MS4S09 : Analysing Crime Rate in Atlanta

Francis Afuwah

Analysis Outline

Introduction

The crime rate in Atlanta has been a major source of concern for residents, visitors and potential residence. As the capital city of Georgia, like most cities, Atlanta experiences crime challenges which has been mostly blamed on the city's population density, drug activities and law enforcement. The study of the crime rate is relevant because it provides insights into the overall security of the city which would help individuals to ensure they take measures for personal safety. The analysis of crime rates in the city would also enable individuals know the likelihood of being a victim of crime and how they could reduce the chances. The crime rate in Atlanta differs by neighbourhood. Some neighbourhoods are more susceptible to crimes than the others. Also, there are different types of crimes, from pickpocketing to robbery to rape. While some crimes are violent, others are not. The level of violent crime rates also depends on the neighbourhood in Atlanta. While the law enforcement agencies are enacting measures to ensure safety of residents and non-residents, it is still important to be informed of the crime rates and the trend in crime rates over time. (Alumni Factor, 2023)

A study of the trend in crimes in Atlanta shows that the crime rate has reduced by over 20% in the past decade which shows that the enforcement agencies has been taking steps to combat the crime rates in Atlanta. However, violent crime rate have declined slowly than the overall crime rate. The city still has prevalence in violent crimes (murder, aggravated assault, rape and armed robbery). Several reports have blamed the reduction in market activities on the crime rates in the city, implying that individuals do not feel safe enough to visit. According to the report by Cauda Analytics, the crime rates in Atlanta has seasons where at certain periods, it experiences an increase or decrease. (Cauda Analytics, 2021)

Aim

The aim of this project is to analyse the trend in crime rates in Atlanta, USA and to forecast future values based on the provided data.

Research Questions

This project seeks to answer the following research questions:

a. What is the trend in crime rates in Atlanta, USA

b. What is the nature of seasonality in the crime rates in Atlanta?

c. Which neighbourhoods have the highest crime rates in total?

d. What are the most popular crimes?

e. What is the future trend in crime rate in Atlanta?

Data

The data used for this project was collected from the Wolfram Data Repository and is sourced from the Atlanta Police Department. (Zviovich, 2017) The link for the data can be gotten here:

https://datarepository.wolframcloud.com/resources/dzviovich_Atlanta-Police-Department-Crime-Data-2009-2017

The data contains details information on crime in Atlanta from January 1st, 2009 to February 28th 2017. The data was subset for the purpose of this project to contain data from January 1st, 2009 to December 31st 2016 which is a period of 8 years.

The data was cleaned by removing missing values which also included entries with a blank space. (Datar et al., 2019)

The data contained information on the crime type, index number, the assigned number of the crime, the date the crime occurred, the beat, the npu, the neighbourhood the crime happened and the longitude and latitude for the exact location of the crime. However, only columns relevant for analysis were selected and they are: index number, the type of the crime, the date the crime occurred and the neighbourhood of the crime.

The date column was transformed from a string data type to a date object. Since the original time series contained daily occurrence of crime, the data was aggregated to find the total number of crimes that occurred monthly. The analysis and forecast also followed a monthly analysis rather than a daily analysis.

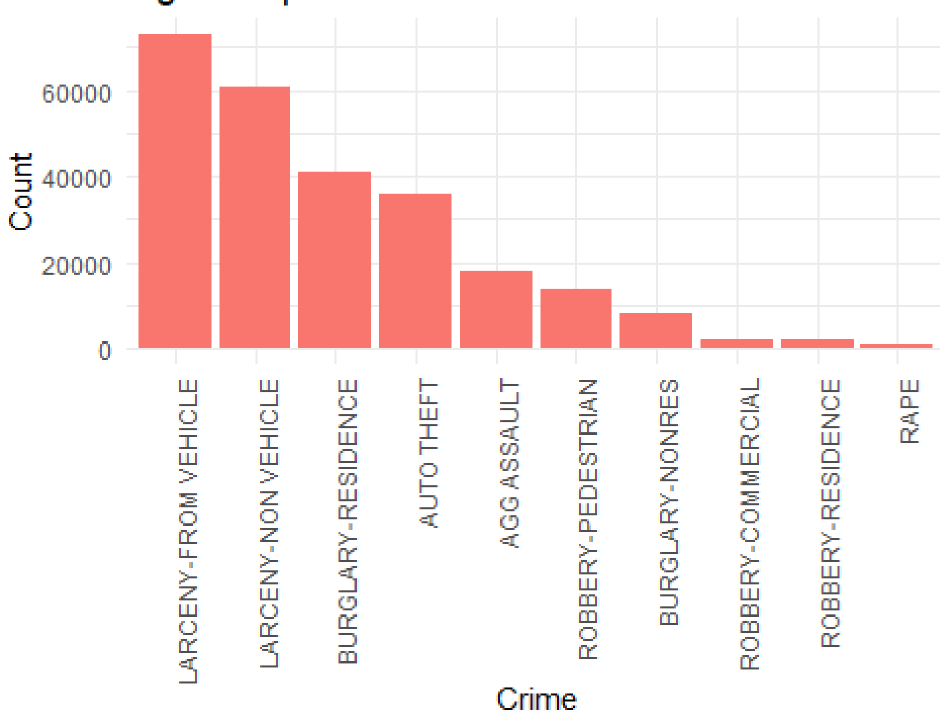
Exploratory Data Analysis

The data is explored to uncover insights about crimes in Atlanta and to discover patterns that would be useful in the model fitting and model forecasting sections of the time series analysis.

Table 1: Top 10 neighbourhoods with the highest crime rates in Atlanta

```
##        neighborhood count
## 80         Downtown 19487
## 149         Midtown 13563
## 165 Old Fourth Ward  7978
## 223        West End  6855
## 136           Lenox  4767
## 210    Sylvan Hills  4357
## 107      Greenbriar  4348
## 109      Grove Park  4206
## 219       Vine City  4173
## 159  North Buckhead  4164
```

Fig. 1: Top ten Crimes in Atlanta



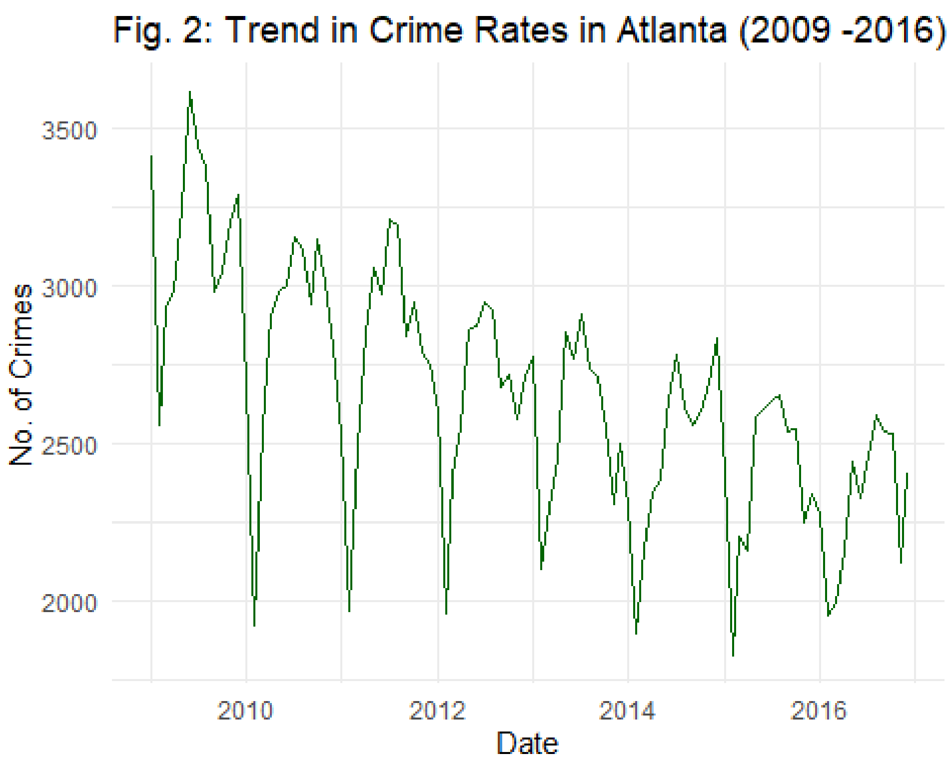


Fig. 3 : Decomposed Plot

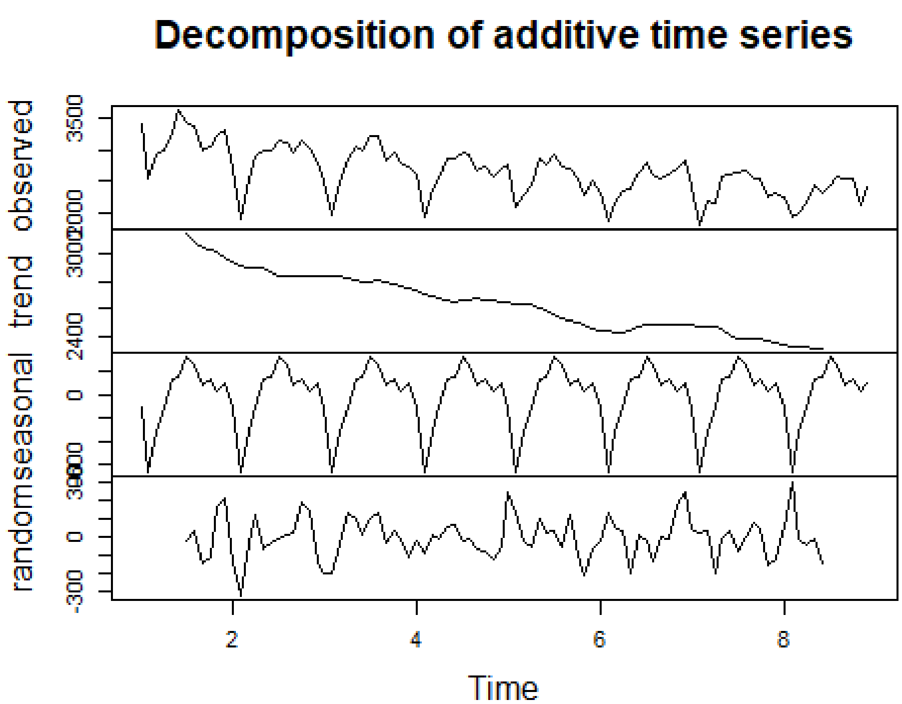


Table 1 shows the analysis of the neighbourhoods with the highest crime rates in Atlanta. The table shows that the neighbourhoods with the highest number of crimes is Downtown with over 19000 crimes recorded from 2009 to 2016, this is followed by Midtwon, Old

Fourth Ward, West End and Lenox. Between these five neighbourhoods, over 52,000 crimes were reported and recorded by the Atlanta Police Department.

The analysis also analyses the most popular crimes in Atlanta, USA and this is visualized using a bar chart which shows the top 10 crimes in the USA city. From the chart, the most popular crimes are: Larceny, Burglary, Robbery & Theft, Aggravated Assault and Rape. It can be observed that most of the crimes mentioned above are violent crimes and a significant proportion of the crimes also involve theft in various forms: armed and unarmed.

The analysis goes further to investigate the trend in crime rates in Atlanta which is visualized in the line chart in Figure 2 which shows the trend in crime rates from 2009 to 2016. From the chart, crime rates has gradually reduced over the years. The chart also showed that there is an evident seasonality context in the data. The crime rates seem to reduce towards the beginning of the year (February to April) and then starts to increase again from mid-year to the end of the year. For further analysis of the trend in the data, a decomposition plot is created (Figure 3) which shows the trend, error and seasonal components of the data. The decomposition plot confirms that there is a downward trend in the data and there is a seasonality with a period of s=12. Due to the presence of trend and seasonality, there is a high chance that the data is not stationary, hence before modelling, the data is differenced to remove the trend and seasonal effect and to allow for proper and accurate forecasting.

Model Fitting

Introduction

From the decomposition plot's seasonal component, it is evident that the data is seasonal with a period of 12. Hence, the most appropriate model to use in this case is the SARIMA model. The SARIMA model is a special type of ARIMA model which is used for seasonal time series. In addition to the non-seasonal components present in the ARIMA model, SARIMA also contains the seasonal component. The non-seasonal part consists of the following components:

1. Autoregressive (AR) component whose order is represented as p.
2. Moving Average (MA) component whose order is represented as q

3. Integrated (I) component which is the number of times the time series is differenced to achieve stationarity. It is represented as d

Since the SARIMA accommodates seasonality, it also has the seasonal part which is similar to the non-seasonal part. The seasonal part consists of the following components:

1. Seasonal Autoregressive component whose order is represented as P
2. Seasonal Moving Average component whose order is represented as Q
3. Seasonal Integrated component which is represented as D.
4. The Seasonal Period which is represented as s.

A default SARIMA model is denoted by: SARIMA (p,d,q)(P,D,Q,s)

The order of a SARIMA model is generally identified by four main parameters

a. The PACF plot is used to find the order of the Autoregressive component

b. The ACF plot is used to find the order of the Moving Average component

c. Differencing (where necessary) is used to convert a non-stationary time series to a stationary one and this is used to find the order of the Integrated component

d. The seasonal component of the decomposition plot is used to find the period of seasonality.

Like most modelling procedures, time series is also iterative. The Alkaline Information Criteria (AIC) is one of the means by which the performance of several SARIMA models can be tested to find the best of the available models which is usually the model with the lowest AIC score. (Cryer et al., 2011)

Table 2: Time Series Data from 2009 to 2016 (January to December)

```
##    Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec
## 1 3410 2558 2933 2979 3254 3617 3442 3379 2982 3046 3209 3291
## 2 2688 1922 2552 2904 2980 2999 3153 3118 2942 3148 3012 2808
## 3 2523 1970 2487 2846 3056 2974 3216 3195 2842 2948 2786 2751
## 4 2600 1959 2397 2551 2859 2874 2952 2924 2681 2721 2578 2711
## 5 2775 2105 2296 2458 2853 2768 2913 2741 2712 2543 2306 2502
## 6 2309 1895 2166 2344 2385 2640 2785 2618 2561 2606 2692 2836
## 7 2406 1826 2203 2160 2586 2608 2635 2657 2539 2548 2253 2337
## 8 2273 1955 1994 2161 2445 2325 2459 2590 2536 2533 2119 2405
```

Fig. 4: Plot of differenced time series

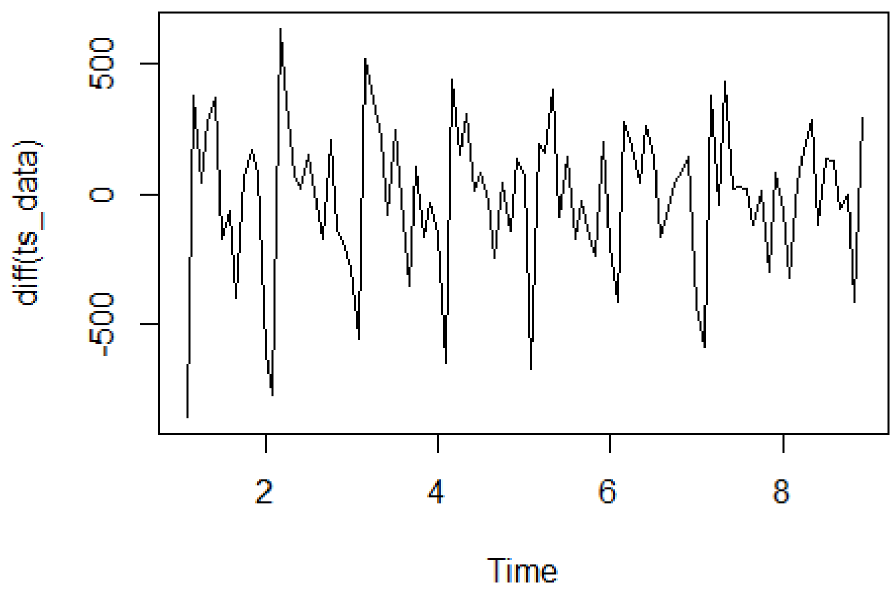
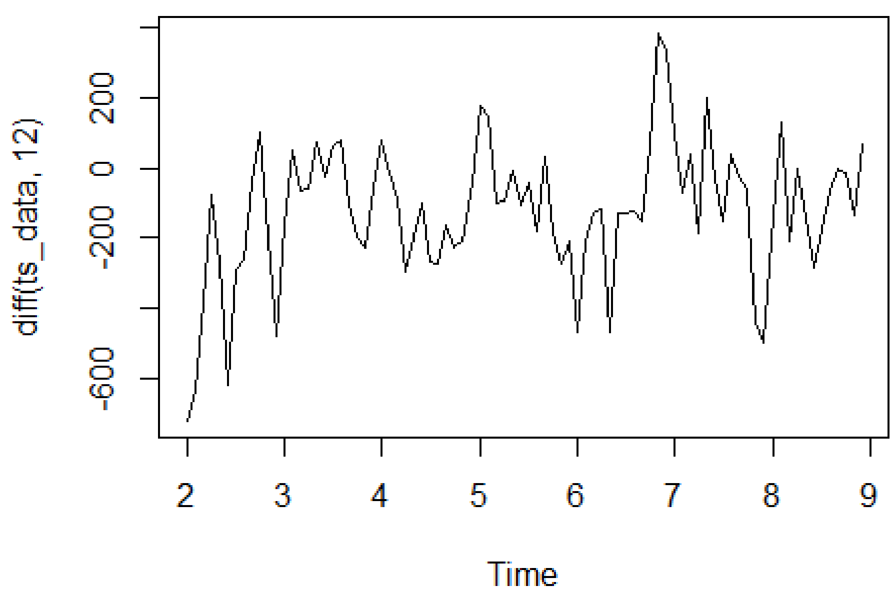


Fig.5: Plot of TS (Seasonality Removed)



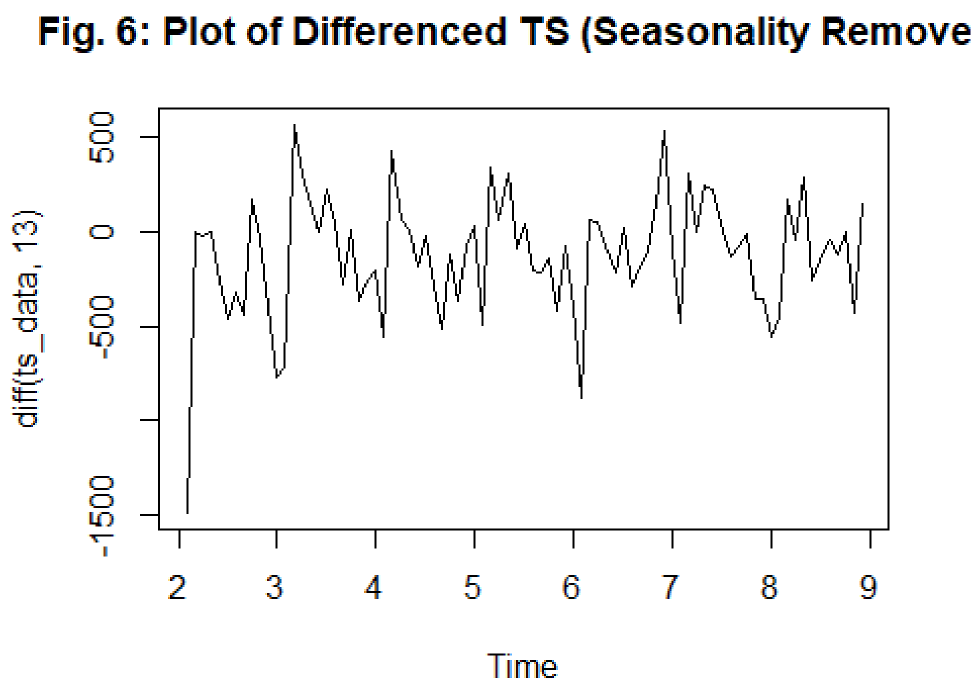


Fig. 6: Plot of Differenced TS (Seasonality Remove
diff(ts_data, 13)
500
0
-500
-1500
2
3
4
5
6
7
8
9
Time

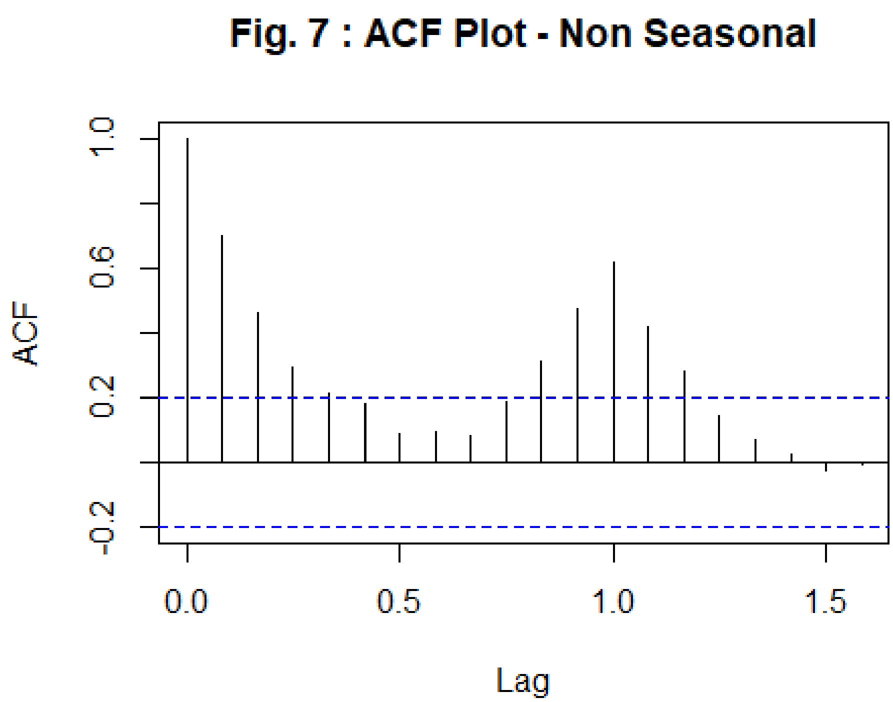


Fig. 7 : ACF Plot - Non Seasonal
ACF
1.0
0.6
0.2
-0.2
0.0
0.5
1.0
1.5
Lag

Fig. 8 : PACF Plot - Non-Seasonal

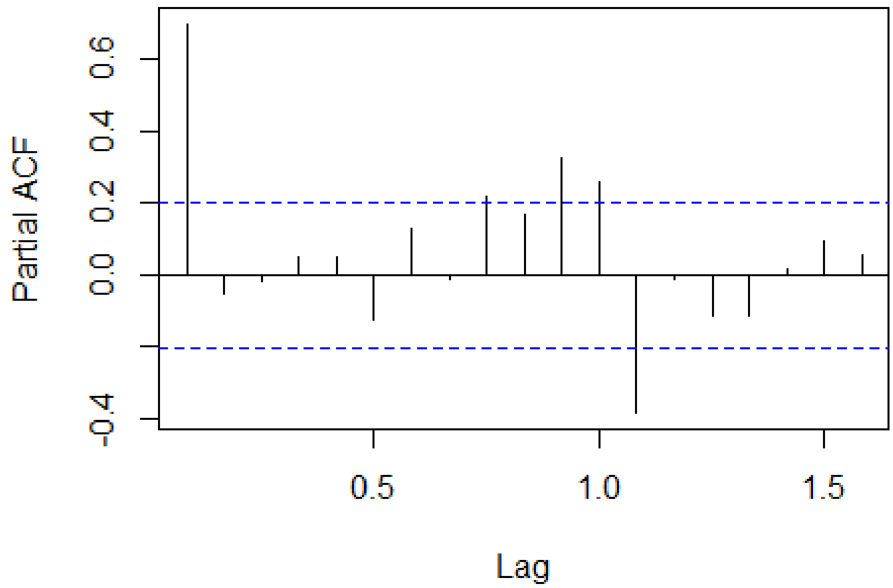
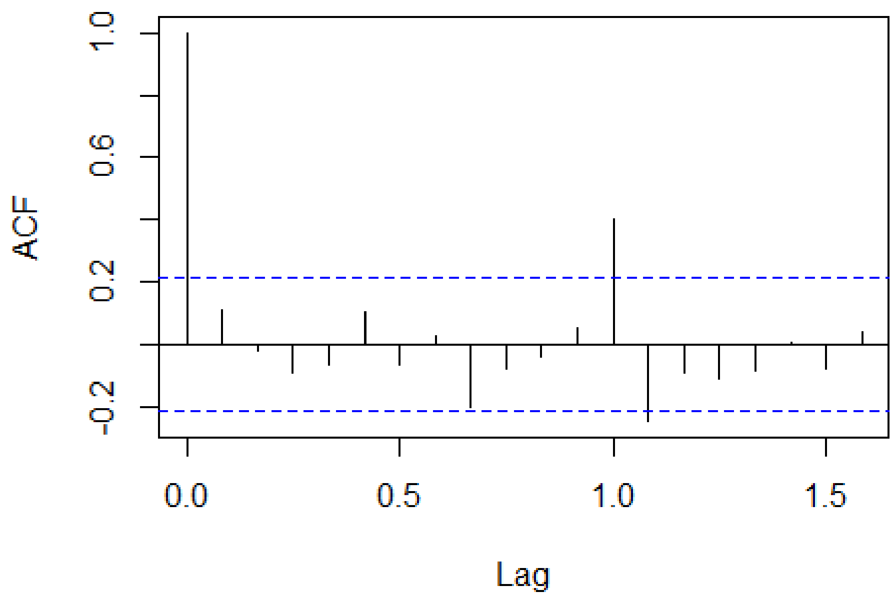


Fig.9 ACF Plot - Seasonal



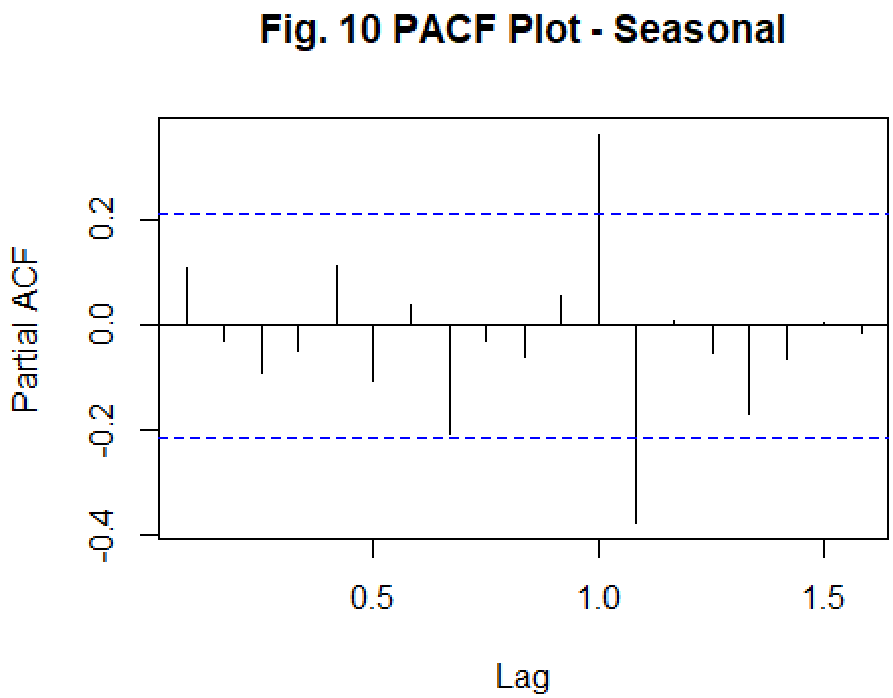


Table 3 : Comparison of Model AIC Scores

```
##     Name    Value
## 1 model1 1101.976
## 2 model2 1088.454
## 3 model3 1119.532
## 4 model4 1327.639
## 5 model5 1084.098
## 6 model6 1327.639
```

Discussion

For the non-seasonal component, both the ACF and PACF plots cut off at lag 1.0. In an attempt to remove the seasonality component, the time series was differenced 12 times. As seen in Figure 5, however, the time series still showed signs of non-stationarity, hence, it was differenced once more (differenced 13 times) to remove the stationarity (Figure 6) and the time series for the seasonal components is then plotted. Six SARIMA models with different model order are tested using the Alkaline Information Criteria (AIC) to compare their various AIC Scores (Table 3). From the result, the best models are model 2 and model 5 which have an AIC Score of 1099.454 and 1084.098 respectively.

Model Forecasting

Introduction

In order to analyse the performance of the models, the data was split into the training and test sections. The data for 2016 (January – December) was separated as the test data while the time series data from 2009 to 2015 was used to train the models. The best performing models from the model fitting section were model 2 and model 5. Model 2 had an AIC score of 1088.4 while Model 5 had an AIC score of 1084. Using the Root Mean Square Error (RMSE) as a metric, the performance of the predictions from each model was compared with the actual result in the test dataset.

```
## [1] "The RMSE of the model: ARIMA (0,1,1)(0,1,1,12) is 177.2"

## [1] "The RMSE of the model: ARIMA (1,1,1)(1,1,1,12) is 176.4"
```

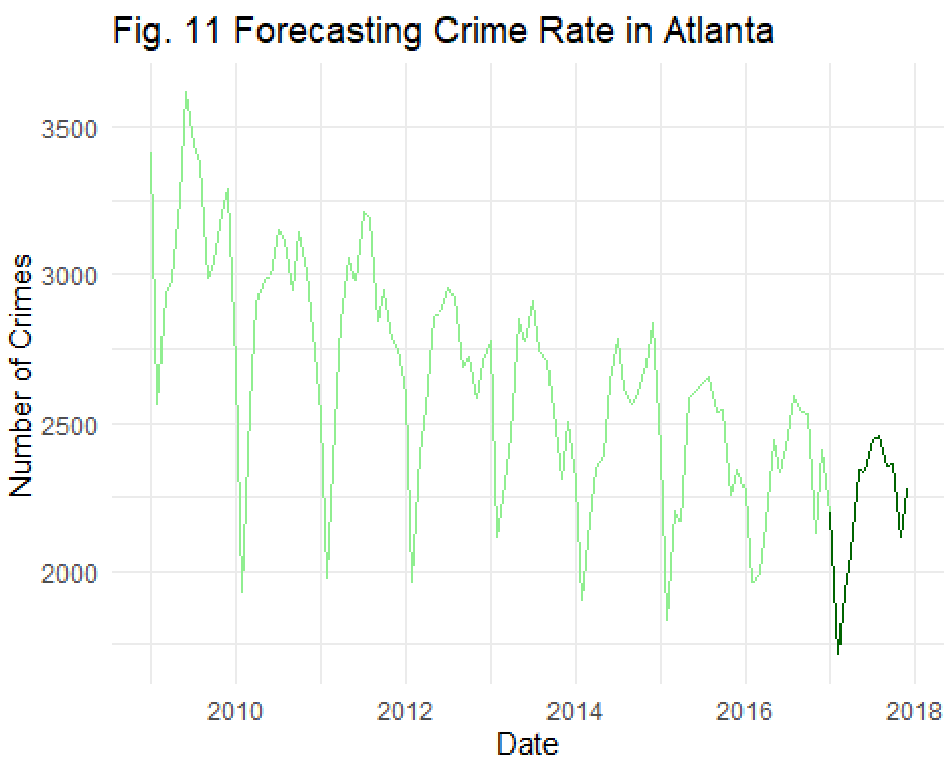


Table 4: Forecast for 2017

```
##          Date Forecast Result
## 1  2017-01-01        2195.513
## 2  2017-02-01        1711.246
## 3  2017-03-01        1930.302
## 4  2017-04-01        2059.251
## 5  2017-05-01        2342.161
## 6  2017-06-01        2332.425
## 7  2017-07-01        2439.631
## 8  2017-08-01        2454.416
## 9  2017-09-01        2354.579
## 10 2017-10-01        2360.341
## 11 2017-11-01        2104.491
## 12 2017-12-01        2280.608
```

Discussion

From the result, the RMSE of model 5 was 176.1 which is the lowest of the two models. The model was then used to forecast the crime rates in Atlanta for 2017 which can be seen in Table 4. Figure 11 shows the trend in past and forecast crime rates in Atlanta. From the chart, crime rate is expected to experience a sharp decline in February 2017 and then start increasing again in March until it reaches its peak in October 2017, and then it declines sharply again in October before gradually increasing again towards the end of the year.

Conclusion

The main goal of this project was to analyse the trend in crime rates in Atlanta from 2009 to 2016 and to forecast the crime rates for the next 12 months (January 2017 to December 2017). The data used for this analysis is originally sourced from the Atlanta Police Department and was downloaded from the Wolfram Cloud Data Repository. The data contains the various crime types, the neighbourhoods they occur and the date the crime was recorded by the Police.

The decomposition plot in Figure 3 shows the different components of trend, seasonality and error of the time series data. The plot showed the presence of a trend and seasonality effect in the data. Hence, the data is differenced before modelling and observed. Due to the presence of seasonality, the SARIMA model is the most ideal model to be used. The order of the SARIMA model was gotten by using the ACF and PACF plots. After observing the ACF and PACF plots and iterating to get the best model with the help of the AIC scores, it was discovered that the SARIMA (1,1,1)(1,1,1,12) model was the best of the tested models with an AIC Score of approximately 1084.

The data was split into training and test sets and 12.5% of the data (January 2016 to December 2016) was used to test the performance of the models selected during the model fitting phase. The SARIMA (1,1,1)(1,1,1,12) model had the best result with a root mean square error of 176.4. The data was then used to forecast the crime rates for January to December 2017. The result of the forecast was visualized in Figure 11. From the chart, it is predicted that there would be a reduction in crime rates in February 2017, however, the crime rates would start to increase again until it reaches its peak in October 2017.

Future work

In addition to the analysis done in this project, additional data can be collected and the following can be incorporated for future work:

a. A comparative analysis between violent crimes and non-violent crimes can be done to analyze the trend in violent crimes and non-violent crimes over time. The analysis could also extend to analyse the neighbourhoods with a higher case of violent crimes.

b. A comparative analysis to compare the crime rates in Atlanta to the other cities in Georgia and the United States as a whole. The comparative analysis would also compare the crime types (violent and non-violent) for these regions.

c. A correlation analysis of the relationship between socio-demographic factors and crime rates. Factors such as population, poverty level, educational level, drug activity, etc. of each neighbourhood in Atlanta could be compared to the crime rates in the neighbourhood.

d. An analysis of the periods of the day and days of the week where crime rates are more prominent in the city.

e. A sentiment analysis of the perception of residents, visitors and non-residents on crime rates in Atlanta and the efforts of the law enforcement agency to combat crime rates in the city.

References

Dataset Reference:

- Zviovich, D., (2017). Atlanta Police Department Crime Data (2009 – 2017). Available at: https://datarepository.wolframcloud.com/resources/dzviovich_Atlanta-Police-Department-Crime-Data-2009-2017

Other References:

- Cryer, J.D. and Chan, K. (2011) *Time series analysis with applications in R*. New York: *Springer*.
- Datar, R. and Garg, H. (2019) *Hands-on exploratory data analysis with R: Become an expert in exploratory data analysis using R packages*. Birmingham: Packt Publishing.
- Alumni Factor, (2023). *Atlanta Crime Rate: Surprising Statistics Revealed.* Available at: https://alumnifactor.com/2023-atlanta-crime-rate-surprising-statistics-revealed
- Cauda Analytics, (2021). *How have crime patterns in the city of Atlanta changed over the last decade?* Available at: https://www.caudaanalytics.com/atlanta-crime-analysis