Boston Crime Analysis

AUTHORS: - Akshay Krishnan, Nitheesh Koushik Gattu, Parv Thakkar, Smit Dar

1) SUMMARY

The data set chosen consists of all crimes reported by the Boston Police Department in Boston from the year 2015 to date. The dataset contains information on the type of crime committed, the time when the crime was committed, location of the crime, whether a shooting was involved and so on. This project predominantly has two aims, one is to find out how the factors such as area and Time affect the crime rate in the city, and the other is to predict the count of crimes that could happen given all the factors. Based on the dataset, trends and patterns related to the crimes being committed in Boston have been explored in this report. This report includes the methods used to pre-process the data, exploratory data analysis and multiple machine learning approaches taken to predict the monthly number of crimes in certain areas of Boston.

The pre-processing of the data involved removing null values, imputing missing data, data cleaning and variable creation. Next, exploratory data analysis was carried out to find important trends and patterns in the data. Finally, time-series models were built to predict the number of crimes that could occur per month in the entire city of Boston as well as in individual Boston neighbourhoods.

This will help residents travel safely and will also aid the law enforcement allocate resources in an efficient manner.

2) METHODS

2.1) Data Pre-processing

There are a total of 8 datasets that contain crime related information reported by the Boston Police Department from 2015-2022. The **first step** of pre-processing would be to create a dataset that is a combination of all these 8 datasets. After combining all the datasets, the resultant dataset has **597,702** rows and **17** columns.

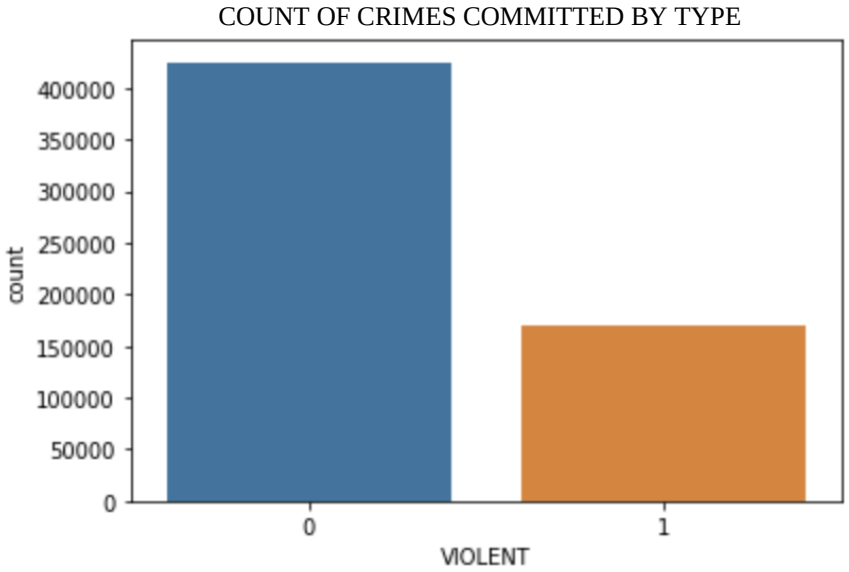
The **second step** of pre-processing would be to find the amount of missing data and perform some sort of imputations. **10** out of the **17** columns in the dataset do not have any NULLS in them. The `*LOCATION*` column has close to **23,000** rows containing invalid longitude and latitude data, hence they were replaced with NULL values. After this replacement, the `LOCATION` variable has **15%** of NULL values. Instead of removing these rows, we were able to deduce an approximate location based on the `DISTRICT` and `REPORTING AREA`. **59%** of the rows in the `*SHOOTING*` column and close to **41%** of the rows in `*OFFENSE_CODE_GROUP*` and `*UCR_PART*` contain NULL values. The remaining 3 columns have less than **4%** of NULL values in them. The `*SHOOTING*` column was converted to a flag variable with two values - 1(Shooting took place) or 0(No shooting involved). The NULL values in this column were converted to 0 as we could assume that no shooting was reported. `*UCR_PART*` variable was removed as it does not play a key role in our analysis and most of the same information could be obtained through the `*OFFENSE_CODE_GROUP*` variable. NULLS in the `OFFENSE_CODE_GROUP` were imputed based on the variables `*OFFENSE_CODE*` and `*OFFENSE_DESCRIPTION*`.

The **final step** of pre-processing was to create further variables required for EDA and modelling. Our project depended heavily on the location information (street name, district name, area name etc.). Thus, We needed to obtain the addresses and not just have the latitude and longitude data. This location data was gathered using Python's **Geopy** package with which we obtained the accurate location information for each crime. For example, if a row with just latitude and longitude information such as (42.34003269410177, -71.08923607501121) is fed into our function the output would be generated as follows - *{'amenity': 'Northeastern University', 'house_number': '360', 'road': 'Huntington Avenue', 'suburb': 'Fenway-Kenmore', 'city': 'Boston', 'county': 'Suffolk County', 'state': 'Massachusetts', 'postcode': '02115', 'country': 'United States', 'country_code': 'us'}*. In addition to this we created a flag for Violent and Non-Violent Offences based on the type of offence.

2.2) Exploratory Data Analysis

We performed exploratory data analysis on the data and plotted heat maps representing the distribution of crimes across various areas in Boston. Heat maps showing the distribution of violent and non-violent crimes committed in 2021 across Boston were also plotted.

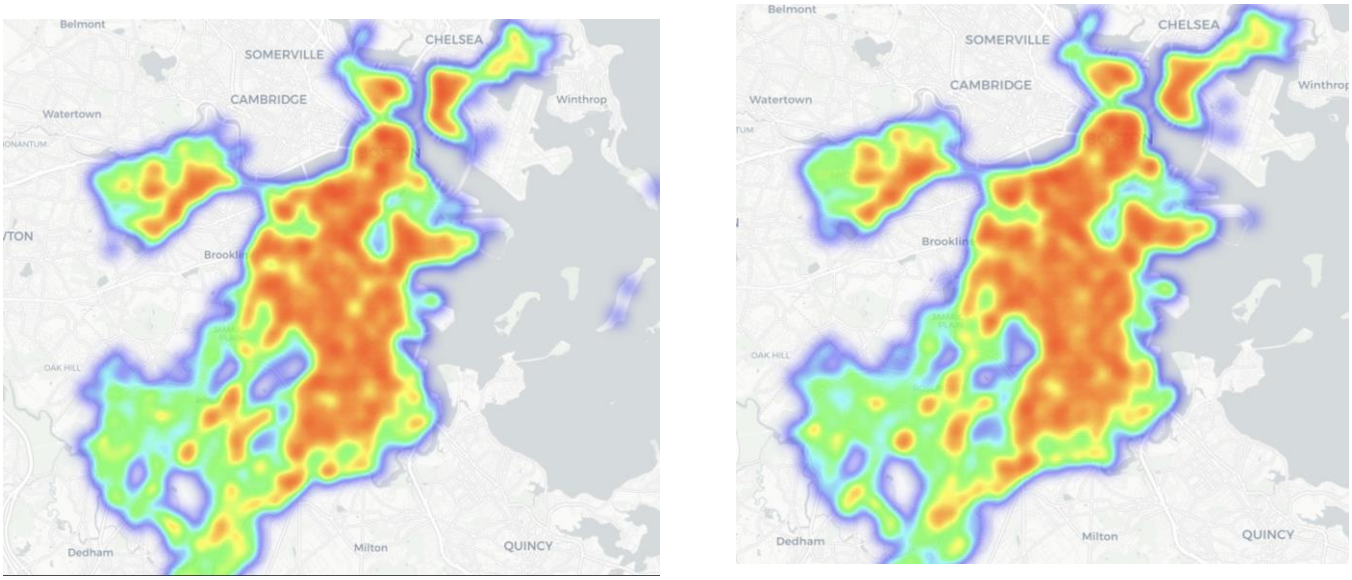
1. There were about 170,000 Violent crimes committed between the years 2015-2021 while the number of Non-Violent crimes committed were 420,000 over the same span of time.



[Fig 2.2.1: The figure shows the count of violent and nonviolent crimes for the year 2015 – 2021]

2. Furthermore, we have analysed the violent and non-violent crimes committed in different areas using the **Folium** library. Fig 2.2.2 represents the crime distribution in Boston for the years 2020(left) and 2021(right).

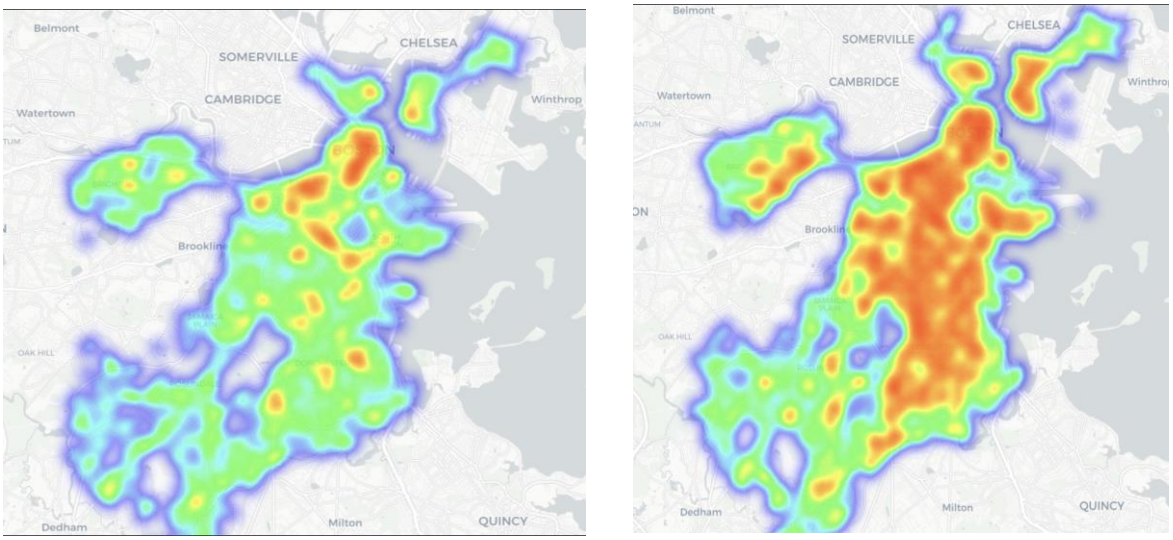
CRIME HEAT MAP IN 2020 VS CRIME HEAT MAP IN 2021



[Fig 2.2.2: The heat map on the left-hand side shows the crime distribution in 2020 and the one on the right-hand side shows the distribution in 2021. Crime distribution seems to be similar for both the years]

3. Fig 2.2.3 represents the Violent(left) vs Non-Violent(right) crime distribution for 2021. Most of the violent crimes take place around Downtown Boston whereas the non-violent crimes are uniformly distributed over the city of Boston except for a small part of southern Boston

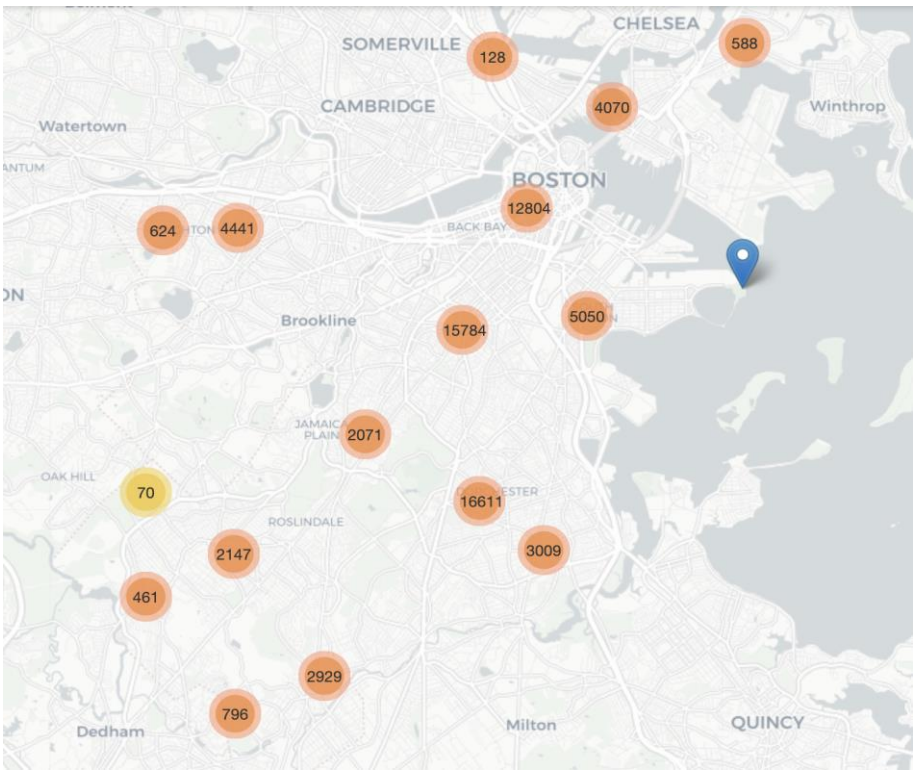
VIOLENT CRIME HEAT MAP VS NON-VIOLENT CRIME HEAT MAP



[Fig 2.2.3: The plot on the left shows the distribution of Violent crimes in 2021 while the one on the right shows the distribution of Non-Violent crimes in 2021. Violent crimes are concentrated more towards Downtown Boston]

4. Fig 2.2.4 represents the number of crimes taking place in various areas of Boston. The cluster map denotes the areas with low number of crimes in yellow and the areas with higher number of crimes in red.

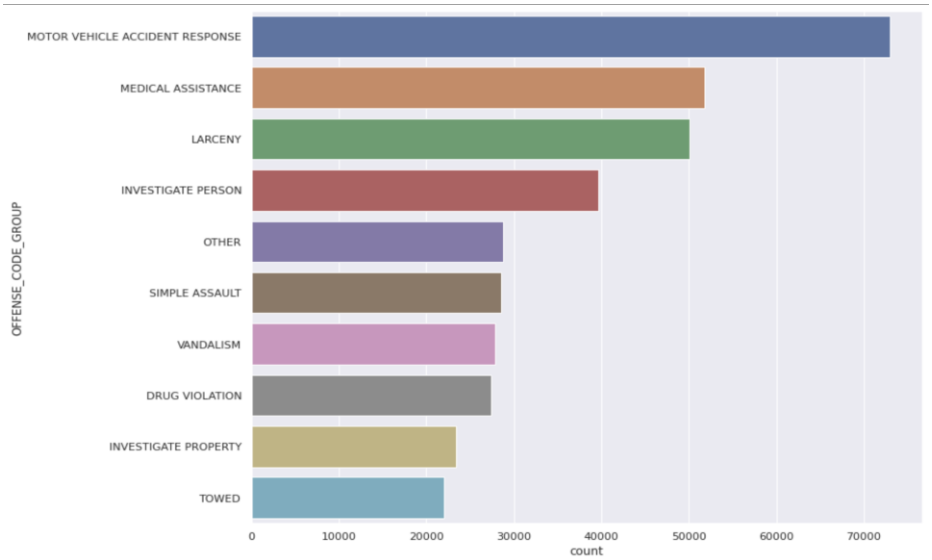
REGIONWISE COUNT OF CRIMES



[Fig 2.2.4: This cluster map plot shows us the count of crimes region wise. This shows that the crimes are highly concentrated in the Downtown Boston and the Dorchester area]

5. Fig 2.2.4 represents the most occurring type of crimes and their count between the years 2015-2021. The most crimes take place as a response to motor vehicle accidents which is approximately 75000, followed by medical assistance on 52000 and then followed by larceny at 50000. The other top crimes include investigating a person or property and assaulting or vandalising.

COUNT OF DIFFERENT TYPES OF CRIMES

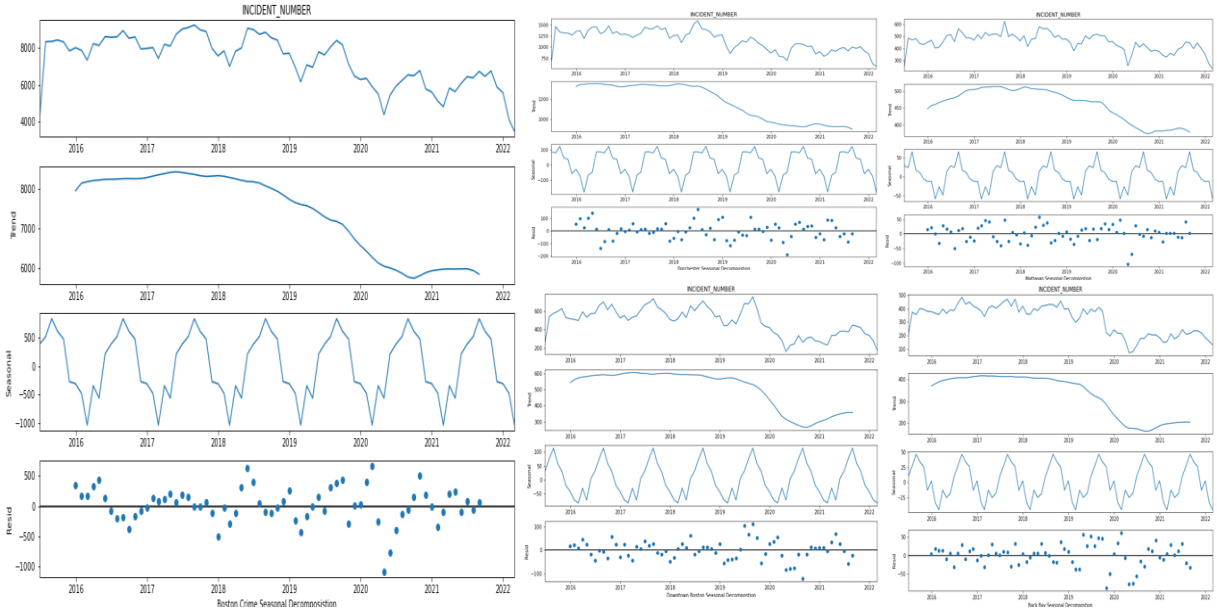


[Fig 2.2.5: The above graph shows the top 10 types of crimes and their counts from 2015 to 2021. We can see that most violations are related to Motor Vehicles]

2.3) Modelling

In this project we have decided to use two models - ARIMA (Auto-Regressive Integrated Moving Average) and SARIMA (Seasonal Auto-Regressive Integrated Moving Average) to predict the crimes by month. There is a clear seasonal trend while plotting the monthly number of crimes from 2015 onwards. There is a peak in the mid of the year and a dip in the trend chart in 2020 would be due to COVID 19. This pattern is not only observed in Boston, but also across most of the suburbs. The presence of these three components (Seasonality, Trend, and changing variance) lead to a non-stationarity of Time-Series.

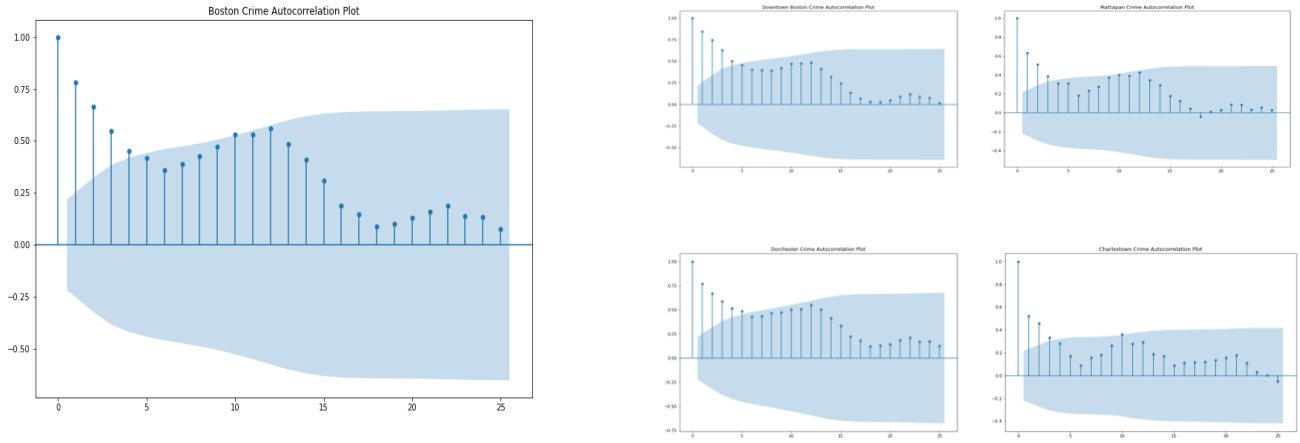
SEASONAL DECOMPOSITION



[Figure 2.3.1: Seasonal Decomposition of Boston and its multiple suburbs (Mattapan, Back Bay, Downtown Boston and Dorchester clock-wise from top left) monthly crime-count]

To be thorough, we also plotted Autocorrelation (a plot of slopes when the data is plotted against itself with respect to different lags, for example slope is 1 when the data is plotted against itself with 0 lags) plots for Boston in general and all the suburbs. We see an increase in the slope for every 12 months.

AUTOCORRELATION PLOT

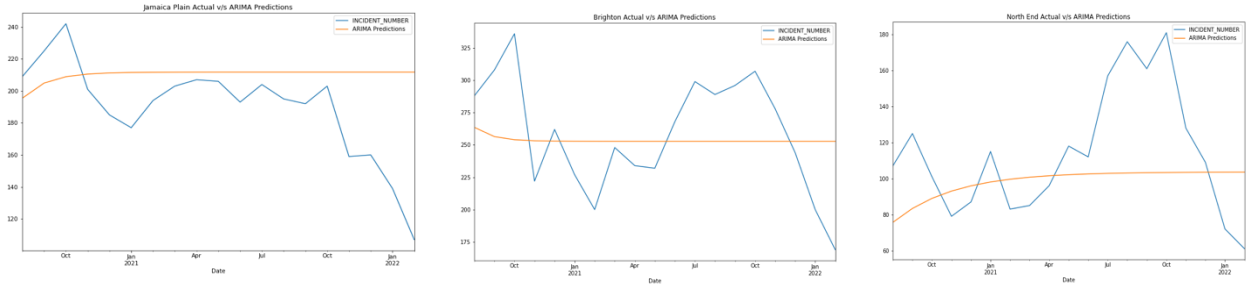


[Figure 2.3.2: Autocorrelation Plots of Boston and its multiple suburbs (Mattapan, Charlestown, Dorchester and Downtown Boston clock-wise from top left) monthly crime-count]

To avoid any mistakes, we need a mathematical affirmation that the data is non-stationary. For this, a test called Dicky-Fuller Test is performed. The p-value obtained from this test is used to figure out whether the data is non-Stationary or not. For the suburbs that are stationary, ARIMA model is used to fit and for the data that isn't a SARIMA model is used. For comparison, the non-stationary data is made stationary and ARIMA is applied.

ARIMA Model

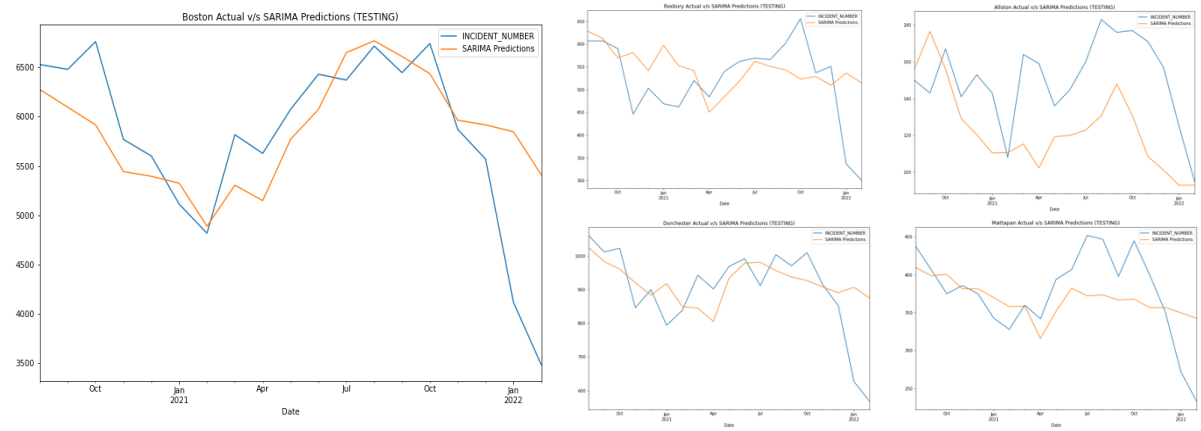
For each suburb that is stationary we fit an ARIMA model according to suggested p, q, and d values by using a step wise fit function by providing maximum p, q, and d values and the model with lowest AIC is chosen. The average Mean Absolute Percentage Error of 21.23% with a minimum of 14%, which is not that good.



[Figure 2.3.3: ARIMA Model on test set of suburbs Jamaica Plain, Brighton, North End with 17.8%, 14.6% and 25.0% MAPE]

SARIMA Model

When SARIMA models are fit to the non-Stationary data, the average Mean Absolute Percentage Error is 15.6% with a minimum of 9.6% error rate at one of the suburbs. When the SARIMA model is applied to the Boston crime count in general, we achieved a 9.604% MAPE, which is good.

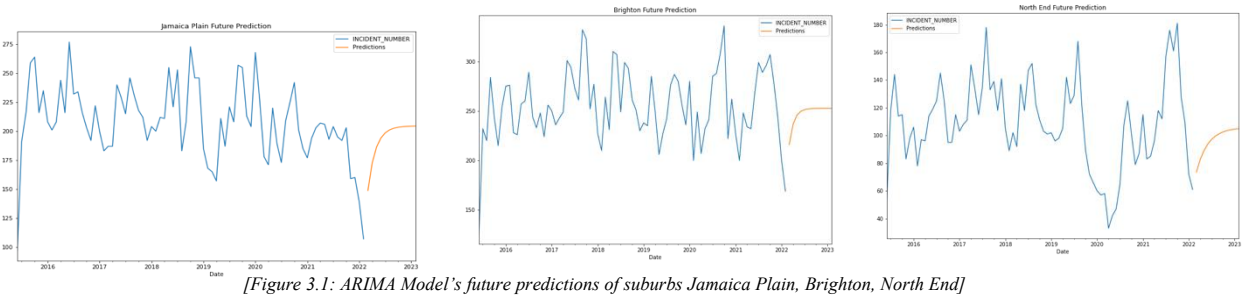


[Figure 2.3.4: SARIMA Model on test set of Boston and its suburbs Roxbury Allston, Mattapan and Dorchester (clockwise from Rightmost) with 9.604%, 15.5% 19.85%, 10.92% and 10.21% MAPE]

3) RESULT

3.1 ARIMA Results

The future predictions of the ARIMA model are calculated and plotted with the actual data.



[Figure 3.1: ARIMA Model's future predictions of suburbs Jamaica Plain, Brighton, North End]

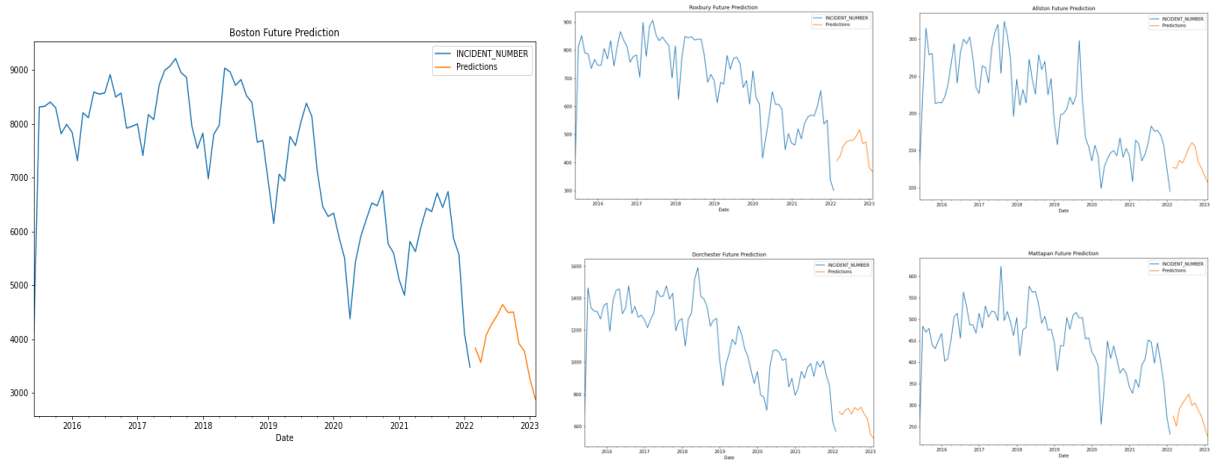
The average MAPE of the ARIMA models in regions with stationary data is 21.3%.We can clearly see that the ARIMA model "gives up" after a certain period so it it's safe to say that we can only predict for short intervals with this model.

	Suburb	ARIMA Orders	MAPE
0	East Boston	(1, 1, 0)	27.499171
1	Jamaica Plain	(1, 0, 0)	17.815626
2	West End	(1, 1, 0)	20.548798
3	Brighton	(1, 0, 0)	14.656604
4	West Roxbury	(1, 1, 1)	21.866094
5	North End	(1, 0, 0)	25.051668

[Figure 3.2: ARIMA Model's MAPEs is Stationary Suburbs]

3.2 SARIMA Results

The future predictions of the SARIMA model are calculated and plotted with the actual data.



[Figure 3.3: SARIMA Model's future predictions of Boston and its suburbs Roxbury Allston, Mattapan and Dorchester (clockwise from Rightmost)]

The average MAPE of the SARIMA model is 15.6%. It's quite clear that the SARIMA model is performing better than ARIMA. If we applied a ARIMA after making the data stationary then we obtain a 25.06% MAPE which is not as good as the SARIMA model.

	Suburb	ARIMA Orders	Seasonal Orders	MAPE
0	Dorchester	(0, 1, 1)	(1, 0, 0, 12)	10.210630
1	Roxbury	(1, 1, 1)	(1, 0, 0, 12)	15.508902
2	Mattapan	(0, 1, 1)	(1, 0, 1, 12)	10.921512
5	Allston	(0, 1, 2)	(1, 0, 1, 12)	19.851660
6	South End	(3, 1, 2)	(1, 0, 2, 12)	21.261254
7	South Boston	(1, 1, 1)	(0, 0, 1, 12)	16.965859
8	Hyde Park	(0, 1, 1)	(0, 0, 0, 12)	17.010737
9	Mission Hill	(0, 1, 1)	(0, 0, 0, 12)	11.738196
10	Back Bay	(0, 1, 0)	(0, 0, 0, 12)	14.868947
12	Beacon Hill	(0, 1, 0)	(0, 0, 0, 12)	21.341781
13	Charlestown	(0, 1, 1)	(1, 0, 0, 12)	18.556223

[Figure 3.4: SARIMA Model's MAPEs in Non-Stationary Suburbs]

4) DISCUSSION

How are we helping people take better-informed decisions?

- ***Police*:** These predictions could help the Boston Police Department be more alert and active in months of predicted high crimes
- ***Residents*:** Residents can also use it as a tool to keep themselves safe during predicted high crime months and take necessary measures to avoid conflict in areas with predicted high crimes
- ***Housing*:** Residents, especially senior residents or residents with infants, can also benefit from having prior knowledge before investing in housing in a predicted high crime area and search for housing in a relatively safer zone.
- ***Students*:** Boston receives a large influx of international and out of state students. The analysis and predictions could help the non-resident students navigate safer paths and avoid any risky routes.

Future Work

- In the future we can look to improve the project by making it more accurate and time specific. For a specific place, at any particular day and time a person can assess how safe it is to travel to that location. This ideology would be similar to that of a weather forecast.
- The performance of the SARIMA model can be improved by adding further exogenous variables.
- We can also add a feature wherein the people are notified of a crime that has taken place in their vicinity and the severity of it as well.

5) STATEMENT OF CONTRIBUTIONS

NAME	CONTRIBUTION
AKSHAY KRISHNAN	Data pre-processing, report and presentation generation
NITHEESH KOUSHIK GATTU	Data modelling, report and presentation generation
PARV THAKKER	Eda, insights & visualizations, report and presentation generation
SMIT DAR	Eda, insights & visualizations, report and presentation generation

6) REFERENCES

(i) https://www.hindawi.com/journals/complexity/2022/4830411/
(ii) https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-newsyste
(iii) https://people.duke.edu/~rnau/411arim.htm
(iv) https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/
(v) https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775

7) APPENDIX

Codes are present in https://github.com/akrishnan96/Boston-Crime-Analysis