Crime Reports in Austin

Group 6 - Final Report

August 2022

Sam Currans (Distance): Programming, Analysis; Math B.S.; Statistics M.S. (sam_currans@tamu.edu)
Geo Lee (Distance): Writing, Analysis; Statistics B.S.+M.S. (geo9931@tamu.edu)
Isaac Ke (Distance): Programming, Analysis; Statistics B.S.+M.S.(isaacke9@tamu.edu)
Kyle Dennis (Local): Writing, Analysis; Statistics M.S. (kyle.dennis@tamu.edu)
Cameron Thomas (Local): Programming, Writing; Statistics M.S.(ctaggie18@tamu.edu)

1 Goal of Project

Our focus for this project is to explore any trends or cycles in Austin's crime rate and identify any crime specific patterns (e.g. the relationship between time of year and crime count). Using that information we will utilize various time series analysis techniques to smooth the data to uncover trend and seasonality, transform to stationarity, generate ACF and PACF plots, formulate appropriate ARMA/SARIMA models, iteratively fit and tune the models, and then forecast future crime rates in Austin.

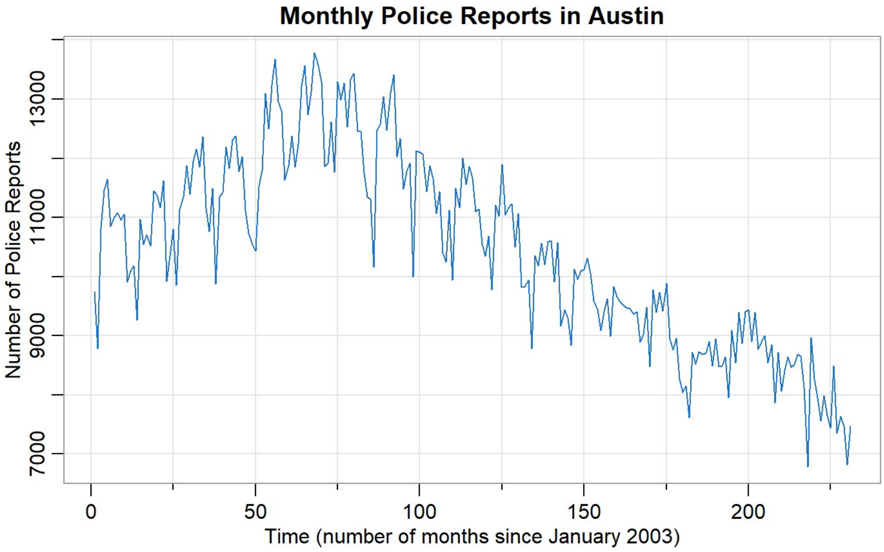

Figure 1: Monthly police reports in Austin from January 2003 to March 2022

2 Data Introduction

The data was obtained from the Austin Police Department (APD). It includes all instances since January 1, 2003 where APD responded to a criminal incident and wrote a police report. Each entry includes the date of the incident, the highest offense, whether it involved family violence, clearance status, clearance date, and various spatial data such as GPS coordinates and council district. The dataset has been updated every Monday since January 1, 2003 and at the time of writing was last updated June 6, 2022. For our particular aims of our project, we specifically look at the count of the monthly police reports from January 2003 to March 2022 (231 observations).

The initial time series plot (figure 1) of the data shows a few things. First, there is a cyclic annual seasonality; each year the crime rates rise and fall in a similar pattern. Second, the time series is highly volatile. Within a year, the number of crime rates fluctuate a lot. Third, there are some spikes of very high or low numbers of police reports on some days, potentially indicating high variability. Finally, there is an initial increasing trend until 2008, and then the total number of police reports begins gradually decreasing to present day.

3 Analysis of Cycle and Trend - Kernel Smoothing

To validate our hypothesis of a potential annual cycle and various increasing/decreasing trends in Austin crime reports, kernel smoothing was performed on the original time series. A normal kernel was used with a bandwidth of 5 (figure 2) to investigate the cyclic behavior. This revealed an annual rise and fall of police reports demonstrated by the hills that span 12 months. Further exploring the annual cyclic behavior, figure 3 shows the cumulative crime report counts for each month over the span of the data set. A sharp fall in police reports is seen in February and November, while a sharp increase is seen in January and March. In general, the spring and summer months have the greatest number of police reports.

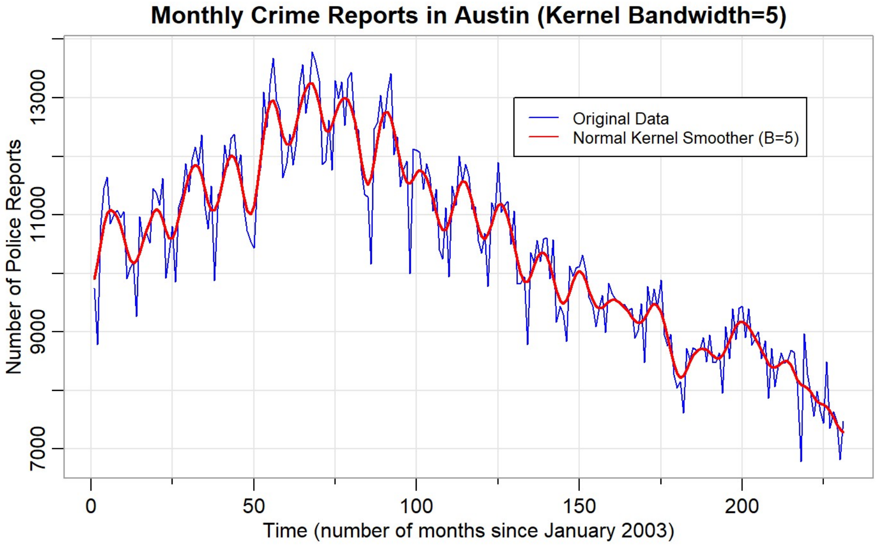

Figure 2: Normal kernel smoother on crime data using bandwidth=5

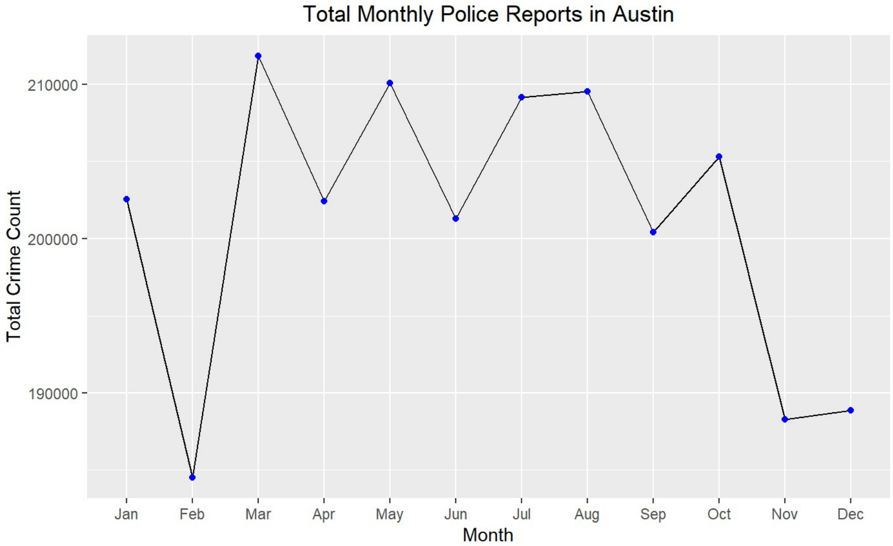

Figure 3: Line plot of the total number of crime reports in each month over the time range of the data set

In figure 3, the number of crime reports starts off low at the beginning of the year, rises in the spring through fall seasons, then drops back down toward the end of the year. This can be explained by the commonly observed correlation between temperature and crime rates. As the weather warms up, it draws more people outside, and thus with it criminal activity and opportunities for such felonies. As the weather gets colder in the winter months, people are drawn inside, and thus crime rates decrease.

In addition, we analyzed the overall trend of crime by using a normal kernel with an increased bandwidth of 30, shown in figure 4. The red line indicates that the number of monthly police reports steadily increased from 2003 to 2008, where it hit a peak. Then after 2008, the reports have been gradually decreasing, falling below the initial level in 2003.

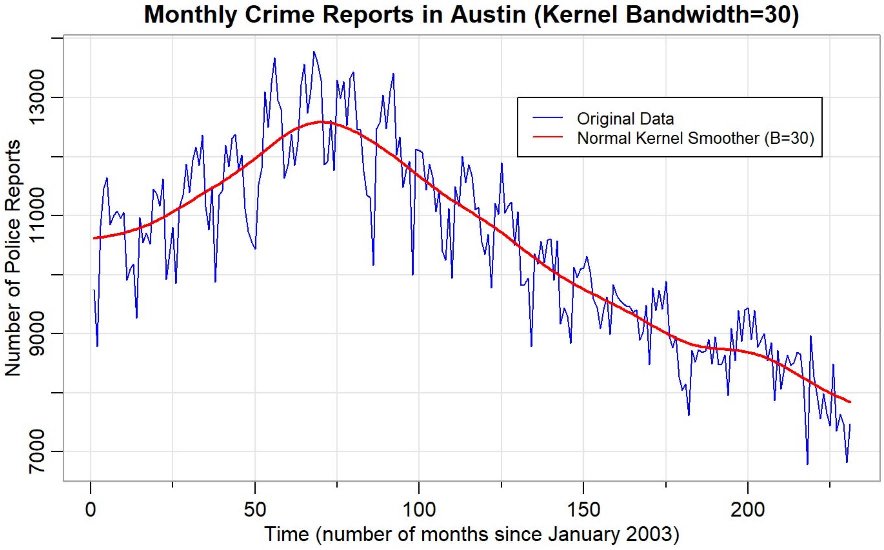

Figure 4: Normal kernel smoother on crime data using bandwidth=30

4 Investigation of Stationarity

As seen in figure 1, since the monthly data demonstrates an initial increasing trend followed by a decreasing trend, it appears to be nonstationary as the mean function is dependent on time. On the other hand, the variances appear to be stable across the time series as the ratio of the variance from the 1st half to the 2nd half of the series is 0.93, which is reasonably close to 1.

5 First Order Differencing

To transform the data to stationarity, we differenced the data (figure 5). Specifically, we computed the first order difference, which calculates the change between consecutive observations of the series by the equation $\nabla x_t = x_t - x_{t-1}$. The resulting time series plot appears much closer to stationarity than the untransformed data since the mean function appears to be a flat line around 0. As such, the mean function does not seem to depend on time. Furthermore, the variance appears to be constant and not dependent on time.

The kernel smoothing line in figure 5 also indicates that the cyclical nature of the data still appears in the first-order differenced data. This seasonality component will be investigated in section 9.

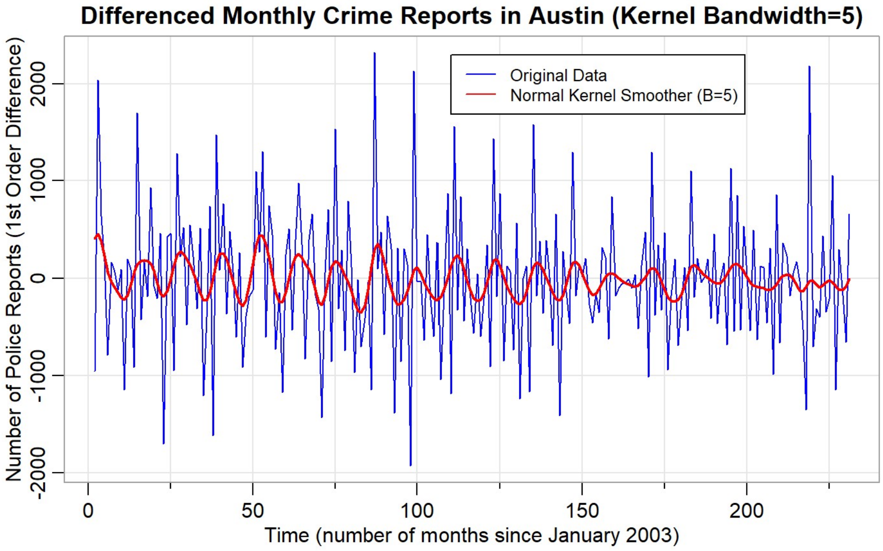

Figure 5: Transformed stationary time series after first-order differencing. The normal kernel smoothing line (b=5) is overlaid in red.

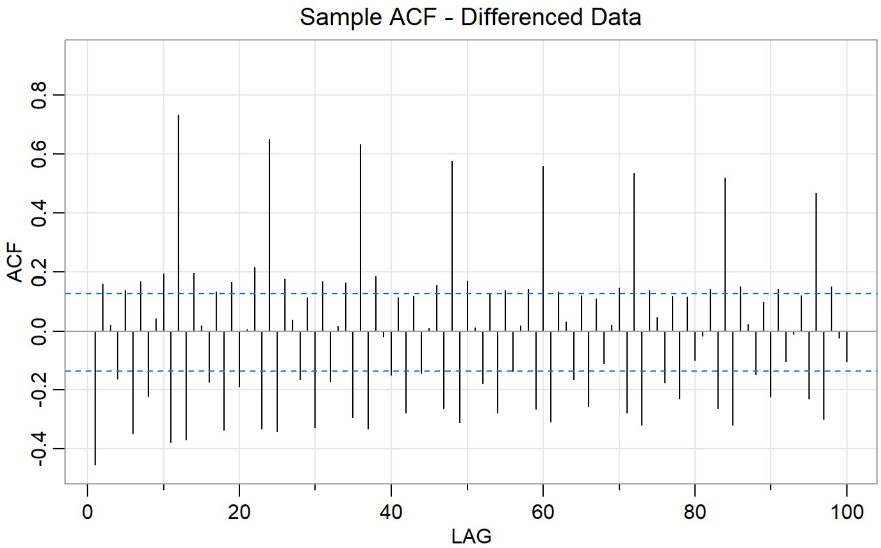

Figure 6: Sample ACF plot of differenced stationary data

The resulting sample ACF (figure 6) plot shows strong autocorrelation in the differenced data. A cyclic pattern is clear every 12 months, with the strength of the autocorrelation gradually decreasing over time. The sample correlations tend to decrease as the lag increases with no significant sample correlations after the first 100 lags, save for the seasonal component every 12 months. Overall, the annual cycle of crime reports is preserved when the data is transformed to stationarity.

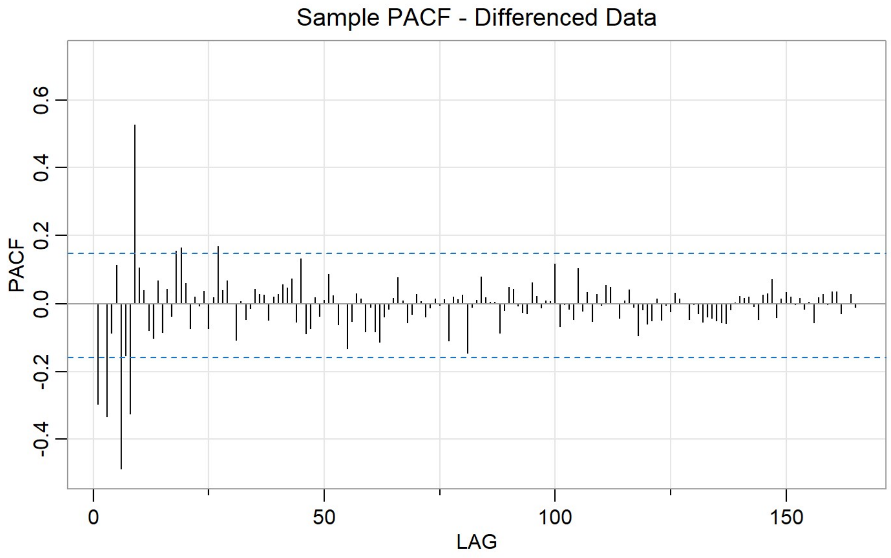

Figure 7: Sample PACF plot of differenced stationary data

In the PACF plot (figure 7), the first eight lags of the partial autocorrelation are large and significant. After the eighth lag, the partial autocorrelation quickly falls to zero. Using this information, we concluded that a moving average component is not present in the time series, but an autoregressive component is, leading us to develop an initial AR(p) model.

6 Model 1: AR(p) Model Formulation and Fit

Initially, an ARMA(p,q) model was considered. To choose p and q, the ACF and PACF plots were analyzed. As discussed in the previous section, for the differenced data, the correlogram in figure 6 gradually tails off and takes a while to become zero. On the other hand, the PACF plot (figure 7) has large significant partial autocorrelations until lag 8, and then it suddenly drops off and stays nearly 0. This combination is indicative of two things. One, there is no need to incorporate a moving average component to the model; hence, q=0 in the MA portion of the ARMA model. Second, p=8 in the AR portion of the ARMA model. Combining these two pieces of information yields an AR(8) model.

We then fit the model using two estimation methods. Initially, we estimated the AR(8) model parameters using the Conditional Sum of Squares Maximum Likelihood (CSS-ML) method resulting in the following model (equation 1) with an intercept:

$$x_t = -18.1683 - .4985x_{t-1} - .4158x_{t-2} - .6515x_{t-3} - .3570x_{t-4} - .3714x_{t-5} - .6553x_{t-6} - .3967x_{t-7} - .4269x_{t-8} + w_t$$

Next, we estimated the coefficients using the Yule Walker (YW) method without an intercept (equation 2), yielding the following autoregressive equation:

$$x_t = -0.385x_{t-1} - 0.269x_{t-2} - 0.545x_{t-3} - 0.214x_{t-4} - 0.227x_{t-5} - 0.559x_{t-6} - 0.263x_{t-7} - 0.325x_{t-8} + w_t$$

Although the coefficients obtained using the Yule Walker method are all smaller (in absolute value) than the corresponding coefficients obtained using the CSS-ML method, all of the estimated coefficients are negative in both models. The standard errors of the coefficients are slightly smaller using the CSS-ML method, however the differences are slight.

7 AR(p) Model Diagnostics

To further check the model assumptions and validity of the fit, a plot of the residuals was created, as shown in figure 8. It can be seen that the residuals roughly resemble a random scatter, but it can be argued that is does not completely appear to be white noise. Thus, the fit can be improved, as will be addressed in subsequent sections. For reference, the AIC for this AR(8) model was 15.55512, and the BIC was 15.7046.

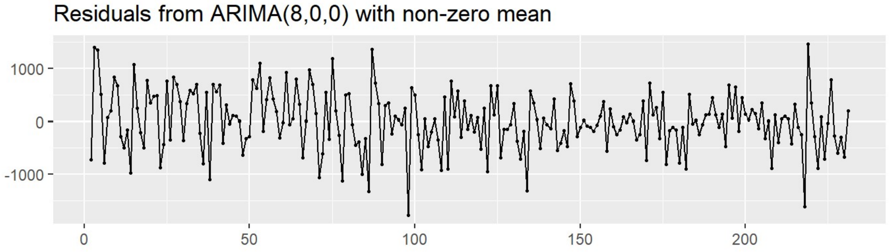

Figure 8: Residual plot from AR(8) model

8 AR(p) Model Forecast

Next, we forecasted the next year (12 months) of first order monthly differenced crime reports in Austin, as seen in figures 9 and 10. The purple highlighted regions depict 95% confidence bands around the predicted first order differences. Both the CSS-ML (equation 1) and YW (equation 2) fitted AR(8) models produce similar forecasts. The predictions rise and fall in a manner similar to the previous observations. Even such, the variability seen in the CSS-ML and YW predictions are smaller than in the data. The estimated predicted differenced count of monthly crime reports is summarized in figure 11.

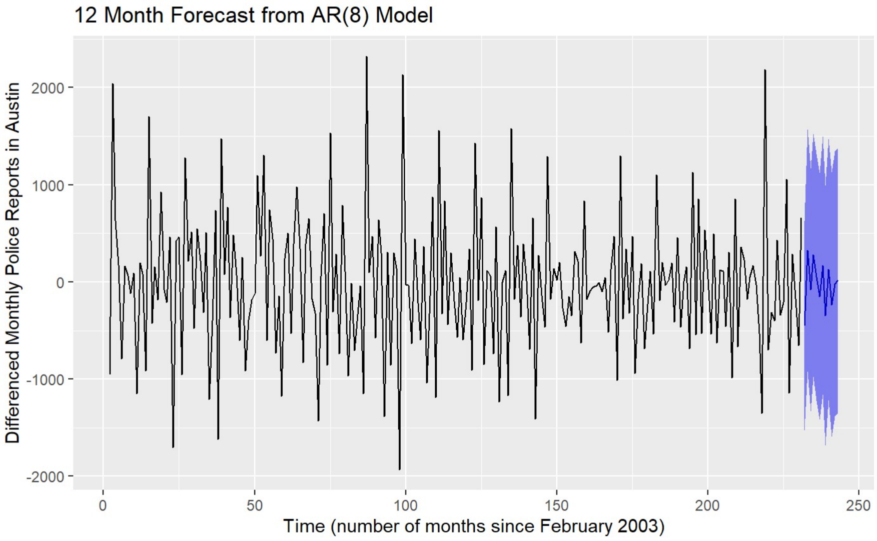

Figure 9: 12-month forecast of Austin crime reports using the CSS-ML fitted AR(8) model (equation 1)

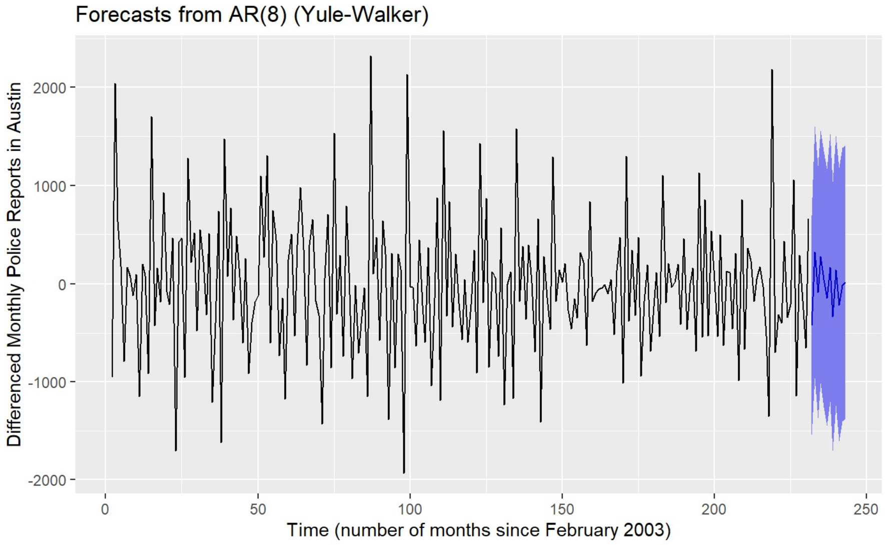

Figure 10: 12-month forecast of Austin crime reports using YW fitted AR(8) model (equation 2)

Months	CSS-ML Prediction	YW Prediction
1	45.36	-37.75162
2	331.68	412.61106
3	-391.03	-338.08991
4	154.70	165.16352
5	25.622	56.30820
6	-231.86	-198.53630
7	-78.55	-18.45310
8	-394.94	-393.02501
9	332.58	278.81329
10	-51.72	-128.57304
11	261.01	243.85150
12	-166.13	-109.43167

Figure 11: Table of 12 month predictions for the differenced monthly crime reports for the CSS-ML and Yule Walker fitted AR(8) models

9 Seasonal Differencing

As discussed in section 5, the first order difference transformed the data to stationarity, but still preserved the annual cyclic component of the series. Looking at figure 6 of the ACF plot of the first order differenced data, it is clear that the data cycles every 12 months. Thus, we seasonally difference the data using S=12, or the operator $\nabla_{12}\nabla x_t$. The resulting time series plot shown in figure 12 appears stationary with mean and variance function independent of time.

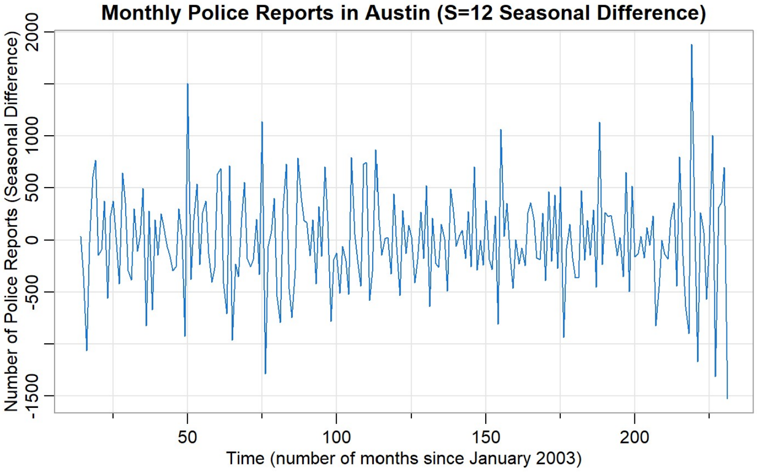

Figure 12: Transformed stationary time series after S=12 seasonal differencing (first and twelfth order differencing ($\nabla_{12}\nabla x_t$))

10 Model 2: SARIMA Model Formulation and Fit

After applying this twelfth-order difference to the first-order differenced data ($\nabla_{12}\nabla x_t$), we considered a SARIMA(p,d,q)x(P,D,Q) model. Since the resulting time series was stationary, we plotted the sample ACF and PACF of the seasonally differenced data to formulate potential parameters for the SARIMA(p,d,q)x(P,D,Q) model.

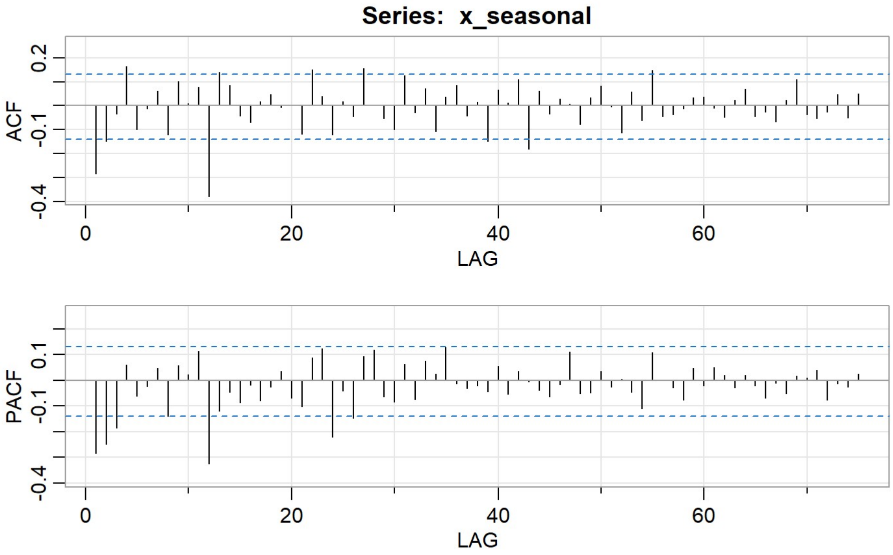

Figure 13: ACF and PACF of seasonally differenced data

Analyzing the sample ACF plot in figure 13, it appears that the autocorrelations for the seasonal (S=12) lags drop off suddenly after 1 (or lag 12). In the PACF, it appears the partial autocorrelations for the seasonal lags gradually tail off. Thus, an SMA component of Q=1 appears to be present, with no SAR component (P=0). Furthermore, for both the ACF and PACF plots, the autocorrelations in between the seasonal lags appear to be nonsignificant, so for the sake of parsimony, p and q for the regular ARMA portion were both chosen to be 0.

Thus, an initial SARIMA(0,1,0)x(0,1,1) was fit using the *sarima()* function in the *astsa* package in R. The AIC was 14.8487, and the BIC was 14.87975, which both indicate that this SARIMA fit improves on the previous AR(8) model (compare to the AIC/BIC in section 7).

To further investigate, the diagnostic plots were generated, as shown in figure 14. The residuals appear somewhat suspect, as they do not completely seem to be white noise. The ACF plot of the residuals indicates that there are a few significant autocorrelations. Furthermore, the p-values for the Ljung-Box Q statistic are all significant at the $\alpha = 0.05$ level. Thus, we concluded that the residuals are correlated and do not satisfactorily resemble white noise.

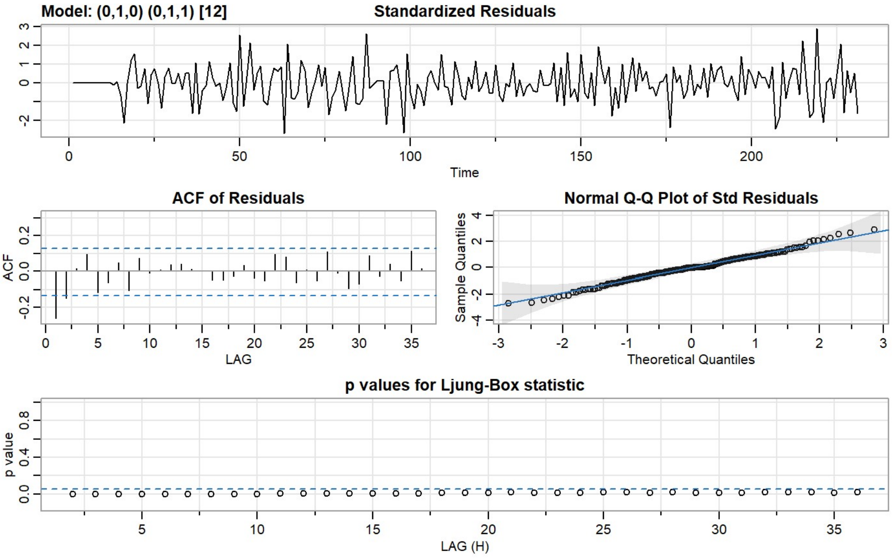

Figure 14: Diagnostic plots for the SARIMA(0,1,0)x(0,1,1) model

11 SARIMA Model Comparison and Best Fit

To further tune our model, we fit many other SARIMA models, varying the parameters for both the ARMA(p,d,q) portion and the SARIMA(P,D,Q) component. The criteria we used to compare models was AIC/BIC, the number of insignificant coefficients (with $\alpha = .05$ threshhold), and the residual diagnostics. A summary of the model fits is provided in the table in figure 15.

Model	**AIC**	**BIC**	**Insignificant parameters (0.05 level)**
(1,1,1) x (1,1,0)	14.87698	14.93908	0
(1,1,1) x (1,1,1)	14.73027	14.80789	1
(1,1,0) x (1,1,1)	14.79243	14.85453	1
(0,1,0) x (1,1,1)	14.85434	14.90092	1
(1,1,0) x (1,1,0)	14.96156	15.00813	0
(1,1,1) x (0,1,1)	**14.72220**	**14.78430**	**0**
(2,1,1) x (1,1,1)	14.73896	14.83211	1
(2,1,0) x (1,1,1)	14.74509	14.82271	1
(2,1,1) x (2,1,1)	14.74475	14.85343	4

Figure 15: Comparing multiple SARIMA models

The SARIMA$(1,1,1) \times (0,1,1)_{12}$ model fit the data the best, as it had the lowest AIC and BIC of 14.7222 and 14.7843, respectively. In summary, this model contained an AR component with p=1 and an MA component with q=1. The seasonal component included a twelfth-order seasonal difference and just a SMA (seasonal moving average) with Q=1. The coefficients for this final best model are summarized in the table shown in figure 16.

	Estimate	SE	t.value	p.value
ar1	0.3345	0.1372	2.4382	0.0156
ma1	-0.7169	0.1025	-6.9965	0.0000
sma1	-0.7205	0.0561	-12.8439	0.0000

Figure 16: SARIMA$(1,1,1) \times (0,1,1)_{12}$ coefficient estimates and their standard errors

All coefficients have p-values that are less than $\alpha = 0.05$, and are thus highly significant. The residual diagnostic plots (figure 17) indicate that the model assumptions are satisfied. The normal QQplot very strongly follows a straight line, indicating that normality is satisfied. Furthermore, there are no significant autocorrelations in the ACF plot of the residuals. Also, for all lags 0 through 36, the p-value for the Ljung-Box Q statistic for testing goodness of fit for white noise are all greater than $\alpha = 0.05$. Thus, we conclude that the residuals are uncorrelated, normal, and are white noise.

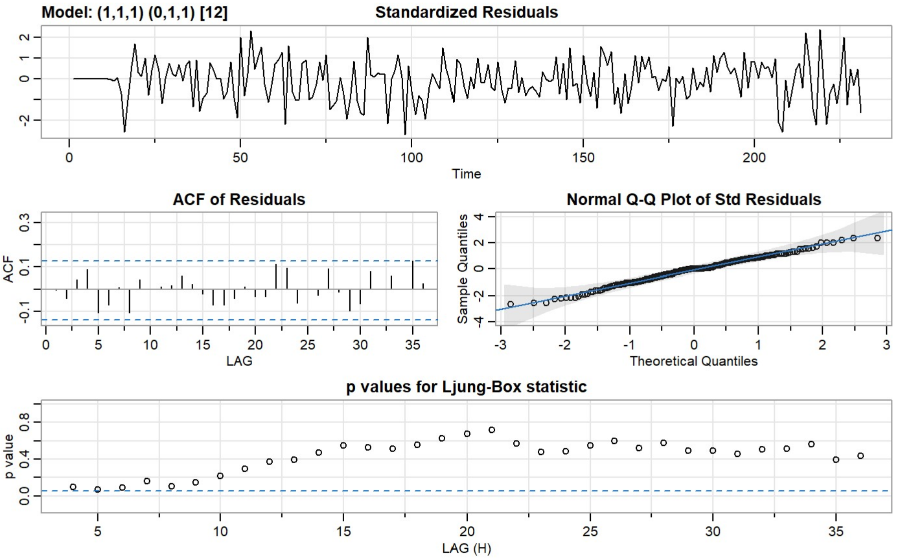

Figure 17: Residual diagnostic plots for the best model: SARIMA$(1, 1, 1) \times (0, 1, 1)_{12}$

12 SARIMA Model Forecast

Using our final SARIMA$(1, 1, 1) \times (0, 1, 1)_{12}$ model, we generated a 24-step ahead forecast to predict the next 2 years of monthly police reports in Austin, from April 2022 through March 2024. Figure 18 shows the forecast in red, with 95% prediction bands in gray.

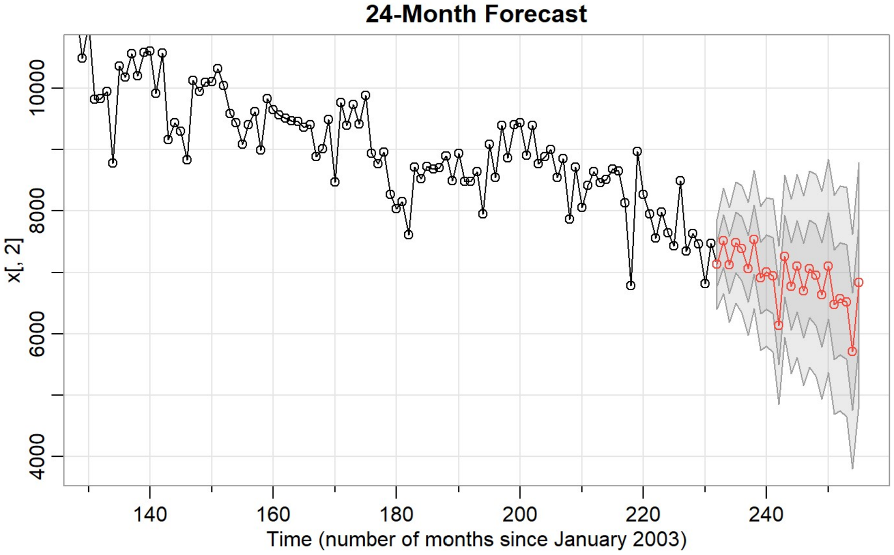

Figure 18: Best model (SARIMA$(1, 1, 1) \times (0, 1, 1)_{12}$): 24-month forecast of Austin crime rates from April 2022 through March 2024.

We note three things. One, the forecast continues the overall downward trend in the number of police reports. Crime rates are predicted to continue falling to record lows, since data collection started in 2003. Secondly, The rise and fall of crime count within a calendar year is captured in the predictions. Zooming into each year of predictions, the forecasted rise and fall of police reports follow the annual cycle similar to the data pattern illustrated in figure 3. Lastly, 95% prediction bands get wider and wider the further out we predict. This makes sense given that the standard error of m-step ahead predictions are a direct function of m, the distance from the end of the original time series.

A table of the forecasted monthly police reports in Austin is summarized in the table in figure 19. Numerically, we see that the predicted number of crime reports decreases gradually, while their standard errors increase.

Month	**Predicted # of police reports**	**S.E**
April 2022	7133.423	366.1102
May 2022	7511.827	430.2973
June 2022	7125.846	466.1453
July 2022	7485.078	494.0209
August 2022	7385.154	518.7823
September 2022	7064.051	541.9055
October 2022	7537.990	563.9192
November 2022	6906.556	585.0530
December 2022	7005.931	605.4327
January 2023	6946.504	625.1429
February 2023	6140.484	644.2486
March 2023	7266.120	662.8032
April 2023	6776.986	711.2688
May 2023	7105.701	744.1984
June 2023	6703.100	772.1394
July 2023	7056.773	797.9876
August 2023	6954.991	822.6670
September 2023	6633.266	846.5119
October 2023	7106.996	869.6656
November 2023	6475.493	892.2065
December 2023	6574.845	914.1877
January 2024	6515.410	935.6515
February 2024	5709.387	956.6332
March 2024	6835.023	977.1644

Figure 19: Table of forecasted monthly police reports in Austin (with their standard errors) from April 2022 through March 2024.

13 Summary and Conclusion

Starting in 2003, monthly police reports in Austin rose until it hit a peak in 2008, where it has been decreasing ever since. There is also a distinct annual seasonal cycle of crime rates, where warmer months have more crime than colder months. The original data was clearly non-stationary, so a first-order difference was applied to transform it to stationarity. After multiple different models were applied, with parameters estimated from sample ACF and PACF plots, the final model chosen was a $\mathrm{SARIMA}(1,1,1) \times (0,1,1)_{12}$ model. This fit had the lowest AIC and BIC amongst the candidate models, and its diagnostic plots and tests indicated that the residuals were white noise and uncorrelated. This model produced insightful forecasts, as illustrated in figures 18 and 19. It is predicted that over the next two years, Austin's monthly crime report counts will decrease into the 6000's, while still observing the pattern of increased rates in warmer months and lower rates in colder months. Overall, Austin appears to be on track to become a safer city, increasingly free of criminal activity.