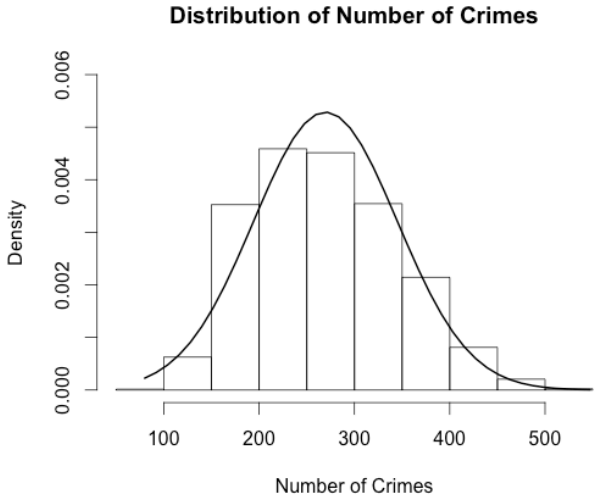
Modeling Chicago's Violent Crime

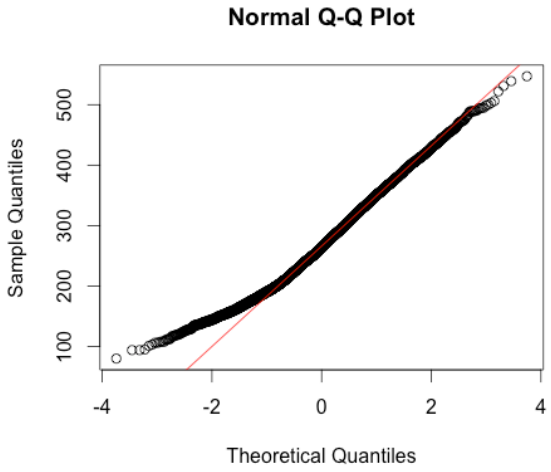
Jennifer Boyce

Violent crime is a fact of life in Chicago. As law enforcement agencies develop strategies to combat crime, it's essential to understand the underlying patterns of occurrences in order to predict future crimes. Time series models offer insight that can help further this understanding.

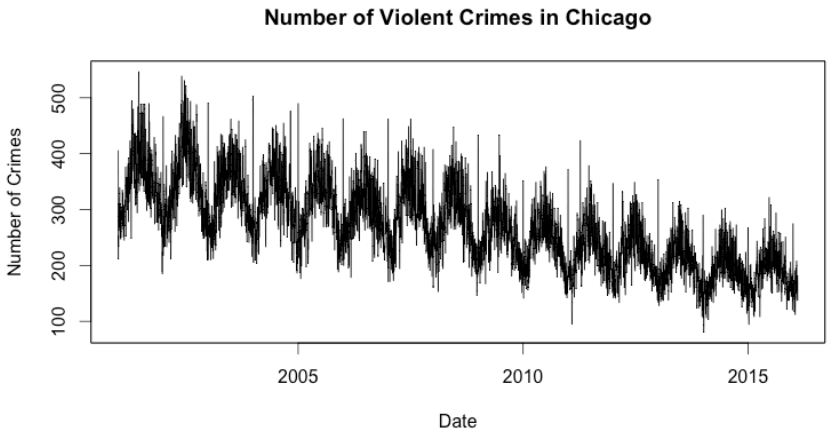
Exploratory Analysis

The data for this study include all incidences of violent crime reported by the Chicago Police Department from January 1, 2001 through February 8, 2016. "Violent crimes" include assault, battery, homicide, robbery, and sexual assault. In total, 1,492,123 crimes were reported. The daily incidence of crime was relatively normally distributed, with some deviation from normality in the tails. The number of violent crimes per day ranged from a high of 547 to a low of 80, with a median of 264 and mean of 270.

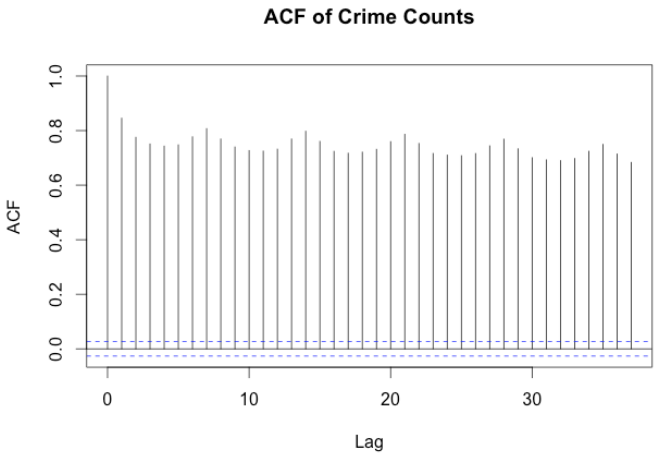


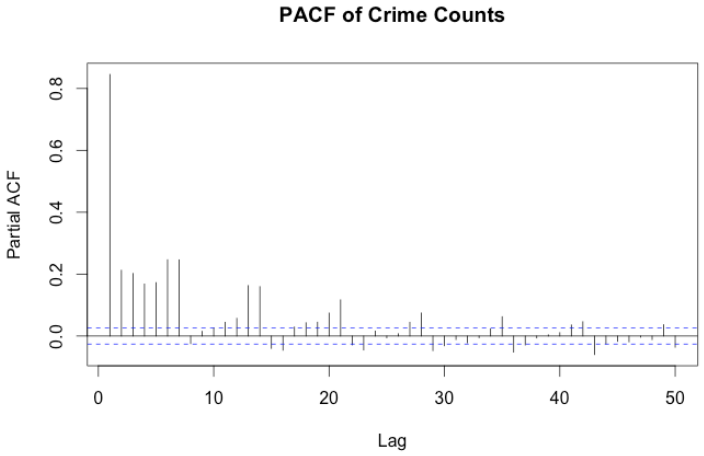


Despite news headlines that implied skyrocketing violence, the crime rate in Chicago showed a steady decrease from 2001 to 2016. Surprisingly, the dates with especially high crime rates do not represent days with significant events in the city; they are simply "normal" days.



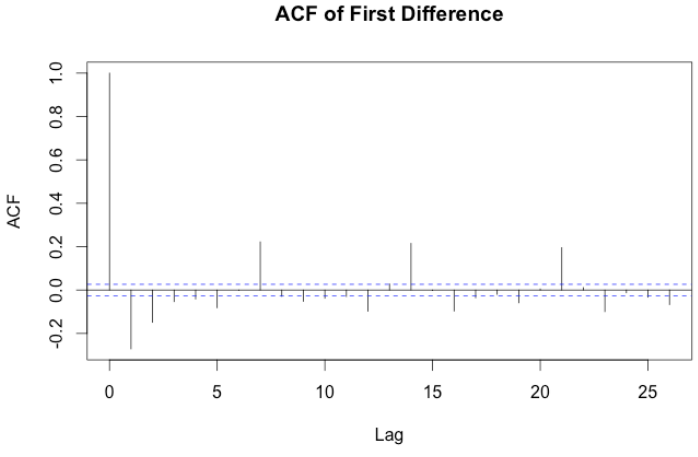
The ACF plot displays serial correlation, with values decaying very slowly toward zero. Box Ljung tests at lags 3, 5, and 7 also indicate serial correlation. Seasonality is apparent in the plot of the overall time series, as well as in ACF and PACF plots. Dickey Fuller test results indicate unit root nonstationarity.

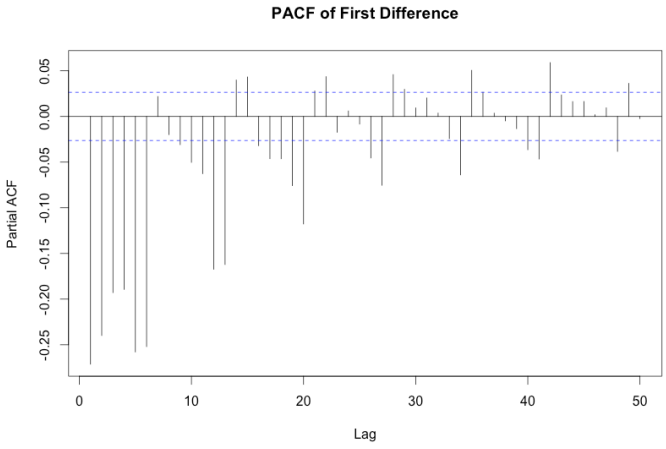




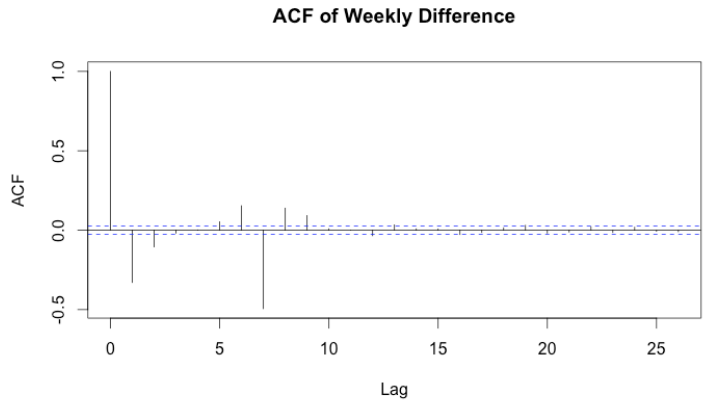
First differencing and seasonal differencing (7-day) produced a stationary time series.

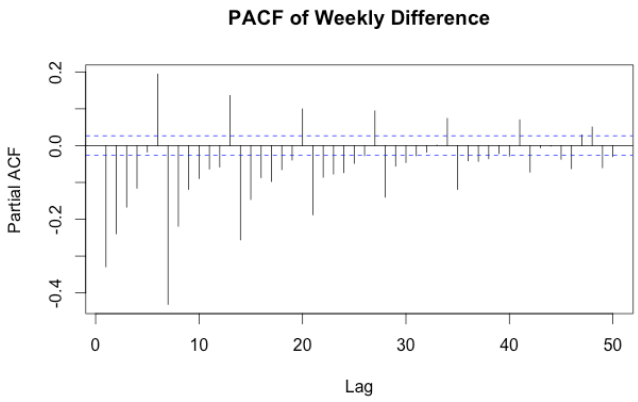
First Differencing





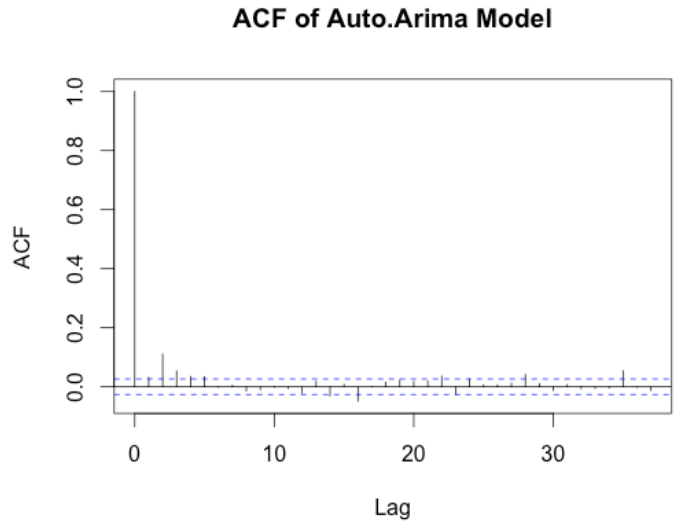
Seasonal Differencing

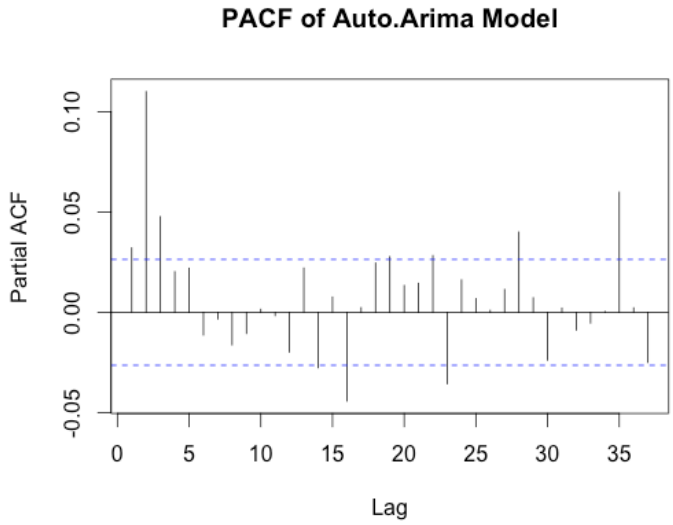




Model Fitting

An initial model was fit using R's auto.arima function, with BIC as the evaluation criterion. The optimal model was identified as ARIMA(6,1,1)(0,0,2)[7]. Ultimately, though, model diagnostics showed the model to be problematic. Both ACF and PACF plots indicated that the model did not adequately describe the dataset. BIC values were significantly higher than alternate models. Furthermore, Box Ljung tests indicated that residuals were not white noise.



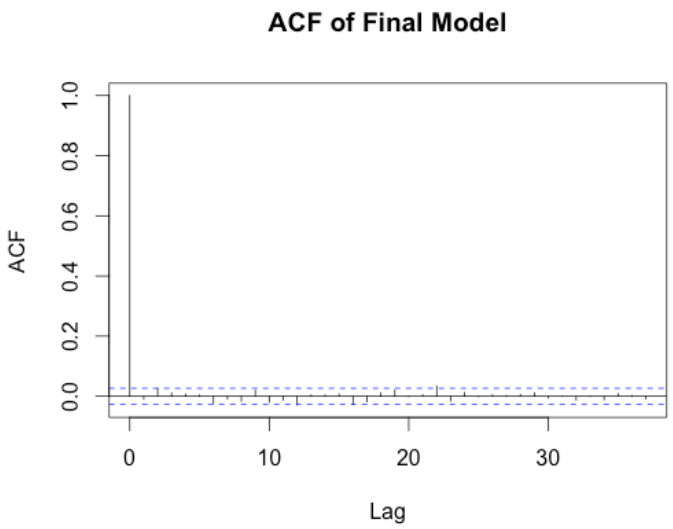


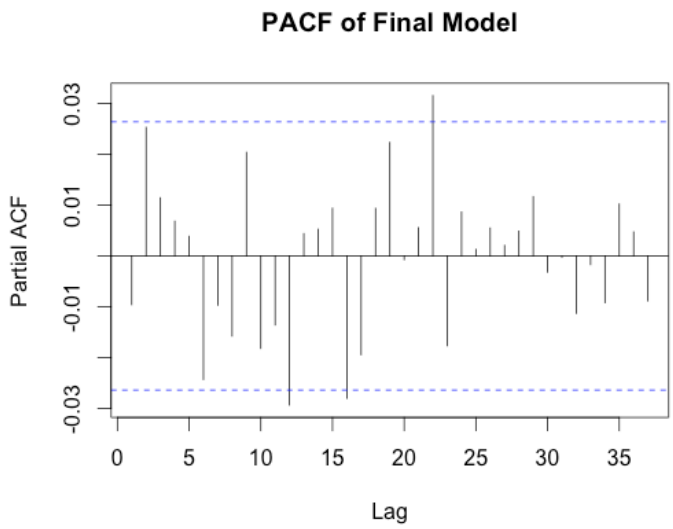
A number of alternate models were evaluated, and overall performance was assessed by comparing BIC values and exploring residuals. The final model that was selected was an ARIMA(8,1,1)(0,1,2)[7].

Residuals and Diagnostics

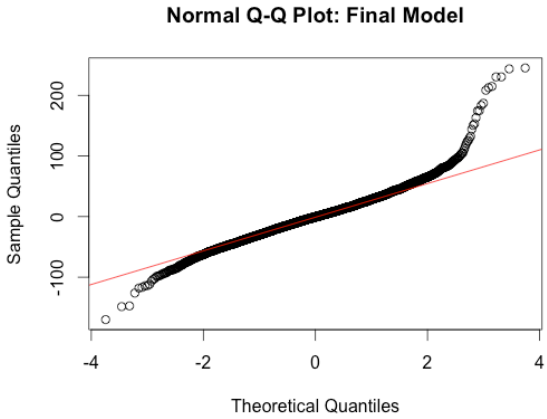
The ACF plot indicates that the residuals for the final model are white noise, although some outliers are

apparent.





This finding is confirmed by the Box Ljung test, which shows that residuals are white noise at lags 3,6 and 10. The QQ plot does indicates that the residuals are not normally distributed, with substantial deviation from normal in the tails. Although this deviation from normality is problematic, the model still has utility.



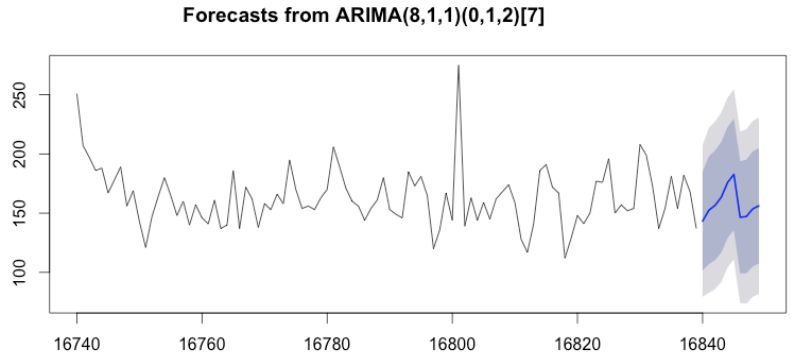
Lag	Box Ljung Test Statistic
3	0.09
6	0.1
10	0.1

Forecasting

The final model was used to forecast daily crime incidences up to ten days ahead. When plotted, the forecasts fit the overall pattern of past observations, although the predictions show more gradual changes in values than those observed.

Back testing was used to assess the overall quality of the model. Results indicated that predictions deviate from the true crime rate by 13.9%.

	Forecast	Low	High
2/9/16	143	102	185
2/10/16	152	107	198
2/11/16	157	110	203
2/12/16	164	117	210
2/13/16	176	129	223
2/14/16	183	136	230
2/15/16	147	99	194
2/16/16	147	99	195
2/17/16	154	105	202
2/18/16	156	107	205



	Forecast Error
RMSE of Out-of-Sample Forecasts	24.8
Mean Absolute Error of Out-of-Sample Forecasts	19.1
Mean Absolute Percentage Error	13.9%
Symmetric Mean Absolute Percentage Error	10.1%

Results and Discussion

Although imperfect, an ARIMA(8,1,1)(0,1,2)[7] can be used to predict the daily rate of violent crime in Chicago. The model captures the overall pattern of crime rate, and provides a realistic picture of the expected crime incidence. The forecasts, however, are fairly imprecise, with an overall error rate of 13.9%. While the model has some utility, the relatively high error rate means that the model would have limited use for operational needs such as police force staffing levels.

While this study offers one model that can be used to describe and predict Chicago's crime rate, more research is needed.