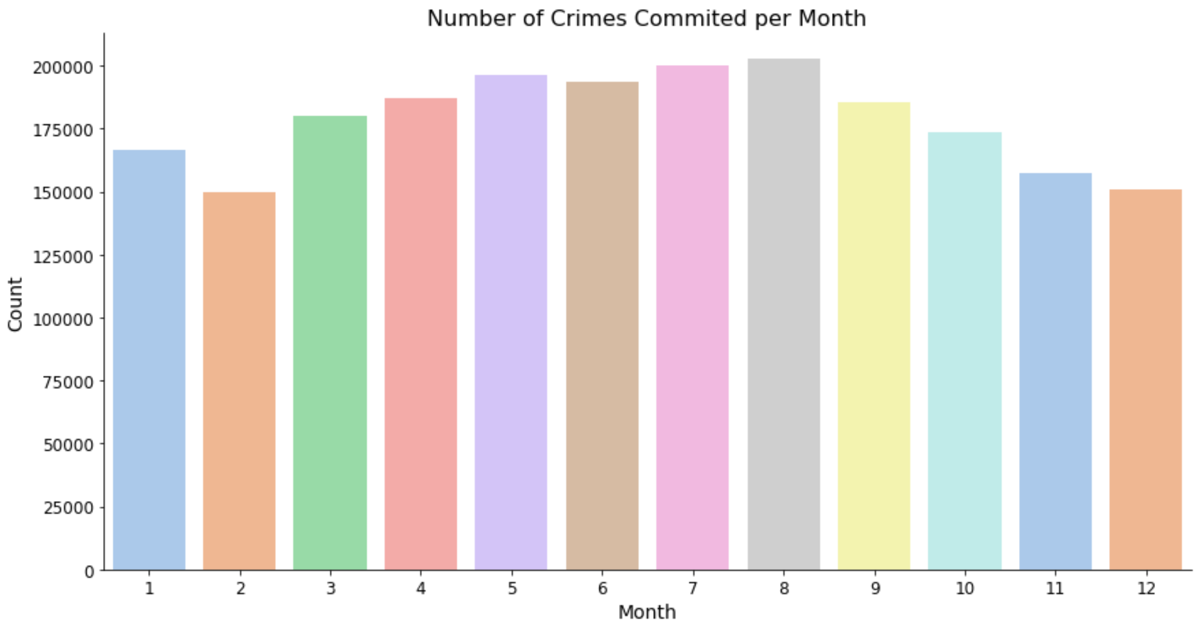



Figure 2. Number of crimes committed per Month.

Figure 3. Number of crimes committed per hour. In the third figure we can see that the number of crimes happening reaches a peak at 16:00h and is at its lowest point around 06:00h. Which is probably due to the sleeping schedule of people? Furthermore, it might be interesting to look at the different types of crime happening around each hour.

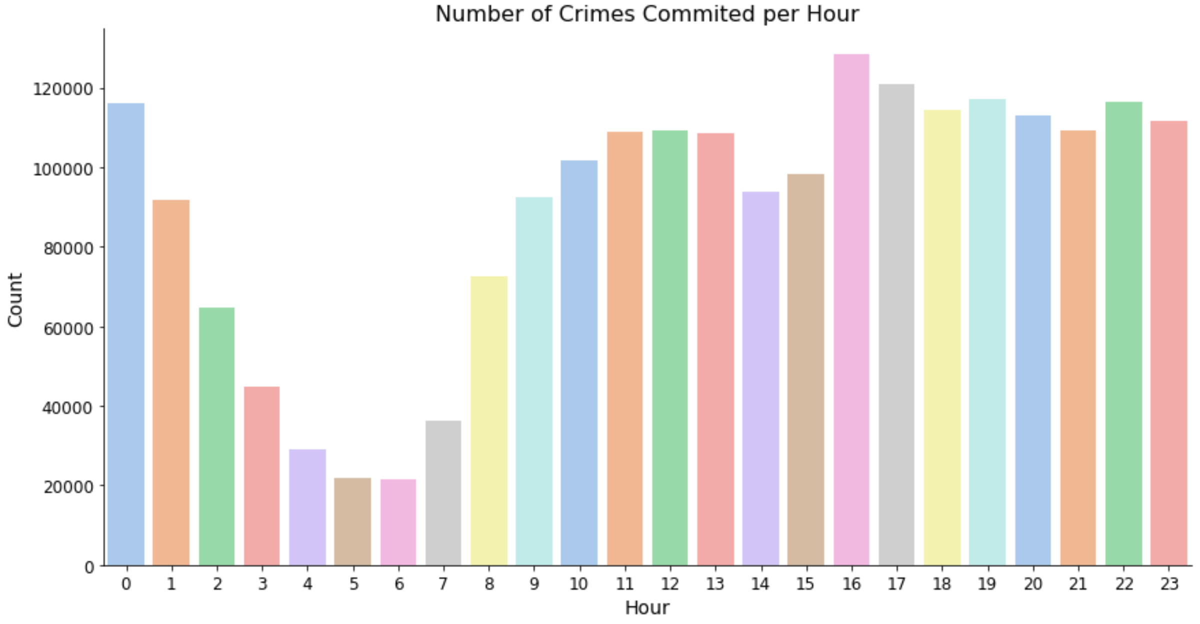



Figure 3. Number of crimes committed per hour.

Figure 4. Number of Times a Specific Crime was Committed. In figure 4, the number of times each crime is committed is visualized. Where the named category Assaults and Theft are crimes that are committed the most within Philadelphia. As shown in the figure, the most committed crime is theft.

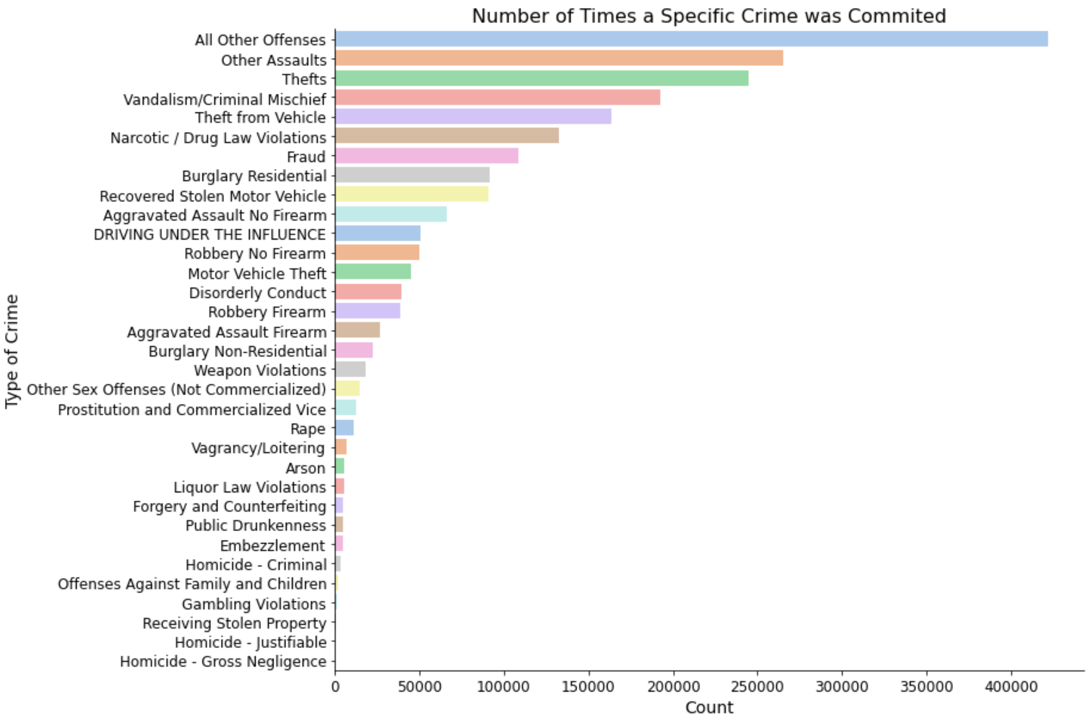



Figure 4. Number of Times a Specific Crime was Committed.

Figure 5. Number of Times a Specific Crime was Committed. In the last figure, figure 5, the number of times a crime is committed in each district is visualized. From this we can see that most crimes happen district 11 and the least crimes are in district 22.

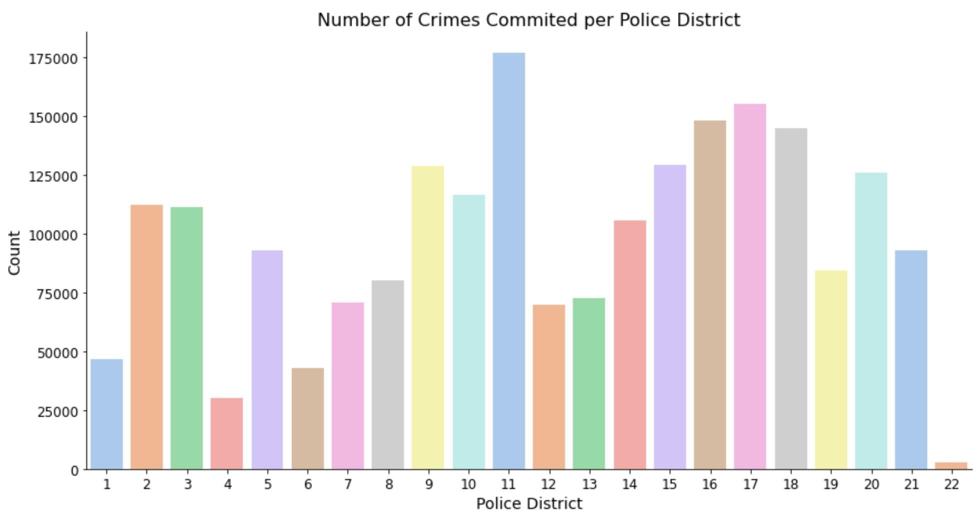



Figure 5. Number of Times a Specific Crime was Committed.

Figure 6. it is a bivariate plot that shows the crime types versus police districts. Theft is most common crime type with over 7000 in each district. There were over 20000 thefts in police districts 2, 3, 6, 9, 14, 15, 18, 22, and 24. There were more than 20000 incidents covered by 'All Other Offenses' - the second most common category, in districts 12, 15, 19, 22, and 24. The 'Other Assaults' category was the third most common incident type. There were over 10000 incidents in 13 out of 21 districts. There were over 15000 crimes of this category in police districts 14, 15, 22, 25, and 35. The 15th district, in the near northeast, saw over 22000 "other" assaults in the 2006 through 2015 period. There were more than 17000 Vandalism crimes committed in the 15th district. The top five police districts for vandalism were districts 3, 12, 14, 15, and 22. Drug Violations took place more in the 19th, 22nd, 24th, 25th, and 35th districts. There were more than 10000 Motor Vehicle Thefts in districts 15, 24, and 25. The 12th district had just below 10000 incidents. The five most common police districts for bulgarlies were the 12th, 14th, 15th, 22nd, and 24th. The worst police districts for murder were the 12th, 19th, 22nd, 24th, 25th, 35th, and 39th with over 200 each for the ten-year period 2006 through 2015. There were over 150 homicides in the 14th, 15th, and 18th districts. The worst police districts for rape were the 12th, 14th, 15th, 22nd, 24th, 25th, and 35th, with over 700 each in the ten-year period. The most robberies occurred in the 15th, 22nd, 24th, 25th, and 35th districts. Aggravated assaults were most prominent in the 12th, 15th, 22nd, 25th, and 35th districts.

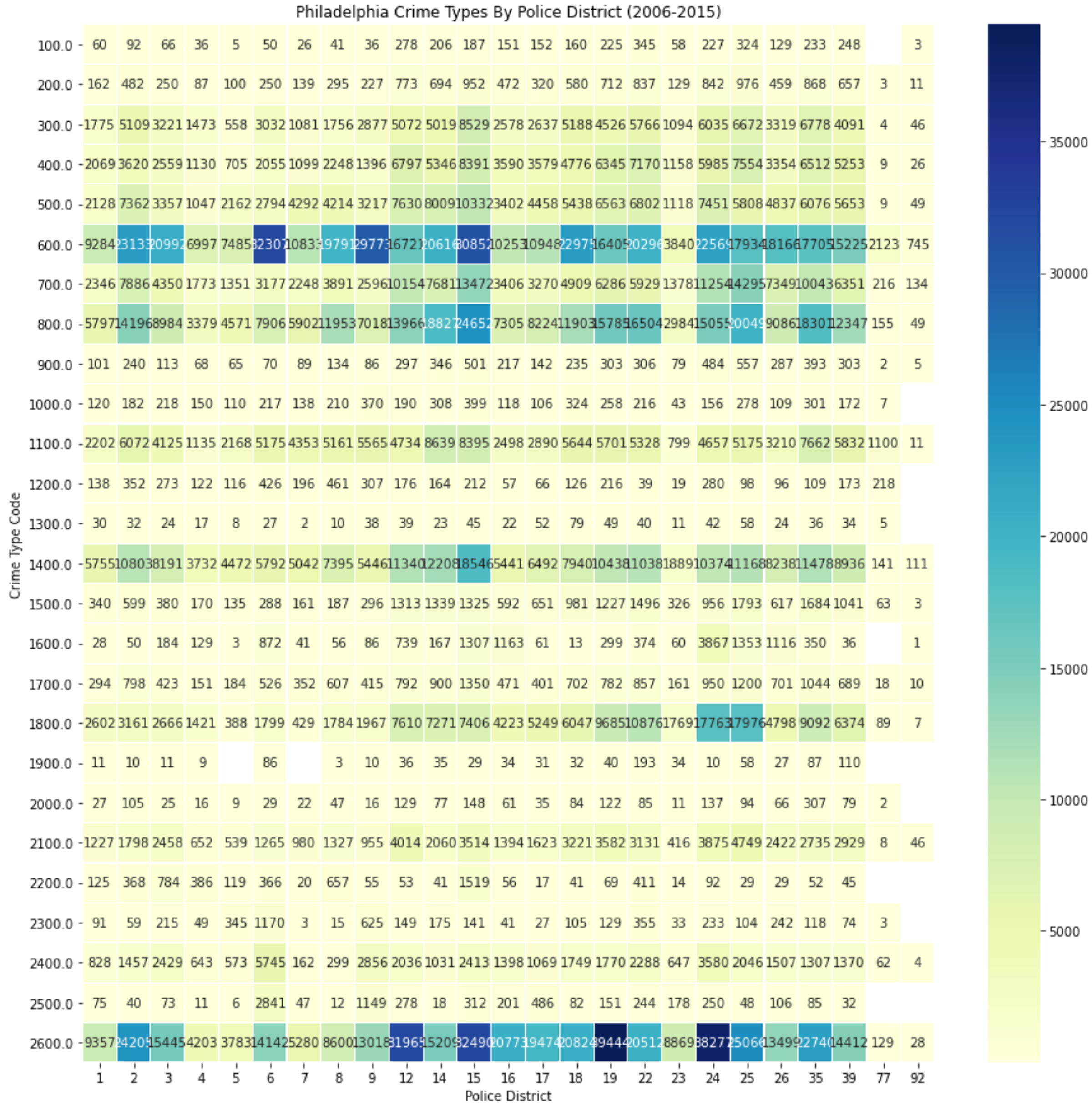



Figure 6. A bivariate plot that shows the crime types versus police districts.

Figure 7. a bivariate figure that shows the types of crimes committed in each year. The top crime type was not a factor in the declined crime rate. Thefts remained steady from 2006 through 2014 and finally fell from 37197 in 2014 to 36063 in 2015. The 'All Other Offenses' category fell steadily from 2006 through 2011 but increased again through 2015. However, there were over 40000 such incidents in 2006, which helps explain its higher incident count. The 'All Other Assaults' category saw a slight decline in the ten-

year period with some fluctuations. Vandalism, Motor Vehicle Thefts, and Burglaries fell in the 2006 through 2015 period, driving the overall counts downwards. The number of Vandalism incidents fell by over 8000 between 2006 and 2015. Motor Vehicle Thefts saw a rapid decline as well - from 20284 in 2006 to 8845 in 2015! Burglaries fell rapidly between 2013 and 2015 (from 10286 to 7998). Drug violations also fell steadily from 2020 to 2015. The highest number of drug violations was 15757 in 2008; the lowest number was 7516 in 2015. The fewer number of Vandalism, Motor Vehicle Thefts, Bulgaries, and Drug Violations helped drive down the crime rate. Here is the low-down on the worst (but not as common) crime types. There were more than 300 murders per year from 2006 through 2012. The murder counts declined by 25% from 2012 to 2013! The counts stayed below 300 through 2015. Rapes declined from 2006 through 2011 but increased to their highest level of 1298 in 2015. Rapes increased by 43% between 2012 and 2013, the same time murders fell the most! Robberies decreased from 10732 in 2006 down to 6728 in 2015. Aggravated assaults also fell from 10176 in 2006 down to 7308 in 2014 with slight increase in 2015. The biggest decrease in crime occurred between 2008 and 2009, driven by decreases in nearly every category

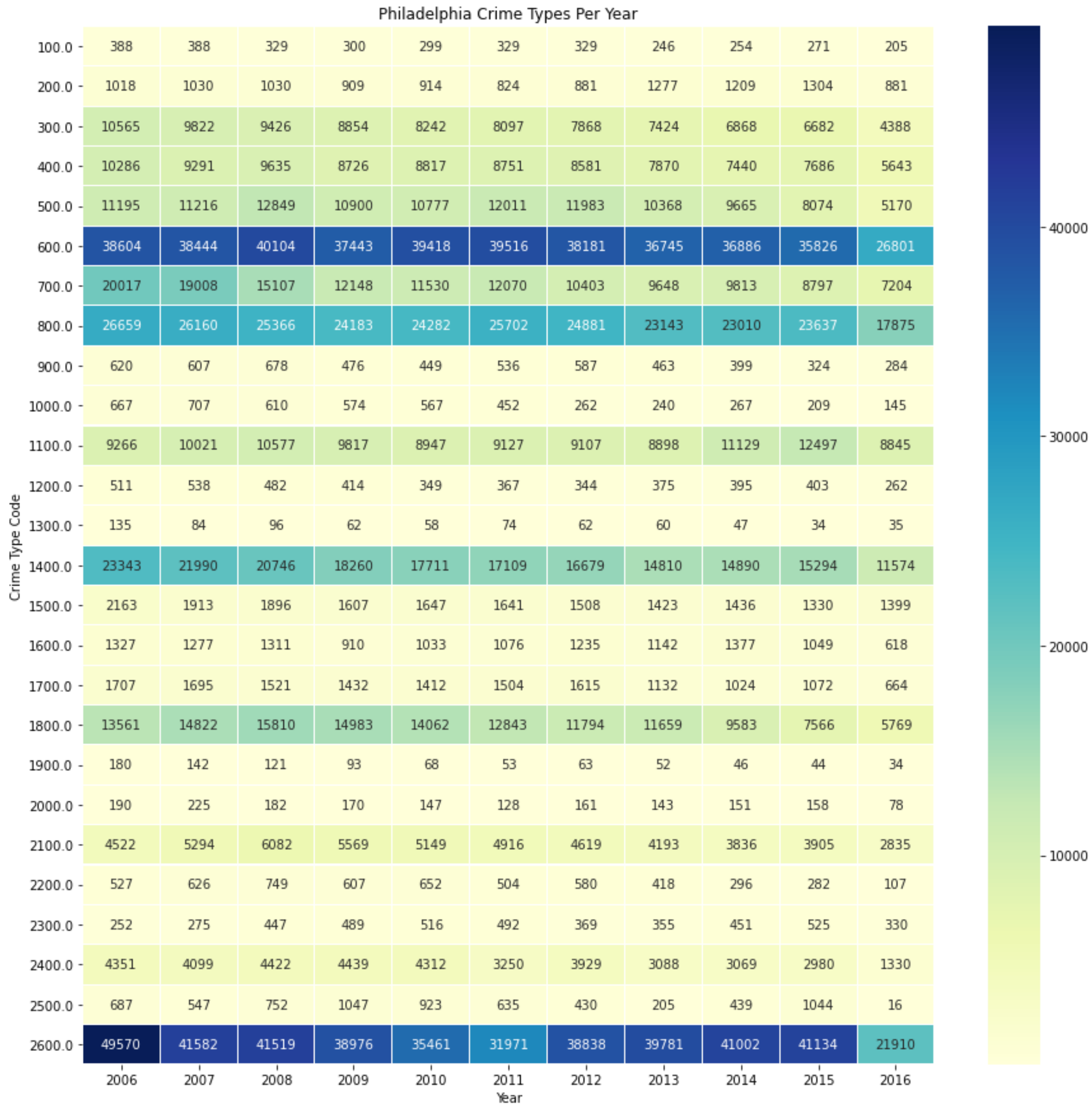



Figure 7. A bivariate figure that shows the types of crimes committed in each year.

Figure 8. a plot shows the crime counts by police district per year. The observation from the plot shows the decrease in the number of criminal incidents from 2006 through 2015 in all but one police district. Highlights include the following: The 3rd district had its incident count fall nearly every year from 13812 in 2006 to 7883 in 2015, a 43% decline. The 5th district incident count fell by nearly 50% in the ten-year period from 3668 to 1948. The 8th district saw a decline from 7878 in 2006 down to just 4840

in 2015. The 12th and 14th districts saw the counts fall by over 3000 each. The 17th district seen its numbers fall by over 40%. The 22nd district saw the incidents cut by 33%. The 25th saw close to 40% decline. The 35th district saw about a 23% decline in crime incidents. The 19th district saw an increase in the count.

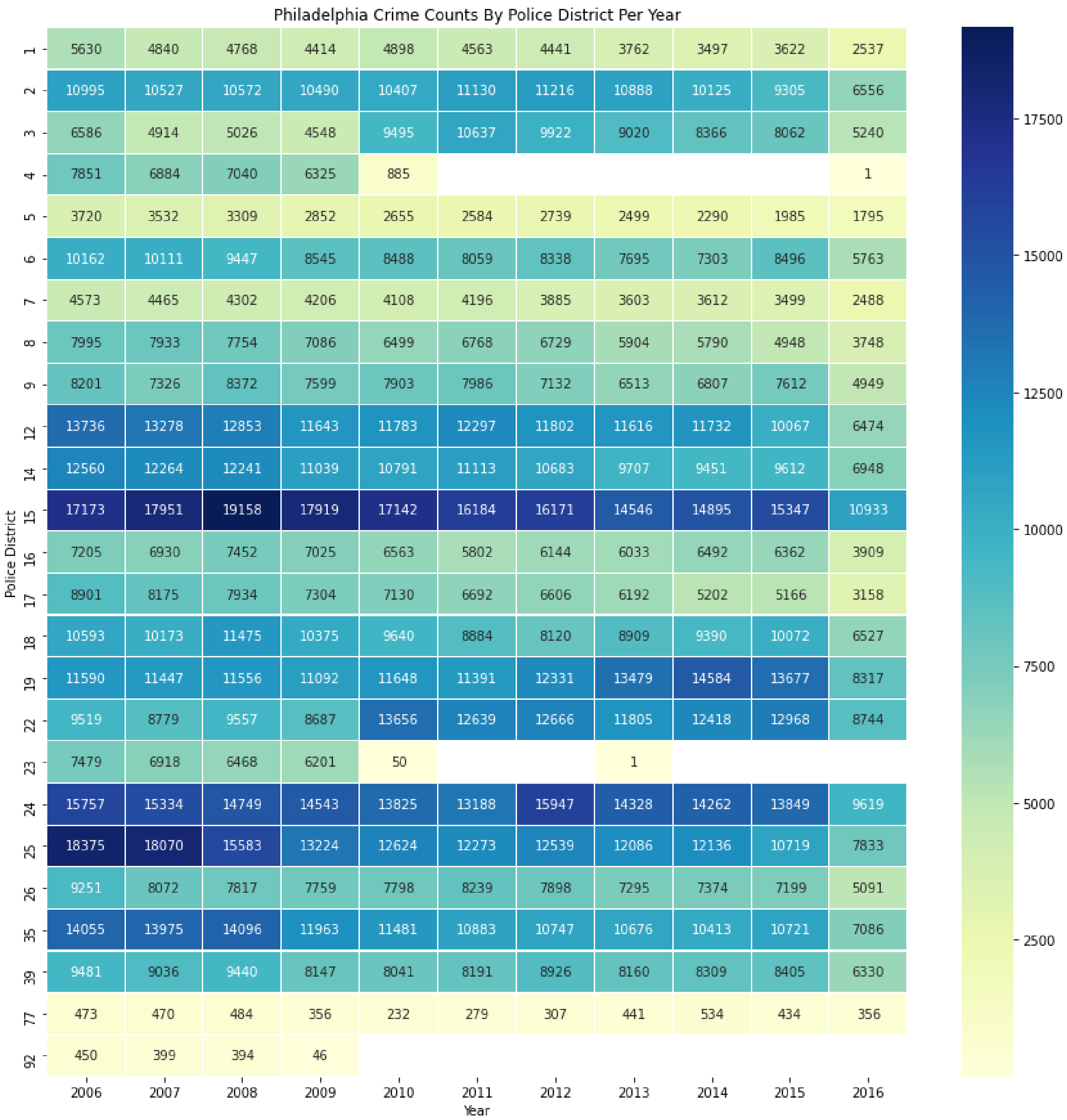



Figure 8. A plot shows the crime counts by police district per year.

Figure 9. a plot that shows the location of each crime that was determined by latitude and longitude coordinates. As shown in the figure, the points make up the map of Philadelphia.

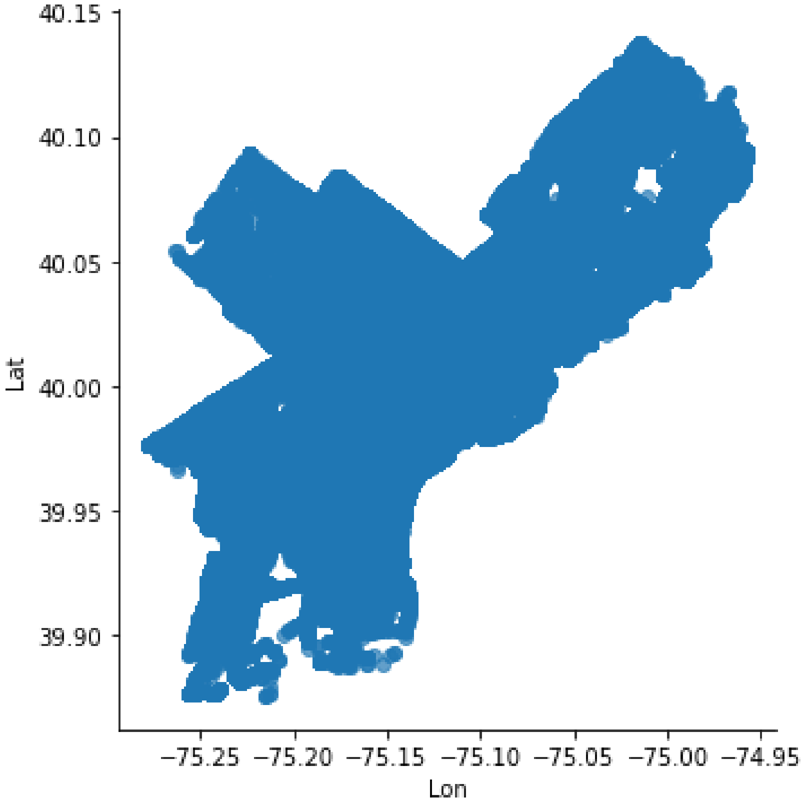



Figure 9. A plot that shows the location of each crime that was determined by latitude and longitude coordinates.

Figure 10. the same figure as figure 9 but the points are colored with respect to the crime type.

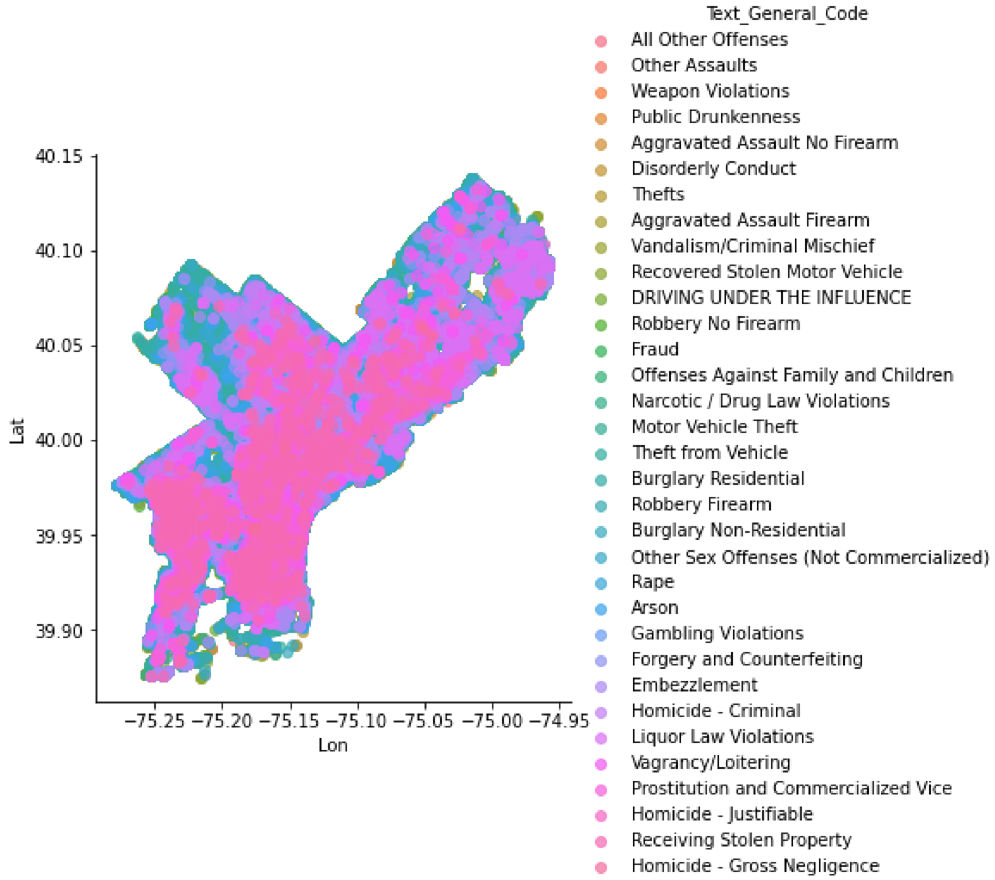



Figure 10. A plot that shows the location of each crime that was determined by latitude and longitude coordinates, points are coloured with respect to crime type.

Figure 11. a live heatmap that shows the number of crimes in each zone over the city of Philadelphia along with pins that represents the police districts in city.

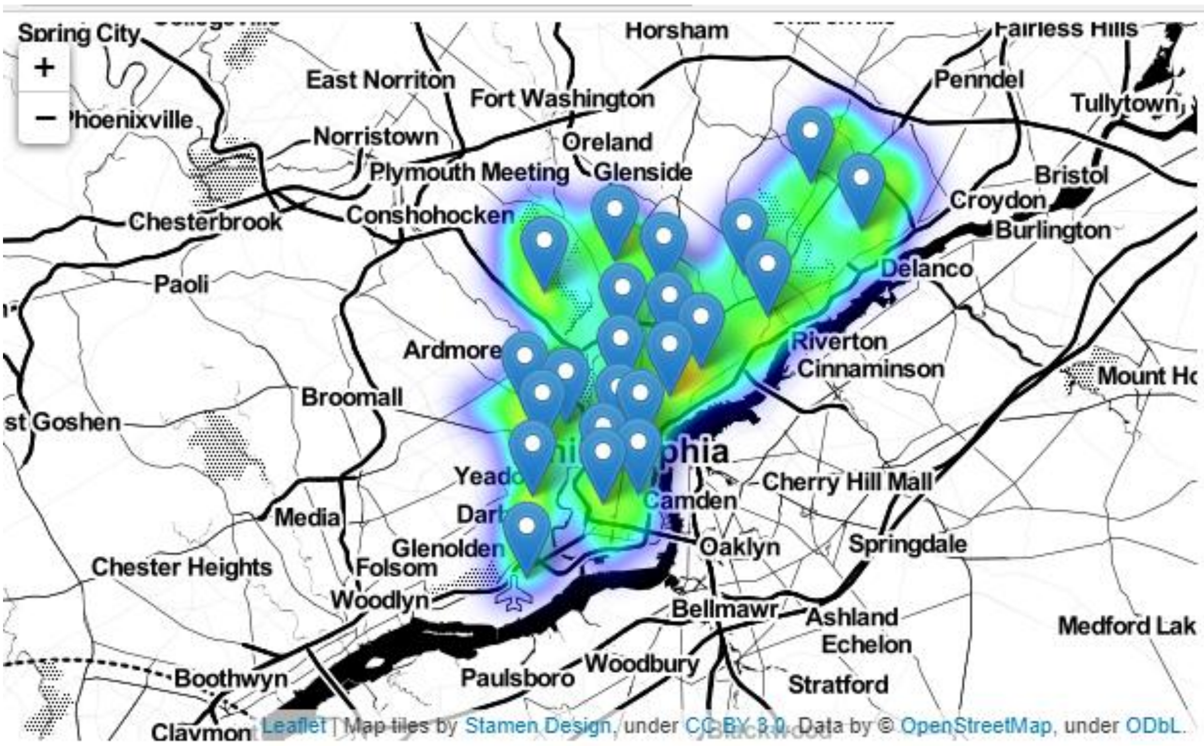



Figure 11. A live heatmap that shows the number of crimes in each zone over the city of Philadelphia along with pins that represents the police districts in city.

Figure 12. the same heatmap above but with out the district pins to give a clear view on the numbers of the crimes in each place in the city.

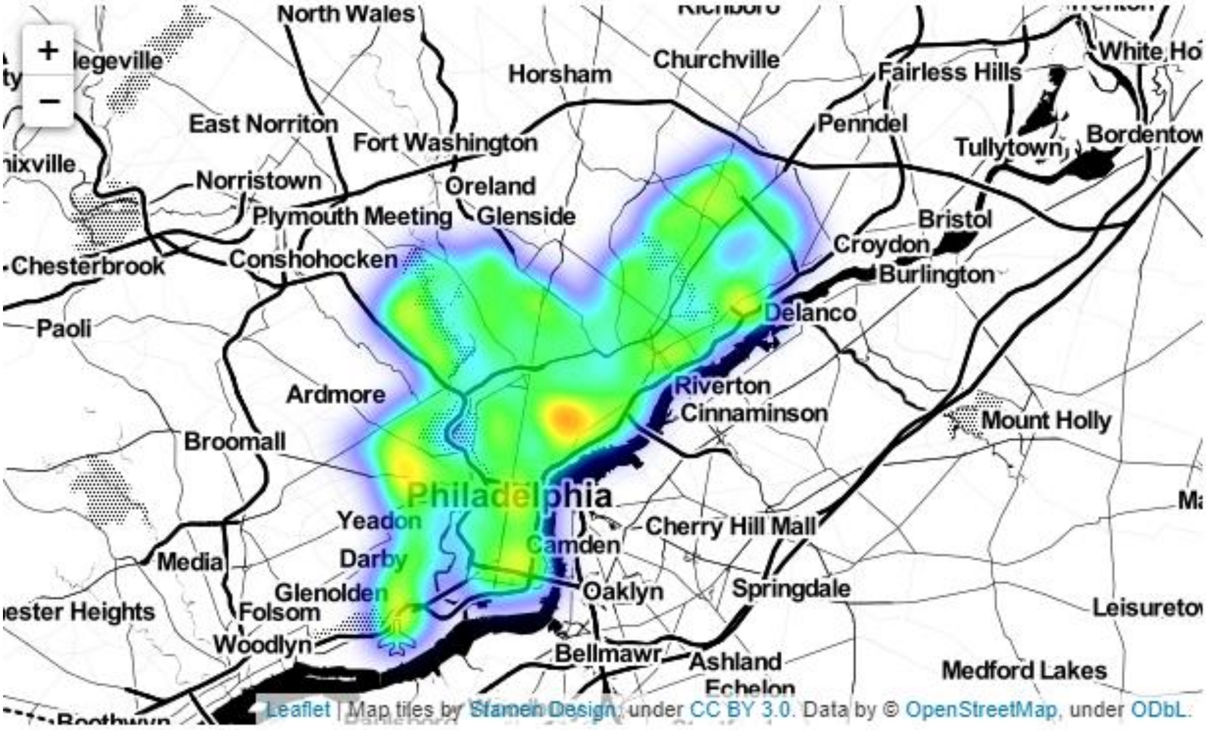



Figure 12. A live heatmap that shows the number of crimes in each zone over the city of Philadelphia.

Figure 13. is a pair plot which shows different relationships with the features, and it is colored on the classes we made.


Figure 13. A pair plot which shows different relationships with the features, and it is coloured on the classes we made.

Figure 14. shows histograms for the features to further see the distribution of the feature values among the datasets.

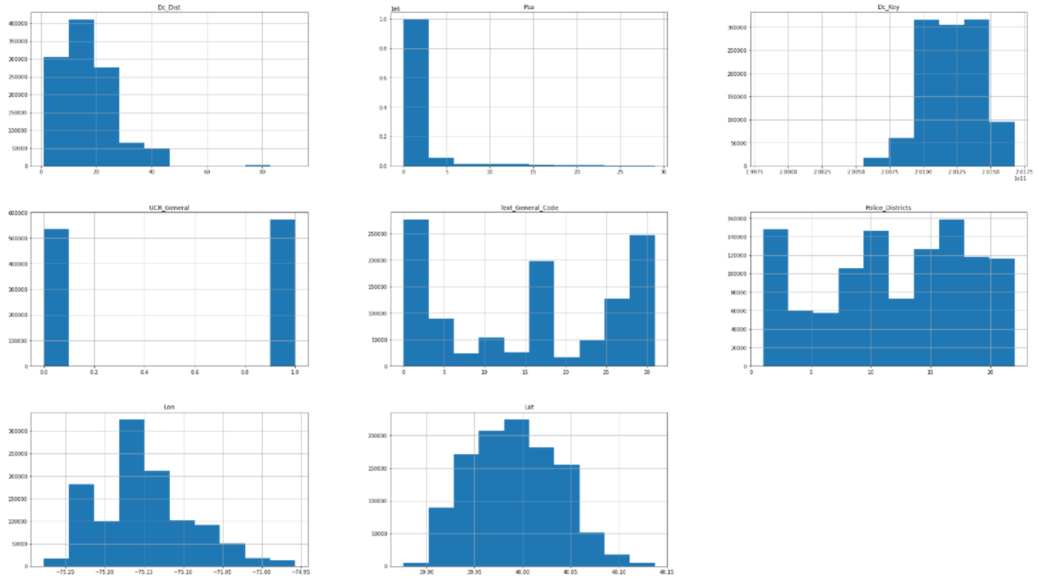



Figure 14. Histograms for the features to further see the distribution of the feature values among the datasets.

Figure 15. shows the correlation matrix which gives insights about how strong the features are related to the target value which is in our case the UCR_General. The figure shows that the text_general_Code is the feature with the strongest positive correlation with the target value.

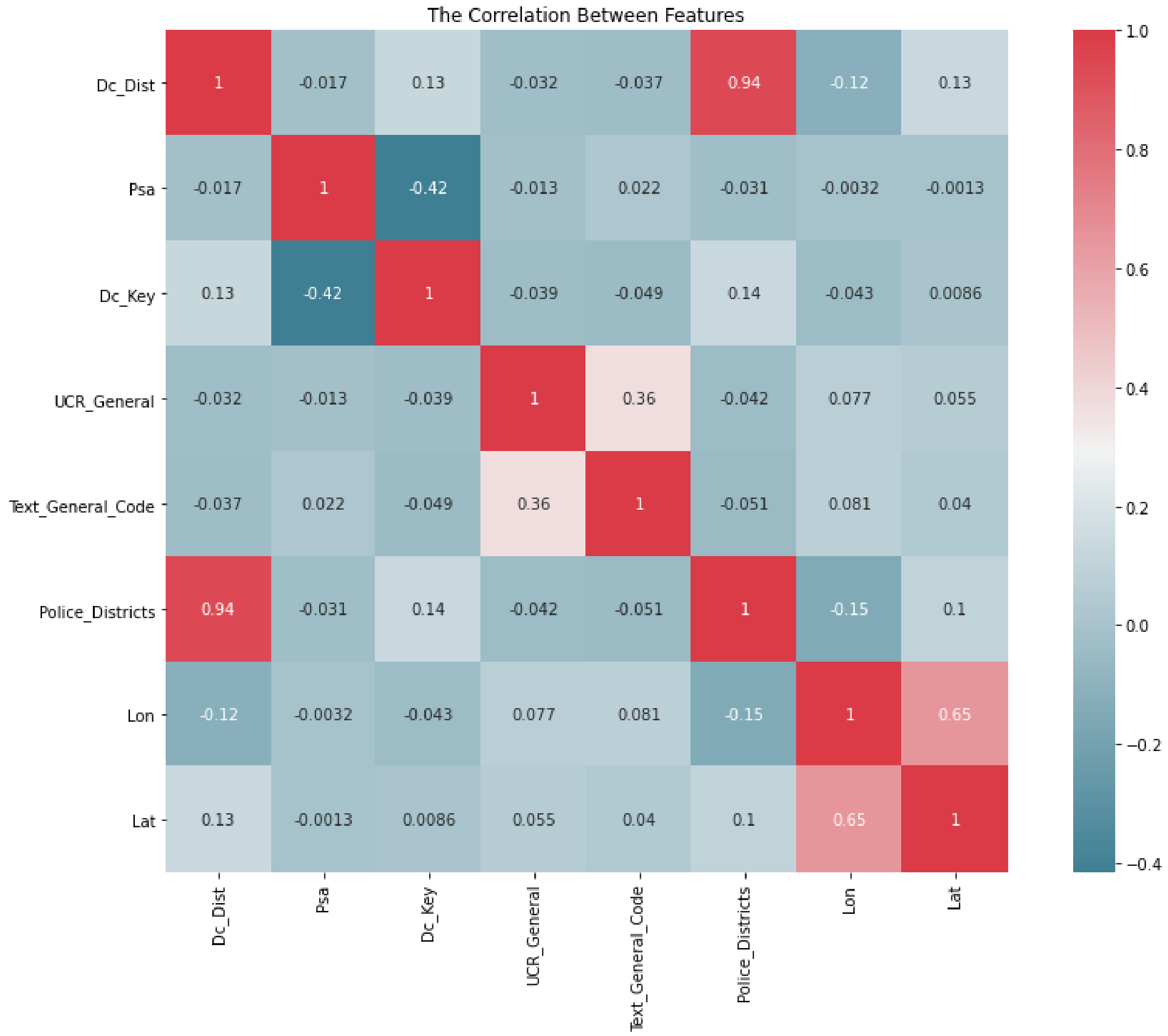



Figure 15. The correlation matrix which gives insights about how strong the features are related to the target value which is in our case the UCR_General.

## 2.3 Time Series Visualization & EDA

Time series data visualization have multiple visualization types to visualize your data in a format that makes the most sense for your use case. Most known types are line graph, and bar graph.

A line graph is the simplest way to represent time series data. It helps the viewer get a quick sense of how something has changed over time. A line graph uses points connected by lines (also called trend lines) to show how a dependent variable and independent variable changed:

- An independent variable, true to its name, remains unaffected by other parameters.
- The dependent variable depends on how the independent variable changes.

For temporal visualizations, time is always the independent variable, which is plotted on the horizontal axis. Then the dependent variable is plotted on the vertical axis. In this case a line graph time series analysis visualization has been used on “crime incidents” dataset for determining total crimes, non-violent crimes, and violent crimes were occurred in Philadelphia during a specific time, analyzed time spans were by day, by month, and by year as represented in figure 16,17, and 18.

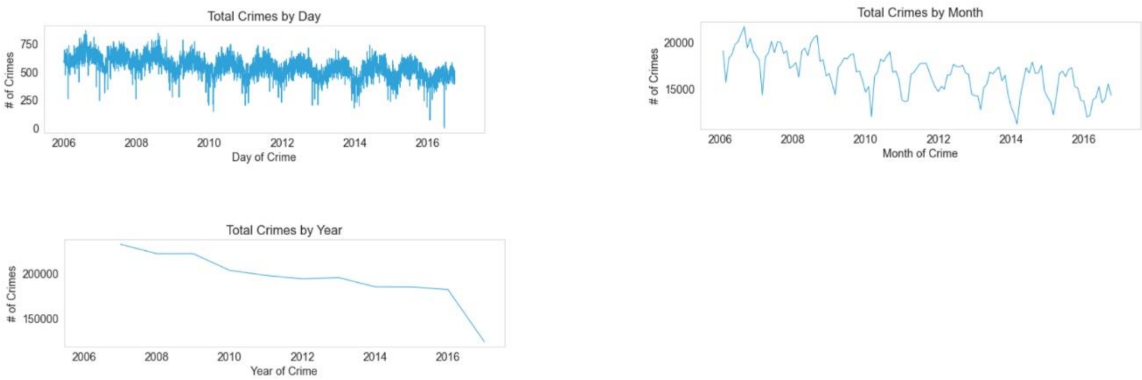



Figure 16. Number of Crimes During Period of Time.

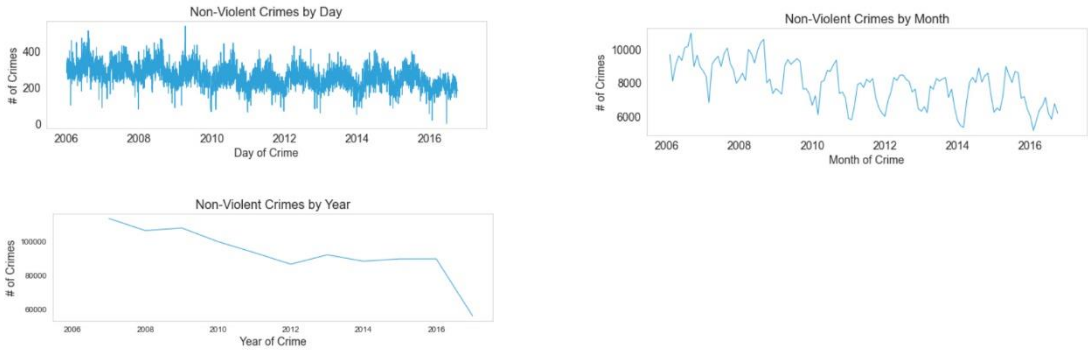



Figure 17. Number of Non-Violent Crimes During Period of Time.

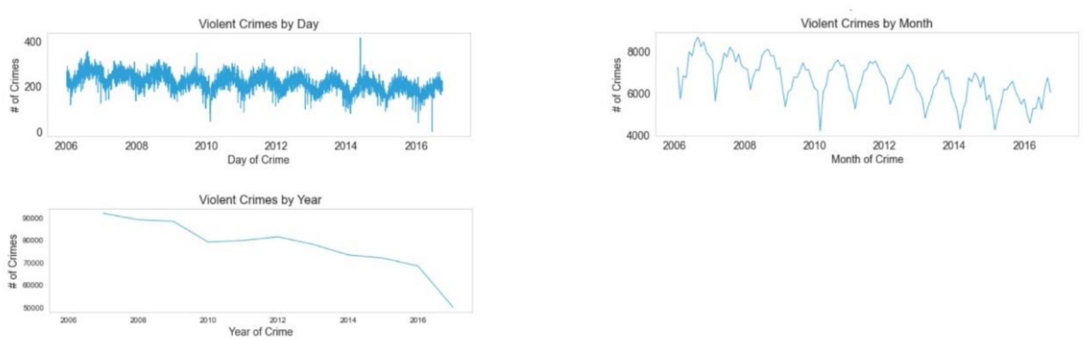



Figure 18. Number of Violent Crimes During Period of Time.

Figure 19 shows the crime prime time per year which means is the average number of crimes per day versus the month. The figure shows a clear pattern The summer months (6,7,8) show the highest number of crimes per day whereas the winter months (1,2,12) have the lowest. And the months during spring (3,4,5) lead up to the summer high whereas the months during fall (9,10,11) lead down to the winter low.

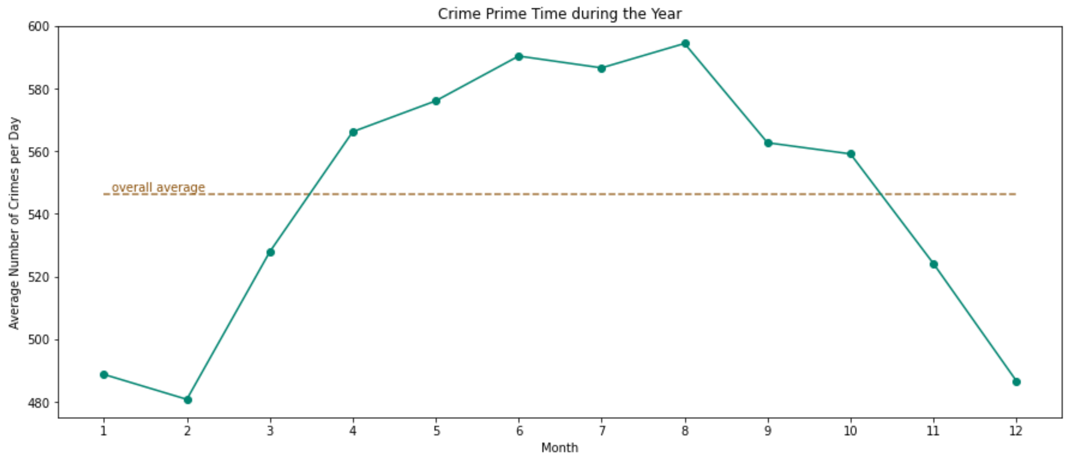



Figure 19. The crime prime time per year which means is the average number of crimes per day versus the month.

Figure 20. shows different categories of crimes based on when this crime was committed during the day. There are basically 3 types of crimes with regards to when they usually occur during the day. First, there are the "Biological Rhythm" crimes, and

they follow the pattern discussed above. Then, there are the "9-to-5" crimes which occur during the day and not so much at night. The last type are "Night Shift" crimes which happen mostly during the night and not so much during the day. There is one special case in the "Night Shift" category, namely "Prostitution and Commercialized Vice". It differs from the other crimes in that there is a clear bump around the lunch break from around 10 to 13 o'clock.

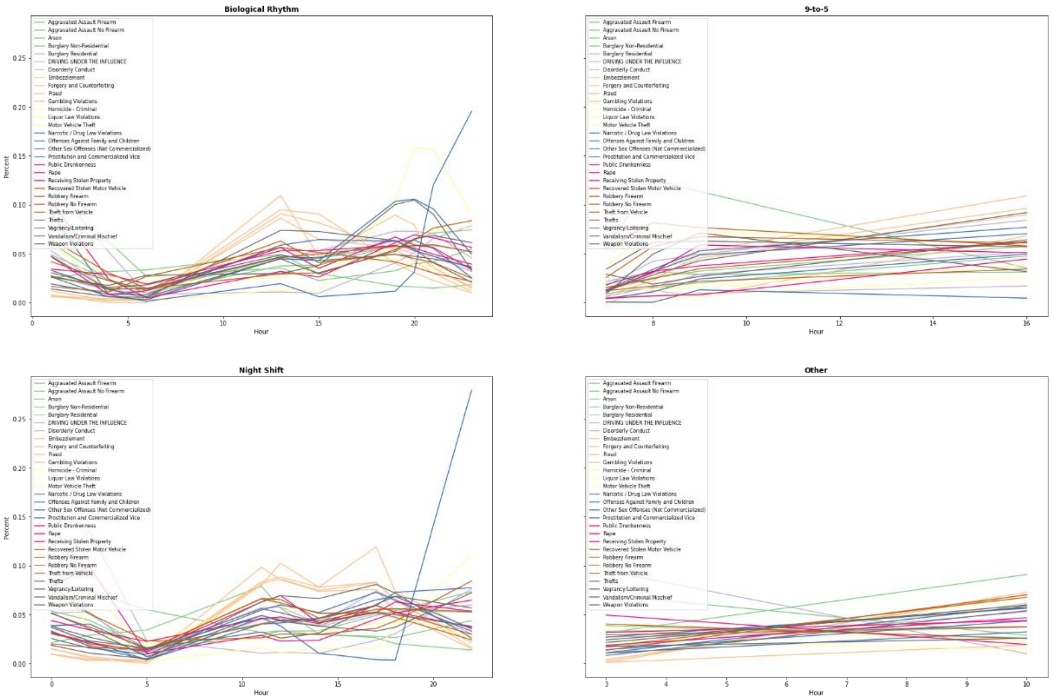



Figure 20. Different categories of crimes based on when this crime was committed during the day.

Figure 21. shows the average number of crimes that are committed in special days like Christmas, $4^{th}$ of July etc. The figure shows clear drops on the 4th of July and Christmas. Thanksgiving, however, doesn't have a fix date but instead its date always changes between the 22nd and 28th of November. And as one can see, there is a visible dent in that period. So, criminals seem to celebrate holidays, too.

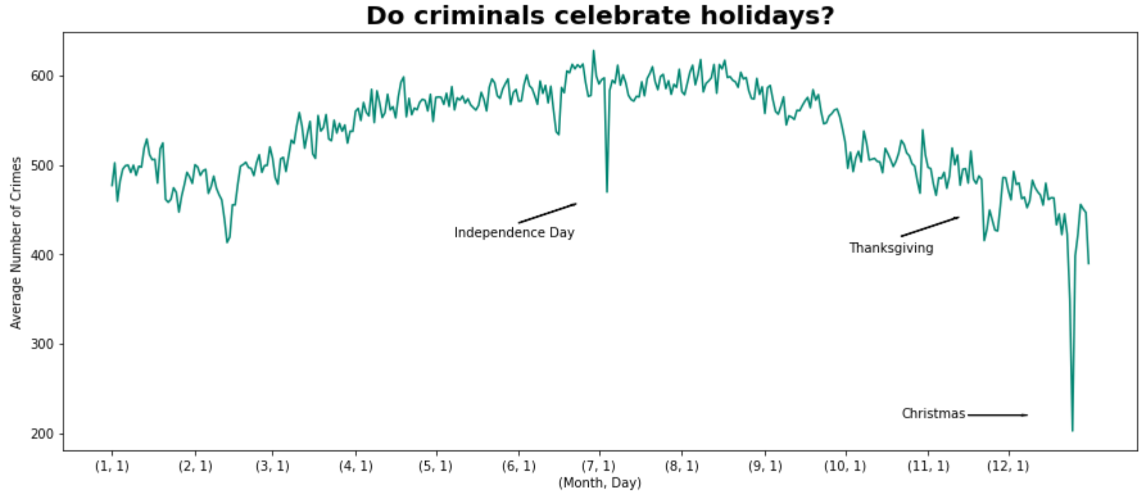



Figure 21. The average number of crimes that are committed in special days.

There are many reasons which cause our forecasted values to fall in the wrong direction, that's why it's significant to be knowing the patterns which are followed by the values with time. Basically, a time series consists of four components which are level, trend, seasonality, and noise. These components are classified into two types which are systematic that includes components of the time series that have consistency or recurrence and can be described and modeled, and second type is non-systematic which contains Components of the time series that cannot be directly modeled. The variation of those components causes the change in the pattern of the time series.

Let's dig deeper into these components:

- Level: What is the overall trend in the data? More formally It is the main value that goes on average with time.
- Trend: The trend is the value that causes increasing or decreasing patterns in a time series. Which could be described in this case as: how does crimes fluctuate between seasons?
- Seasonality: This is a cyclic event that occurs in time series for a short time and causes the increasing or decreasing patterns for a short time in a time series.

- Noise / Residuals: When removing trends and seasonality what does the data look like? In other words, these are the random variations in the time series.

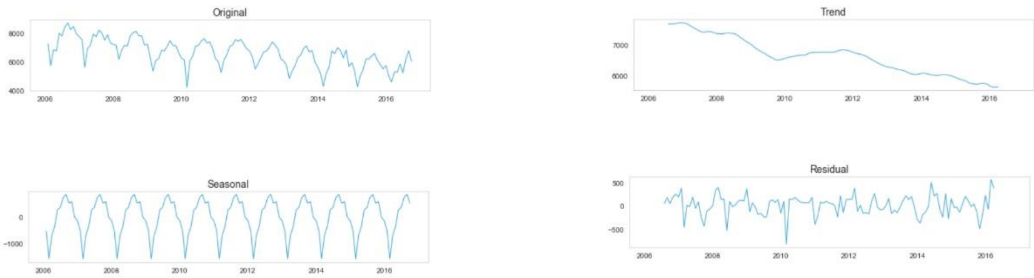


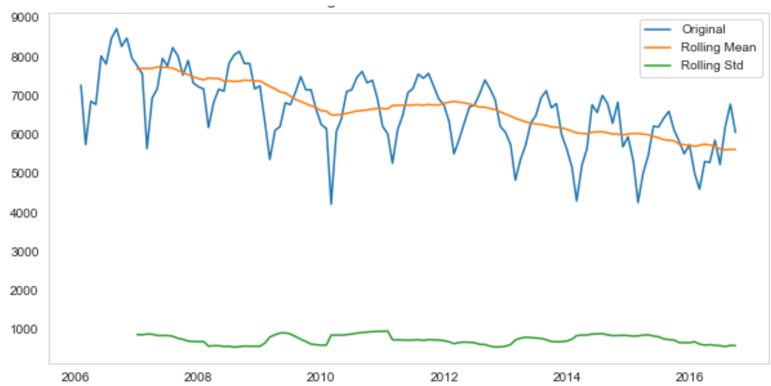

Figure 22. Time series components for violent crimes in annual manner.

# 3. Time Series

## 3.1 Preprocessing (Making data stationary)

Time series stationarity used for making sure that the predicted values accuracy is high because statistical modeling methods assume or require the time series to be stationary to be effective. So, in order to effectively use any statistical modeling technique, it is imperative to check the presence of any trend or seasonality in it. In other words, time series could be stationarity throughout freed from trend or any seasonal effects. There are three methods for testing whether time series is stationary or not: Visually, Statistical Measures, and Statistical Tests. One of these statistical tests is called dickey-Fuller test which is a test to see whither the data is stationary or not by calculating the critical values and the T-statistic value and compares between them. The less the value of T-statistic is the closer the data to stationarity. We first calculated the decomposition of the seasonal data and calculated the difference of it and the monthly data and the difference between the seasonal and monthly data, then we performed the stationarity test by plotting the mean and std of each of the calculated new data while comparing between the T-statistic value that is a result of augmented dickey-Fuller test and the critical values obtained also from the dickey-

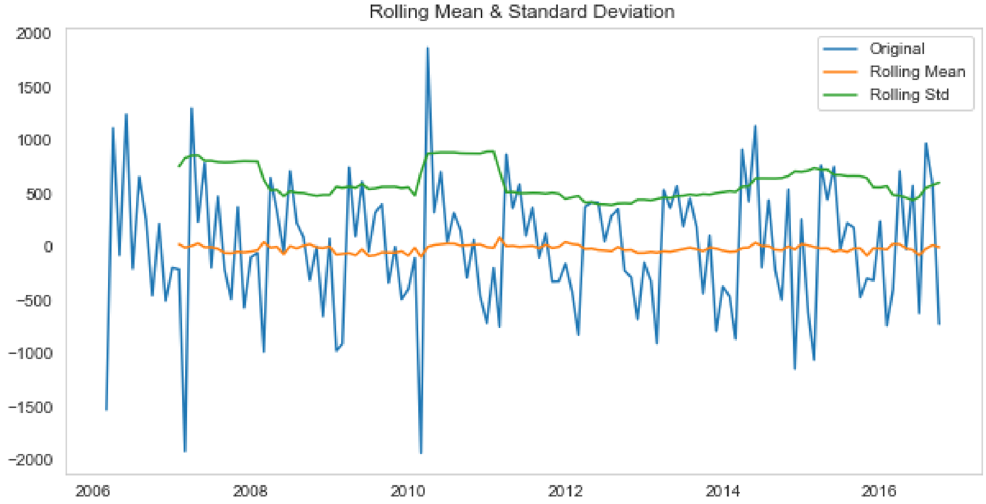
Fuller test. Figure 23. shows the first test we had which was on the original data. The original data fails the augmented Dickey-Fuller Test (T-Statistic > Critical value 1%)




```
Results of Dickey-Fuller Test:
Test Statistic          -0.938033
Critical value (1%)     -3.488535
Critical value (5%)     -2.887020
Critical value (10%)    -2.580360
```

Figure 23. The first test we had which was on the original data where it fails the augmented Dickey-Fuller Test.

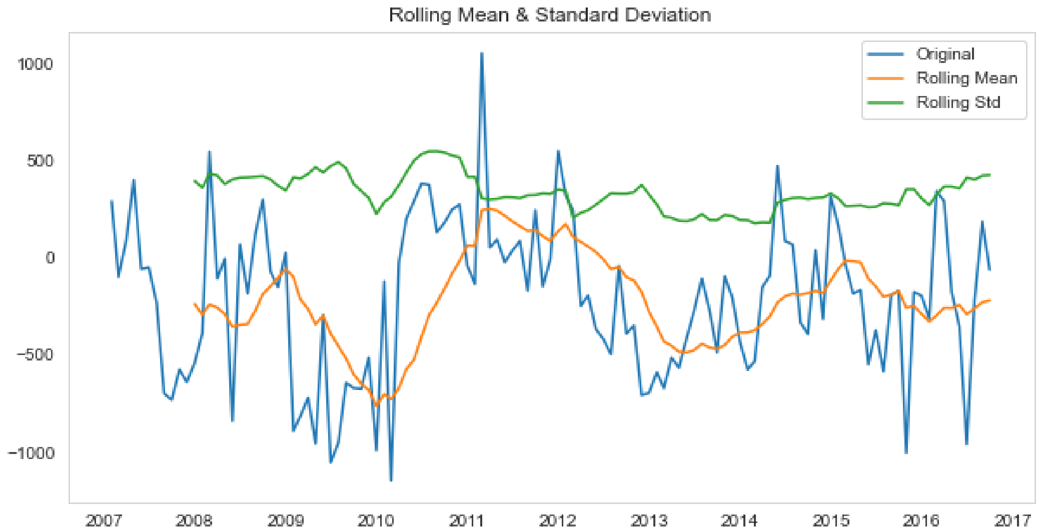
We now try the test on the monthly difference (difference in time of 1) data and see the results. Figure 24 shows the results of the augmented Dickey-Fuller Test on the monthly difference data. The data is now below 1% critical value.




```
Results of Dickey-Fuller Test:
Test Statistic          -3.890816
Critical value (1%)     -3.488535
Critical value (5%)     -2.887020
Critical value (10%)    -2.580360
```

Figure 24. The results of the augmented Dickey-Fuller Test on the monthly difference data.

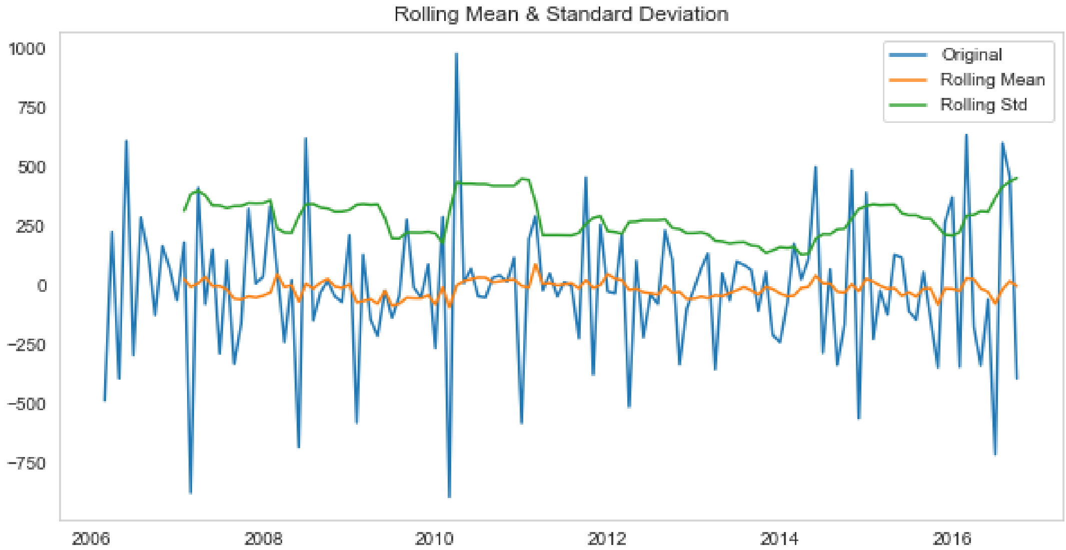
We now try to use the seasonality data by applying difference in time of 12 and see the results. The results show a good result, but the monthly difference gave a better result. Figure 25 shows the results.




```
Results of Dickey-Fuller Test:
Test Statistic          -3.569863
Critical value (1%)     -3.494220
Critical value (5%)     -2.889485
Critical value (10%)    -2.581676
```

Figure 25. The results of using the seasonality data by applying difference in time of 12.

Finally, we tried seasonality from our seasonal decomposition and use a difference in time of 1 to see the results. Figure 26 shows the result which is much better than the previous results with the T-statistic is much lower than the critical 1% value.




```
Results of Dickey-Fuller Test:
Test Statistic          -7.589421
Critical value (1%)     -3.484667
Critical value (5%)     -2.885340
Critical value (10%)    -2.579463
```

Figure 26.The results of using seasonality from our seasonal decomposition and use a difference in time of 1.

ARIMA (Autoregressive Integrated Moving Average) is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. In other words, ARIMA is a method for forecasting or predicting future outcomes based on a historical time series. It is based on the statistical concept of serial correlation, where past data points influence future data points. A statistical model is autoregressive if it predicts future values based on past values. As occurs in our case, an ARIMA model might seek to predict crimes places in Philadelphia and times with its type based on its past performance or crimes that happens based on past periods.

An autoregressive integrated moving average model is a form of regression analysis that gauges the strength of one dependent variable relative to other changing variables. The model's goal is to predict future securities by examining the differences between values in the series instead of through actual values.

An “ARIMA” model could be described by describing its components which are AR “Autoregression”, I “Integrated”, and MA “Moving average”. Let’s dig deeper into the components of “ARIMA”:

- AR (Autoregression): refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
- I (Integrated): represents the differencing of raw observations to allow for the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).
- Moving average (MA): Incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Let’ determine parameters of each component in “ARIMA” model:

- P: The number of lag observations in the model (AR terms), also it is known as the lag order.
- d: The number of times that the raw observations are differenced, also known as the degree of differencing.
- q: The size of the moving average window (MA terms), also known as the order of the moving average.

Now after understanding “ARIMA” model well and its components, it is very important to determine whether our incidents crimes time series is stationary or not before feeding it to the model in order to create the model with the best parameters. The process of determining whether data is stationary or not will be achieved through using PACF and ACF.

ACF is an auto-correlation function which gives us values of autocorrelation of any series with its lagged values. In simple terms, it describes how well the present value of the series is related with its past values. A time series can have components like trend, seasonality, cyclic and residual. ACF considers all these components while finding correlations hence it’s a ‘complete auto-correlation plot’.

PACF is a partial auto-correlation function. Basically, instead of finding correlations of present with lags like ACF, it finds correlation of the residuals (which remains after removing the effects which are already explained by the earlier lag(s)) with the next lag value hence ‘partial’ and not ‘complete’ as we remove already found variations before we find the next correlation. So, if there is any hidden information in the residual which can be modeled by the next lag, we might get a good correlation and we will keep that next lag as a feature while modeling. Remember while modeling we don’t want to keep too many features which are correlated as that can create multicollinearity issues.

In our “Philadelphia crime incidents” experiment after experimenting the stationarity of the data which results’ are shown in the figures [23-26]. We noticed that the seasonal decomposition and use a difference in time of 1 gave the best stationary score which will lead to the best ARIMA model parameters. We plotted the ACF and PACF for the data and the results are shown in figures 27, 28. And as noticed the data looks stationary which means it is ready to be fed to the ARIMA model.

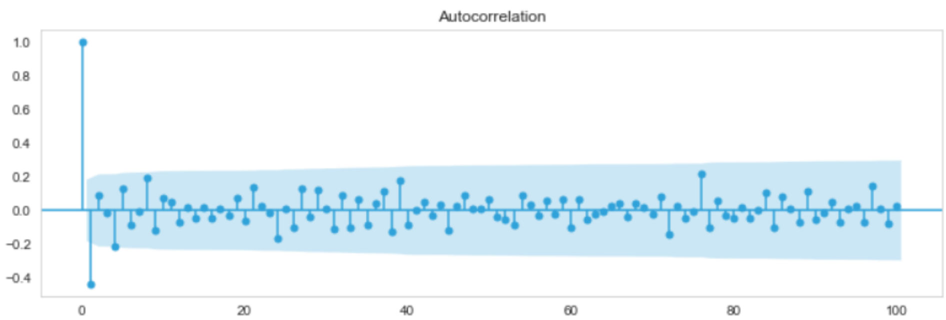



Figure 27. Autocorrelation (ACF).

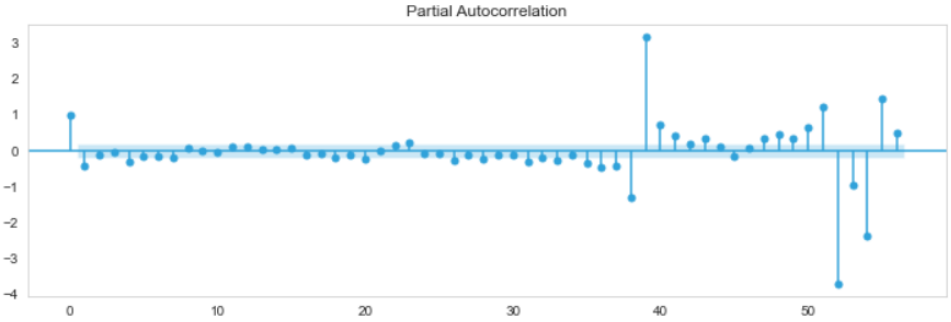



Figure 28. Partial Autocorrelation (PACF).

### 3.2 ARIMA Forecasting Model (Autoregressive Integrated Moving Average)

After the model training, its performance will be evaluated using the MAPE (Mean Absolute Percentage Error) evaluation metric which is a measure of prediction accuracy of a forecasting method in statistics. Three experiments were conducted using the MA, AR, and combined AR+MA models respectively. The first model (MA) gave a MAPE score of 7.9% (shown in figure 29), and both the (AR) and the (AR+MA) models gave a MAPE score of 9.6% (shown in figure 30, 31) which makes the MA model is the best model among the 3.

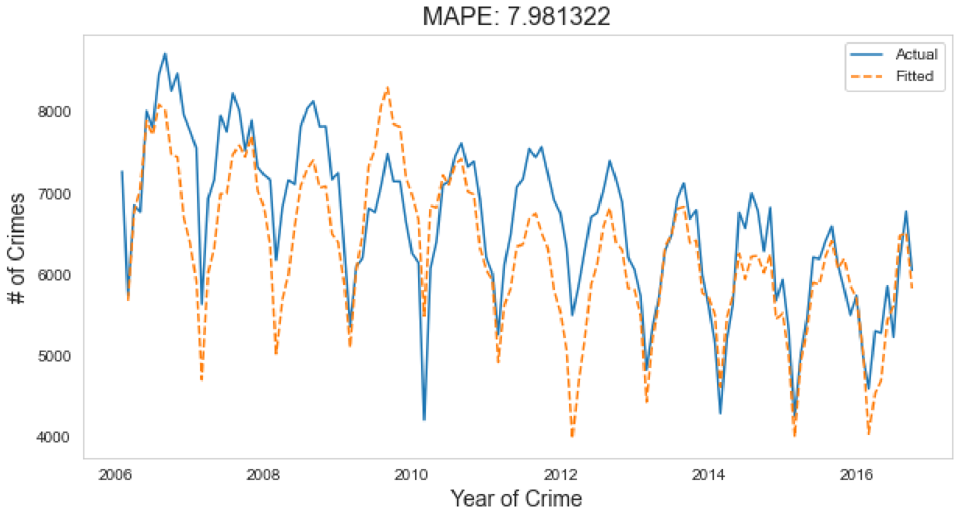



Figure 29. MA model results.

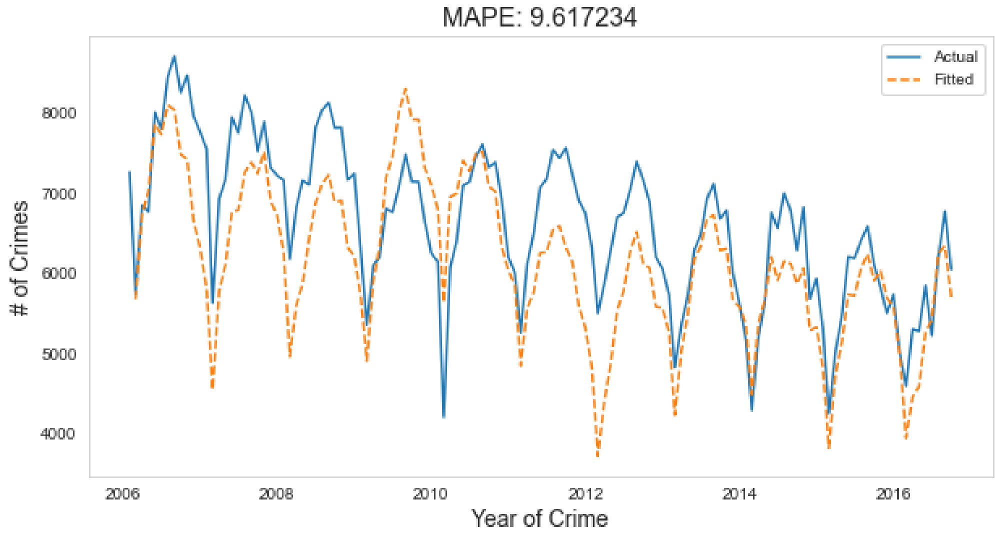



Figure 30. AR model results.

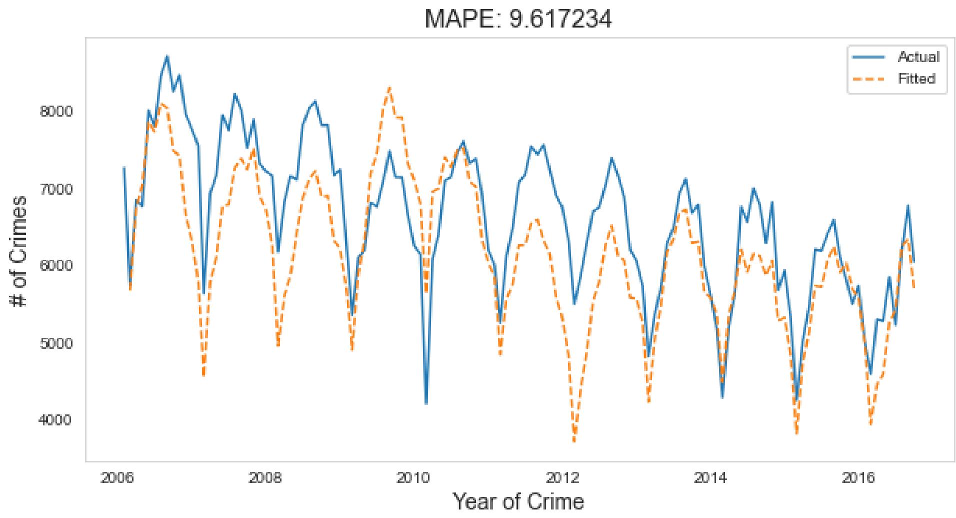



Figure 31. AR+MA model results.

# 4. Violent / non-violent Crimes Classification

## 4.1 Preprocessing

To start the preprocessing for the classification models, all the features that were in datetime format were dropped as they are not needed in classification. And as the data is huge (2.2 million), the classification will take so much time due to computational limits; so half of the data has been taken. The second step in the preprocessing was to encode all categorical data using label encoding method from sickit learn. As there were no feature with distinguishable classes, we used the UCR_General as out target value by splitting the range of values in 2 classes using a threshold of 800 as any value above it is class 0 (non-violent) and any value below it is class 1 (violent). Then we made sure that out target value is integer then we performed the train-test split with ratio of 70% for training and 30% for testing. And the final step was to standardize the data because it contained a variance values so the models would get distracted and not learn well.

### 4.2 Machine Learning Models Training & Evaluation

To start with, a helper function was created to help us in the modeling which contains the fitting procedure and after the model fits it calculates all evaluation metrics needed which are the accuracy, validation accuracy, precision, recall, F1-score, confusion matrix, and balanced accuracy score. The models that were chosen for this experiment are K-Nearest Neighbor, Support Vector Machine, and Random Forest. The first experiment was done using the KNN model and the model performed extraordinary and achieved a 100% in all evaluation metrics both training and validation which makes the data so perfect and easy to classify due to the huge number of records and the very highly correlated feature with the target value which is the Text_General_Code. All other experiments which include the other 2 models gave the same result which is 100% in all evaluation metrics. Also, to back-up our words, we used the feature importance technique from Random Forest model which shows us what is the feature that has the highest influence on the target value and the output of this function was the same as out assumption which is the Text_General_Code feature has an importance of 0.8 (Figure 32) which is very high and that is why the results of the models are 100% because the models can easy classify using this very important very highly correlated feature. We plotted the learning curves in order to see the behavior of the models during the training phase. The comparison of learning curves of the models is shown in Figure 33.

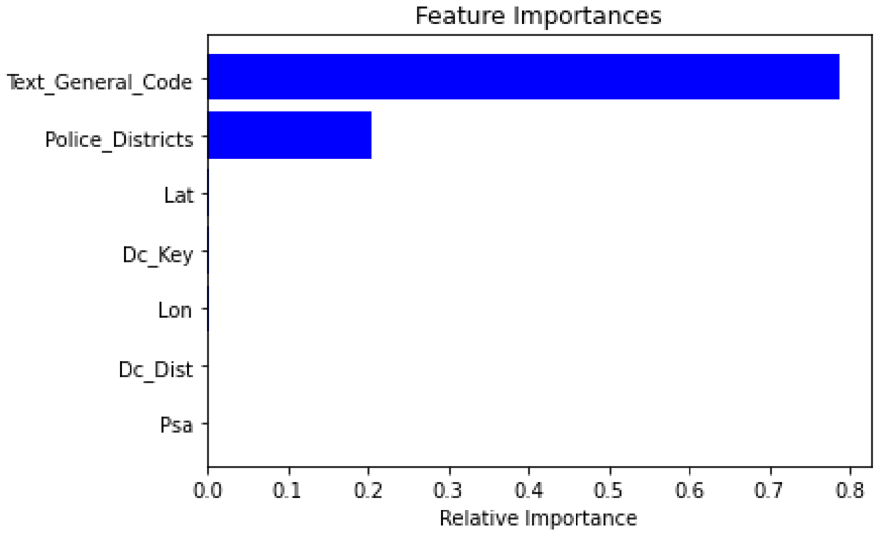



Figure 32. Feature importance from Random Forest.

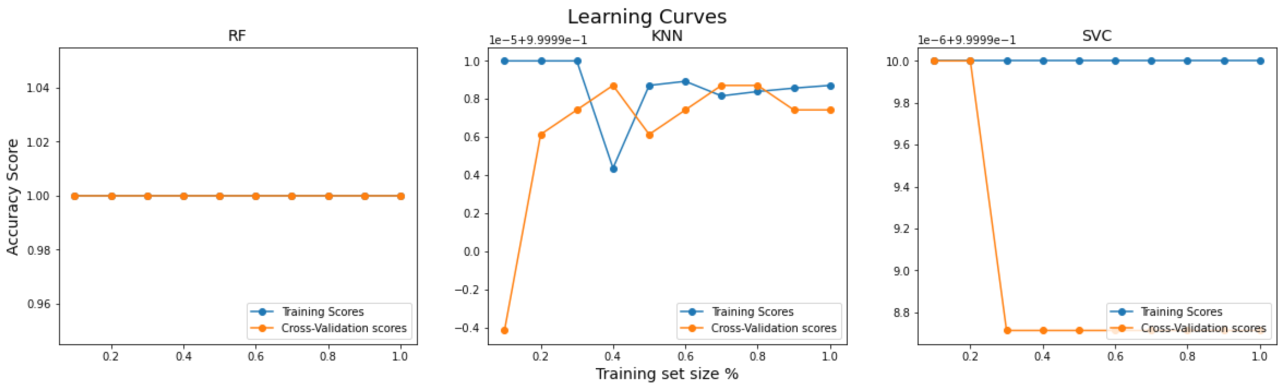



Figure 33. Learning Curves Comparison between the applied models.

### 4.3 Neural Network Implementation

To see whither the results that we obtained from the machine learning models were true or not, we implemented a Multi-Layer Perceptron (MLP) on the dataset to compare its performance to the machine learning models. The architecture of the network was an input dense layer with 1000 neurons and ReLU activation, a hidden layer with 500 neurons and ReLU activation, and an output layer with sigmoid activation. The optimizer used was Adam optimizer and the loss function was binary cross entropy function because this is a binary classification problem. We fit the data to the model with batch size of 500 and validation split of 20% for 4 epochs. After the training was done, the MLP gave the same results as the machine learning models which is 100% accuracy in both training the validation with loss of 0. Figure 34 shows the results of the MLP.

```
Neural Network Accuracy is 1.0000
Neural Network Loss is 0.0000
True Postive  : 161029  False Postive  : 0
False Negative  : 0     True Negative  : 171942
```

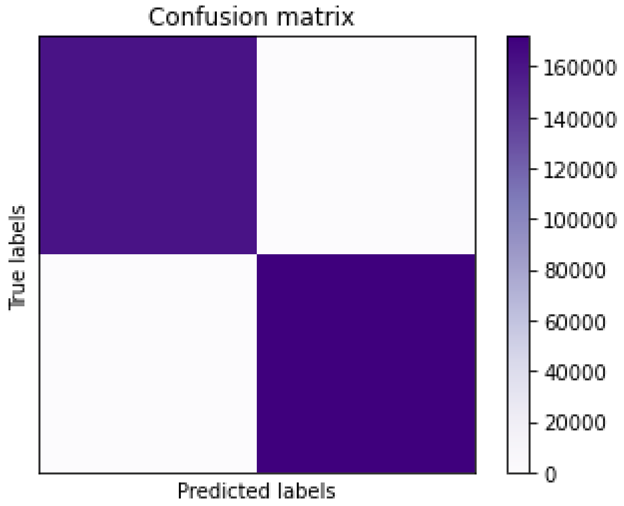



Figure 34. Results of the MLP.