# MATH 546 Project Report

Chicago Crime Analysis

Karan Bhandari: A20469335 Jasleen Bhatia: A20495939

05/05/2022

## 1. ABSTRACT

Crime is an integral aspect of our society whether as a victim or an offender, everyone has been witnessing a crime. In our study, we analyze crime data, and we chose the "Chicago Criminal dataset," which contains crime episodes in Chicago from 2001 to 23rd March 2022. We looked at crime patterns over time.

## 2. OBJECTIVE

The main objective of this project is to analyze and compare the patterns of Chicago crime based on historical patterns by using statistical methods. Our focus is to build a model that allows a benchmarking comparison and serves as a reference for future research on this subject.

## 3. RESEARCH SCOPE

To analyze Chicago crime data and the effects of external factors , we want to answer the following questions at the end of this project:

- What is security status of the city?
- Which month has the highest crime rate on average?
- Is there any influence of the external factors on crime?
- Is a predictive model useful in anticipating crime?

## 4. EXPLORATORY DATA ANALYSIS

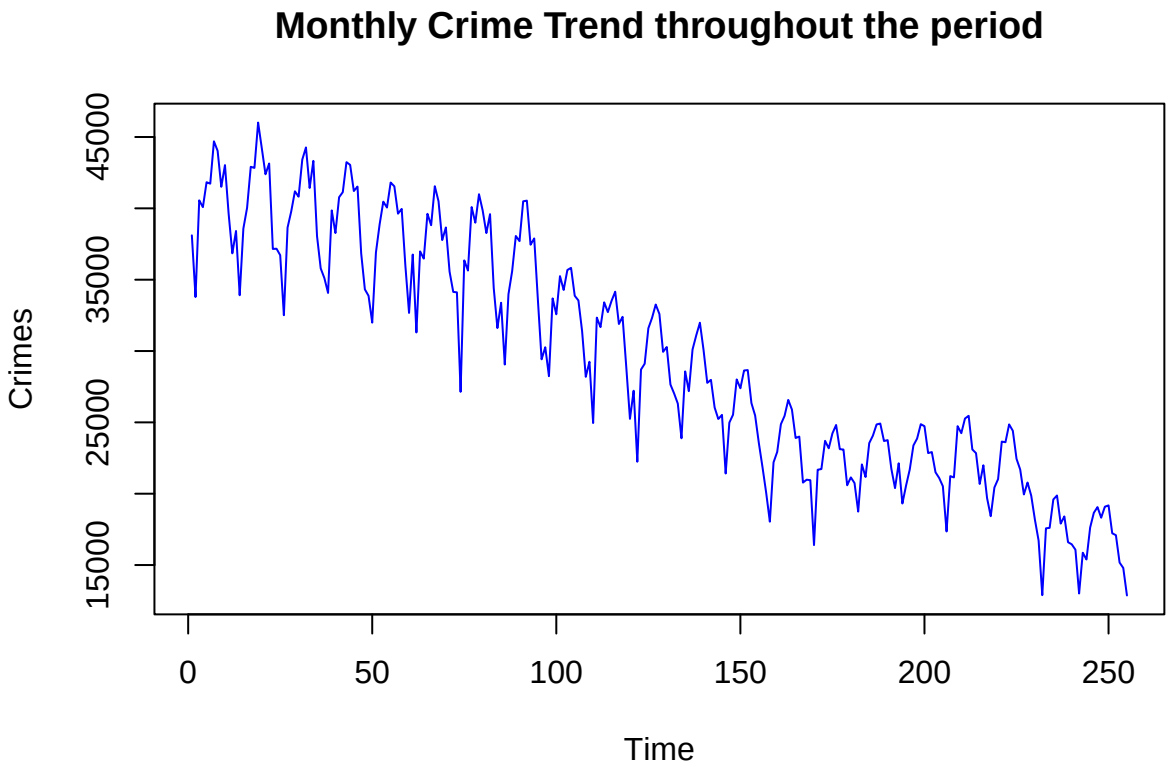
The initial step consists of reworking the data into a lot of comprehensible formats to feed it as an input to the model. Knowledge preparation may be a wrangle technique wherever knowledge transformation happens, i.e., inconsistent, incomplete knowledge with errors is converted to a clear format. In this pre-processing phase, we imported Chicago Crime data collected from (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2). The most crucial task in this stage is to identify the exact features necessary for model building During data preparation, we discovered that just the date feature is necessary for model building. Because the dataset is large, we are utilizing the fread function for quicker readings rather than the read.csv method. The original dataset contains information on each reported event, such as the case number, the date the incident happened, the kind of crime, if an arrest was made, geographic details, when the case was last updated, and so on. The monthly number of criminal occurrences will be the primary subject of this report. As a result, we chose one feature from the list of 22 for data modeling. Because the data contains a record for each crime recorded, we must first aggregate the data across monthly incidences reported. We just used the Date column from the data to aggregate the data on months. We next sort the data in ascending order once it has been collected.

```
##    year month Crimes
## 1: 2001     1  38110
## 2: 2001     2  33783
## 3: 2001     3  40562
## 4: 2001     4  40085
## 5: 2001     5  41832
## 6: 2001     6  41730
```

```
##
##  Range of Time:  Jan 2001 , Mar 2022
```

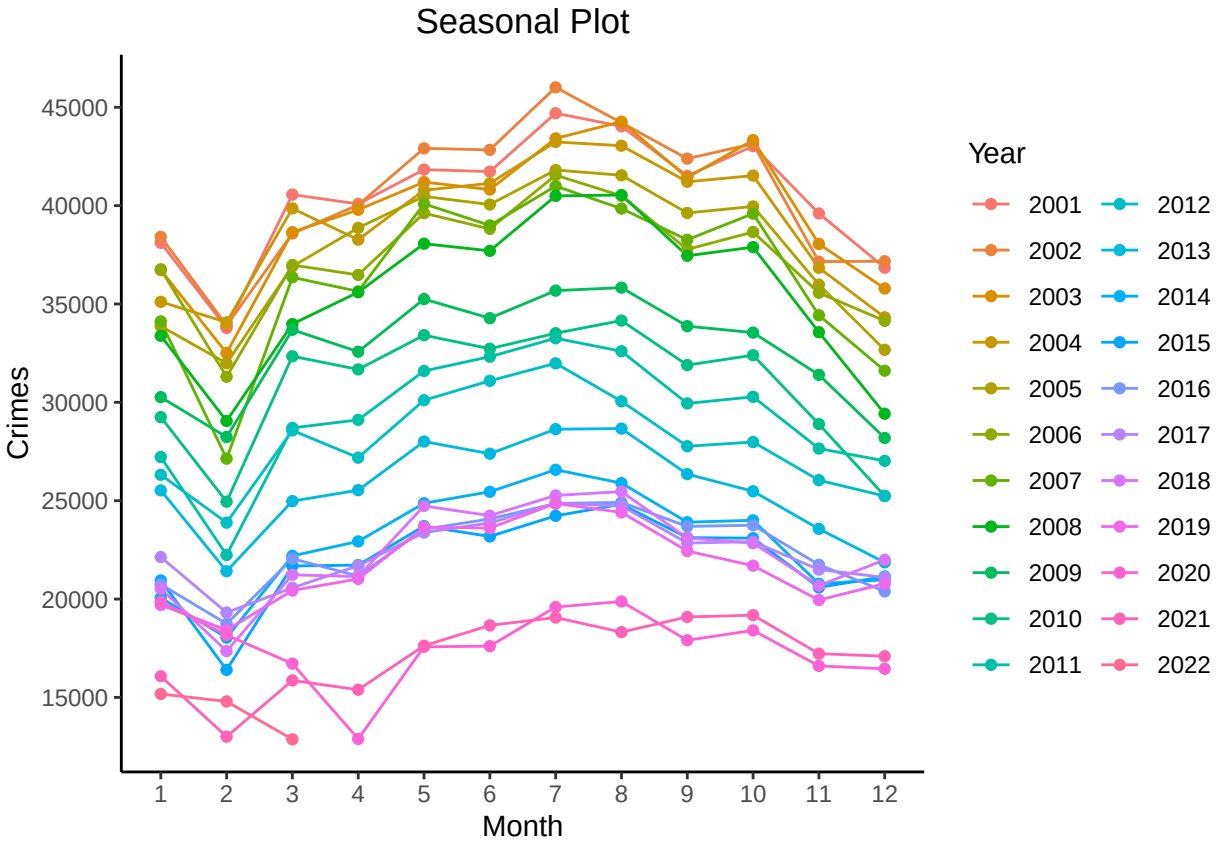
### 4.1. Visualization of Monthly Crime Trend through the period

By examining the time series plot below, we can observe that the series has a declining trend and cycles that appear to represent season impacts. There is also heteroscedasticity and the variance seems to be decreasing with time.



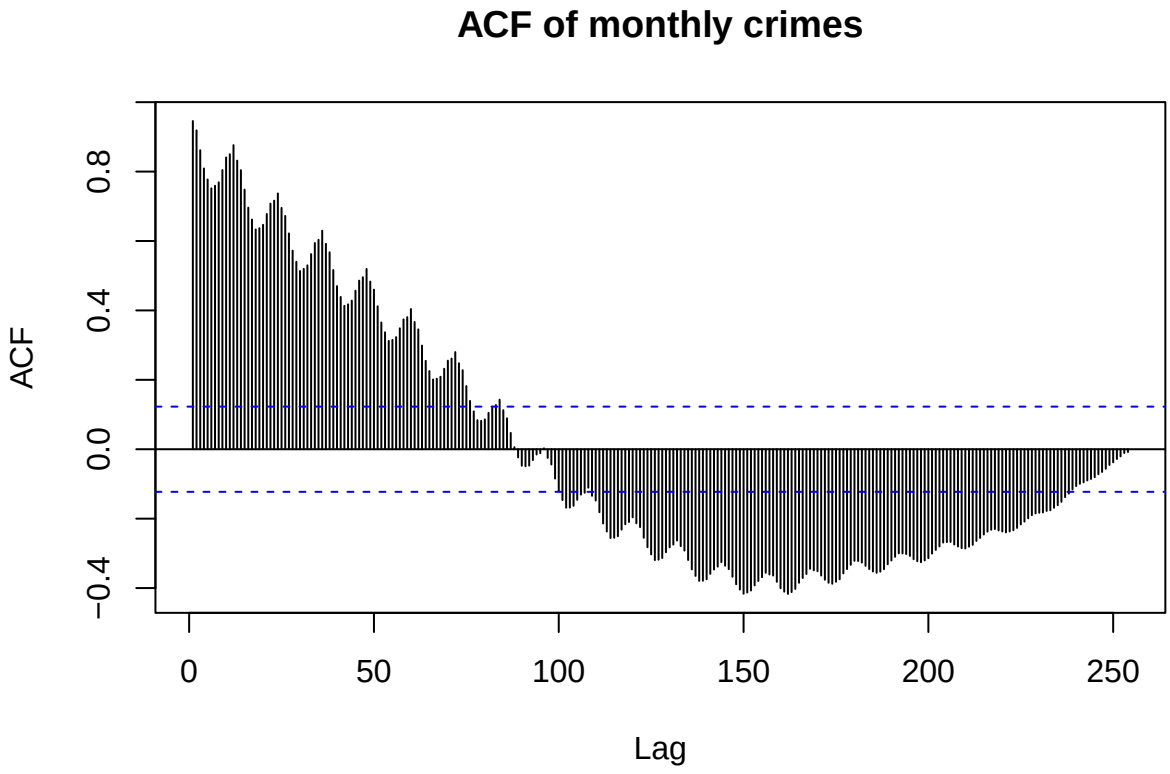
The seasonal plots described below indicates the underlying seasonal pattern, which might be valuable in spotting years when the pattern changes. In this plot, we can observe that the crime rate has continuously declined over the years. It is also worth noting that volatility has decreased in recent years, and the months with low crime rates in recent years have been considerably lower than in prior years. Furthermore, we see an increase in crime in the fall and a decrease in crime throughout the winter, and peaks during the fall.

The lockdowns enforced during the commencement of the COVID-19 pandemic can explain the substantial fall in crime in the first few months of 2020. We can say that the security status of the city is improving comparing to prior year.



## 5. SEASONALITY

The autocorrelation plot below indicates that there is the presence of cycles. Since it does not decrease at a constant rate when the lag increases. This shows that there is a strong correlation between the values that are a specific period apart. Looking at the seasonal plot we think that this cycle in the autocorrelation plot has a period of 12 and that it is not just white noise.




To formally validate our claims that the ACF is not just white noise, and a series of correlated values, we perform the Ljung-Box test. The p-value at lag 12, is much lower than 0.001 so we can reject the null hypothesis of the absence of serial correlation.

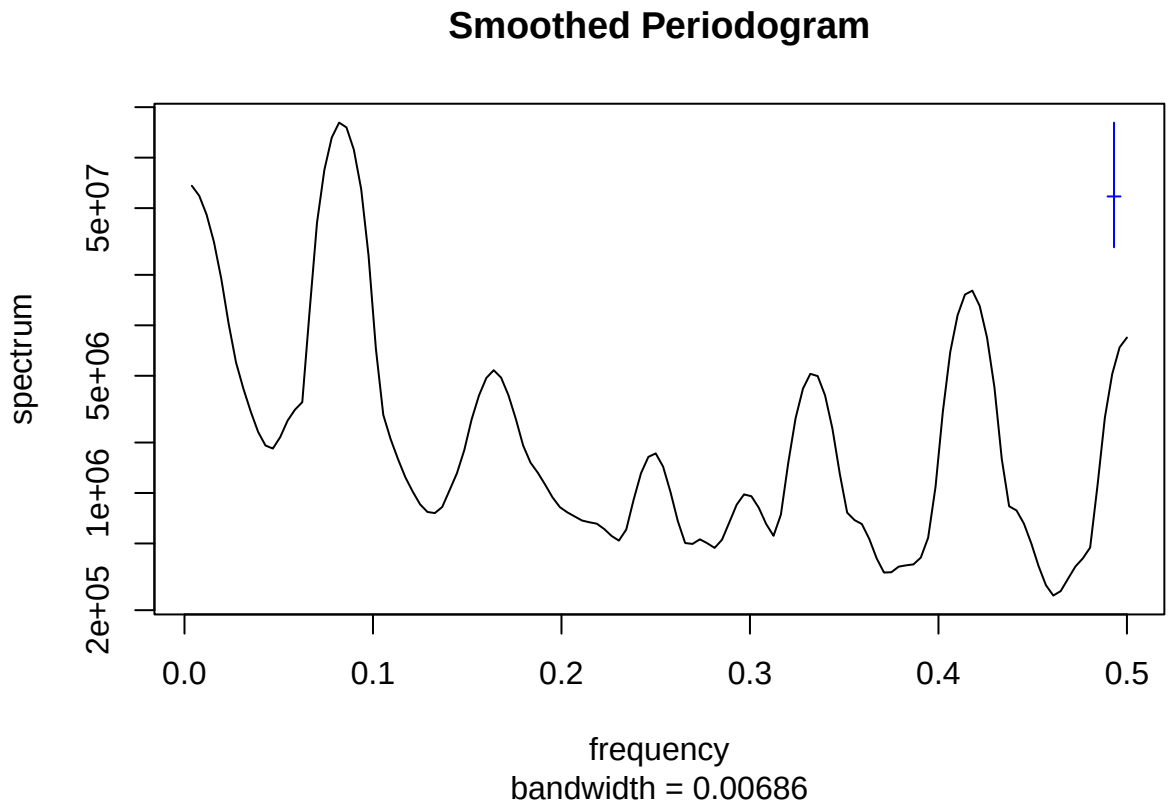
```
##
##  Box-Ljung test
##
## data:  monthly_crimes$Crimes
## X-squared = 2192.5, df = 12, p-value < 2.2e-16
```

### 5.1 Frequency Analysis

We can see that the data has a length of 12 months by looking at the monthly plot.To confirm our hypothesis, we used findfrequency().

```
## Frequency of original data 12
```

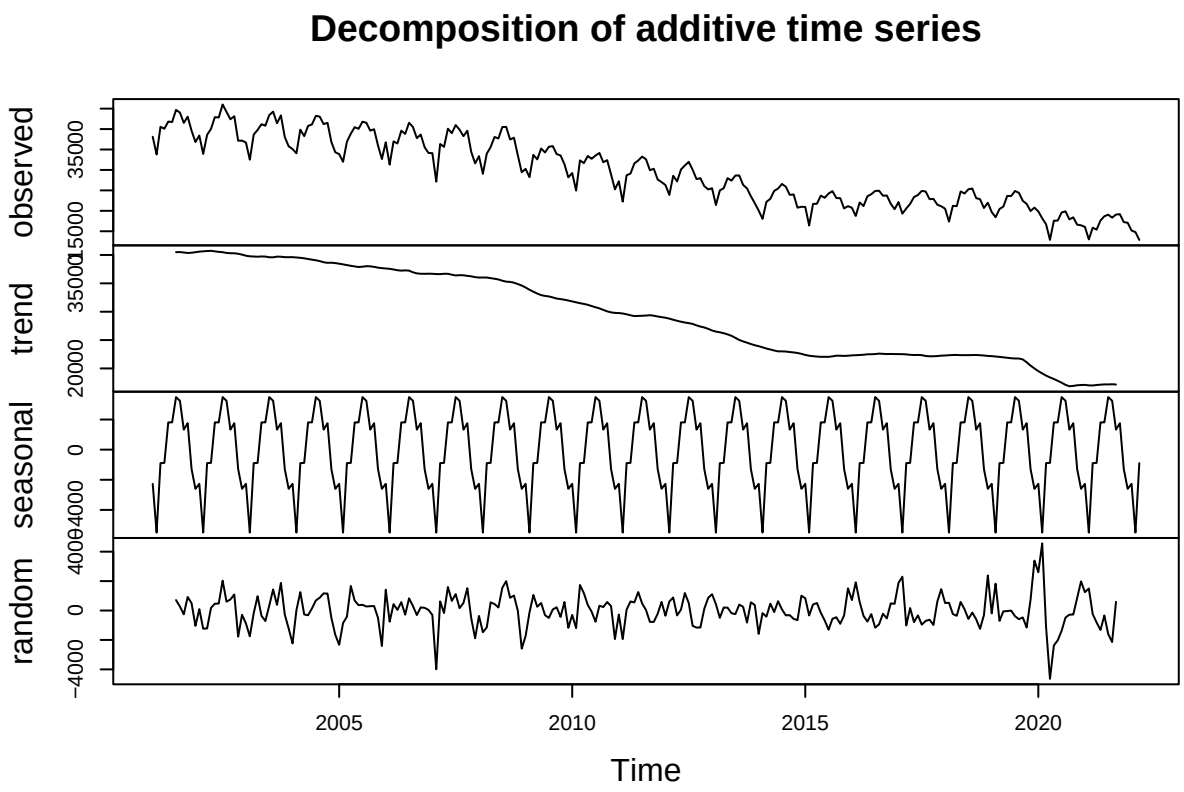
Calculating the frequency of the data using spectrum yields similar results.




```
## Frequency of data using spectrum is 12.1904761904762
```

### 5.2 Summarizing the Seasonality

We analyzed the crime rate from a frequency domain perspective to gain information from a different angle i.e. using decompose over a frequency of 12. We may look at the data plot and break down monthly crimes into a trend, noise, and cycle components.



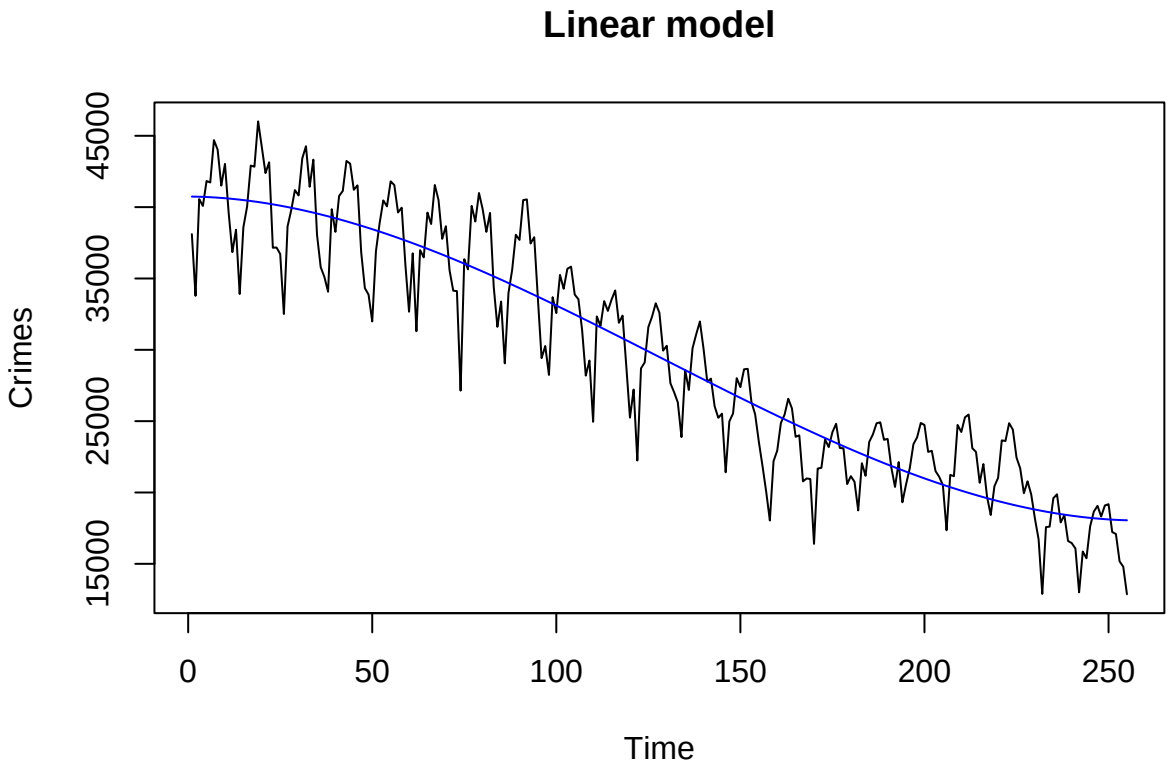
# 6. DATA MODELLING

## 6.1 Linear Model Fit

We first tried to fit a linear model to see how well the linear model can perform and if it is able to explain the data. Since the data shows a cubic trend, we fit the data with a cubic polynomial.

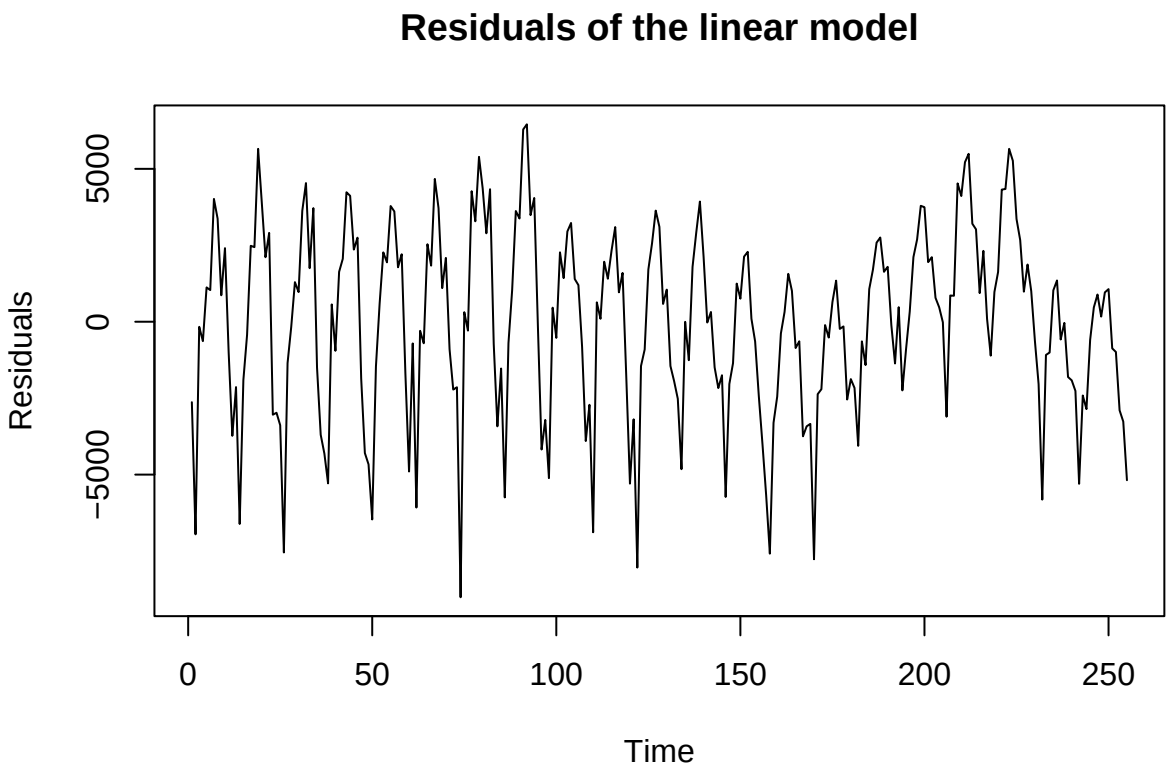
$$\hat{Y} = \hat{\beta}_1 \cdot month + \hat{\beta}_2 \cdot month^2 + \hat{\beta}_3 \cdot month^3$$

```
##
## Call:
## lm(formula = monthly_crimes$Crimes ~ month + I(month^2) + I(month^3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9011.0 -1913.1   318.4  2129.3  6455.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.074e+04  7.757e+02  52.527  < 2e-16 ***
## month       -2.195e+00  2.619e+01  -0.084    0.933
## I(month^2)  -1.001e+00  2.375e-01  -4.216 3.47e-05 ***
## I(month^3)   2.592e-03  6.098e-04   4.250 3.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3051 on 251 degrees of freedom
## Multiple R-squared:  0.8701, Adjusted R-squared:  0.8685
## F-statistic: 560.3 on 3 and 251 DF,  p-value: < 2.2e-16
```

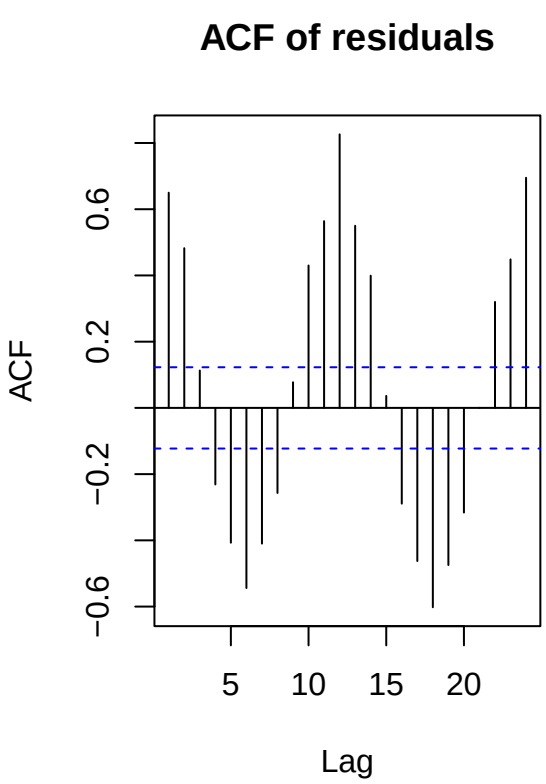


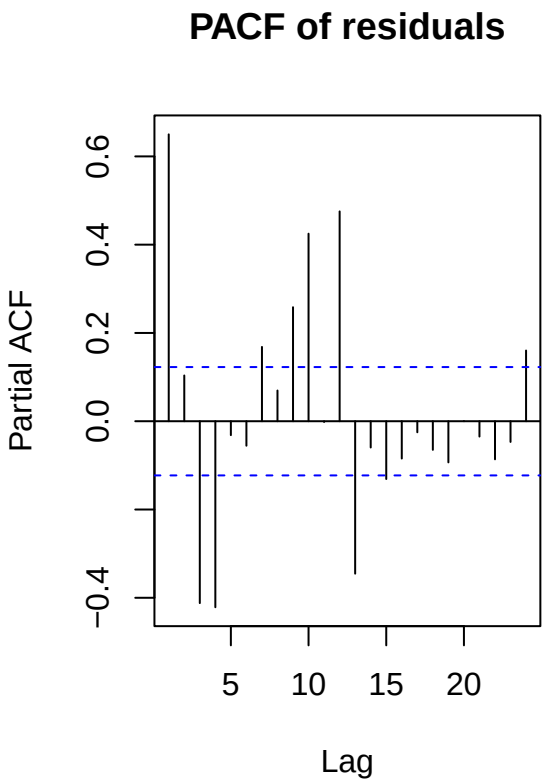

Plotting the fitted model on the original data reveals that the model can fit on the trend and that the residuals show no evidence of trend. We do a Cox-Stuart trend test to formally confirm this assertion. We do not reject the null hypothesis of there being no trend in the residuals since the p-value is substantially more than 0.05.

```
##
##  Approximate Cox-Stuart trend test
##
## data:  cubic_fit$residuals
## D+ = 69, p-value = 0.1645
## alternative hypothesis: data have a increasing trend
```



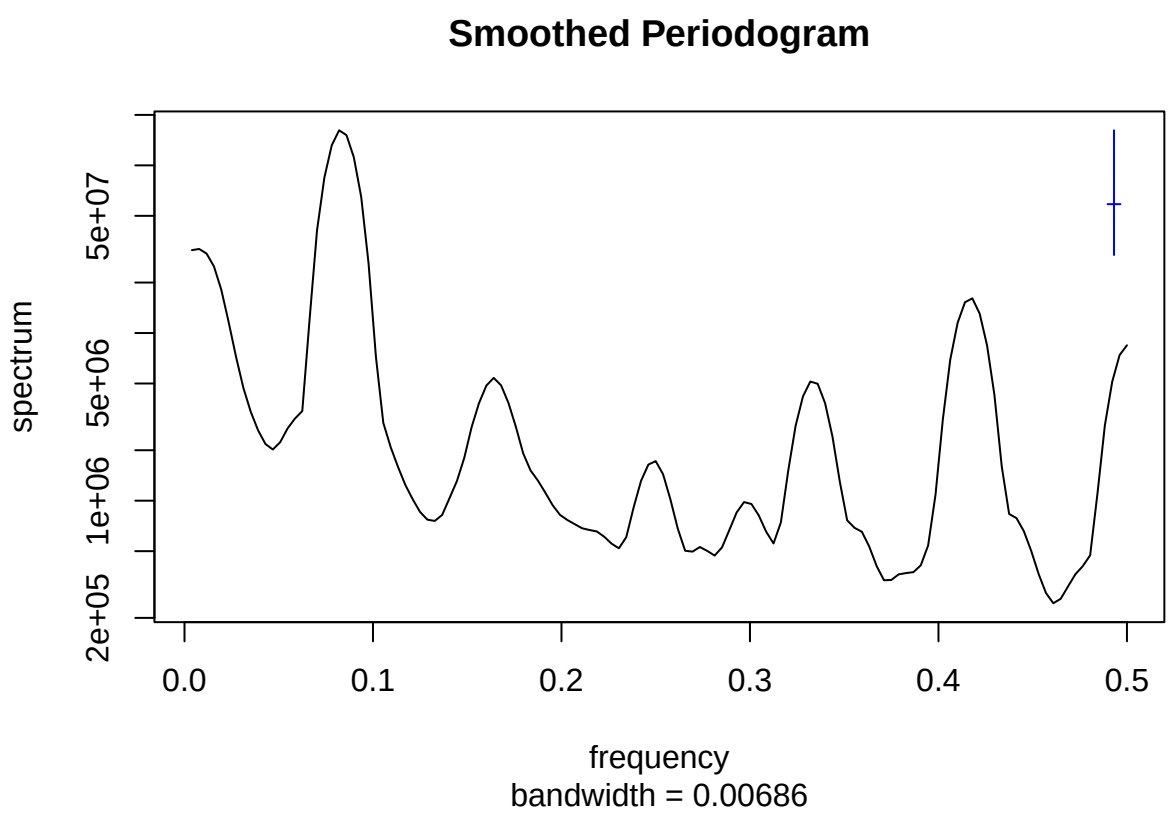

Looking at the residuals, we can say that they are not random and the model cannot explain the seasonality in the data. Examining the residuals ACF and PACF reveals that there is still some unexplained association.




Using the findfrequency function, we determine the frequency period of the residuals.

```
## Frequency of the residuals using findfrequency() is :  12
```




```
## Frequency of data using spectrum is 12.1904761904762
```

### 6.2 Fitting an SARMA Model

Fitting a SARIMA model with P = 2 and Q = 0, keeping the period 12, we try out different values of p and q to get the best fitting model for the data.

Table 1: AIC scores of diffenrent SARIMA models

| | MA0 | MA1 | MA2 | MA3 | MA4 |
|---|---|---|---|---|---|
| AR0 | 4596.349 | 4537.298 | 4521.273 | 4459.432 | 4451.560 |
| AR1 | 4393.572 | 4363.700 | 4362.478 | 4364.190 | 4363.193 |
| AR2 | 4371.967 | 4360.932 | 4365.168 | 4365.790 | 4364.403 |

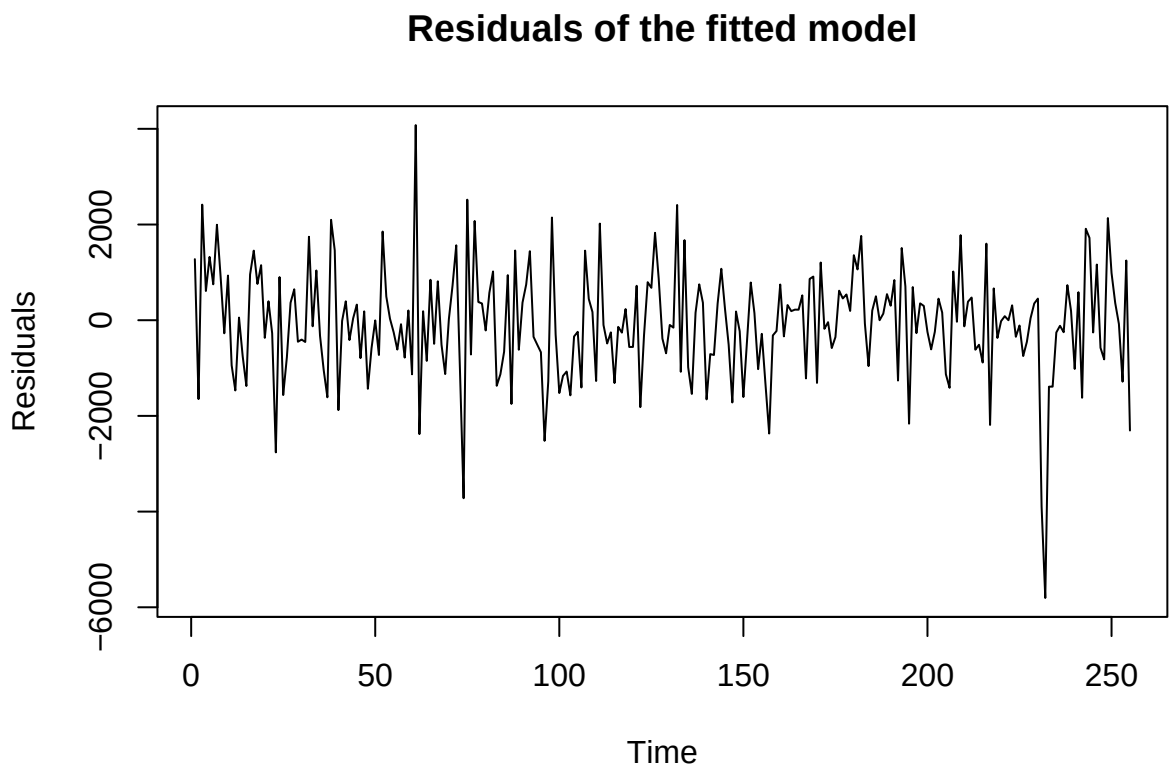
Looking at the AIC values, we can say that the best model is ARMA(2, 0, 1) as it gives the lowest AIC values.

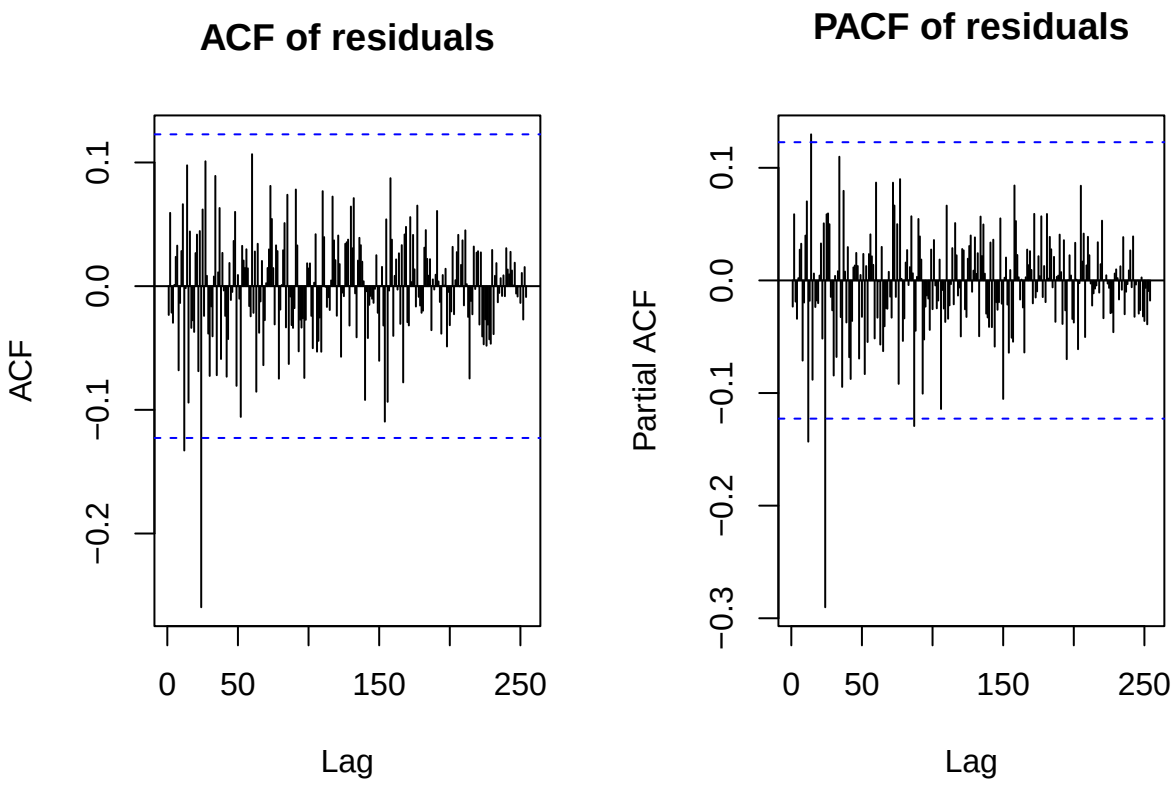
### 6.3 Final Model

```
## Series: monthly_crimes$Crimes
## ARIMA(2,0,1)(2,0,0)[12] with non-zero mean
##
## Coefficients:
##          ar1      ar2      ma1    sar1    sar2      mean
##       1.2846  -0.2964  -0.7465  0.4852  0.4380  22783.00
## s.e.  0.1198   0.1144   0.0883  0.0596  0.0612  13116.64
##
## sigma^2 = 1384122:  log likelihood = -2174.55
## AIC=4363.11   AICc=4363.56   BIC=4387.9
##
## Training set error measures:
##                      ME     RMSE      MAE        MPE     MAPE      MASE
## Training set -68.15399 1162.564 871.5876 -0.4832773 3.360924 0.4568239
##                     ACF1
## Training set -0.0232643
```

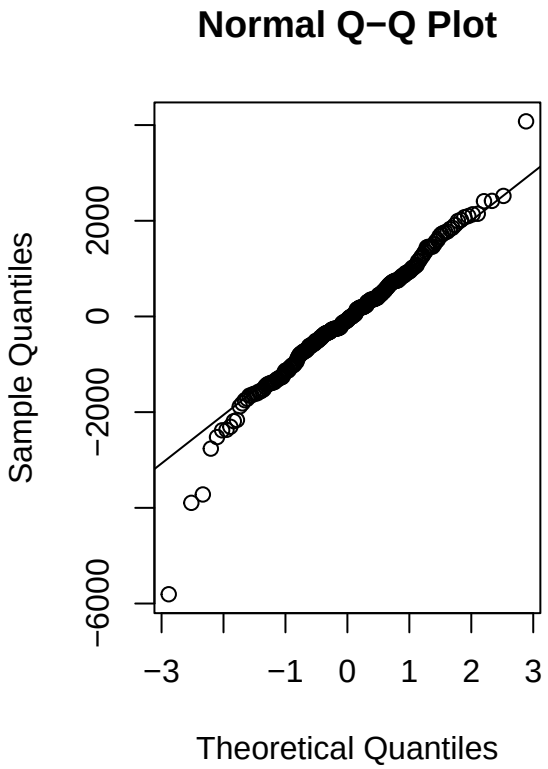
$$(1 + 1.2846B - 0.2964B^2)(1 + 0.4852B^{12} + 0.438B^{24})(\hat{Y_n} - 22783.00) = (1 + 0.7465B)\hat{\epsilon_n}$$

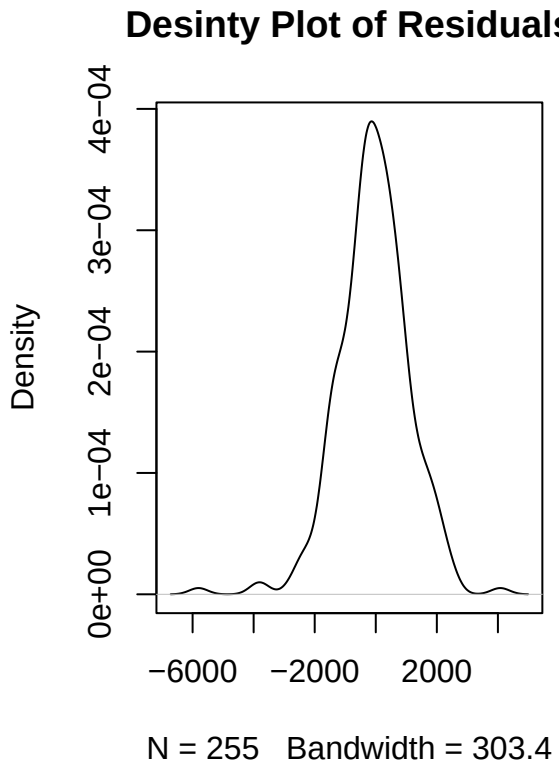
Plotting the model's residuals, we don't observe any kind of trend or pattern. To ensure that our model is not missing anything important, we examine the ACF and PACF of the model residuals.






Looking at the residuals and their ACF and PACF, we can see that the model can explain seasonality, with the exception of occasional spikes in them that may be attributed to random noise. To confirm that the residuals do not include anything meaningful, we are using qqplot.




We may deduce from the Q-Q plot and the density map of the residuals that the model is not missing anything substantial and that the residuals are near to normal.

### 6.4 Forecasting

Before we began forecasting, we separated our data into test and train sets. We can evaluate the performance of the out-of-sample data set by scoring it with the forecast() function and comparing the results to the actual values.

We fit our ARIMA(2,0,1)(2,0,0)(12) model to training data and predicted the numbers for the following 12 months, which is our test data.

We concluded that the actual and predicted values are quite close, and the model predicts the values as expected with a minimal error rate.

```
## Series: train_data$Crimes
## ARIMA(2,0,1)(2,0,0)[12] with non-zero mean
##
## Coefficients:
##          ar1      ar2      ma1    sar1    sar2      mean
##       1.2874  -0.2963  -0.7539  0.5464  0.3789  16703.41
## s.e.  0.1175   0.1120   0.0875  0.0645  0.0660  25545.68
##
## sigma^2 = 1365267:  log likelihood = -2071.35
## AIC=4156.7   AICc=4157.18   BIC=4181.15
##
```

```
## Training set error measures:
##                     ME     RMSE      MAE         MPE     MAPE      MASE
## Training set -56.1778 1153.931 861.5471 -0.4402686 3.198112 0.4412085
##                      ACF1
## Training set -0.02311261
```

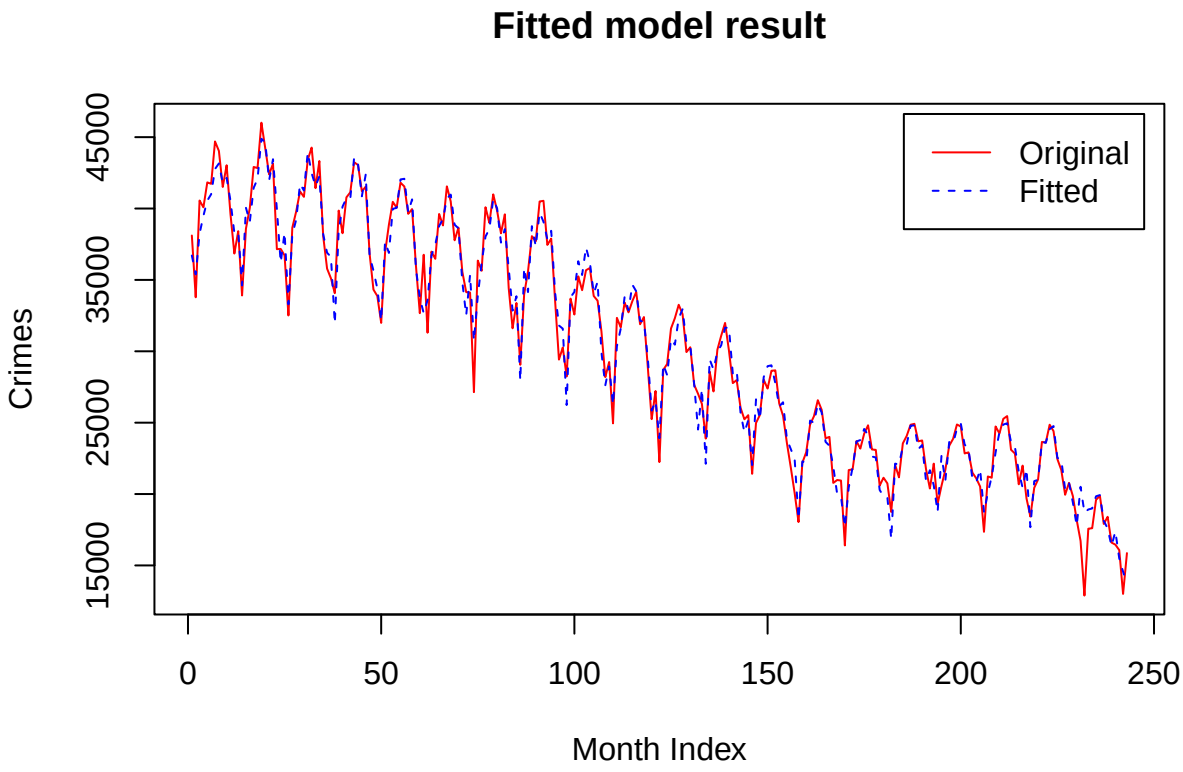
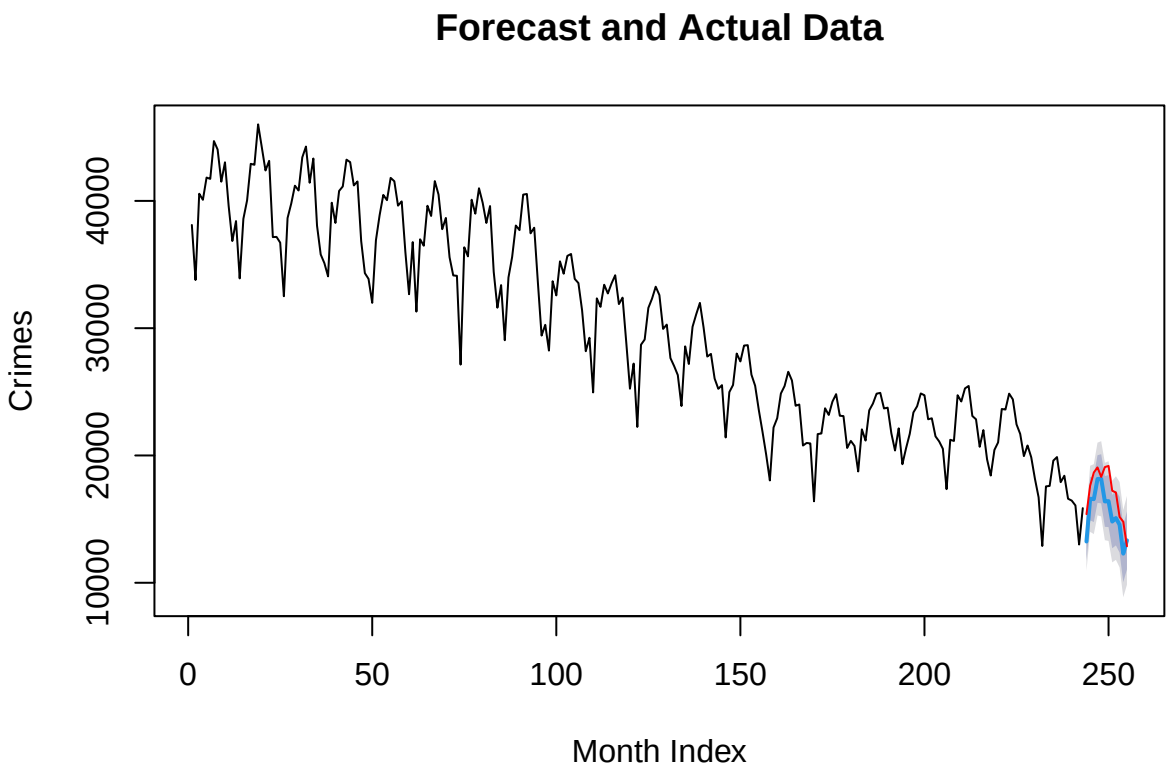
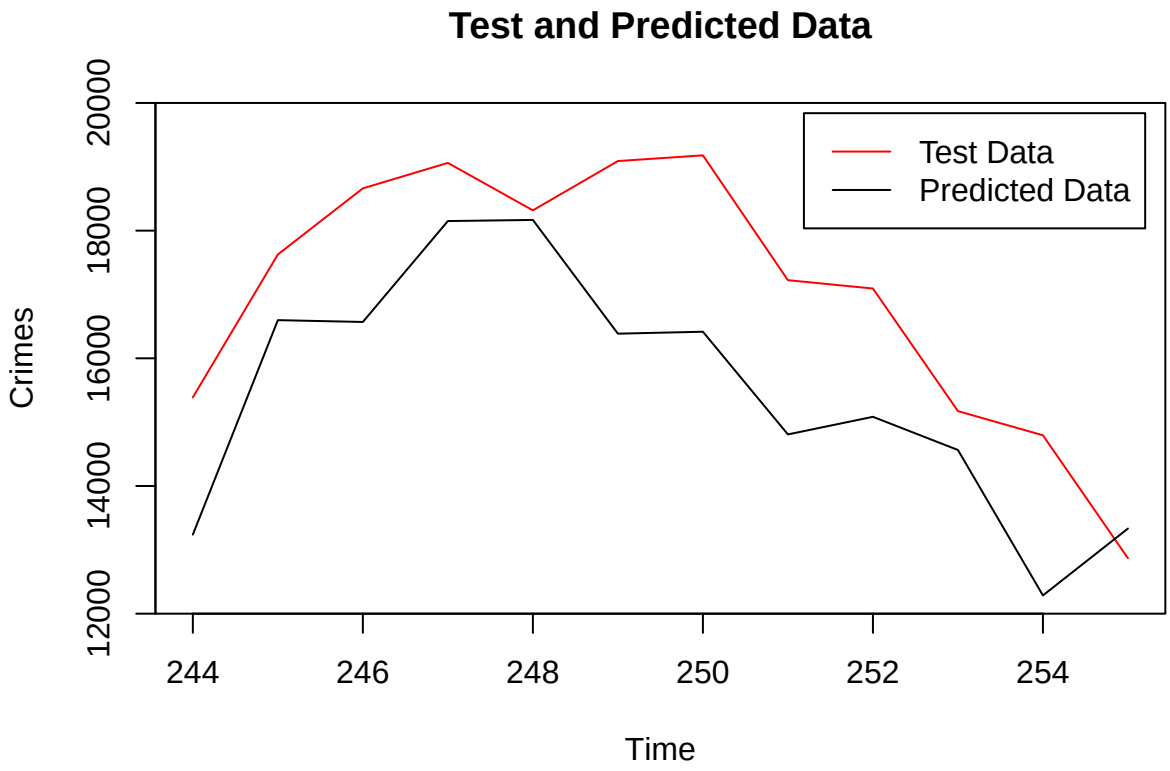



Table 2: Forecast for the next 12 months

| | Point Forecast | Lo 80 | Hi 80 | Lo 95 | Hi 95 |
|---|---|---|---|---|---|
| 244 | 13238.99 | 11741.56 | 14736.41 | 10948.873 | 15529.10 |
| 245 | 16597.93 | 14900.75 | 18295.11 | 14002.322 | 19193.55 |
| 246 | 16569.21 | 14774.16 | 18364.27 | 13823.909 | 19314.52 |
| 247 | 18150.09 | 16282.36 | 20017.82 | 15293.645 | 21006.53 |
| 248 | 18167.69 | 16236.52 | 20098.86 | 15214.217 | 21121.16 |
| 249 | 16385.78 | 14396.06 | 18375.50 | 13342.764 | 19428.79 |
| 250 | 16416.41 | 14371.63 | 18461.19 | 13289.188 | 19543.63 |
| 251 | 14807.66 | 12710.72 | 16904.60 | 11600.671 | 18014.65 |
| 252 | 15083.37 | 12936.85 | 17229.89 | 11800.554 | 18366.19 |
| 253 | 14564.05 | 12370.30 | 16757.79 | 11209.004 | 17919.09 |
| 254 | 12286.15 | 10047.34 | 14524.96 | 8862.189 | 15710.12 |
| 255 | 13333.45 | 11051.58 | 15615.32 | 9843.634 | 16823.27 |


Forecast and Actual Data
Crimes
10000
20000
30000
40000
0
50
100
150
200
250
Month Index

Taking a closer look at the predictions and test data. We can conclude that the model is quite significant and capable of predicting drops and peaks in crime rates accurately.




```
##                ME     RMSE      MAE      MPE     MAPE
## Test set 1573.101 1884.419 1650.676 9.027218 9.630072
```

## 7. OTHER INFLUENCING FACTORS

The external factors plays important role in crime statistics. Here, we are taking the weather data and analyzed the patterns in Chicago. By examining, we can see the strong correlation between weather and crime over annual time periods, which in turn allows for more insight into the decision-making process behind the crime.

The below mentioned graph signifies the direct relationship with the crimes communed. We can deduce that when the heat index was higher especially in the month of June, July and August, rate of crime were higher compared to other months. Overall, crime rates were highest in the warmest months of the year.During the year's colder months, the contrast of high versus low rates of crime on more comfortable versus cooler temperature days was more striking.

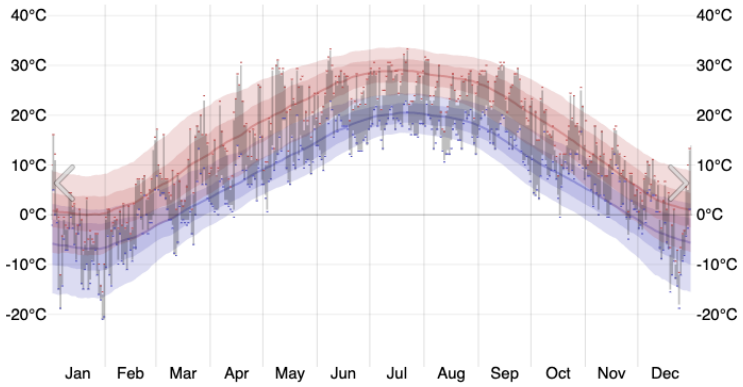



Figure 1: Chicago weather in 2004

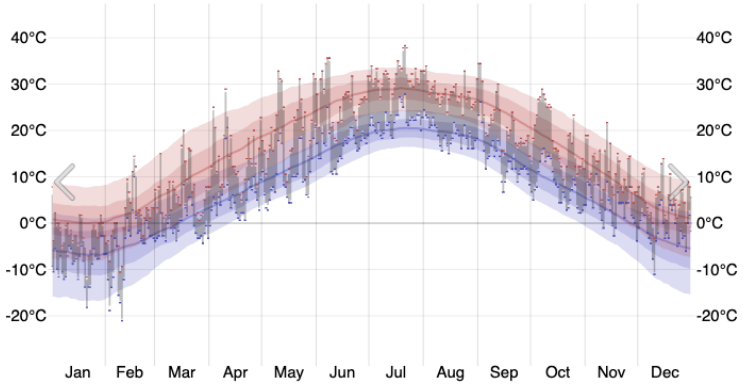



Figure 2: Chicago weather in 2011

## 8. CONCLUSION

This data analytic research provided us a scientific perspective on the city of Chicago's security and crime rate. According to the analytically results and visualization, the frequency of occurrence of the crime trend repeats every 12 months. It is also worth mentioning that volatility has lessened in recent years, and months with low crime rates have been significantly lower in recent years than in previous years. Furthermore, we examined the pattern of high peaks in criminal activity in Chicago throughout the months of June, July, and August. The lockdowns imposed during the start of the COVID-19 pandemic can explain the significant drop in crime in the first few months of 2020. While ARIMA models forecast using the whole time series, exponential smoothing approaches give more weight to the most recent observations. However, in this dataset, a SARIMA model produced the best results. The SARIMA model with ARIMA(2,0,1)(2,0,0)(12) well explains the 12-month crime rate predictions , which exhibits a trend.

## 9. FUTURE WORK

1. Investigate data with other external factors such as unemployment, and holidays, as well as the impact of these factors on crime rates as the weather obviously has an impact on the quantity of crimes, we do not believe it is the sole significant element impacting the outcomes.
2. Create a better model with more helpful features like crime type, location, and arrest rate.
3. To experiment with different data formats like AVRO or Parquet to make row/column operations more efficient.

## 10. REFERENCES


1. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2
2. https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/time-series/how-to/partial-autocorrelation/interpret-the-results/partial-autocorrelation-function-pacf/
3. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/decompose
4. https://rdrr.io/cran/forecast/man/findfrequency.html
5. https://www.rdocumentation.org/packages/aTSA/versions/3.1.2/topics/trend.test
6. https://weatherspark.com/y/14091/Average-Weather-in-Chicago-Illinois-United-States-Year-Round
7. https://bookdown.org/yihui/rmarkdown/