Joe Coyne

Dr. Frey

STAT 8444: Time Series & Forecasting

9 May 2025

Time Series & Forecasting Final Report

Introduction

The goal of this project is to try and analyze the different trends associated with car accidents in the United States from 2016 through 2023. Just from anecdotal experience, it seems like nowadays, there are more and more accidents on the road, much more than even from a few years ago, so I wanted to research and model these patterns.

The data used in this analysis originates from two papers by researchers at Ohio State University. In total, there are about 7.7 million accident records from a countrywide car accident dataset that the researchers use. Some of the variables available for analysis include the state the accident occurred in, what time zone the accident occurred in, what the temperature and weather were at the time of the accident, and any features of the road, such as exits, jughandles, roundabout, etc. This data was collected starting from February 2016 all the way through March 2023 using both MapQuest Traffic and Microsoft Bing Map Traffic. The APIs (Application Programming Interfaces) from these sites are captured using various entities like departments of transportation (whether that be local, state, or federal), law enforcement agencies, and traffic cameras and sensors. The data from said APIs was pulled every 90 seconds from 6:00AM – 11:00PM, and every 150 seconds from 11:00PM – 6:00AM. For the purposes of this project and to aid in the time series analysis, I consolidated the data into 2,572 different data points based on the unique dates of accidents being recorded and noted how many accidents there were per day.

Data Overview and Exploration

Before creating any models for the data, we first look at an initial time series plot, as seen in Figure 1 of the Appendix. At first glance, there seems to be a dip in the total number of accidents around June 2020, with a stark jump back up in September 2020. My initial thinking is that this could be a COVID effect, although it does take a dip 3 months after COVID lockdowns hit the United States. In March of 2021, there seems to be a return to a similar trend that was observed prior to the jumps in 2020. And finally, there is a steep decrease in accidents in 2023, although this could be because only 3 months of data had been collected at the time of this study, and maybe some data is recorded later in the year for previous dates (for example, perhaps data is collected and recorded every 6 months for the entire 6 month period). Overall, we see that there is a clear lack of constant variance. Looking at a distribution of the missing values in this initial time series plot, we see that there are 38 total missing values. Most of these missing values occur at the very beginning and the very end of the data, so maybe this is simply due to recording

errors and not that there were a total of 0 accidents across the United States, which seems extremely unlikely. To combat this issue, these missing values were imputed using the na_interpolation function in R.

I suspected that there would be some weekly pattern in the data, where there would be a different number of accidents on the weekend versus during the week. I decided to zoom in on March 2019 to view the weekly patterns, as seen in Figure 3 of the Appendix. Surely enough, there appear to be more accidents during the week with a lower number on the weekend. This may be because there is less accident tracking on the weekends with some officials being off, or maybe less people are driving to work and school. We then ran an ACF plot on the initial data, as seen in Figure 4 of the Appendix. There is an initial spike in the PACF at lag 1, and then spikes in the ACF every 7 lags, indicating a weekly trend and a seasonal trend of period 7. Finally, a scaled periodogram was plotted, as seen in Figure 5 of the Appendix. We see that the largest value occurs at the frequency 1/7, with the next largest occurring at 2/7, and so on. This provides further proof that there is seasonality of period 7.

To address the constant variance issue, I decided to transform the data both with a log transformation and a differencing of lag 7. The log transformed time series, as seen in Figure 6 of the Appendix, does a good job at reducing the variability by a large margin, although the downward spikes towards the latter half of the plot are a little concerning. Upon further investigation, most of the downward spikes occur when the number of total accidents is less than 10 (thus the log is about 2 or less). This could be because of two reasons: one that there were just an unusually low number of accidents that day, or more likely, some of the recording was done incorrectly. Maybe the total number of accidents was lost, so a placeholder of 1 was used, or perhaps some of the traffic sensors and cameras were down for the day. Another explanation could be that some weekends have less reported accidents if the departments of transportation and/or law enforcement agencies are closed or have limited staffing. Looking next at the differenced time series, which can be seen in Figure 7 of the Appendix, we see that the overall trend has been removed, but there is still a clear violation of constant variance, particularly towards the latter half of the plot. We ran some initial SARIMA functions and models on this data, but further transformed data was used to find better SARIMA models.

SARIMA Model

It was decided that the best transformation of the total number of US accidents was to take the log of total accidents, and then difference these logs at lag 7, this way both the trend is removed and the constant variance assumption is closer to being met. Figure 8 of the Appendix shows the final time series plot. The spikes in variability occur at the same places as before in the log plot, so our conclusions about the investigation of these abnormal spikes and drops will help our understanding here too. Looking next at the ACF plot of this final data, as seen in Figure 9 of the Appendix, there are large values at lag 1 and lag 7 in both plots. There are also large values at the multiples of 7 in the PACF plot. Through trying different models, we come to the conclusion

that a SARIMA(5,0,4,1,0,1) model with s=7 is the best model to analyze the data. The results of this SARIMA model can be found in Figures 10 and 11 of the Appendix. As you can see, there is still the issue of non-constant variance and the Normal Q-Q plot is still very concerning. However, the ACF and Ljung-Box plots look good, and most of the parameters of the model are significant. Even though the sar1 parameter is not significant at the 0.05 level, we chose to keep it in the model because it helped the forecasted observations retain the weekly pattern that was observed earlier. This can be seen in Figure 12 of the Appendix, as we see that the first forecasted value for April 1, 2023 (a Saturday) is smaller than the previous observation.

Regression Model

To see if the drastic changes in trend later in the time series plot could be modeled better than with the SARIMA model, a regression model was run with indicators created to denote the dip in June 2020, the spike back up in September 2020, the return to the original trend in March 2021, and the sharp decrease in 2023. There also seemed to be a change in slope around June of 2022, but this indicator was tested and found to be insignificant.

The full regression model, which can be seen in Figure 13 of the Appendix, includes the following variables: centered time and time squared terms, indicators for the day of the week (ind1 through ind6), and indicators for whether the data was from post-June 2020, post-September 2020, and post-January 2023. Indicators around March of 2021 and June of 2022 were determined to not be significant in this model. Sine and cosine were also tested in the model, but were also found to not be significant. Finally, some simple interaction terms were tested but ultimately found insignificant. The residuals of this model along with the accompanying ACF plot are found in Figures 14 and 15 in the Appendix, respectively. Once again, there is a large spike at lag 1 in the PACF plot, and spikes at lag 7 and all of the multiples of 7 in the ACF plot.

An ARMA(2,1) model was then run on these residuals, which resulted in the diagnostic plots found in Figure 16 of the Appendix, with the coefficients of the model in Figure 17. There is still an issue of non-constant variance, and the Normal Q-Q plot is still very concerning, particularly with the left tail. However, the ACF and Ljung-Box plots look good, and all of the parameters are significant. Overall, the model performs about the same as the SARIMA model from before, as it has just a slightly higher AIC. Looking at the forecasted values in Figure 18 of the Appendix, we see similar 7 day trends, however the dips now occur on a Thursday and Friday instead of a Saturday and Sunday, so the SARIMA(5,0,4,1,0,1) with s = 7 will be the final model.

Further Data Investigation

Next, I wanted to look at the US accident trends at a statewide level to see if there were any interesting trends or patterns related to the state an accident occurred in. Figure 19 shows a map of the continental United States filled by the total number of accidents over the timeframe of this dataset (2016-2023). At first glance, we see that New York and California clearly have the

most total accidents overall, with New York having a total of 2.2 million accidents and California at 1.9 million. Michigan (755,720 accidents), New Jersey (428,285 accidents), and Illinois (364,048 accidents) have the next largest amount of total accidents. One thing that became apparent when investigating the total accidents per state is that there was a lot of missing data from this dataset. Many states have no accident data for multiple years, as you can see in Figures 20, 21, and 22. The states with no accident data are shaded in gray. In fact, Delaware, Maine Mississippi, New Hampshire, North Dakota, Oklahoma, and South Dakota all have 0 reported total accidents over the entirety of the 7+ year study. West Virginia only has 7 reported total accidents over this same time frame. This seems unlikely and virtually impossible to be true, so there is likely major reporting issues with this data.

Limitations and Further Considerations

One of the biggest limitations with this dataset was in regards to the major data collection issues present. There were many dates with either no accidents or 1 accident (likely a placeholder) recorded. As just mentioned, many states had years where no accidents were recorded, and some states had no reported accidents at all. This missingness issue definitely had impact on my modeling and forecasting, as many of the trends are likely not 100% accurate and generalizable to the entire United States. This issue was also likely the cause of the many jumps in accidents, which resulted in the very concerning Normal Q-Q plots. If I were to run this analysis again, I would either find a new dataset that had complete values and a better data collection process, or I would run the SARIMA and regression models only on the states where there was data for each year. If I had more time, I would have also liked to explore more of the other predictors available in the dataset (like temperature, weather conditions, time zone, etc.) and their relationships with the number of daily accidents, as well as some more complex interactions between predictors.

Conclusion

In conclusion, the SARIMA (5,0,4,1,0,1) model with seasonal period s = 7 and no indicators did the best job at predicting total daily accidents in the United States. This analysis also showed that weekdays tend to have more accidents than the weekend, possibly because less people are on the road driving to work or school. In addition, we saw that as time goes on and driving becomes more popular, more accidents will tend to occur. And with the state level, this analysis showed that California and New York seem to have the most accidents in the United States. Overall, states that are more populated and have more drivers will have more accidents.

Appendix

References:

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

https://arxiv.org/abs/1906.05409 ("A Countrywide Traffic Accident Dataset")

https://arxiv.org/abs/1909.09638 ("Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights")

Figure 1:

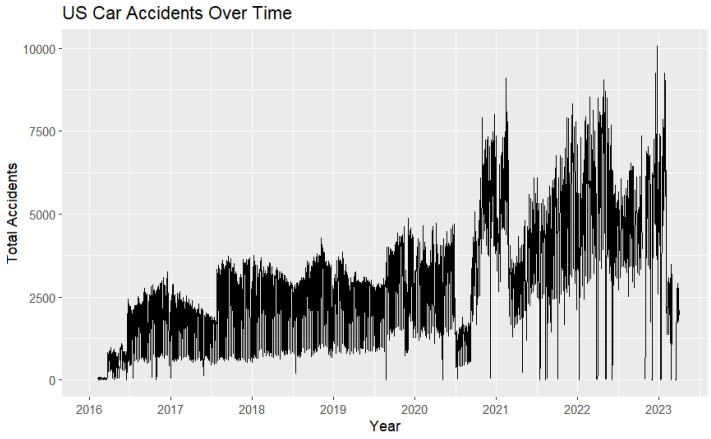

Figure 2:

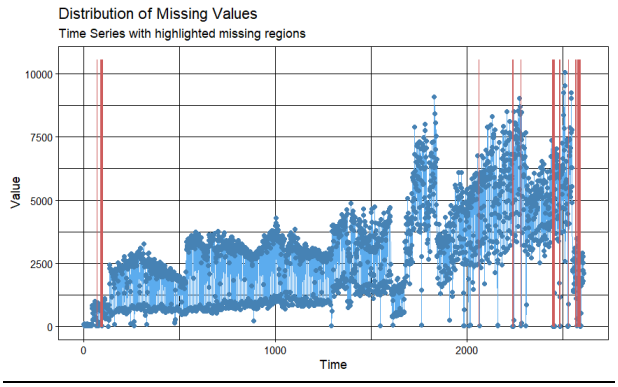

Figure 3:

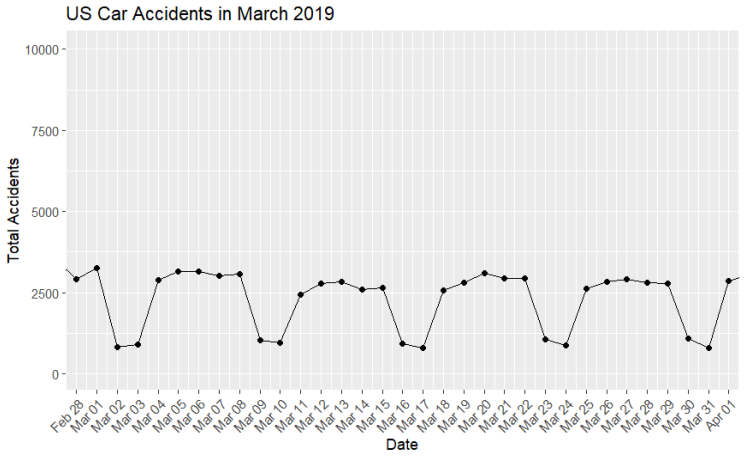

Figure 4:

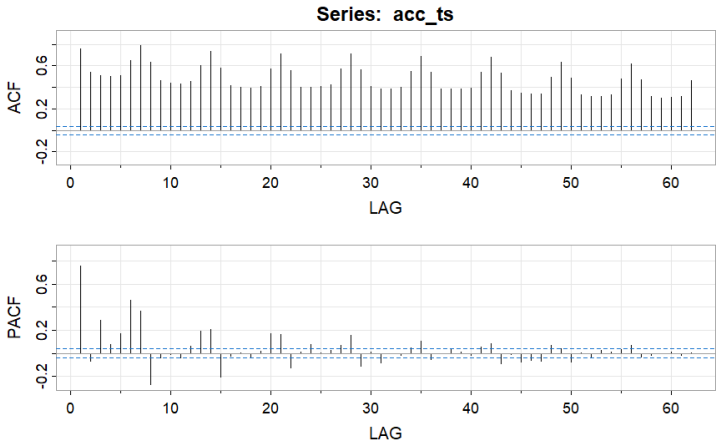

Figure 5:

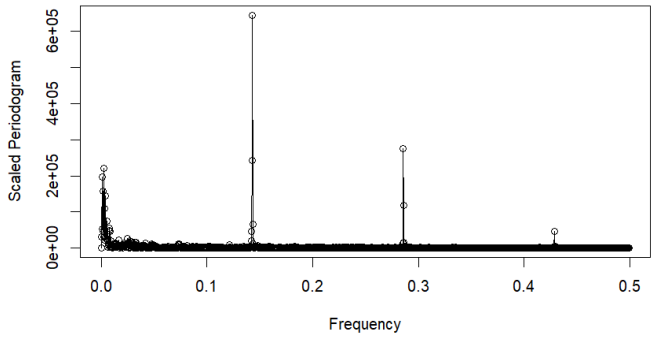

Figure 6:

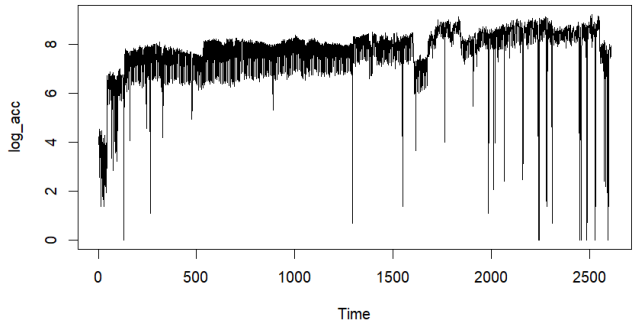

Figure 7:

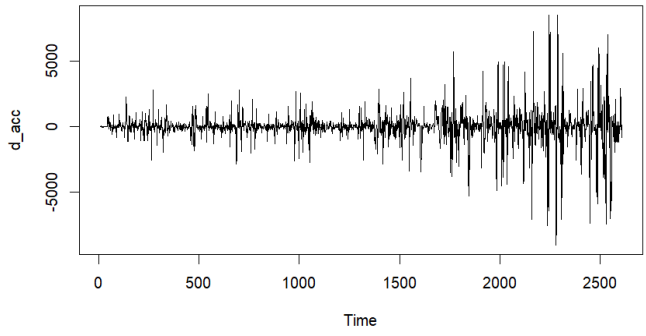

Figure 8:

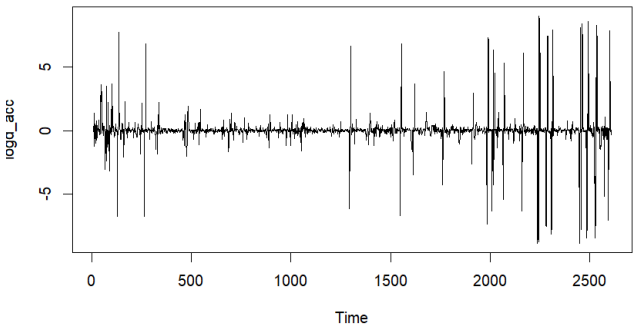

Figure 9:

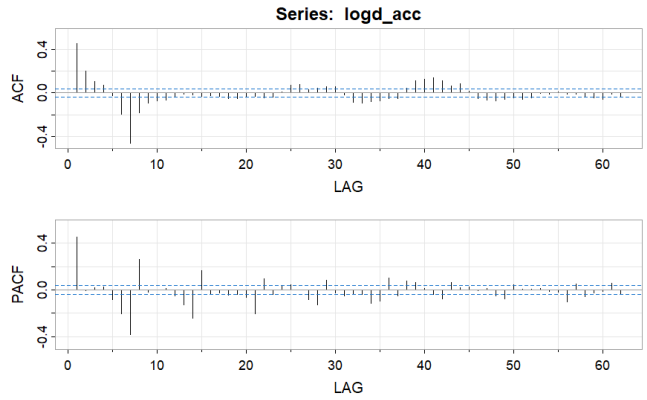

Figure 10:

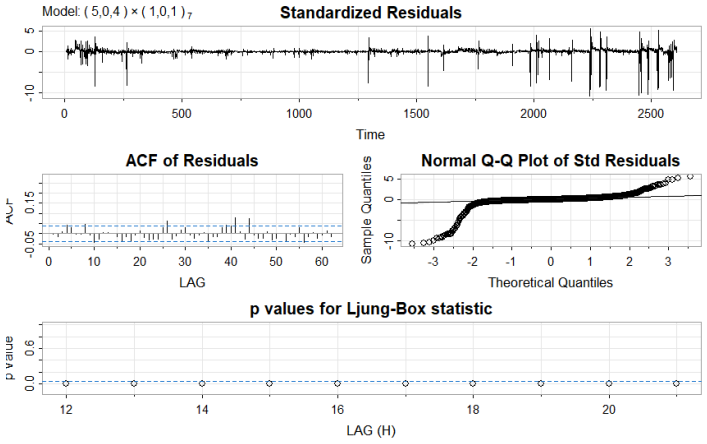

Figure 11:

```
Coefficients:
      Estimate     SE   t.value p.value
ar1    -0.5047 0.3842   -1.3137  0.1891
ar2     0.4992 0.1258    3.9673  0.0001
ar3     1.0922 0.0889   12.2856  0.0000
ar4     0.2843 0.3546    0.8018  0.4227
ar5    -0.3727 0.1837   -2.0290  0.0426
ma1     1.0147 0.3815    2.6594  0.0079
ma2     0.0360 0.0766    0.4702  0.6382
ma3    -1.0538 0.0449  -23.4472  0.0000
ma4    -0.7853 0.3691   -2.1274  0.0335
sar1    0.0031 0.0216    0.1452  0.8846
sma1   -0.9725 0.0047 -209.0583  0.0000
xmean   0.0142 0.0369    0.3864  0.6993

sigma^2 estimated as 0.6571864 on 2590 degrees of freedom

AIC = 2.435631  AICc = 2.435678  BIC = 2.464929
```

Figure 12:

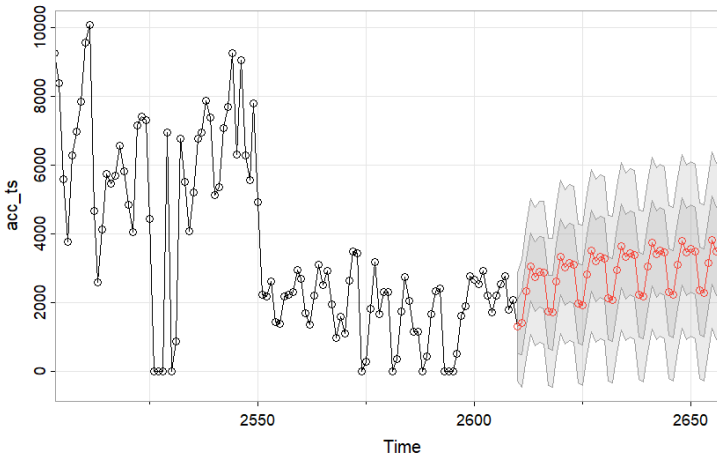

Figure 13:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.220e+00  5.994e-02 120.458  < 2e-16 ***
tc              4.484e-04  6.168e-05   7.271 4.71e-13 ***
tc2            -6.692e-07  5.298e-08 -12.631  < 2e-16 ***
ind1_trim       8.961e-01  7.311e-02  12.257  < 2e-16 ***
ind2            1.030e+00  7.311e-02  14.082  < 2e-16 ***
ind3            9.994e-01  7.311e-02  13.669  < 2e-16 ***
ind4            1.026e+00  7.311e-02  14.038  < 2e-16 ***
ind5            9.221e-01  7.311e-02  12.612  < 2e-16 ***
ind6            1.110e-01  7.316e-02   1.517    0.129
ind_20TRUE     -1.068e+00  1.286e-01  -8.306  < 2e-16 ***
ind_20_sepTRUE  1.458e+00  1.339e-01  10.892  < 2e-16 ***
ind_23TRUE     -6.632e-01  1.301e-01  -5.098 3.69e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9978 on 2597 degrees of freedom
Multiple R-squared:  0.3457,    Adjusted R-squared:  0.3429
F-statistic: 124.7 on 11 and 2597 DF,  p-value: < 2.2e-16
```

Figure 14:

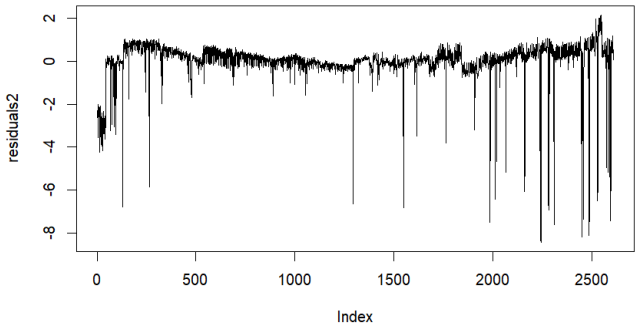

Figure 15:

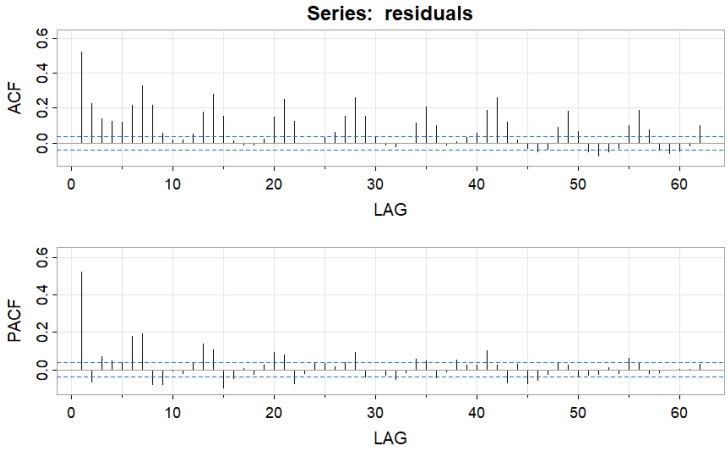

Figure 16:

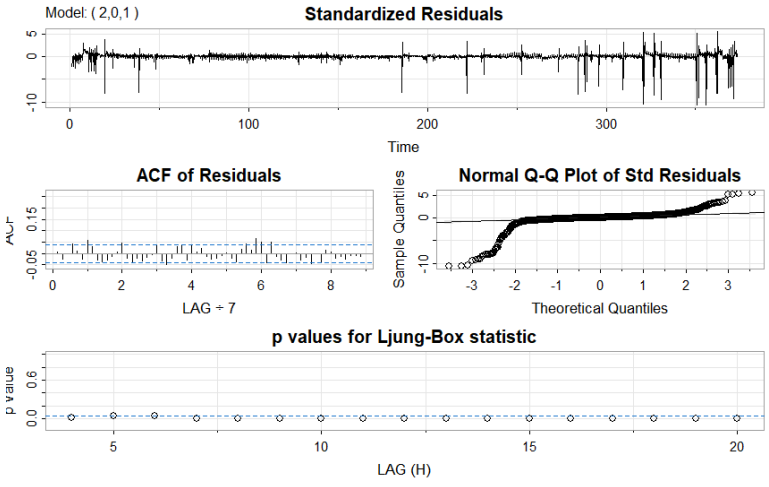

Figure 17:

```
Coefficients:
           Estimate     SE   t.value p.value
ar1          1.4460 0.0195   74.2879  0.0000
ar2         -0.4470 0.0196  -22.8538  0.0000
ma1         -0.9603 0.0064 -149.1637  0.0000
intercept    7.0655 0.3186   22.1796  0.0000
tc           0.0008 0.0003    2.1947  0.0283
ind1_trim    0.8973 0.0490   18.3110  0.0000
ind2         1.0245 0.0587   17.4642  0.0000
ind3         0.9949 0.0619   16.0753  0.0000
ind4         1.0253 0.0619   16.5659  0.0000
ind5         0.9242 0.0587   15.7528  0.0000
ind6         0.1138 0.0490    2.3207  0.0204
ind_20      -1.1220 0.3892   -2.8829  0.0040
ind_20_sep   0.9699 0.3881    2.4991  0.0125
ind_23      -1.0907 0.3801   -2.8696  0.0041

sigma^2 estimated as 0.6749774 on 2595 degrees of freedom

AIC = 2.4571  AICc = 2.457162  BIC = 2.49083
```

Figure 18:

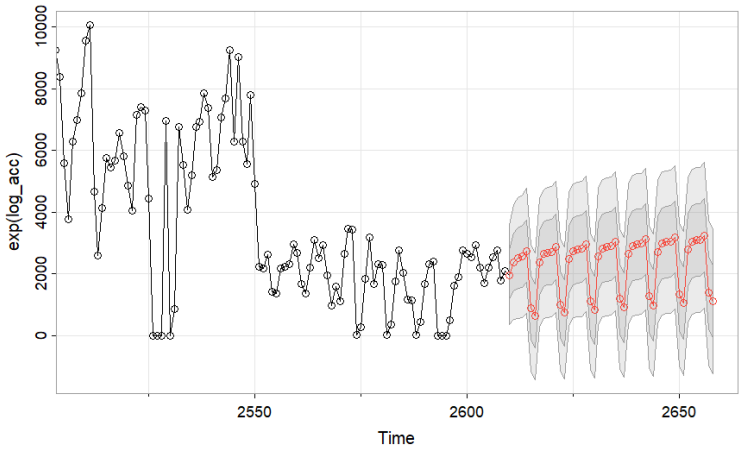

Figure 19:

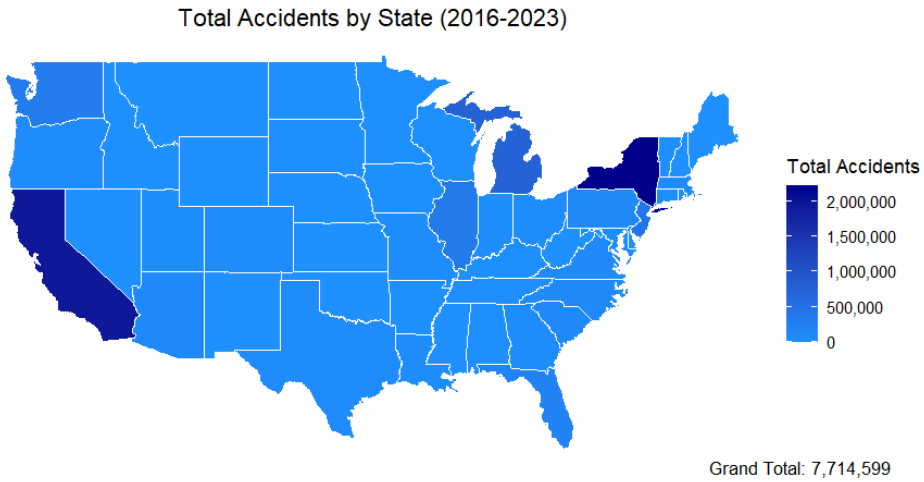

Figure 20:

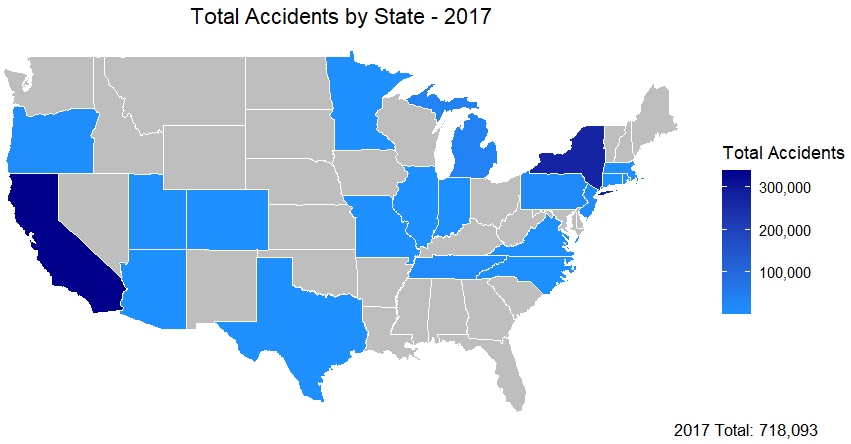

Figure 21: