Time Series Analysis:

Predicting Crime Incident Counts in Boston

MA 585

Wanchen Hong

Dataset: Crime Incident Reports, Boston

https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system

This project is to explore the crime incident counts in Boston and model it in time, with an emphasis on the total crime incident counts and the assault crime counts as an example of violent crime. I developed a seasonal ARIMA and a seasonal Holt-Winters model, with the seasonal Holt-Winters model having a better performance in forecast validation. From the dataset, the total crime counts and assault crime counts have been stable from 2015 to 2019 with a clear seasonal pattern but have a significant drop from the beginning of 2020. The Holt-Winters model forecasts that the total crime count would slowly go up while maintaining a similar level as recent years, while the assault crime count could slowly recover to pre-COVID levels and increase rather quickly.

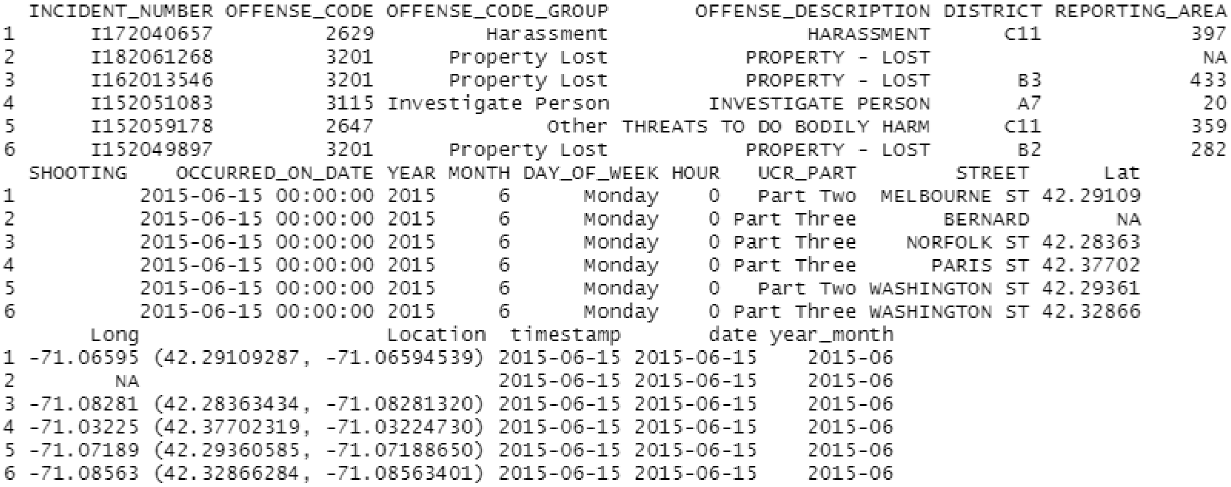
1. Introduction

Boston is one of the major cities in the United States. Many people have settled in this city as their permanent home city. An important factor people consider when choosing a place to live is its safety, therefore, I decided to study how the total number of crimes has evolved during the past few years, using a public crime dataset provided by the Boston government [1], starting in June 2015 and ending in March 2023. I also paid attention to a specific type of crime, assault, to see if there is any difference in time between this highly dangerous type of crime and crime in general. I built two models, seasonal ARIMA and seasonal Holt-Winters, to forecast the monthly crime numbers in 2022 and 2023 using data prior, since 2015.

2. Data Exploration and Preprocessing

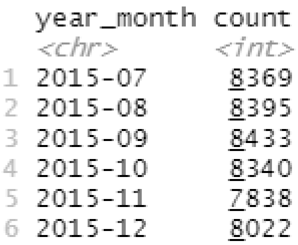
The raw data from the Boston government [1] contains each entry a unique crime incident report of all of Boston, from June 15th, 2015 to March 30th, 2023. This dataset is cleansed and formatted as shown in Appendix 1. Figure 1 below shows a sample of raw and cleansed data.

Figure 1 a). Raw data of crime reports from boston.gov



```
  INCIDENT_NUMBER OFFENSE_CODE OFFENSE_CODE_GROUP       OFFENSE_DESCRIPTION DISTRICT REPORTING_AREA
1      I172040657         2629         Harassment                HARASSMENT      C11            397
2      I182061268         3201      Property Lost           PROPERTY - LOST                      NA
3      I162013546         3201      Property Lost           PROPERTY - LOST       B3            433
4      I152051083         3115 Investigate Person        INVESTIGATE PERSON       A7             20
5      I152059178         2647              Other THREATS TO DO BODILY HARM      C11            359
6      I152049897         3201      Property Lost           PROPERTY - LOST       B2            282
  SHOOTING    OCCURRED_ON_DATE YEAR MONTH DAY_OF_WEEK HOUR   UCR_PART        STREET      Lat
1          2015-06-15 00:00:00 2015     6      Monday    0   Part Two  MELBOURNE ST 42.29109
2          2015-06-15 00:00:00 2015     6      Monday    0 Part Three       BERNARD       NA
3          2015-06-15 00:00:00 2015     6      Monday    0 Part Three    NORFOLK ST 42.28363
4          2015-06-15 00:00:00 2015     6      Monday    0 Part Three      PARIS ST 42.37702
5          2015-06-15 00:00:00 2015     6      Monday    0   Part Two WASHINGTON ST 42.29361
6          2015-06-15 00:00:00 2015     6      Monday    0 Part Three WASHINGTON ST 42.32866
      Long                    Location  timestamp       date year_month
1 -71.06595 (42.29109287, -71.06594539) 2015-06-15 2015-06-15    2015-06
2        NA                             2015-06-15 2015-06-15    2015-06
3 -71.08281 (42.28363434, -71.08281320) 2015-06-15 2015-06-15    2015-06
4 -71.03225 (42.37702319, -71.03224730) 2015-06-15 2015-06-15    2015-06
5 -71.07189 (42.29360585, -71.07188650) 2015-06-15 2015-06-15    2015-06
6 -71.08563 (42.32866284, -71.08563401) 2015-06-15 2015-06-15    2015-06
```

Figure 1 b). Cleaned data in the format of time series, monthly crime counts



```
  year_month count
  <chr>      <int>
1 2015-07     8369
2 2015-08     8395
3 2015-09     8433
4 2015-10     8340
5 2015-11     7838
6 2015-12     8022
```

3. Model Selection

3.1 Data Visualization and Transformation

The plot of the time (Fig. 2) series data does not look quite stationary, with a quite significant drop near the beginning of 2020. It may also include a change in variance due to this drop. However, typical variance stabilization method, such as a logarithm/square root transform as well as Box-Cox transform does not work as shown in Appendix 2. It seems the relative variance in a year does not change since 2020, even though the number dropped. Ways to deal with outliers in time series would be required to solve this problem.

A classical decomposition plot (Fig. 3) and Dickery-Fuller test (Appendix 3a) were used to test the seasonal and trend component of this dataset. The decomposition shows a clear seasonal component with a period of 12 and a trend component, and the Dickery-Fuller test has a result of 0.05387, which fails to reject the null hypothesis that the data is stationary.

Figure 2. Time Series Plot of Total Crime Count

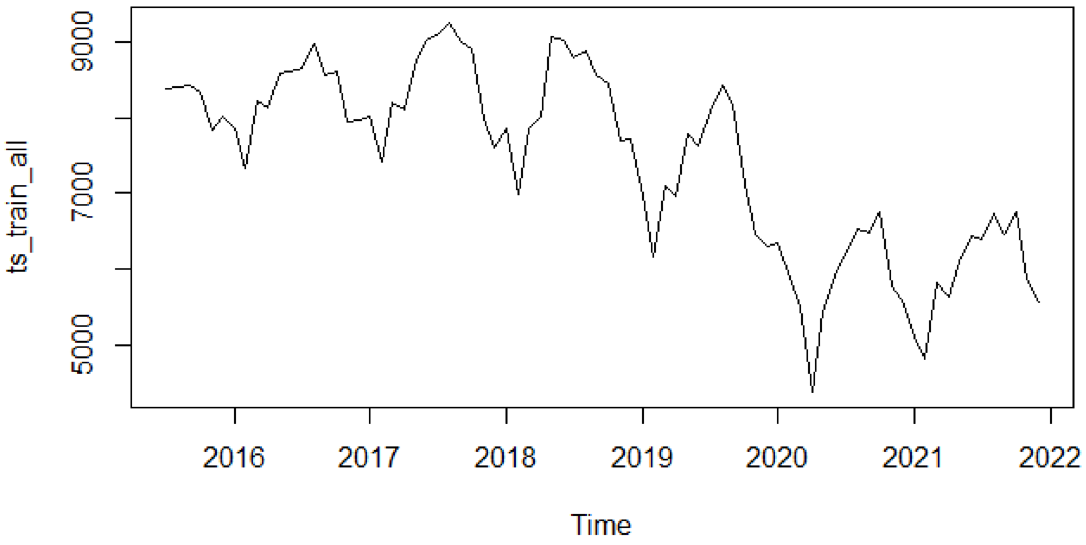
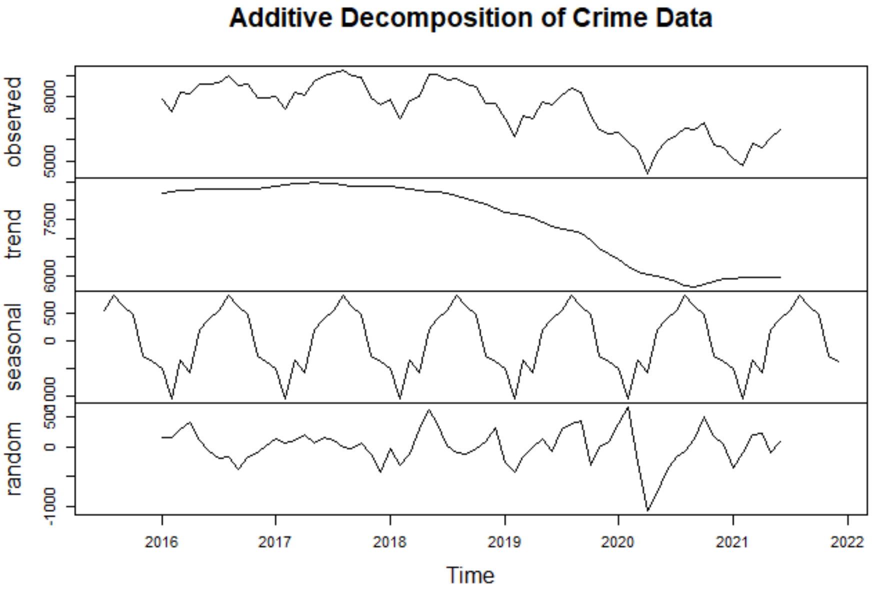


Figure 3. Classical Decomposition Plot of Total Crime Count



The same procedure was used with assault crime counts as well. The result for the assault count (Fig. 4) shows a similar pattern as the total crime count with a significant drop in 2020, but the variance still seems constant within the two separate parts, and different transformation of the data shows the same general result (Appendix 4). The classical decomposition plot (Fig. 5) and Dickery-Fuller test (Appendix 3b) also show a similar seasonal and trend component as the total count with a seasonal period of 12. The Dickery-Fuller test has a result of 0.3831, which fails to reject the null hypothesis that the data is stationary.

Figure 4. Time Series Plot of Assault Crime Count

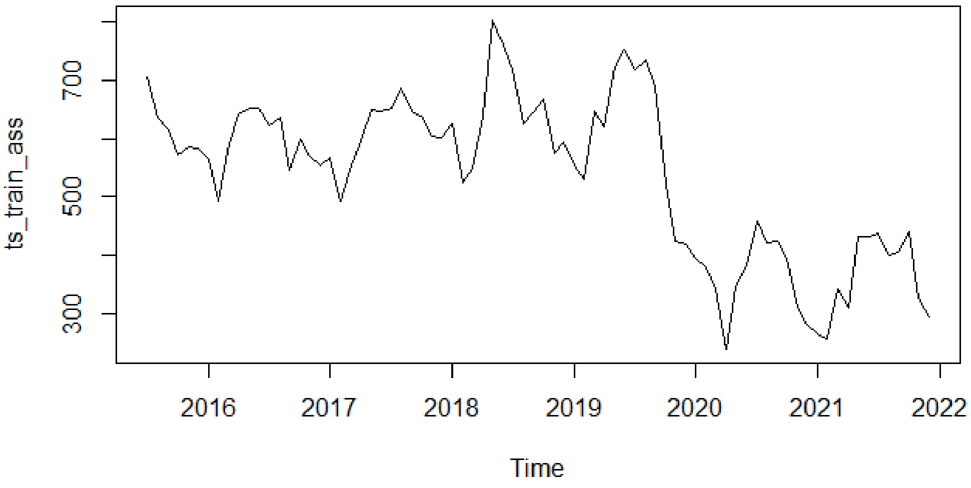
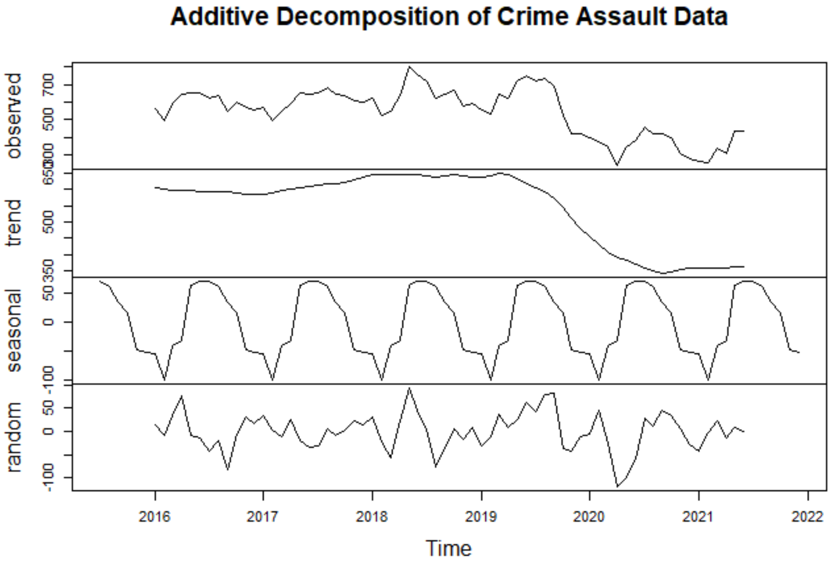


Figure 5. Classical Decomposition Plot of Assault Crime Count



In order to correct for seasonality and trend in both sets of data, I applied a first-order difference in both training datasets. In both sets of data, the differenced data result in stationarity (Fig. 6 a, b) with the Dickey-Fuller test result to be less than 0.01 for the Total Count and 0.02 for the assault count (Appendix 5). Therefore, variance stabilization can be achieved via first-order differencing.

Figure 6 a. Differenced Time Series Plot of Total Crime Count

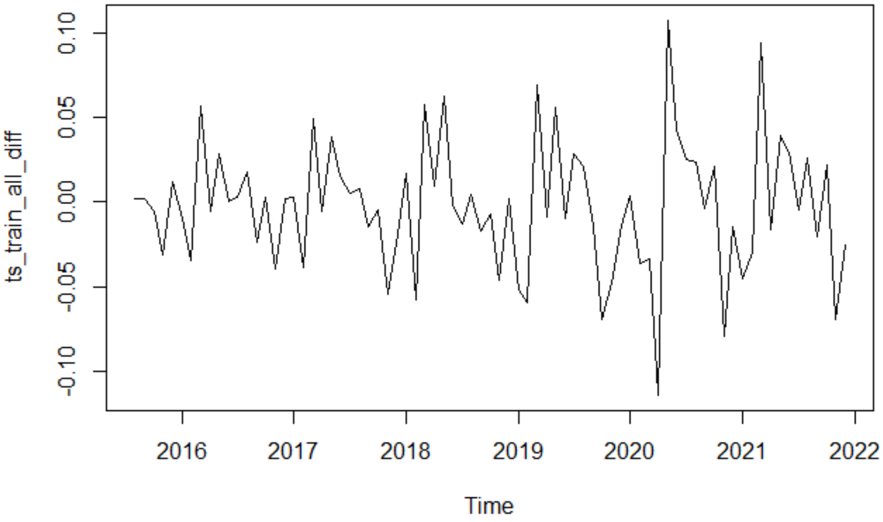
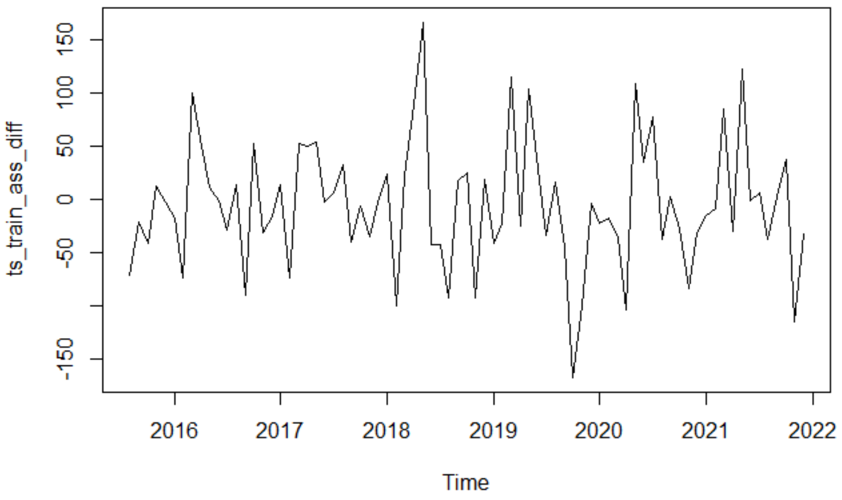


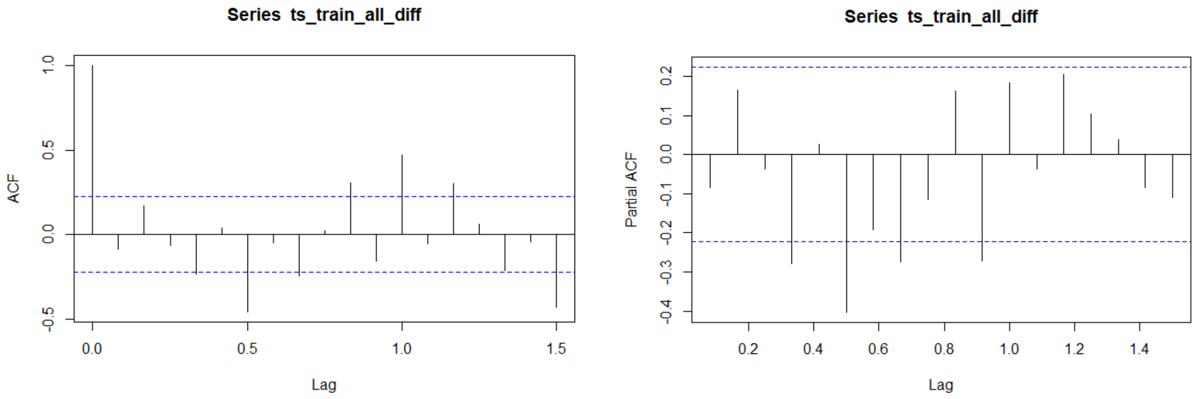
Figure 6 b. Differenced Time Series Plot of Assualt Crime Count



3.2 Model-Based Forecast: SARIMA

To forecast the crime count of the time series, I first attempted a seasonal ARIMA model by fitting a best-fit model to forecast. First, for the total crime count, its ACF and PACF are presented in Fig. 7. From the plots, we can see the ACF cuts off after lag 1, but both the ACF and PACF seem to be a sinusoid decay. Note that the period of decay is exactly 1 year in the ACF plot. Therefore, a low-order ARMA model, with a first-order difference, and a seasonal period of 12 would be reasonable. This coincides with the seasonality observed in the original data.

Figure 7. ACF and PACF of the Differenced Total Crime Count



Since there are many possibilities for the parameters, I tried different combinations of small p, q, P and Q. in the SARIMA (p,d,q) x (P,D,Q) model. This way the parameters would coincide with the observation that it is a low-order ARMA model, and the best-fit model would be discovered readily. The AICc models of all the models are in Table 1. below. We can see that the model with the lowest AICc is SARIMA (2,1,0) x $(0,1,1)_{12}$, which is the model I'll be using to forecast the total crime count.

Table 1. Combinations of SARIMA Models and Their AICc Values

Model	AICc
(2,1,0) x $(0,1,1)_{12}$	957.7877 **
(0,1,0) x $(0,1,1)_{12}$	958.1311
(1,1,0) x $(0,1,1)_{12}$	958.3536
(2,1,1) x $(0,1,1)_{12}$	958.6999
(1,1,1) x $(0,1,1)_{12}$	960.1067
(2,1,2) x (0,1,1)	960.4063
(2,1,1) x $(1,1,1)_{12}$	961.0901
(1,1,2) x $(1,1,1)_{12}$	962.0241
(1,1,1) x $(1,1,1)_{12}$	962.2918
(1,1,2) x $(0,1,1)_{12}$	962.59
(2,1,2) x $(1,1,1)_{12}$	966.0201
(2,1,1) x $(1,1,0)_{12}$	969.2355
(1,1,2) x $(1,1,0)_{12}$	970.4235
(2,1,2) x (1,1,0)	971.0199
(1,1,1) x $(1,1,0)_{12}$	971.737

A similar analysis was done to select the best-fit SARIMA model for forecasting the assault crime count. The combinations of SARMIA models attempted for it is in Table 2 below. The model of SARIMA (2,1,1) x $(0,1,1)_{12}$ again has the lowest AICc and will be selected to forecast the assault crime count.

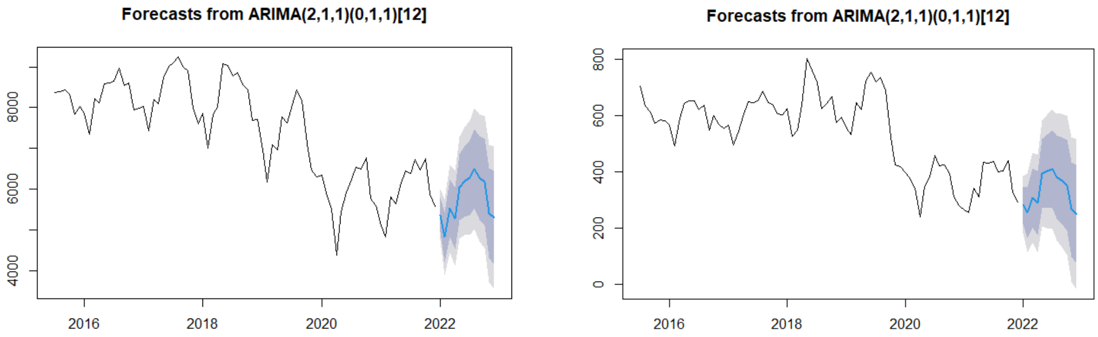
Table 1. Combinations of SARIMA Models and Their AICc Values

Model	AICc
(2,1,0) x $(0,1,1)_{12}$	713.9072 **
(0,1,0) x $(0,1,1)_{12}$	714.1849
(1,1,0) x $(0,1,1)_{12}$	714.384

Model	AICc
(2,1,1) x (0,1,1)$_{12}$	715.2929 **
(1,1,1) x (0,1,1)$_{12}$	715.5987
(1,1,2) x (0,1,1)$_{12}$	715.6353
(1,1,1) x (1,1,1)$_{12}$	717.4908
(2,1,1) x (1,1,1)$_{12}$	717.5796
(1,1,2) x (1,1,1)$_{12}$	717.9604
(2,1,2) x (0,1,1)	718.0237
(2,1,2) x (1,1,1)$_{12}$	720.45
(1,1,1) x (1,1,0)$_{12}$	721.0079
(2,1,1) x (1,1,0)$_{12}$	722.0005
(1,1,2) x (1,1,0)$_{12}$	722.0299
(2,1,2) x (1,1,0)	724.0775

The forecast results are in Figure. 8 below.

Figure 8. Forecast Results using the SARIMA Model
Left: forecast result for total crime count
Right: forecast result for assault crime count



Forecasts from ARIMA(2,1,1)(0,1,1)[12]

Forecasts from ARIMA(2,1,1)(0,1,1)[12]

3.3 Model Diagnostic

The selected models have the lowest AICc values. However, the diagnostic plots in Fig. 9 and Fig. 10 show that the residuals are not quite normally distributed, with significant outliers. This result coincides with the discussion in 3.1 that further techniques to eliminate outliers need to be applied in order to solve this problem. Overall, aside from the few months of data in 2020, the model fits quite well, and the forecast does seem to follow the general pattern in previous years.

One significant issue of the models is that most of the parameters in the model are not significantly different from 0. For the total crime count, only the parameter for SMA(1) is significant. However, removing AR(2) or both AR(2) and AR(1) parameters model did not improve the AICc value (Table 1). Similarly for the assault crime count, again only the parameter for SMA(1) is significant, and removing AR(2) or both AR(2) and AR(1) parameters model did not improve the AICc value (Table 1).

Therefore, SARIMA models might not be the best models for this dataset due to the possible significant outliers in the original data near 2020.

Figure 9. Diagnostic Plots For Total Crime Count

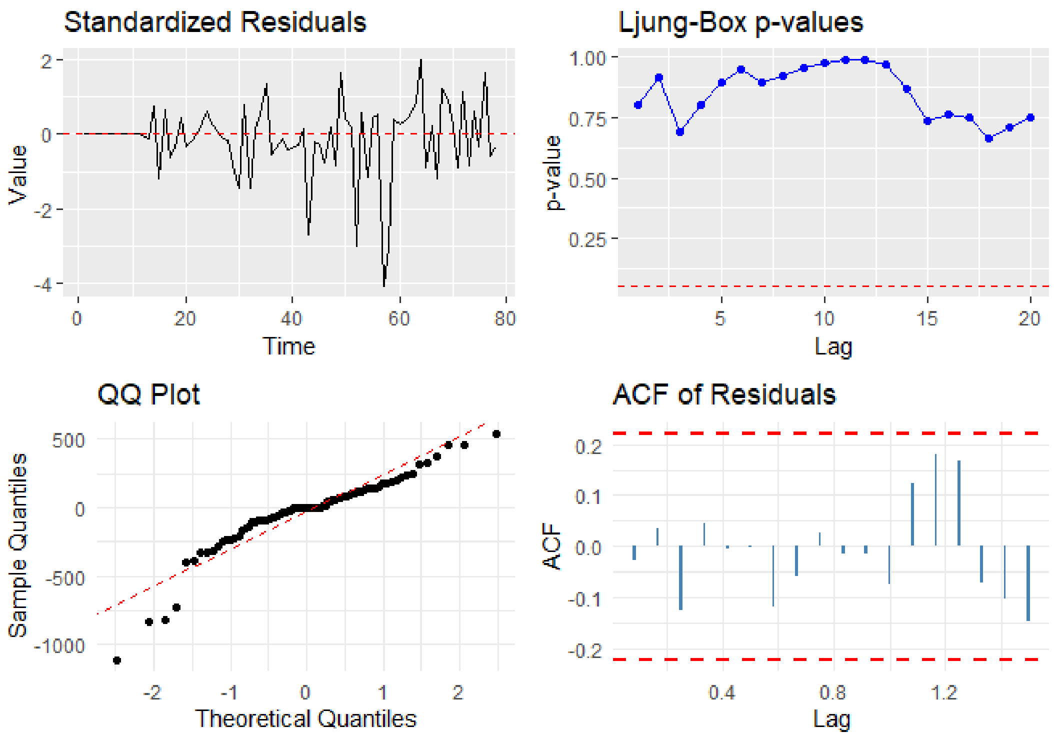
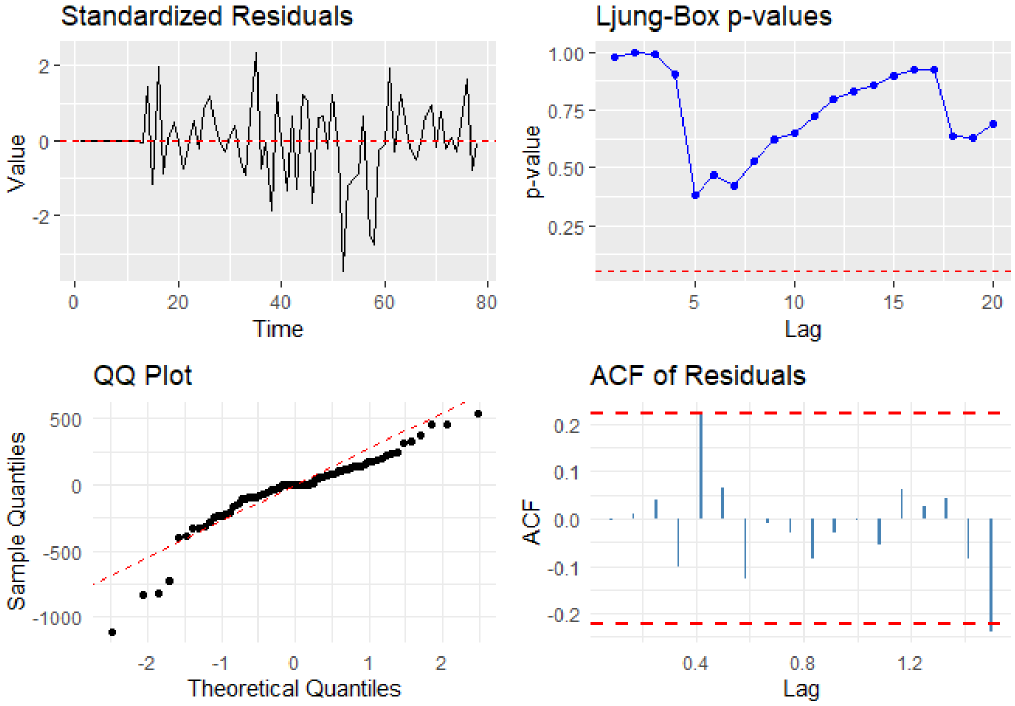


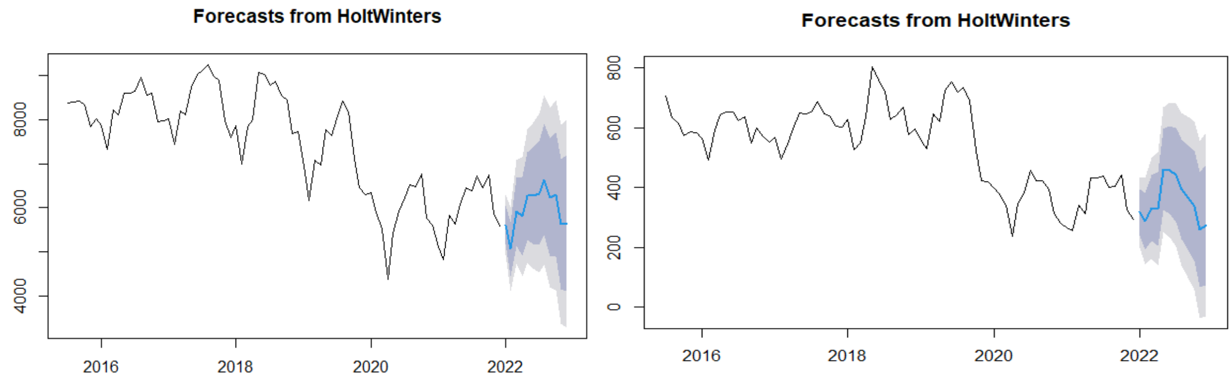
Figure 10. Diagnostic Plots For Assault Crime Count



3.4 Smoothing-based Forecast: Holt-Winters

Since the model-based forecasting using the SARIMA model has clear drawbacks, I attempted to use the smoothing-based Holt-Winters model with lag 12 to forecast the crime count, in hopes that exponential smoothing would capture the pattern more accurately and disregard the possible outliers. The additive version of Holt-Winters was used since the variance seems constant in the entire data, in both sets of data. The forecast results in both sets of data are reasonable (Fig. 11).

Figure 11. Holt-Winters Forecast Results
Left: Total Crime Count; Right: Assault Crime Count



3.5 Model Evaluation

Despite the clear drawbacks of the SARIMA model, the overall shapes of forecast in both models in both sets of data are similar. In order to evaluate the performance of the two models in both sets of data accurately, the models are validated using the crime data from 2022 to 2023 to assess the seasonal estimation. About 15 months of data compared to the 78 months of data used in the training set, which is approximately 16% of the entire dataset. Root mean square error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were used as evaluation metrics. The results below (Table 3, 4) show that the Holt-Winters model provides a much better forecast than the SARIMA model.

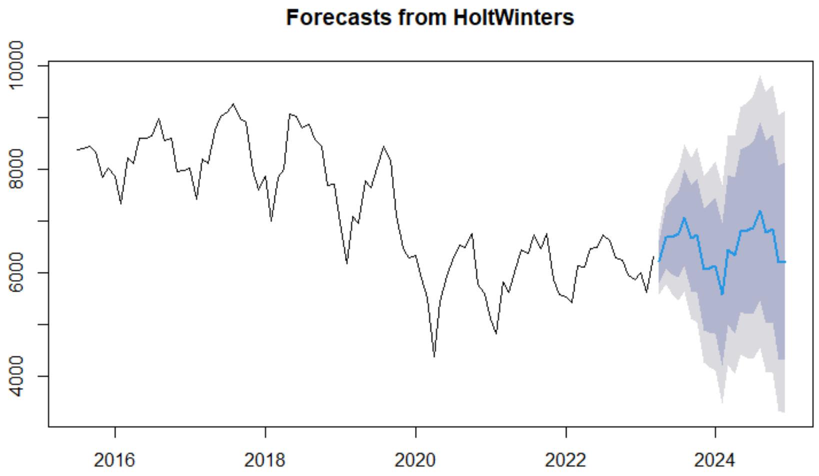
	Holt-Winters	SARIMA(2,1,1) x $(0,1,1)_{12}$
RMSE	238.0239	460.0976
MAE	206.8156	389.8099
MAPE	3.547897	7.074669

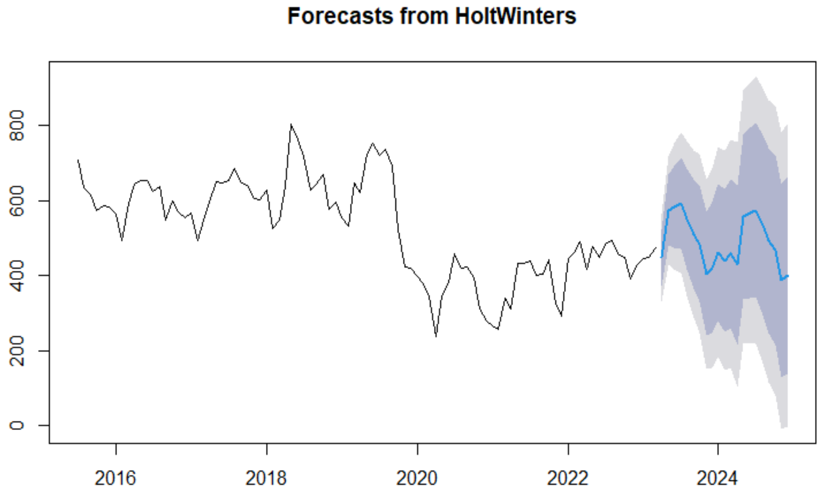
	Holt-Winters	SARIMA(2,1,1) x $(0,1,1)_{12}$
RMSE	45.69045	63.76976
MAE	37.78368	53.0197
MAPE	11.86339	18.05076

3.6 Future Forecast

From the model evaluation above, the Holt-Winters model is more accurate for the chosen dataset. Therefore, I used the Holt-Winters model to forecast the total crime count as well as the assault crime count for the rest of 2023 and 2024 (Fig. 12). From the recent trend, the model suggests that the total crime count would slowly go up while maintaining a similar level as recent years, while the assault crime count could slowly recover to pre-COVID levels and increase rather quickly.

Figure 12. Holt-Winters Future Forecast Results
Top: Total Crime Count Future Forecast
Bottom: Assault Crime Count Future Forecast





4. Conclusion

This time series analysis is to assess the possible change in crime count in time. I used a seasonal ARIMA and a seasonal Holt-Winters model to forecast the total crime count and assault crime count from Jan 2022 to Mar 2023 (Appendix 7). The best-fit SARIMA model was selected using AICc and compared to the Holt-Winters model using RMSE, MAE, and MAPE. The result shows that the Holt-Winters model has a much better performance for this particular dataset. The future forecast using the Holt-Winters model shows that the total crime numbers would likely stay similar to past years, while the total number of assaults, a particularly violent crime, would increase rather more quickly and may recover to pre-COVID level soon. This model can be used to consider datasets in other cities, at different times, of different crime categories, to consider the general safety of the area. Future analysis is needed to handle the significant outliers in 2020 due to COVID to improve the accuracy of the models.

References

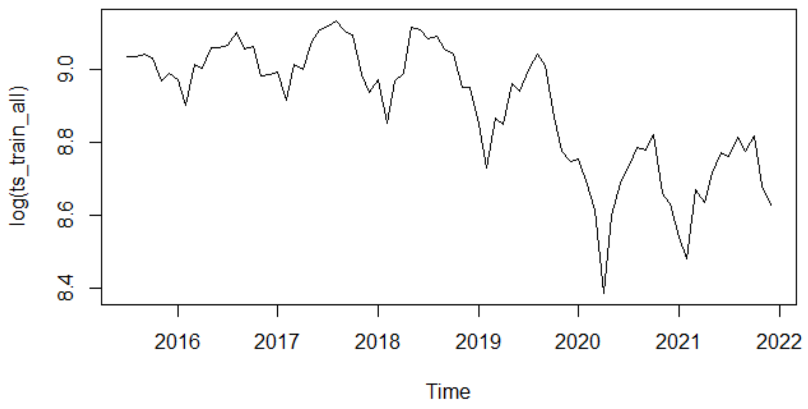
[1] https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system

Appendix

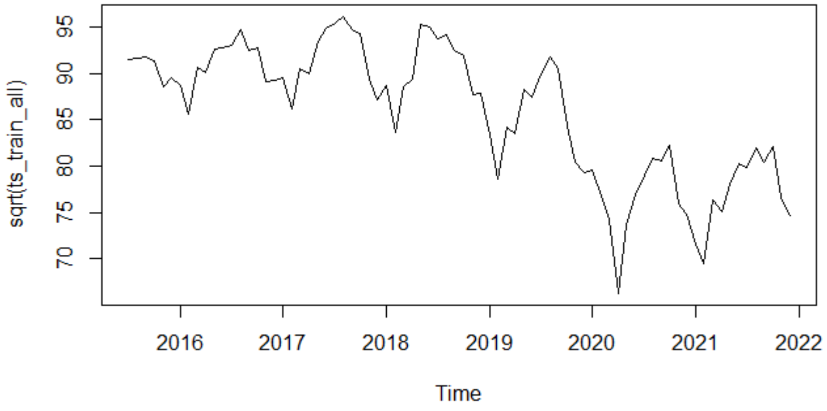
1. Steps to clean and reformat the crime dataset:
   a) The original dataset comes in separate years, so they were combined together at first
   b) The rows are counted by month to retrieve the number of crimes per month, and all other columns are discarded. For assault counts, only crimes with the category “assault” are selected.
   c) Counts for June 2015 is deleted since it is not a full month

2.

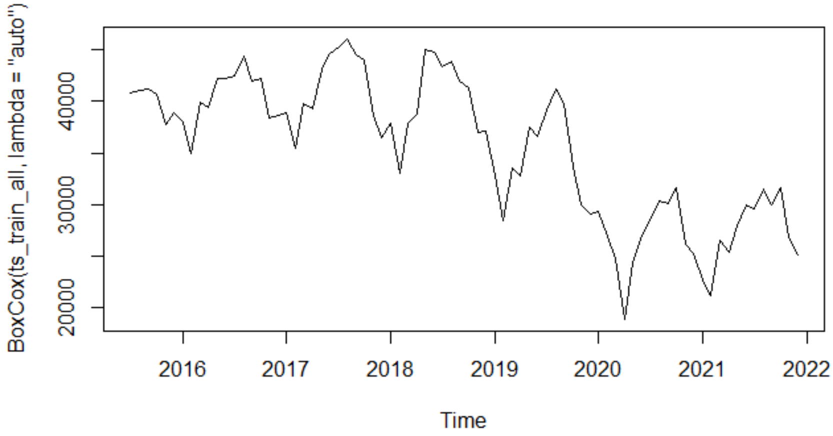
a. Time Series Plot of Log Total Crime Count



b. Time Series Plot of Square Root Total Crime Count

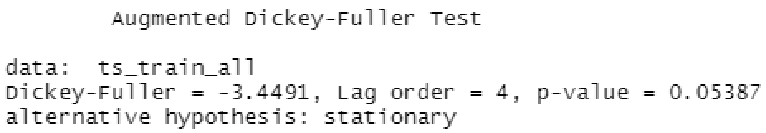


c. Time Series Plot of Box-Cox Transformed Total Crime Count



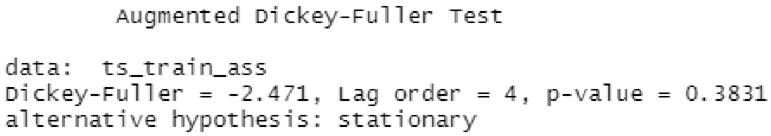
3.

a. ADF test result of Total Crime Count, p=0.05387



```
            Augmented Dickey-Fuller Test

data:  ts_train_all
Dickey-Fuller = -3.4491, Lag order = 4, p-value = 0.05387
alternative hypothesis: stationary
```

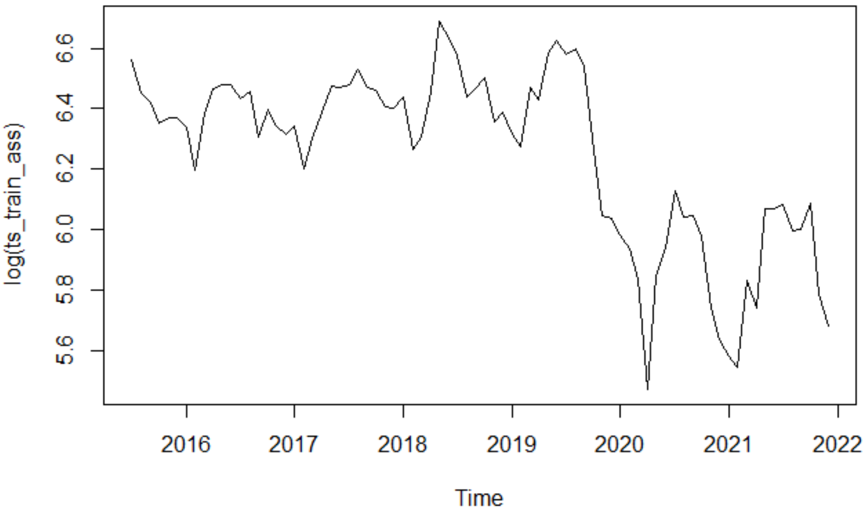
b. ADF test result of Assault Crime Count, p=0.3831



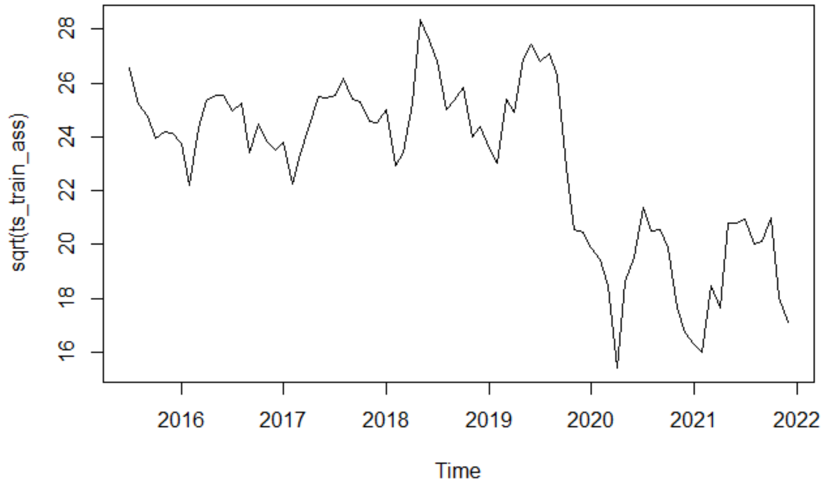
```
            Augmented Dickey-Fuller Test

data:  ts_train_ass
Dickey-Fuller = -2.471, Lag order = 4, p-value = 0.3831
alternative hypothesis: stationary
```

4.

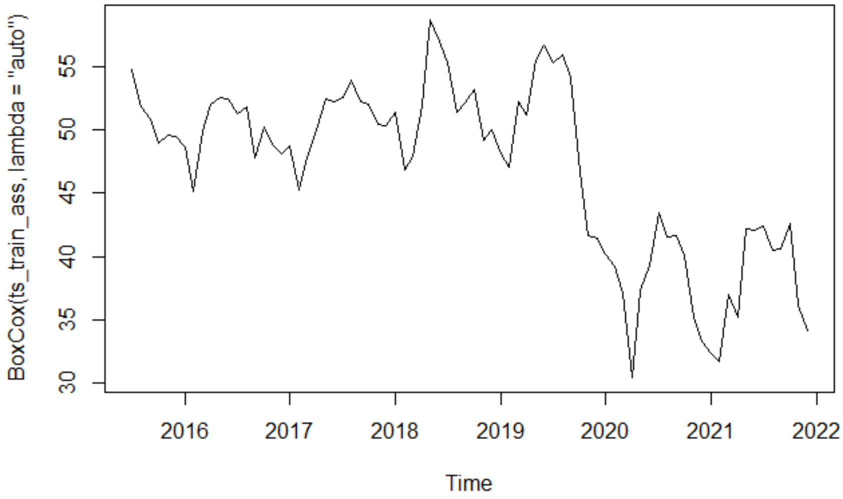
a. Time Series Plot of Log Assault Crime Count



b. Time Series Plot of Square Root Assault Crime Count

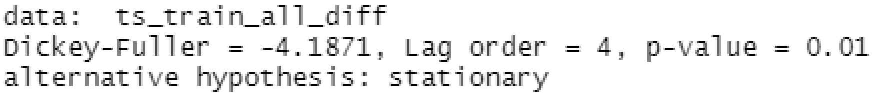


c. Time Series Plot of Box-Cox Transformed Assault Crime Count



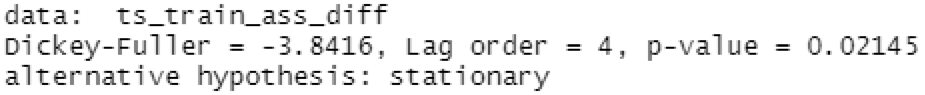
5.

a. ADF test result of Total Crime Count, p=0.01



```
data:  ts_train_all_diff
Dickey-Fuller = -4.1871, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

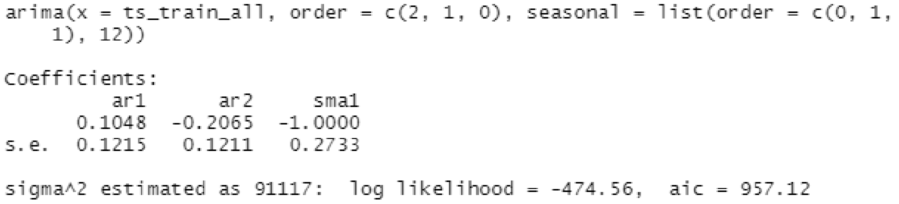
b. ADF test result of Assault Crime Count, p=0.02145



```
data:  ts_train_ass_diff
Dickey-Fuller = -3.8416, Lag order = 4, p-value = 0.02145
alternative hypothesis: stationary
```

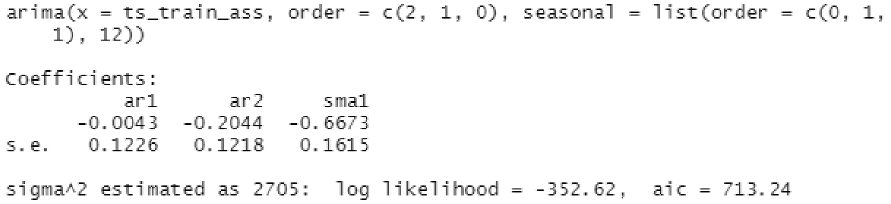
6.

a. Parameter Estimates for Total Crime Count



```
arima(x = ts_train_all, order = c(2, 1, 0), seasonal = list(order = c(0, 1,
    1), 12))

Coefficients:
         ar1      ar2     sma1
      0.1048  -0.2065  -1.0000
s.e.  0.1215   0.1211   0.2733

sigma^2 estimated as 91117:  log likelihood = -474.56,  aic = 957.12
```

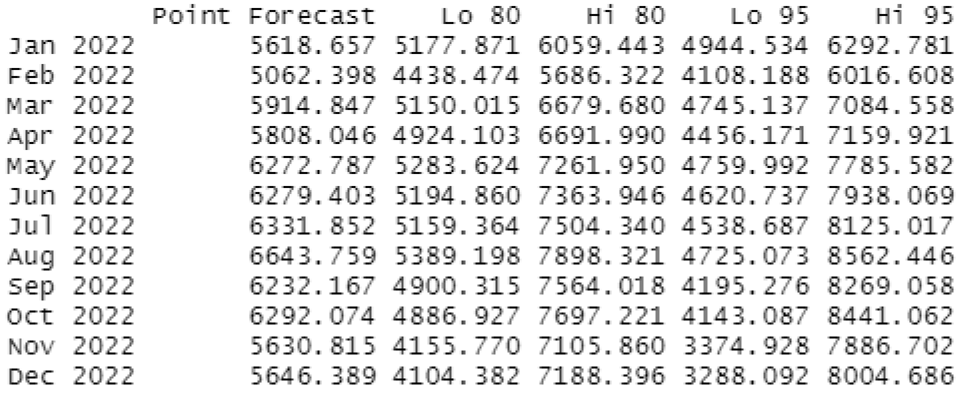
b. Parameter Estimates for Assault Crime Count



```
arima(x = ts_train_ass, order = c(2, 1, 0), seasonal = list(order = c(0, 1,
    1), 12))

Coefficients:
          ar1      ar2     sma1
      -0.0043  -0.2044  -0.6673
s.e.   0.1226   0.1218   0.1615

sigma^2 estimated as 2705:  log likelihood = -352.62,  aic = 713.24
```

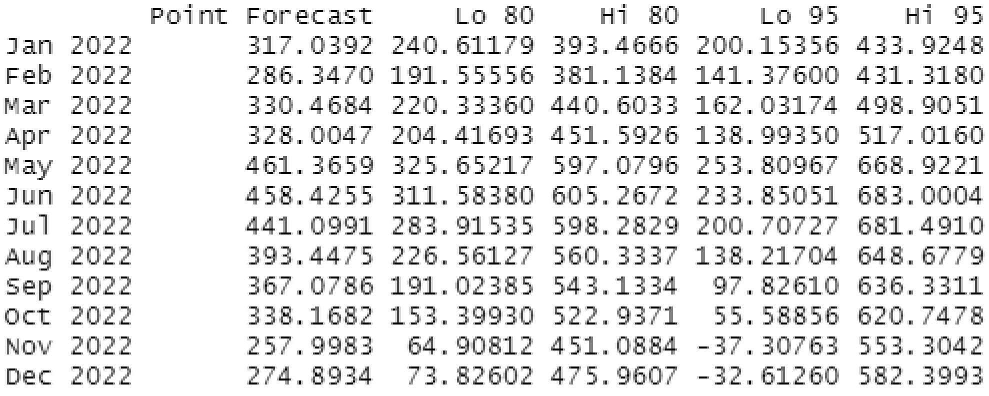
7.

a. Forecasted Total Crime Count, Holt-Winters Model



```
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Jan 2022       5618.657 5177.871 6059.443 4944.534 6292.781
Feb 2022       5062.398 4438.474 5686.322 4108.188 6016.608
Mar 2022       5914.847 5150.015 6679.680 4745.137 7084.558
Apr 2022       5808.046 4924.103 6691.990 4456.171 7159.921
May 2022       6272.787 5283.624 7261.950 4759.992 7785.582
Jun 2022       6279.403 5194.860 7363.946 4620.737 7938.069
Jul 2022       6331.852 5159.364 7504.340 4538.687 8125.017
Aug 2022       6643.759 5389.198 7898.321 4725.073 8562.446
Sep 2022       6232.167 4900.315 7564.018 4195.276 8269.058
Oct 2022       6292.074 4886.927 7697.221 4143.087 8441.062
Nov 2022       5630.815 4155.770 7105.860 3374.928 7886.702
Dec 2022       5646.389 4104.382 7188.396 3288.092 8004.686
```

b. Forecasted Assualt Crime Count, Holt-Winters Model



```
         Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
Jan 2022       317.0392 240.61179 393.4666 200.15356 433.9248
Feb 2022       286.3470 191.55556 381.1384 141.37600 431.3180
Mar 2022       330.4684 220.33360 440.6033 162.03174 498.9051
Apr 2022       328.0047 204.41693 451.5926 138.99350 517.0160
May 2022       461.3659 325.65217 597.0796 253.80967 668.9221
Jun 2022       458.4255 311.58380 605.2672 233.85051 683.0004
Jul 2022       441.0991 283.91535 598.2829 200.70727 681.4910
Aug 2022       393.4475 226.56127 560.3337 138.21704 648.6779
Sep 2022       367.0786 191.02385 543.1334  97.82610 636.3311
Oct 2022       338.1682 153.39930 522.9371  55.58856 620.7478
Nov 2022       257.9983  64.90812 451.0884 -37.30763 553.3042
Dec 2022       274.8934  73.82602 475.9607 -32.61260 582.3993
```