# Vancouver Violent Crimes

Applied Time Series Analysis

Tyler Oh

Dec 08, 2022

## Executive Summary

Violent crimes are widespread across many nations, and Canada is no exception. While several investigations and studies have been undertaken in the past to study crime patterns in the country, few of them have used the crime data available at the VPD (Vancouver Police Department) considered in this report. Our goal is to see whether there is a pattern, seasonality, or cycle of violent crimes in the dataset, and then build a time series prediction model to predict the number of crimes for a given period. Because the COVID-19 epidemic occurred within the forecasted time (late 2019), it would be intriguing to discover whether it influenced the number of violent crimes.

A seasonal ARIMA model was implemented based on the version of the dataset that has been aggregated and summarized monthly. The best model was a seasonal ARIMA model of the form ARIMA(1,1,1)(2,0,0)[12], obtained after performing multiple diagnostics on the residuals of the various versions of the model created. This model was chosen as the framework for prediction and analysis because it produced residuals with the least autocorrelation.

Following the selection of the best model, the next stage was to carry out the actual forecasting for the following 23 months (years Jan 2021 to Nov 2022). The results reveal two significant findings. First, the seasonality feature remained apparent, with some months recording large numbers of violent crimes while others reported lower crime rates, and the seasonality of the predicted values repeated over the two years of prediction. The second observation is that, after 2020, the general trend of the data (previously available dataset and the predicted values) changed and assumed a negative average change. This means that, in general, the number of crimes started increasing after COVID-19 broke (a

nd continued increasing for the next two years) and still retained high crime rates in summer and low rates in other months. (While generally cases increased, some months still had lower crime rates than others. But generally, all months during the COVID-19 time had lower crime rates than the same month pre-pandemic).

## Introduction and Hypothesis

COVID-19 has affected our daily lives, businesses, global trade, and movements. We were curious about how COVID-19 played a role in the crime rate. It is a seasonality of crime rates which is summer fall, winter, and spring that come back every year. The famous statistician Adolph Quetelet stated “first brought attention to seasonal patterns in crime. In 1842 he proposed a “thermic law of delinquency,” stating that violent crimes were more common in hotter climates and seasons and that crimes against violence were more frequent in colder climates and seasons.” It is quite interesting that in certain seasons, there is a drastic increase or decrease in violent crimes.

The purpose of this study is to examine the lag in crime data to better understand crime in our current period. Our focus will be on the number of violent crimes committed every month between 2010 and 2020, to identify trends and fresh insights utilizing the statistical tools we learnt in this course. In this study, we hypothesize that crime occurs on a seasonal basis. As a result, our data will be non-stationary by definition. As a result, the study is classified as time series analysis, and the method utilized is ARIMA (which is a mix of autoregressive (AR) and moving average (MA) analysis, but with the diffe rencing factor, I). Our final objective is to predict the number of violent crimes on average to have the right policy to prevent crimes that we don’t want to happen in the first place.

Some of the key aspects to be addressed in the section are a more detailed overview of the dataset to be used, a description of the time series model used, checking the stationarity of the dataset and transforming the data to a format suitable for ARIMA analysis, performing diagnostics tests on the model, performing predictions and reporting the accuracy of the model.

First and foremost, we begin our research topic by hypothesizing that our model is naturally non-stationary. In our first stage, we confirm our hypothesis by testing whether our data is stationary or not. Once that has been proven, we will choose the appropriate models that will well estimate our data. In the second step, we will examine various models (ARMA, ARIMA, ARIMA seasonal) to determine which model will be best for our predictions. In the final step, we will determine if the series is white noise to establish that we are performing a forecasting analysis. We will forecast crime data using the ARIMA seasonal model and assess the accuracy (lowest RMSE) of our prediction using testing and testing samples whether it meets our white noise hypothesis testing using residual analysis.

## An Overview of the Dataset

As noted in the introductory part of the paper, the dataset is a report of crimes extracted from the Vancouver Police Department (VPD) website. It has 120 observations and four features, namely, the date of the crime, the month of the crime, and violent crimes. The crimes for the two categories are reported monthly, and the modelling will perform predi ctions on those bases. It is also noted that the current analysis will only focus on the violent crimes variable. Additionally, the 120 observation data frame is further reduced to its monthly version, implying that the final version consists of approximately 239 monthly entries as seen in figure 1. Figure 2 displays a time series plot of the dataset, and what we can see from figure 3 is that there are peaks in May, August, and September (Summer season).

## Model Description

ARIMA is the analysis model proposed. The model was chosen since it combines the two characteristics of autoregression (AR) and moving averages (MA), making it more robust while being reasonably straight- forward to execute. Furthermore, the process of obtaining the main seasonality factors is simplified because the R packages have automated methods of obtaining them by looping through several combinations of the seasonality constants p, d, and q and providing the ones that result in a model with the highest estimation of the true sample. In addition, we will fit the data in R using the seasonality model ARIMA (p, d, q)(P, D, Q)[t]. The MLE and AIC of all models will then be compared to see which one explains the most variation in our data.

## Checking Residuals Of Initial Data

The trend is a pattern in data which displays a Long-term increase or decrease in data points over time. Seasonality is cycling in a series that repeats at fixed frequencies (hour of the day, week, month, year, etc.), in a fixed known period and thus a seasonal pattern exists. Please see the appendix for the complete analysis of figures 3 - 11. Both models' histogram and normal-QQ plot are symmetry and lie 45 degrees. However, the biggest d ifference is the coefficient of R-squared, which is nearly twice as different. It means that there is also seasonality in our data rather than only having trends. As a result, seasonality is an important component for our prediction because it displays a symmetry histogram indicating normality, a 45-degree line with data points close to the line in a normal q-q plot. From the ACF plot for the residuals using 12 lags, the data does not appear to be white noise. Also, lags 1 and 12 have a high correlation, indicating seasonality.

## Checking the Stationarity of the Time Series of ARIMA and ARIMA Seasonality Model

A quick Augmented Dickey-Fuller Test, as seen in figure 12, revealed that the original dataset was not stationary, with a P-value of 0.2837 > $\alpha$ = 0.05. The p-value is insignificant statistically. As a result, we fail to reject the null hypothesis. Figure 12 depicted a pattern that suggests there is no white noise process in our data. To make our data a stationary process, we clean it into 12-month periods. However, the data appears to be non-stationary. As a result, we normalize our data. The lags improved significantly, indicating that the data will be easier to evaluate. At the end of the day, the data is naturally non-stationary.

Thus, it was necessary to take another difference, a lag of 1, which solved the problem. The result of an ADF-test score of less than 0.01 shows us that the process is no longer stationary. The results for this second difference’s graph are shown in the procedure below. It can be seen that the means are now approximately centered around zero. However, there is evidence of instances of spikes, indicating that the data could be seasonal. Thus, we can assume the data is now stationary and proceed to implement the ARIMA mode

1. The four plots in Figure 11 show a histogram, Q-Q plot, ACF and PACF  time series of violent crimes and diff(log(violent crimes)). lag 1 and lag 12 are correlated which means there can be some seasonality

## Obtaining the Best ARIMA Model

Figure 15 shows two EACF plots that suggest ARIMA(1,1,1), ARIMA(0,1,1), ARIMA(1,1,0), ARMA(1,2), and ARMA(2,0). ARIMA(1,1,1) is one potential option for our prediction. Please see **figure 16 and 17**..

## Creating and Diagnosing the Model

An ARIMA(1, 1, 1) m     odel was constructed and various diagnostics were performed, primarily on its residuals.  The initial test was to examine if the residuals are autocorrelated, as shown in figure 13.. The ACF graph shows a considerable increase at the 7th and 12th lag. Similar conclusions may be drawn from the PACF graph shown in Figure 14. This suggests that the time series dataset is seasonal. The next subsection examines how the seasonality component was addressed.

Another diagnostic test involved performing the Box Jenkins test procedure on the residuals. A p-value much higher than 0.05 was obtained (0.6459), which would make us fail to reject the null hypothesis as seen in figure 24. In other words, we have strong evidence to conclude residuals are independent. Therefore, these results provide an optimal ARIMA model. However, we know there is seasonality in our data, we will need to finetune it by introducing the seasonality components. To get the best values of p, d, and q, we used the auto.arima() function to allow the software to automatically loop through various combinations of the p, d, and q, and provide the best-selected model (lowest AIC). The best model would have ARIMA(p = 1, d = 1, and q = 1). In other words, the best ARIMA model has a lag of 1, a stationarity difference of 1, and a residual lag of 1.

## Dealing with the Seasonality Components

Upon trying different values for the seasonality component (P, D, Q) for the dataset, a model of the form ARIMA(1, 1, 1)(2, 0, 0) [12] was obtained, and this seemed to solve the seasonality problem. For the seasonal component, AR = 1, seasonal differencing = 1, and the seasonal MA = 1 (corresponding to P, D, Q). Based on these parameters, the residuals ACF and PACF were indicated in Figure 14.

The readings are now within the appropriate margins, with no spikes going outside of this region. Performing the Box Test and obtaining the associated scores might provide more information. This time, the Box. test() value achieved is 0.87 as seen in Figure 19, which is significantly more than 0.05. As a result, we fail to reject the null hypothesis (which is usually that the "model does not exhibit lack of autocorrelation"). To put it another way, we do not have enough evidence to establish that our model has considerable autocorrelation on the residuals. In addition, Figure 17 depicts a graph for testing the data's normality.

Figure 18 shows that, except for the few outliers indicated above, the residuals are approximately normally distributed. ADF testing confirms that the results are substantial. Furthermore, the lowest AIC indicates that it is the best optimal fit model for not just this sample but also others. Since, intuitively, crime exhibits a seasonal trend. Thu s, based on the results described in this section, we conclude that the optimal model for analyzing and predicting time series data is of the form ARIMA(1, 1, 1)(2, 0, 0)[12].

## Forecasting of Violent Crimes - ARIMA Seasonality Model

Since the dataset is reported in months, we want to predict the number of crimes for the next two years (23 months), as shown below. The residuals fairly at zero are shown in Figure 20. The same results can be presented in the form of a graph as seen in Figure 18.

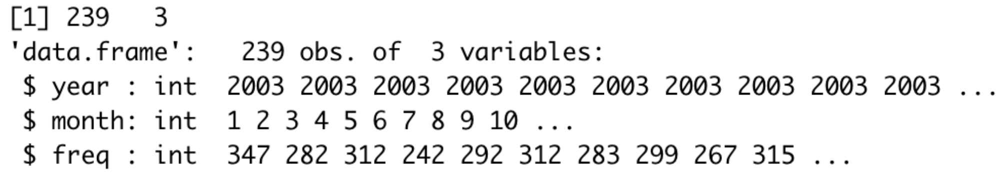
As a result, the study's goal has been achieved since the model may be utilized to make more forecasts for various periods as needed (Figure 21). Furthermore, the model reports a MAPE of the test set of 21.3 (Figure 23), indicating that the predicted values would differ from the actual values for the violent crime data provided by roughly 21% on average. (We can see that the various errors are large, indicating that the model is accurate. For example, RMSE falls below zero, signifying a reasonably well-performing model).

## Conclusion

The project's objective was to develop a prediction model that could be used to predict the number of crimes. Furthermore, we give insights on probable seasonal changes in the number of crimes throughout time and, more precisely, during the various seasons of the year. Furthermore, the findings of the model were to be compared before and during the COVID-19 period. It was feasible to anticipate the number of violent crimes for the following 23 months by using an ARIMA model with seasonality components (the next two years). According to the projected figures, the number of crimes remained high during the summer months and low throughout the rest of the year. Furthermore, the forecasted data revealed a general growing tendency in the number of instances for the predicted years. As seen in Figure 22, this rise (during the following 23 months, from January 2021 to November 2022) might be attributed to socioeconomic reasons such as the COVID-19 epidemic that erupted in late 2019.

# Appendices

## Data Description:



```
[1] 239   3
'data.frame':   239 obs. of  3 variables:
 $ year : int  2003 2003 2003 2003 2003 2003 2003 2003 2003 2003 ...
 $ month: int  1 2 3 4 5 6 7 8 9 10 ...
 $ freq : int  347 282 312 242 292 312 283 299 267 315 ...
```

*Figure 1: Data Frame’s Descriptions*

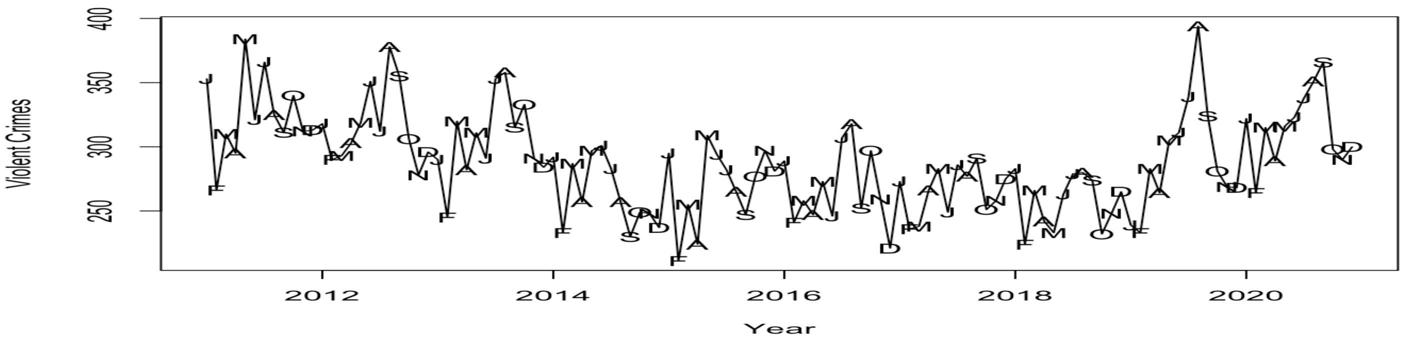


*Figure 2: Violent Crime-Monthly Plot*

## Residual Analysis:

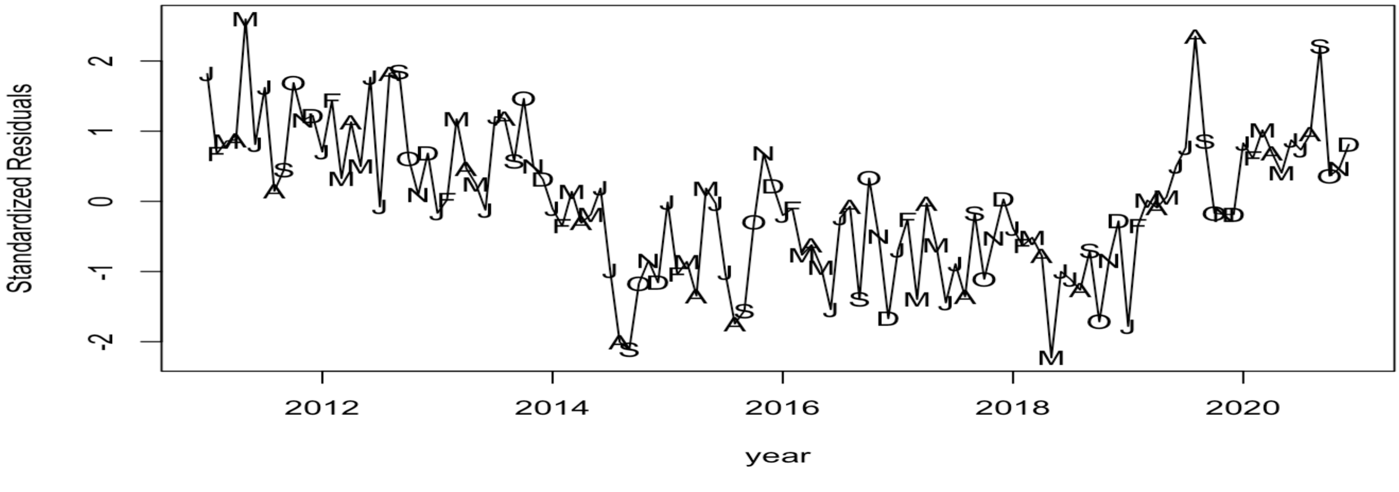
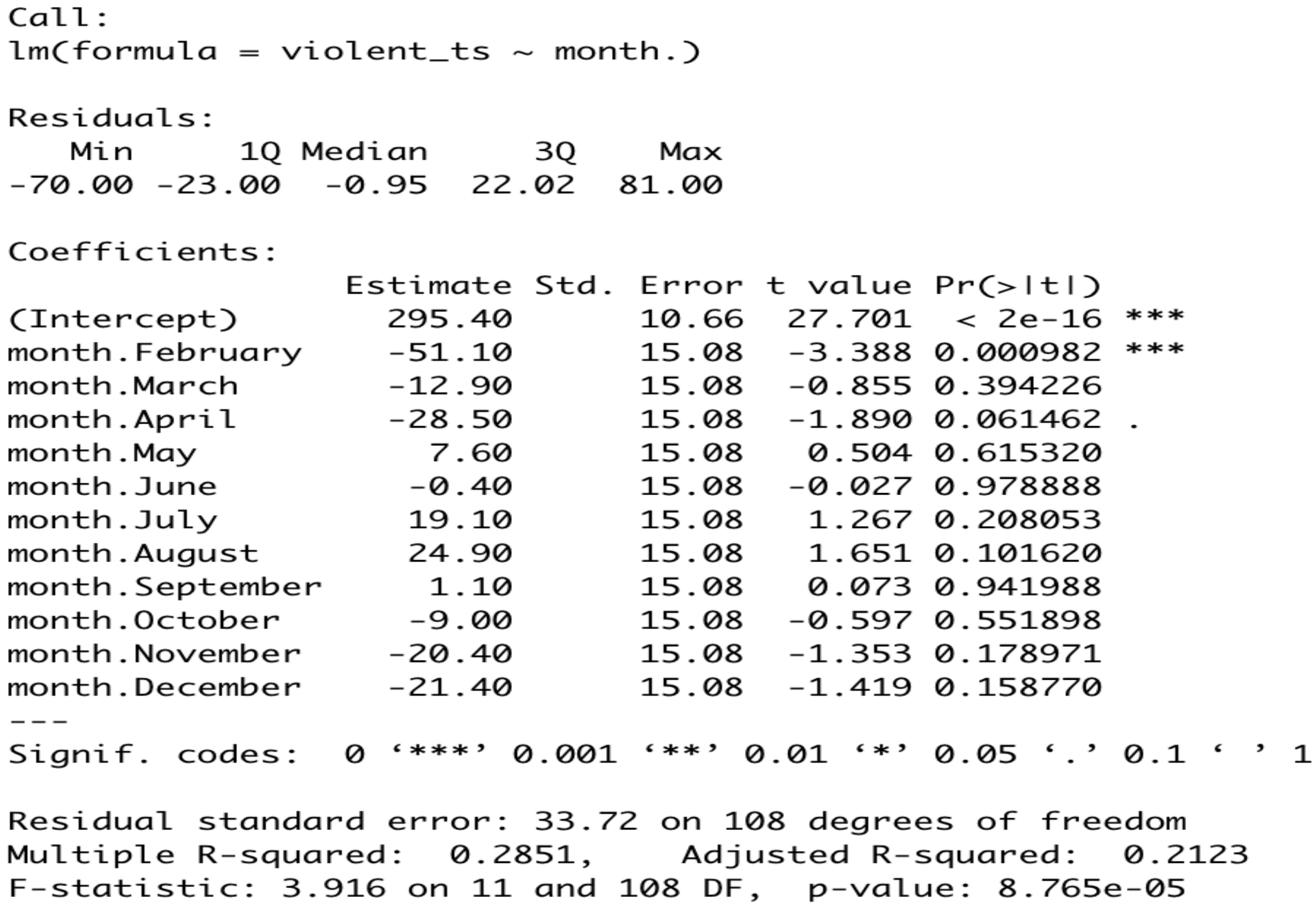


*Figure 3: Standardized Residuals*



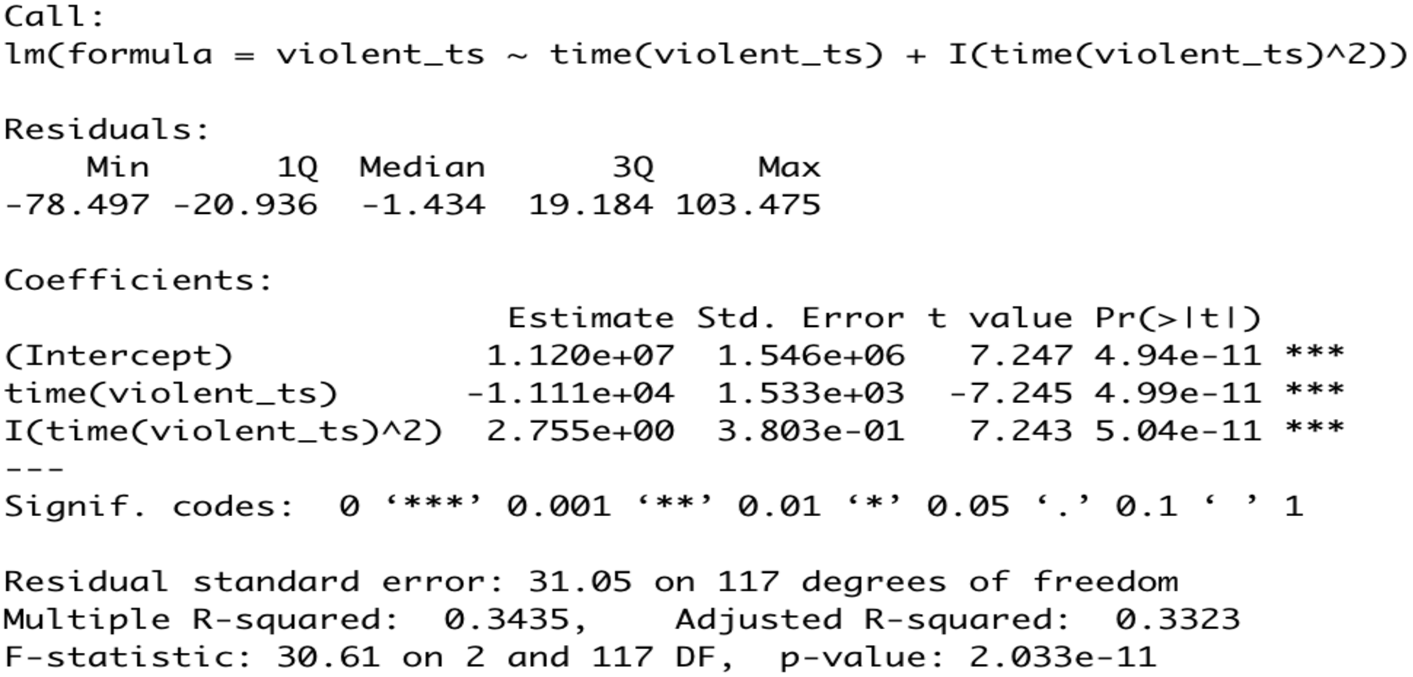
```
Call:
lm(formula = violent_ts ~ month.)

Residuals:
   Min     1Q Median     3Q    Max 
-70.00 -23.00  -0.95  22.02  81.00 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       295.40      10.66  27.701  < 2e-16 ***
month.February    -51.10      15.08  -3.388 0.000982 ***
month.March       -12.90      15.08  -0.855 0.394226    
month.April       -28.50      15.08  -1.890 0.061462 .  
month.May           7.60      15.08   0.504 0.615320    
month.June         -0.40      15.08  -0.027 0.978888    
month.July         19.10      15.08   1.267 0.208053    
month.August       24.90      15.08   1.651 0.101620    
month.September     1.10      15.08   0.073 0.941988    
month.October      -9.00      15.08  -0.597 0.551898    
month.November    -20.40      15.08  -1.353 0.178971    
month.December    -21.40      15.08  -1.419 0.158770    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.72 on 108 degrees of freedom
Multiple R-squared:  0.2851,    Adjusted R-squared:  0.2123 
F-statistic: 3.916 on 11 and 108 DF,  p-value: 8.765e-05
```

*Figure 4 . Summary of the fitted linear model*



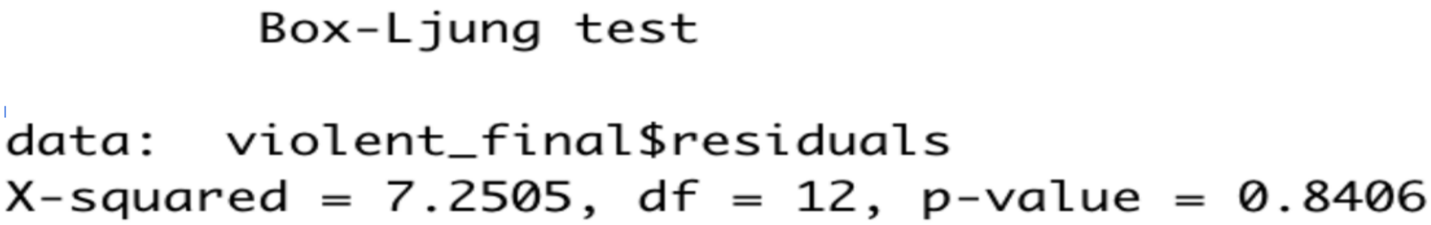
```
Call:
lm(formula = violent_ts ~ time(violent_ts) + I(time(violent_ts)^2))

Residuals:
    Min      1Q  Median      3Q     Max 
-78.497 -20.936  -1.434  19.184 103.475 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            1.120e+07  1.546e+06   7.247 4.94e-11 ***
time(violent_ts)      -1.111e+04  1.533e+03  -7.245 4.99e-11 ***
I(time(violent_ts)^2)  2.755e+00  3.803e-01   7.243 5.04e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.05 on 117 degrees of freedom
Multiple R-squared:  0.3435,    Adjusted R-squared:  0.3323 
F-statistic: 30.61 on 2 and 117 DF,  p-value: 2.033e-11
```

*Figure 5. Summary of fitted trend model*



```
        Box-Ljung test

data:  violent_final$residuals
X-squared = 7.2505, df = 12, p-value = 0.8406
```

*Figure 6: Box-Ljung Test*

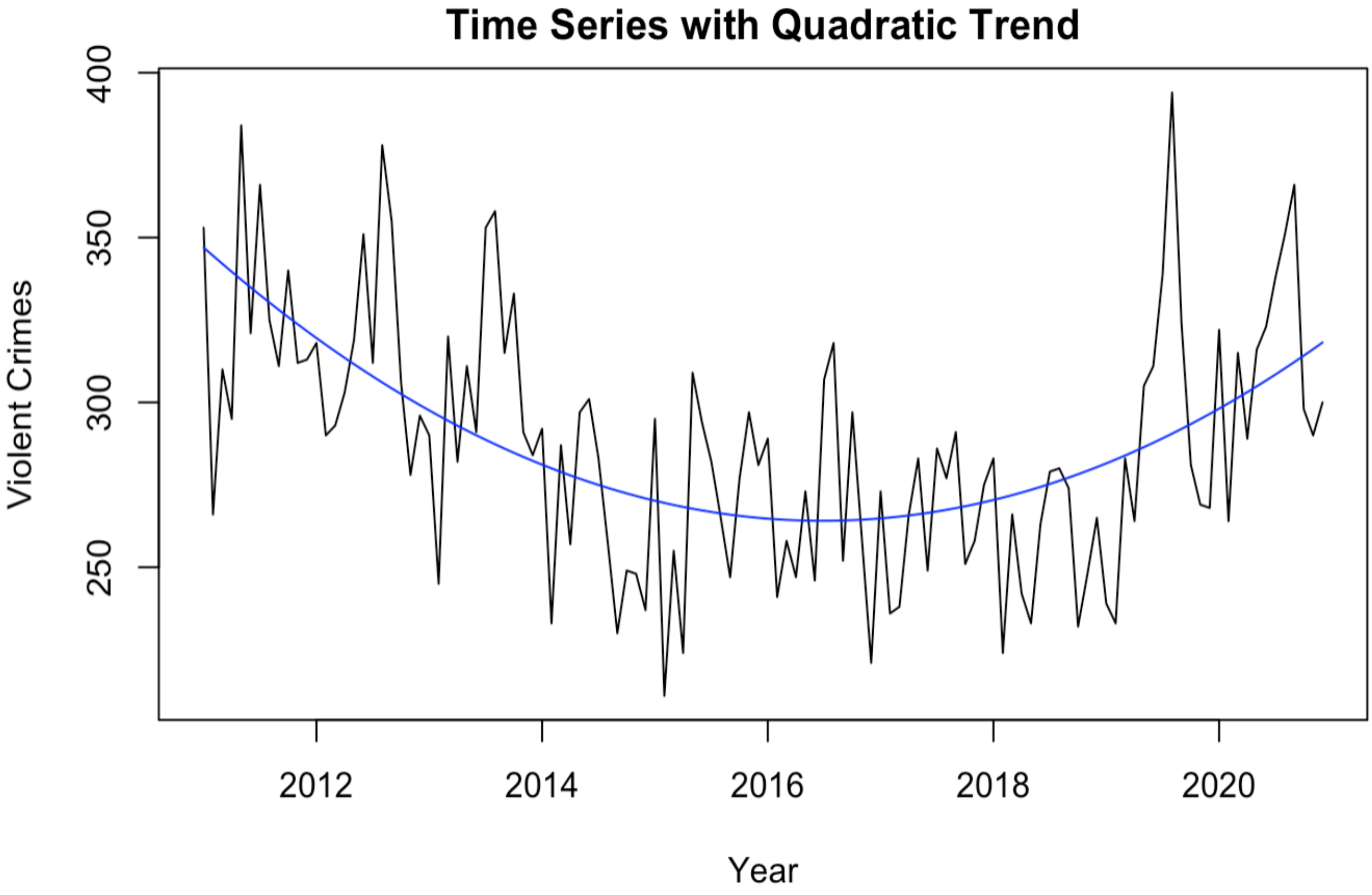


*Figure 7. The plot of time series with a quadratic trend*

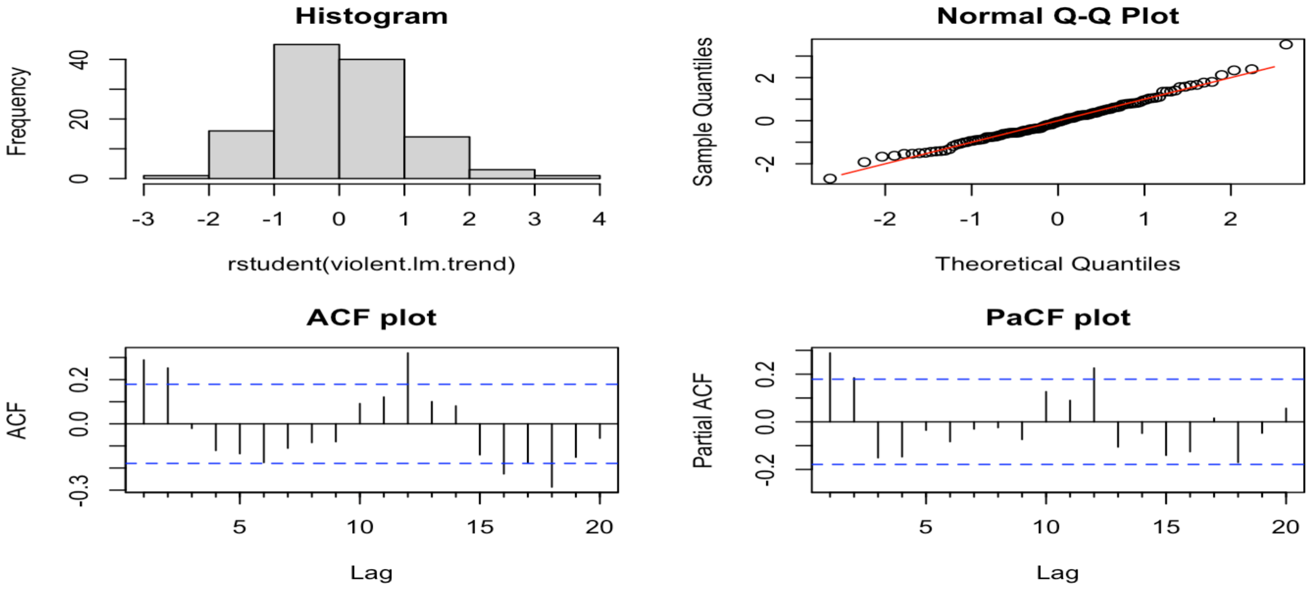
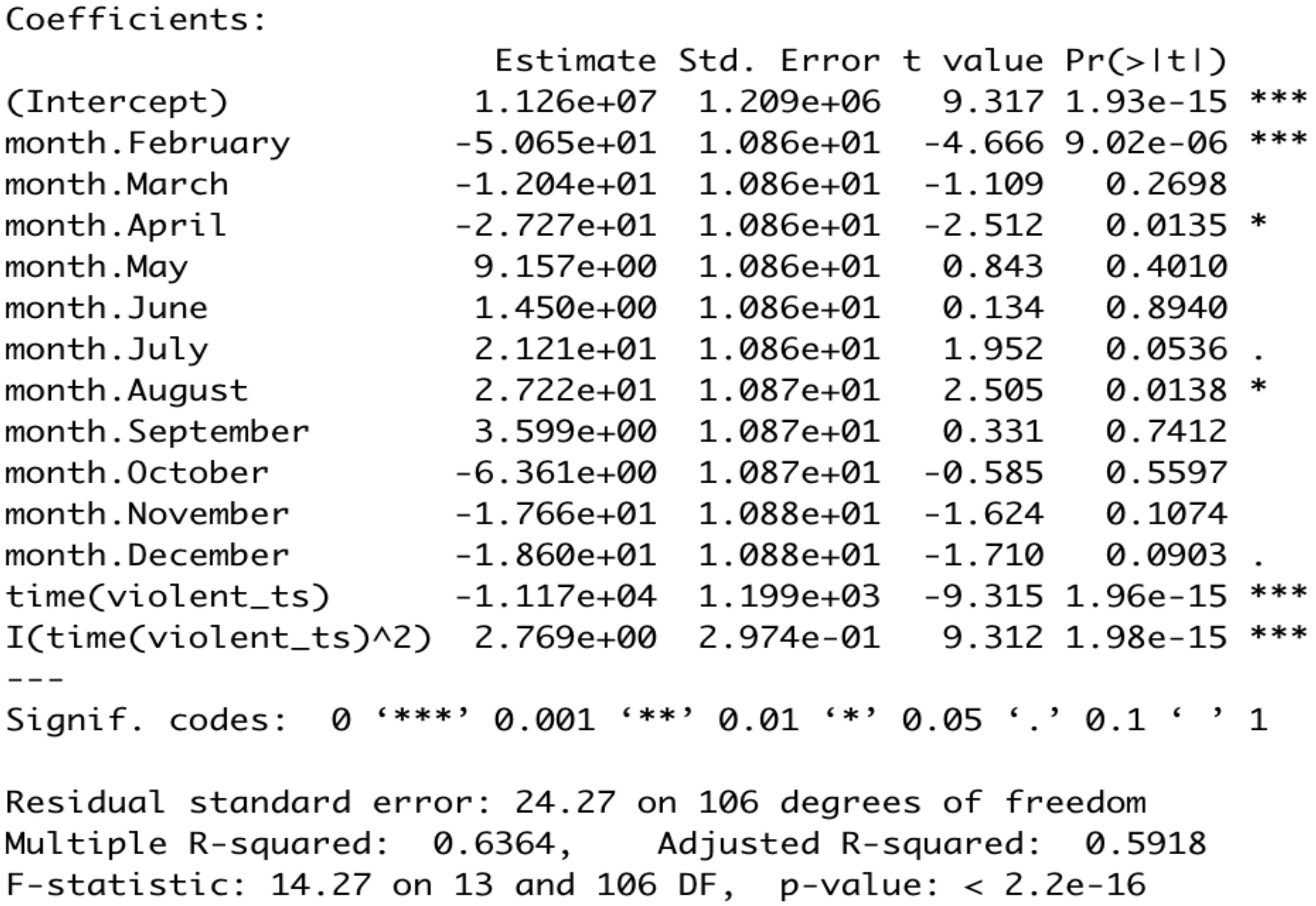


*Figure 8. Residual analysis with quadratic trend model*



```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.126e+07  1.209e+06   9.317 1.93e-15 ***
month.February       -5.065e+01  1.086e+01  -4.666 9.02e-06 ***
month.March          -1.204e+01  1.086e+01  -1.109   0.2698    
month.April          -2.727e+01  1.086e+01  -2.512   0.0135 *  
month.May             9.157e+00  1.086e+01   0.843   0.4010    
month.June            1.450e+00  1.086e+01   0.134   0.8940    
month.July            2.121e+01  1.086e+01   1.952   0.0536 .  
month.August          2.722e+01  1.087e+01   2.505   0.0138 *  
month.September       3.599e+00  1.087e+01   0.331   0.7412    
month.October        -6.361e+00  1.087e+01  -0.585   0.5597    
month.November       -1.766e+01  1.088e+01  -1.624   0.1074    
month.December       -1.860e+01  1.088e+01  -1.710   0.0903 .  
time(violent_ts)     -1.117e+04  1.199e+03  -9.315 1.96e-15 ***
I(time(violent_ts)^2) 2.769e+00  2.974e-01   9.312 1.98e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.27 on 106 degrees of freedom
Multiple R-squared:  0.6364,	Adjusted R-squared:  0.5918 
F-statistic: 14.27 on 13 and 106 DF,  p-value: < 2.2e-16
```

*Figure 9. Fitting adding seasonal component in trend model*

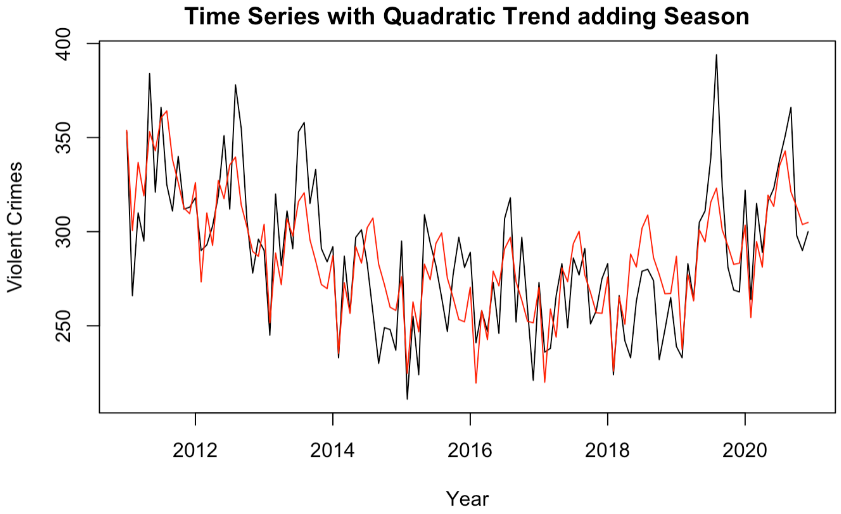


*Figure 10. The plot of the times series with Quadratic Trend adds seasonality*

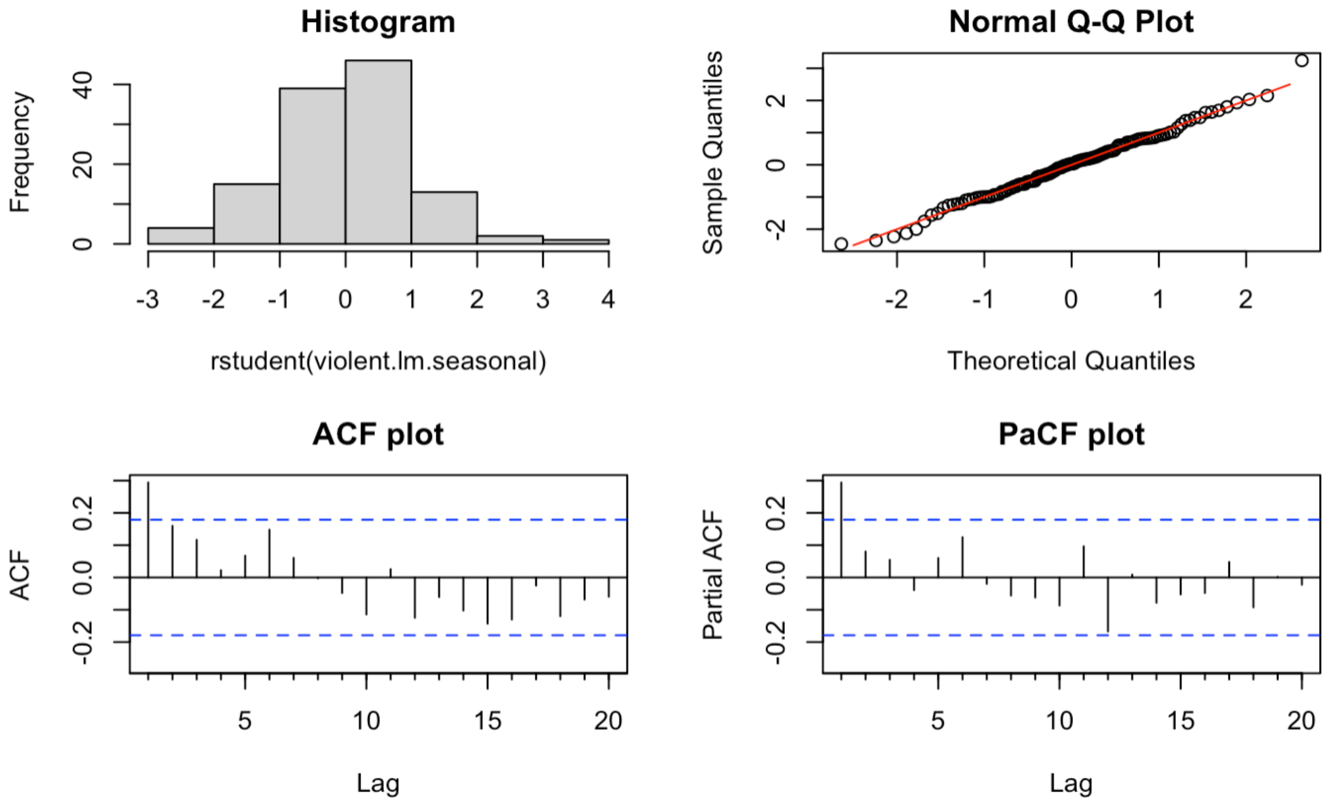
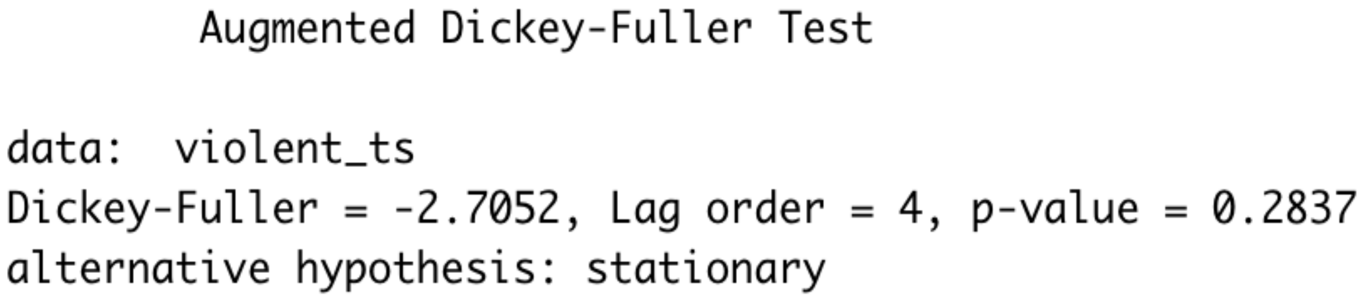


*Figure 11: Checking residuals of seasonal model*

Checking Stationarity and Model Selection:



```
        Augmented Dickey-Fuller Test

data:  violent_ts
Dickey-Fuller = -2.7052, Lag order = 4, p-value = 0.2837
alternative hypothesis: stationary
```

*Figure 12: Augmented Dickey-Fuller Test*

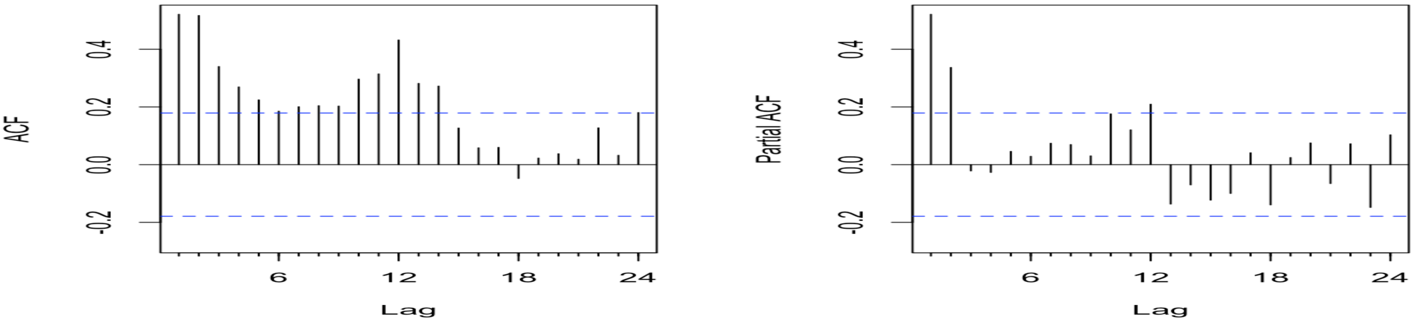


*Figure 13: ACF and PACF Plot of time series*

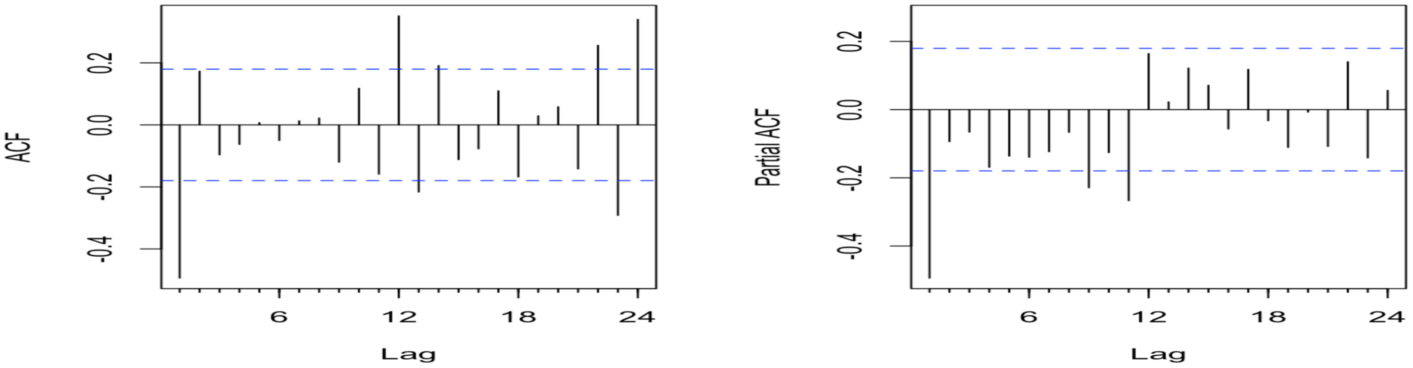
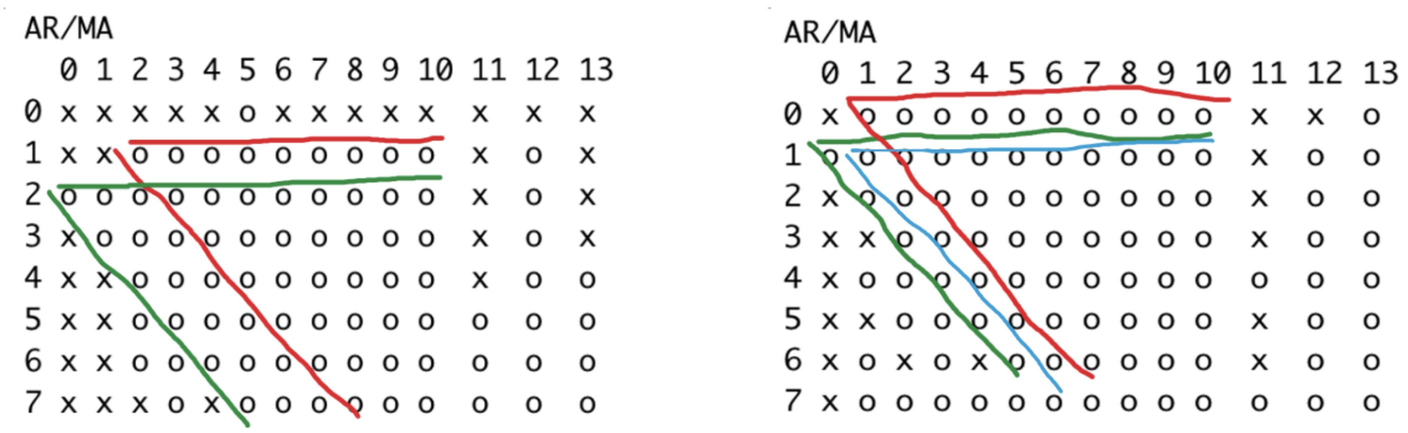


*Figure 14: ACF and PACF Plot of times series with differencing*

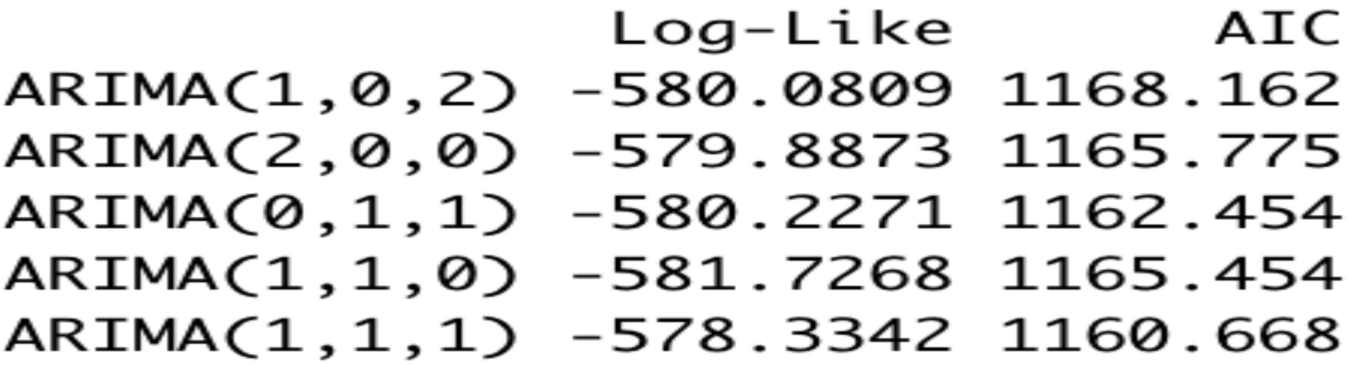


```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x o x x x x x  x  x  x
1 x x o o o o o o o o o  x  o  x
2 o o o o o o o o o o o  x  o  x
3 x o o o o o o o o o o  x  o  x
4 x x o o o o o o o o o  x  o  o
5 x x o o o o o o o o o  o  o  o
6 x x o o o o o o o o o  o  o  o
7 x x x o x o o o o o o  o  o  o
```

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o o o  x  x  o
1 o o o o o o o o o o o  x  o  o
2 x o o o o o o o o o o  x  o  o
3 x x o o o o o o o o o  x  o  o
4 x o o o o o o o o o o  o  o  o
5 x x o o o o o o o o o  x  o  o
6 x o x o x o o o o o o  x  o  o
7 x o o o o o o o o o o  o  o  o
```

*Figure 15: EACF plot Left plot: NO differencing, Right plot: With differencing*

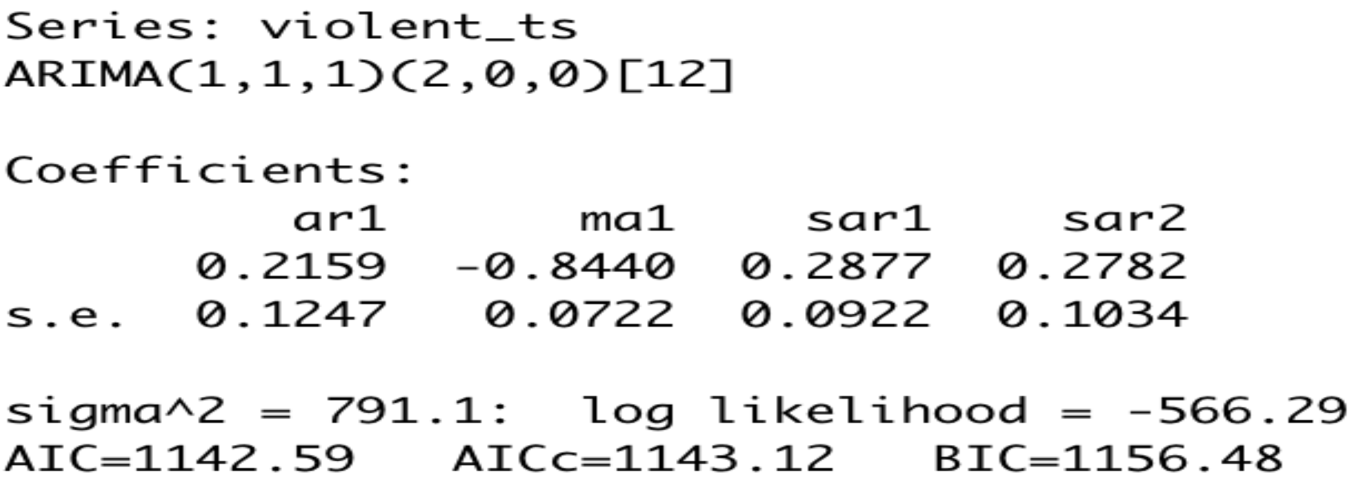
Model Selection and Choosing Parameters:



```
               Log-Like        AIC
ARIMA(1,0,2) -580.0809 1168.162
ARIMA(2,0,0) -579.8873 1165.775
ARIMA(0,1,1) -580.2271 1162.454
ARIMA(1,1,0) -581.7268 1165.454
ARIMA(1,1,1) -578.3342 1160.668
```

Figure 16: Tried fitting with ARIMA model based on EACF

Analysis for Final Model Selection:



```
Series: violent_ts
ARIMA(1,1,1)(2,0,0)[12]

Coefficients:
         ar1      ma1    sar1    sar2
      0.2159  -0.8440  0.2877  0.2782
s.e.  0.1247   0.0722  0.0922  0.1034

sigma^2 = 791.1:  log likelihood = -566.29
AIC=1142.59   AICc=1143.12   BIC=1156.48
```

*Figure 17: ARIMA Diagnostic*

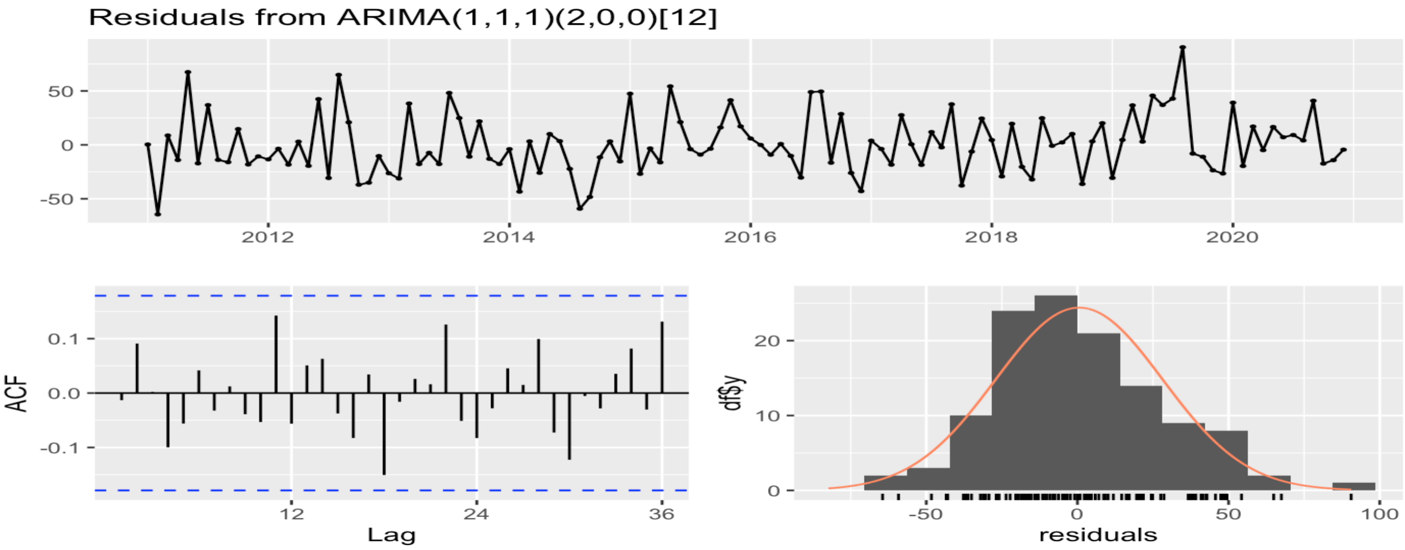
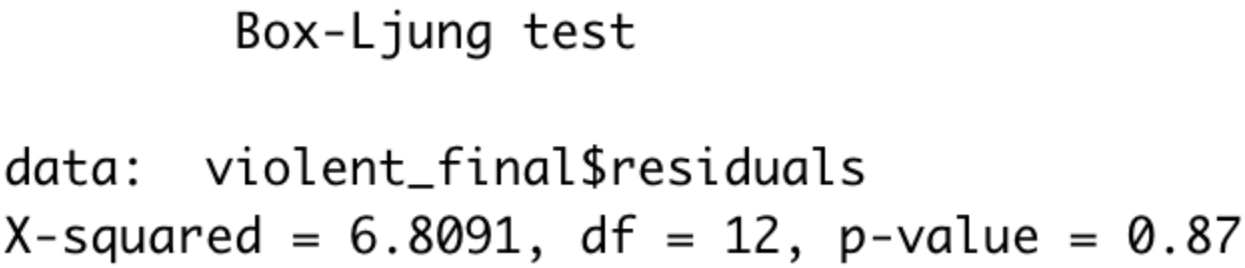


*Figure 18: Checking Residuals - White Noise Assumption Testing*



```
           Box-Ljung test

data:  violent_final$residuals
X-squared = 6.8091, df = 12, p-value = 0.87
```

*Figure 19: Residual Symmetry Distribution (ARIMA Seasonality)*

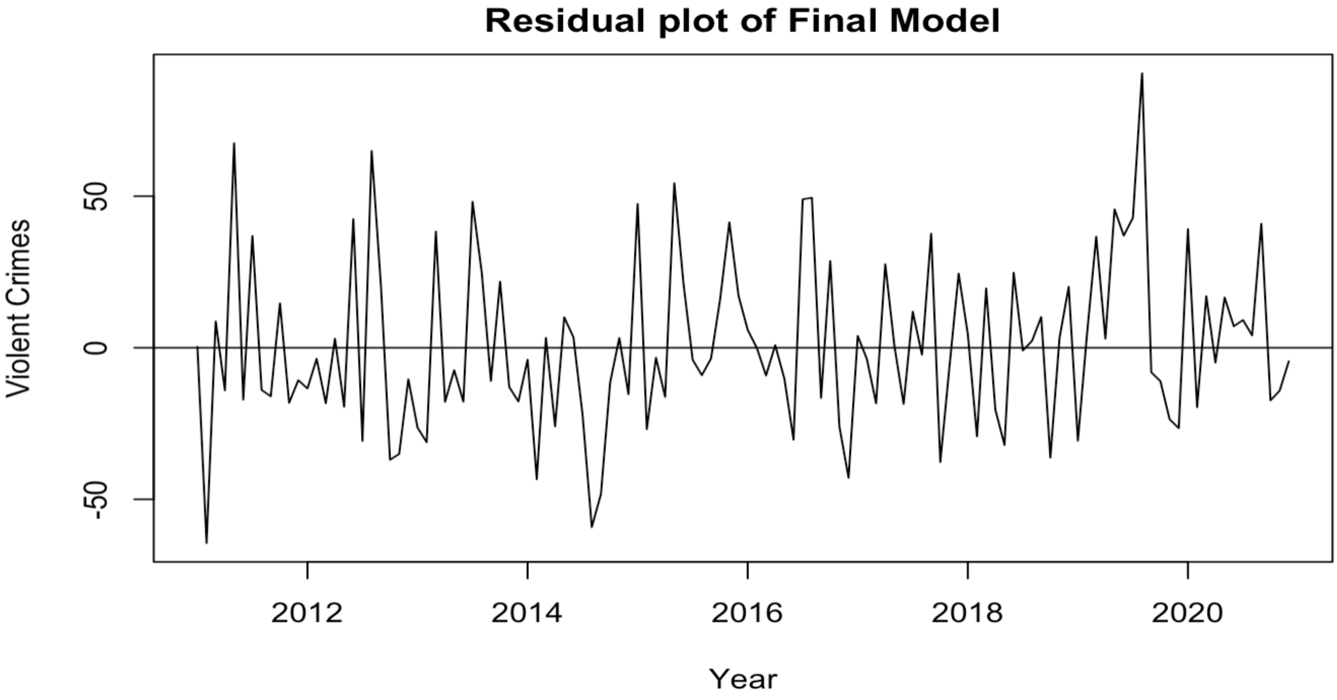
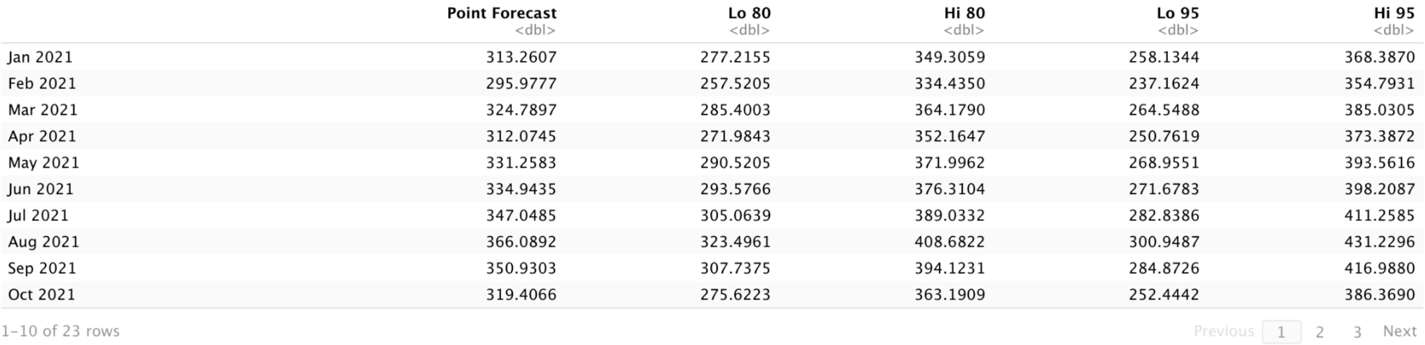


*Figure 20: Checking Mean of Residuals of the final model*



| | Point Forecast | Lo 80 | Hi 80 | Lo 95 | Hi 95 |
|---|---|---|---|---|---|
| | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| Jan 2021 | 313.2607 | 277.2155 | 349.3059 | 258.1344 | 368.3870 |
| Feb 2021 | 295.9777 | 257.5205 | 334.4350 | 237.1624 | 354.7931 |
| Mar 2021 | 324.7897 | 285.4003 | 364.1790 | 264.5488 | 385.0305 |
| Apr 2021 | 312.0745 | 271.9843 | 352.1647 | 250.7619 | 373.3872 |
| May 2021 | 331.2583 | 290.5205 | 371.9962 | 268.9551 | 393.5616 |
| Jun 2021 | 334.9435 | 293.5766 | 376.3104 | 271.6783 | 398.2087 |
| Jul 2021 | 347.0485 | 305.0639 | 389.0332 | 282.8386 | 411.2585 |
| Aug 2021 | 366.0892 | 323.4961 | 408.6822 | 300.9487 | 431.2296 |
| Sep 2021 | 350.9303 | 307.7375 | 394.1231 | 284.8726 | 416.9880 |
| Oct 2021 | 319.4066 | 275.6223 | 363.1909 | 252.4442 | 386.3690 |

1–10 of 23 rows

Previous 1 2 3 Next

*Figure 21: Forecast*

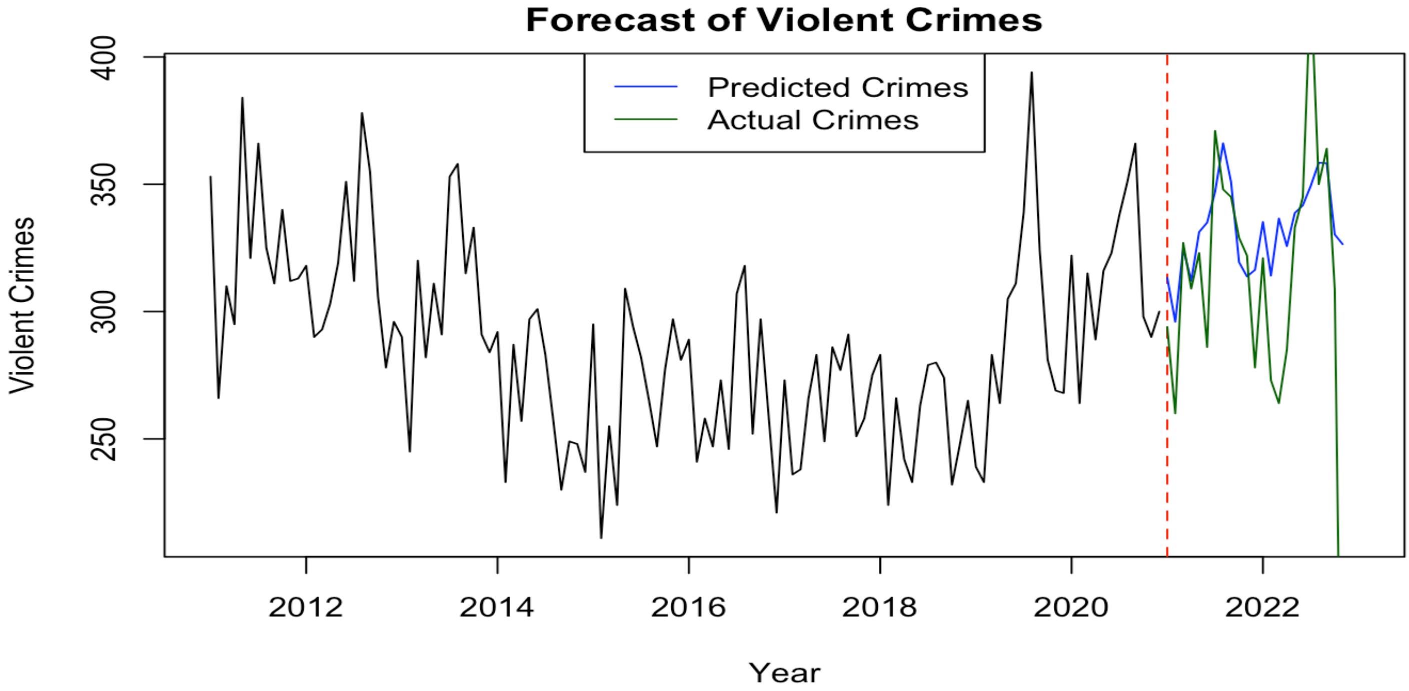
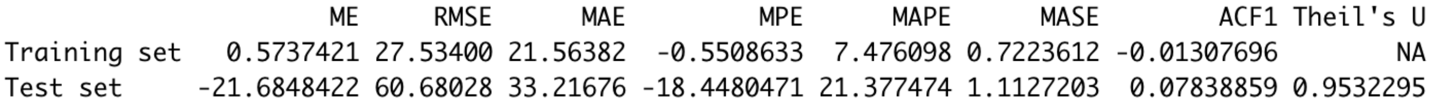
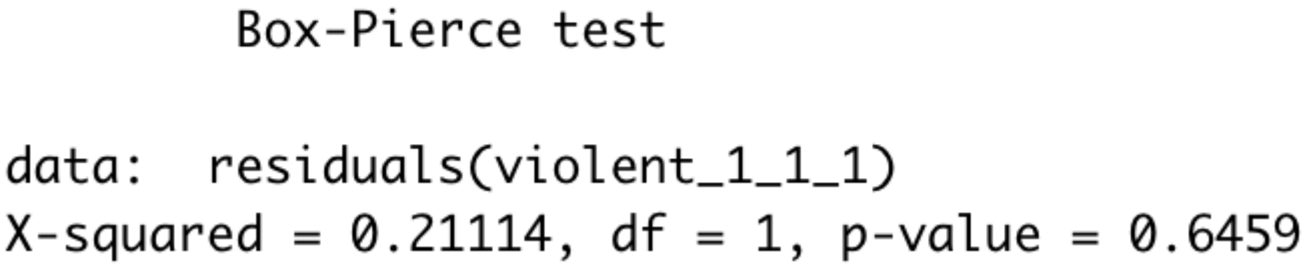


*Figure 22: Plot of forecast*



```
                      ME     RMSE      MAE         MPE      MAPE      MASE        ACF1 Theil's U
Training set   0.5737421 27.53400 21.56382  -0.5508633  7.476098 0.7223612 -0.01307696        NA
Test set     -21.6848422 60.68028 33.21676 -18.4480471 21.377474 1.1127203  0.07838859 0.9532295
```

*Figure 23: Prediction Accuracy Testing*



```
	Box-Pierce test

data:  residuals(violent_1_1_1)
X-squared = 0.21114, df = 1, p-value = 0.6459
```

Figure 24. Box test