US Accident Trends from 2016-2023

By: Joe Coyne

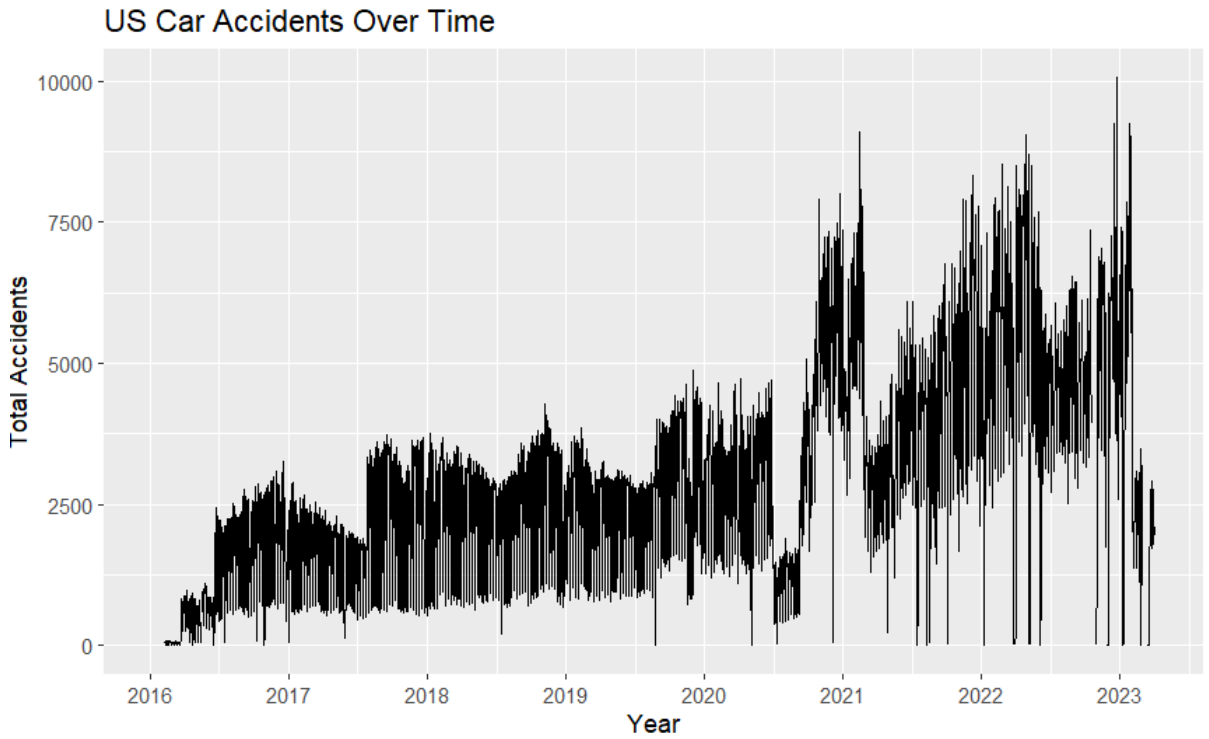
Outline

- Introduction
- Data Overview
- Initial Data Exploration
- SARIMA Model
- Regression Model
- Further Data Investigation
- Conclusion

Introduction

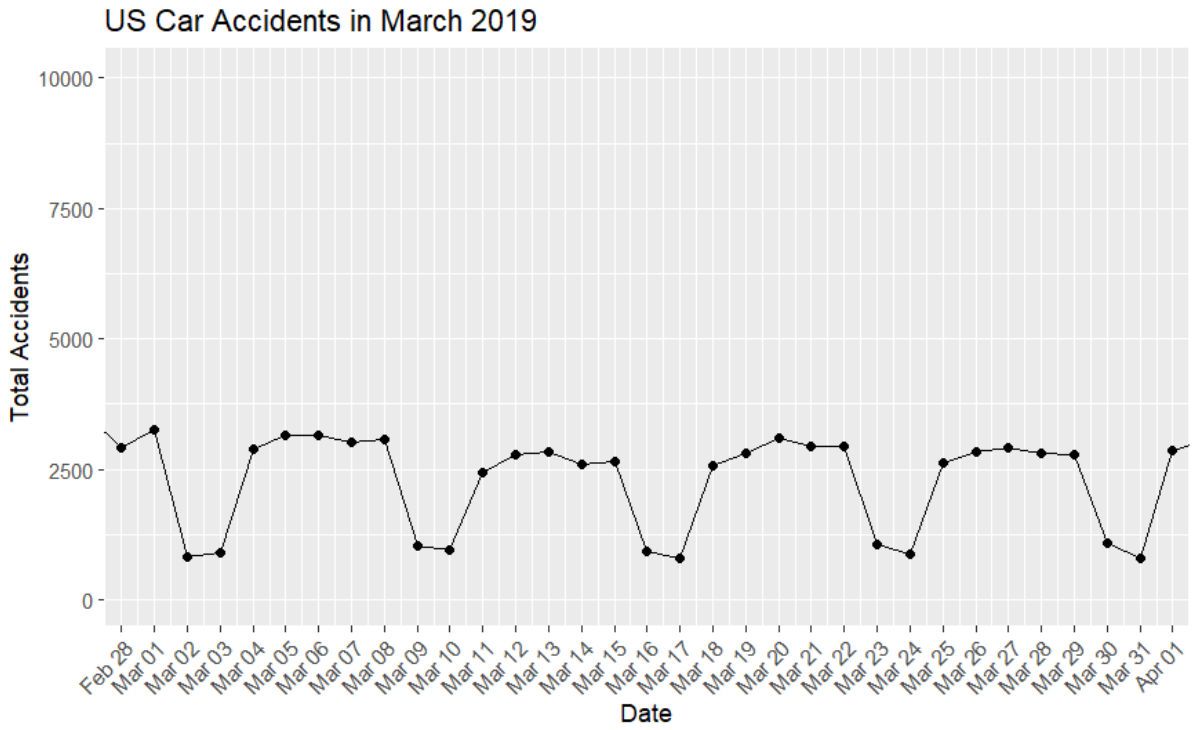
- Dataset originates from two papers by researchers at Ohio State University
- About 7.7 million accident records from a countrywide car accident dataset
  - Some variables include state, time zone, temperature, weather, features of the road (exits, roundabouts, etc.)
- Data collected from February 2016 through March 2023 using MapQuest Traffic and Microsoft Bing Map Traffic
- APIs from these sites are captured using various entities like departments of transportation, law enforcement agencies, and traffic cameras and sensors
- Data was pulled every 90 seconds from 6am-11pm, and every 150 seconds from 11pm-6am

Data Overview



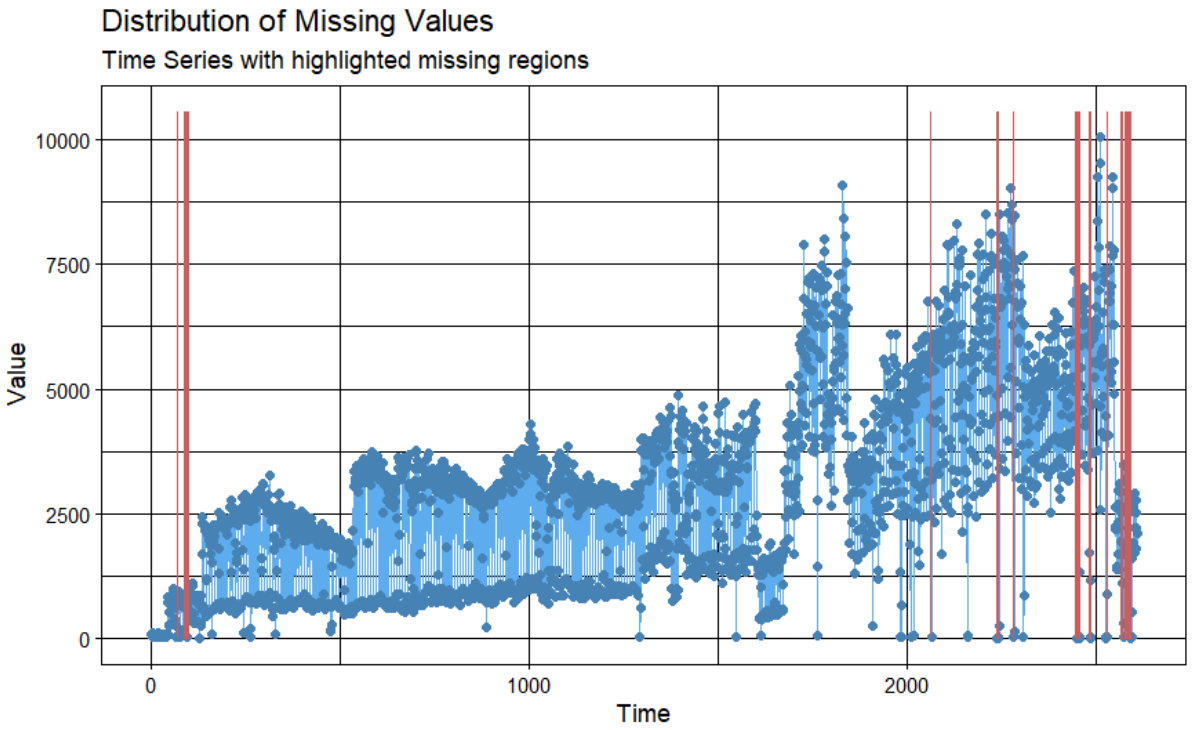
- There seems to be a dip in total accidents around June 2020, with a stark jump in September of 2020
- March of 2021 seems to return to a similar trend as the data before the 2020 jumps
- Finally, there is a steep decrease in accidents in 2023

Data Overview



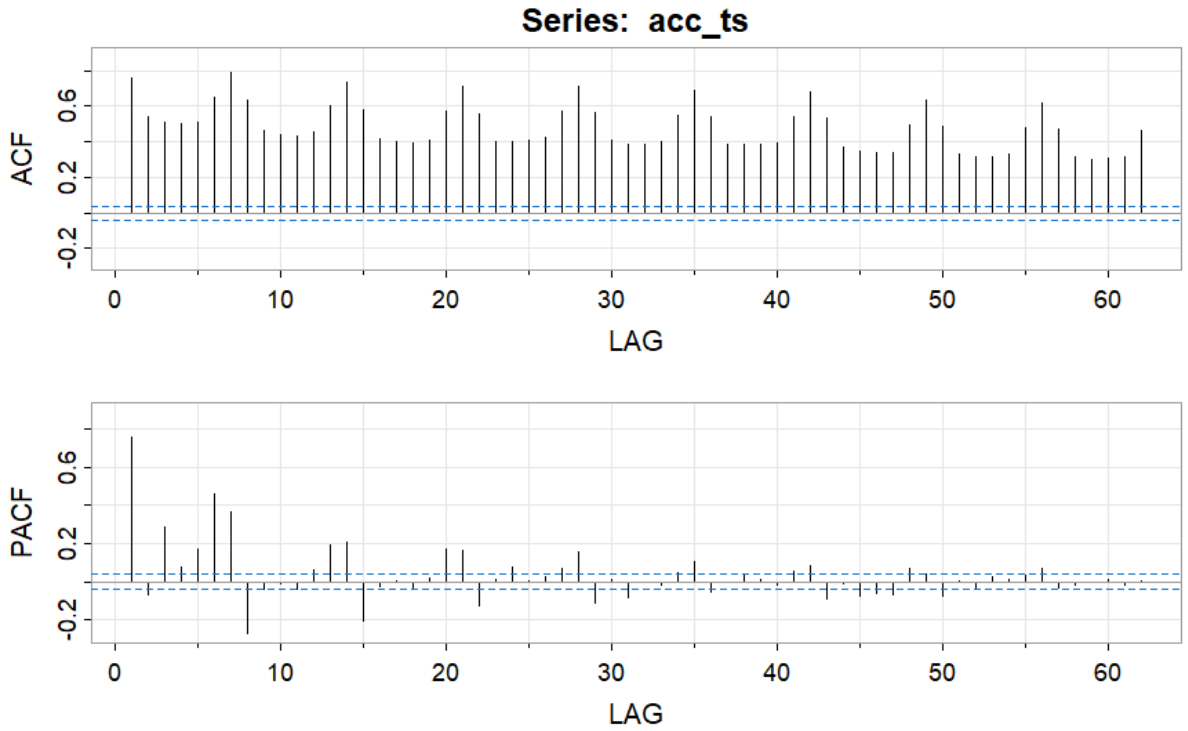
- Looking at the weekly patterns in March 2019, we see a lower amount of total accidents on the weekend than we do during the week

Initial Data Exploration



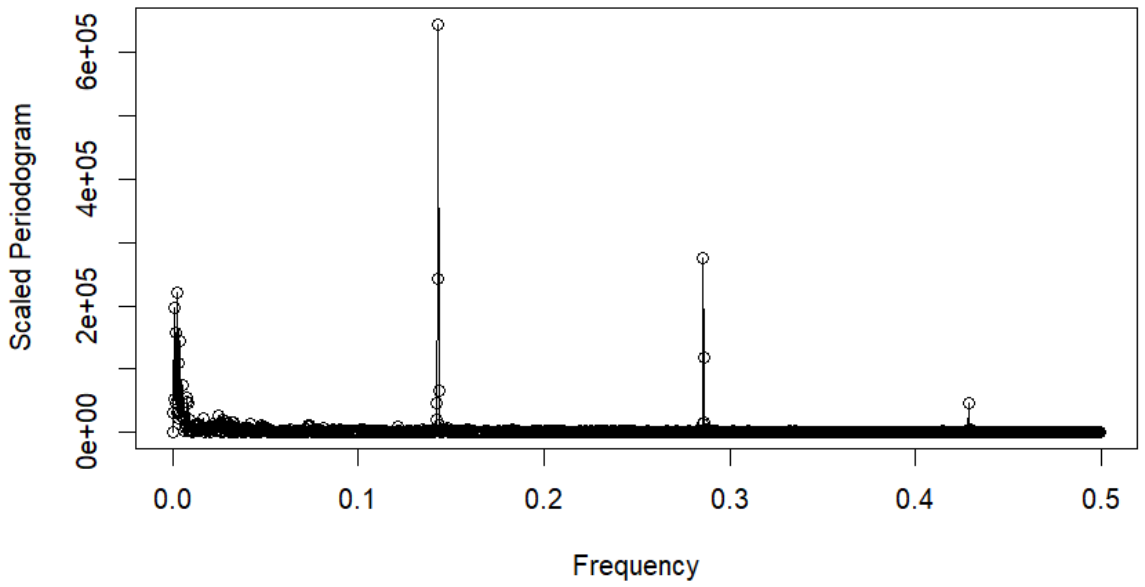
- 38 total missing values

Initial Data Exploration



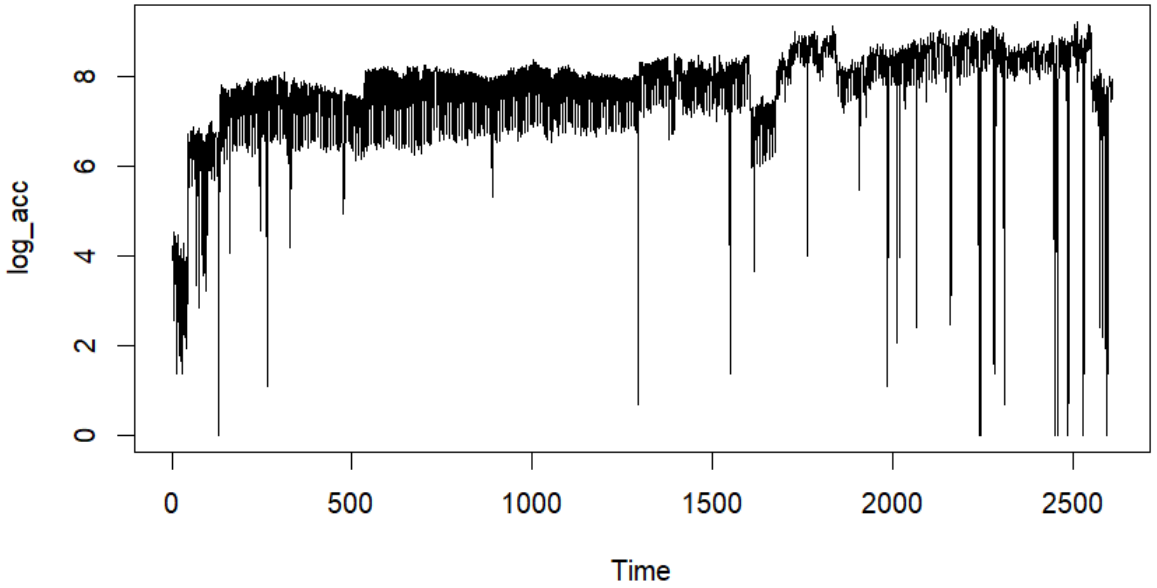
- Initial spike in PACF at Lag 1
- Spikes in ACF every 7 lags, indicating a weekly trend (seasonal trend of period 7)

Scaled Periodogram



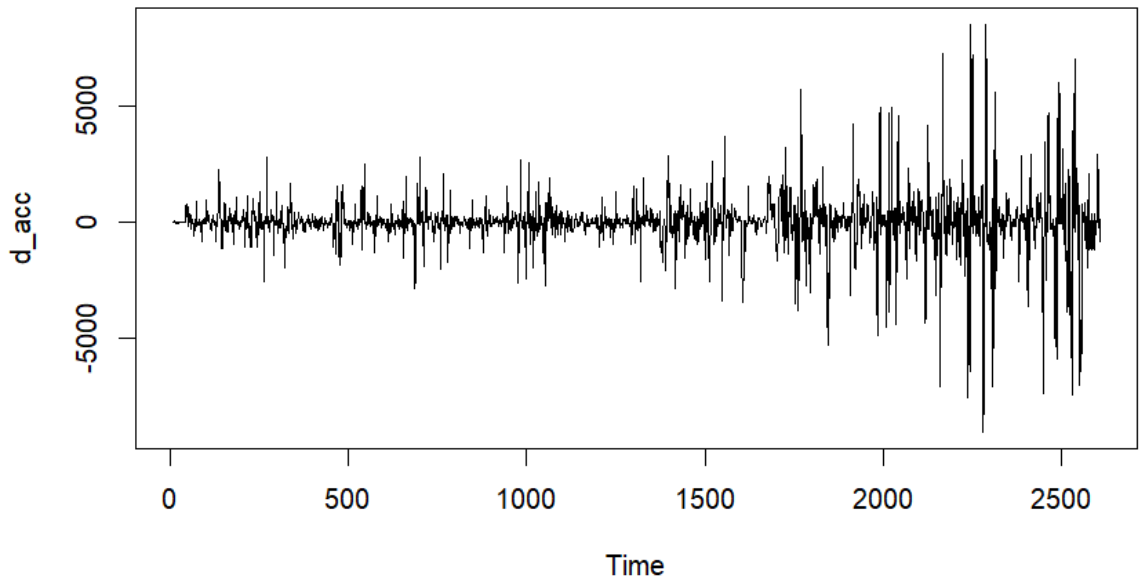
- Looking at the scaled periodogram, we see the frequency of 1/7 have the largest value, with 2/7 having the next highest, and so on
  - Indicates seasonality of period 7

Initial Data Exploration



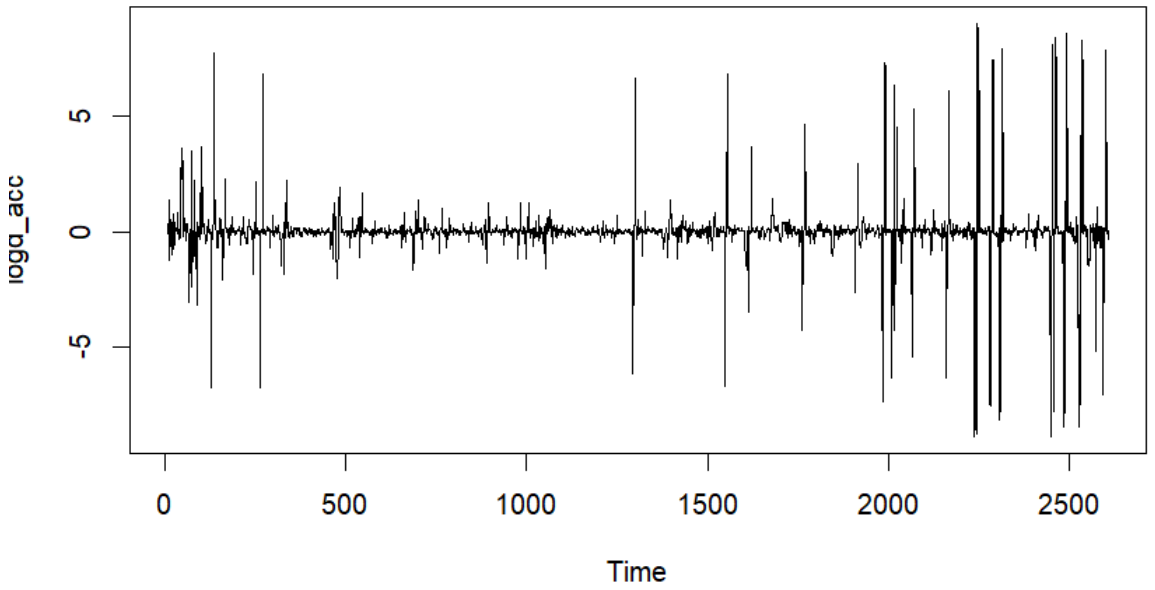
- Log transformation of data reduces variability by a large margin
- Will want to investigate the downward spikes, but a lot better time series plot than the original

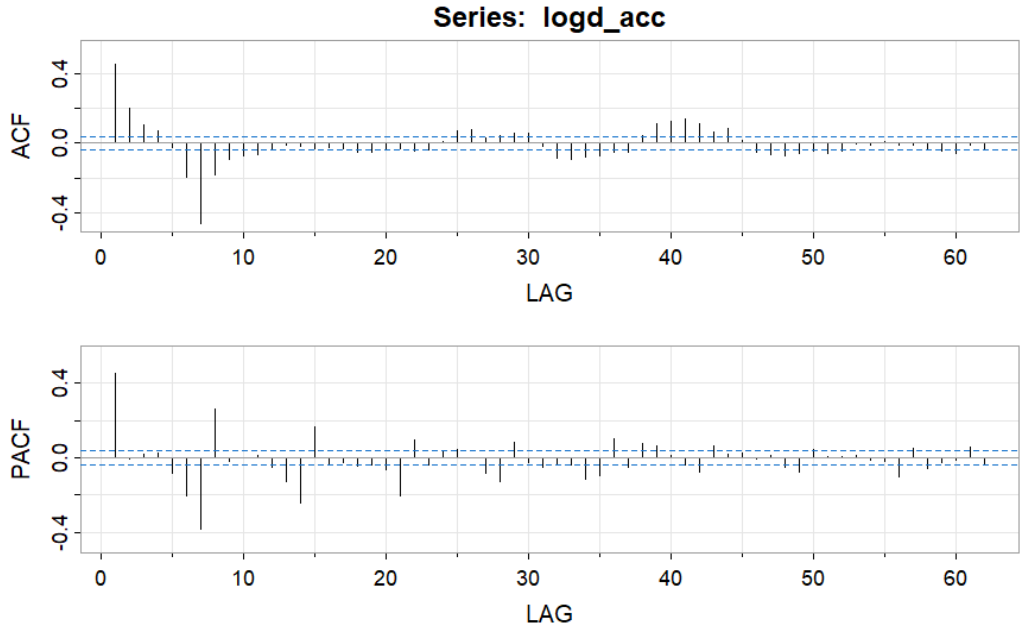
Initial Data Exploration



- Differenced data removes trend, but there is still a clear violation of constant variance
- We ran some initial SARIMA functions on this data, but further transformed data was used to find better SARIMA models

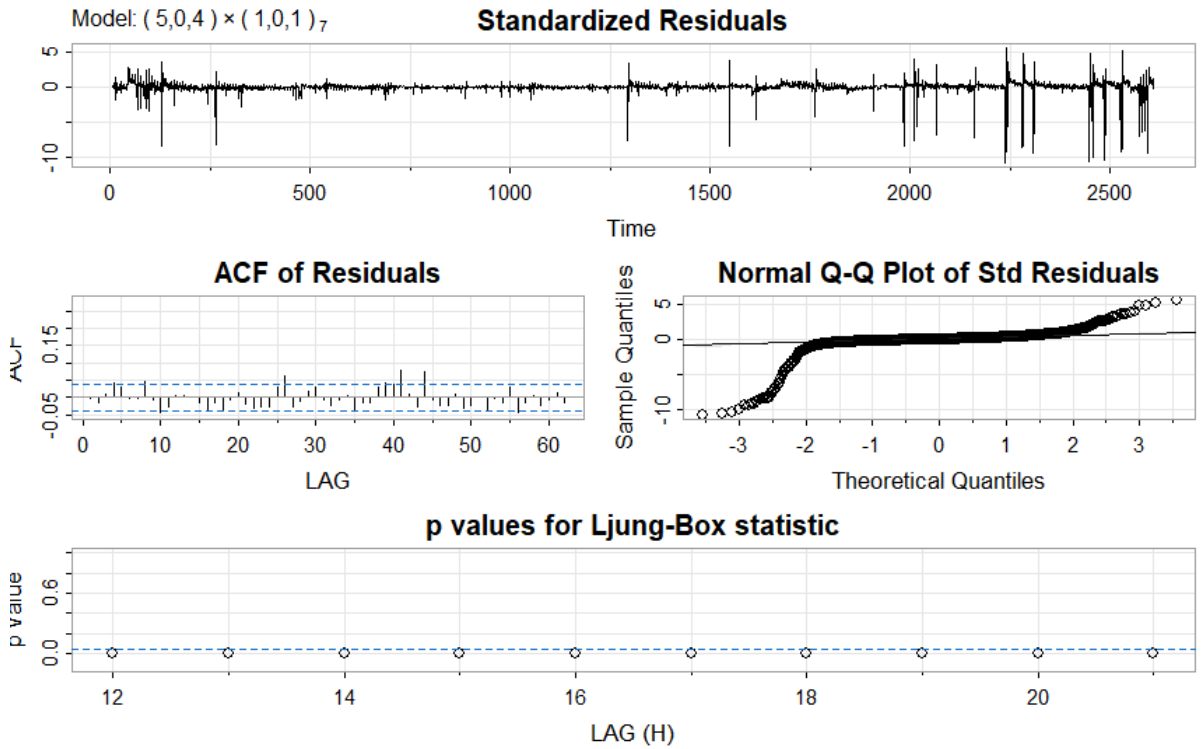
SARIMA Model





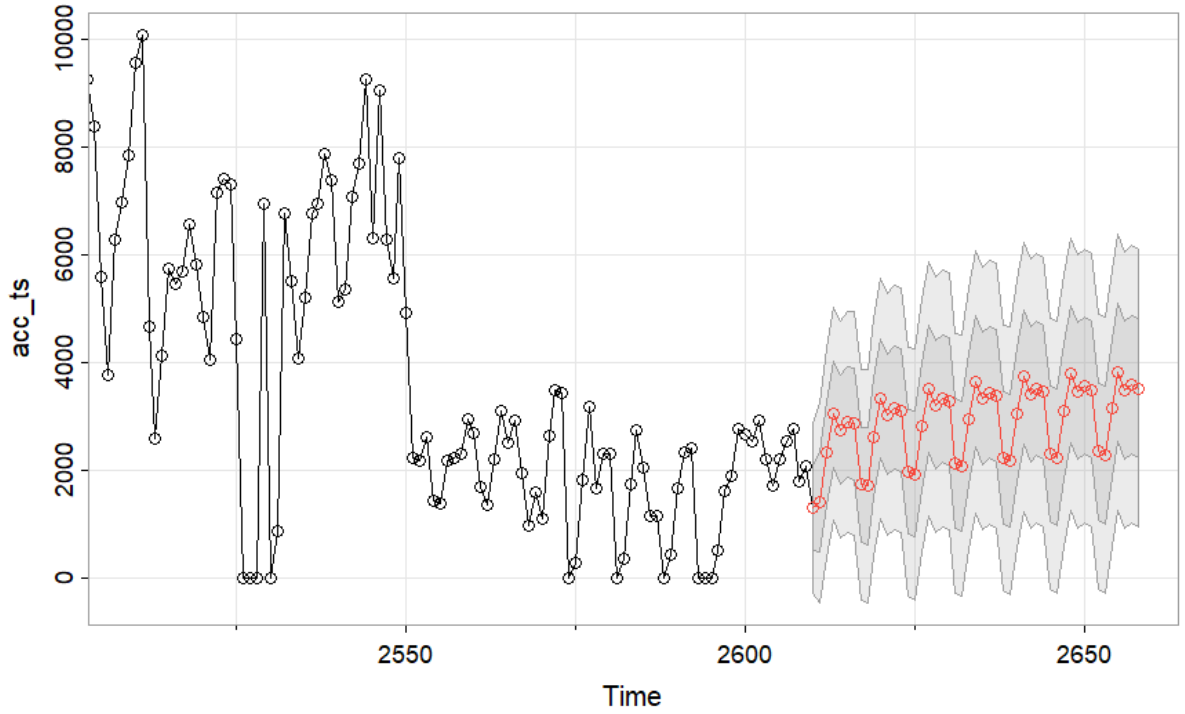
- Differenced the log time series at lag 7 to remove the trend and improve the data's stationarity
- The spikes in variability occur at the same places as before in the log plot, so investigating the previous spikes will help our understanding here too

SARIMA Model




```
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
```

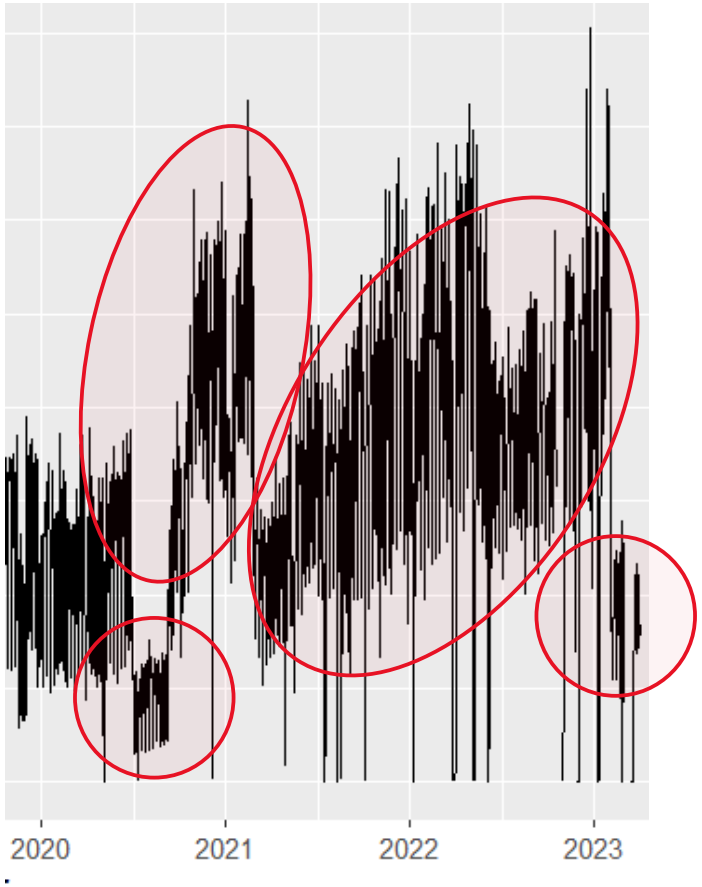
SARIMA Model: Forecasting



- Retains the weekly patterns

```
$pred
Time Series:
Start = 2610
End = 2658
Frequency = 1
 [1] 1302.649 1407.842 2319.887 3042.866 2753.589 2896.365 2866.070 1734.279 1707.528
2613.953 3328.112
[12] 3032.266 3162.496 3112.730 1971.895 1931.199 2820.977 3521.625 3214.922 3334.159
3274.723 2128.146
[23] 2078.861 2957.124 3647.392 3334.395 3446.169 3380.361 2231.577 2176.598 3046.313
3728.990 3412.196
[34] 3518.988 3449.074 2300.210 2241.582 3104.653 3781.549 3462.649 3566.048 3493.533
2345.987 2285.039
[45] 3142.731 3815.034 3495.136 3596.191 3522.059
```

Regression Model



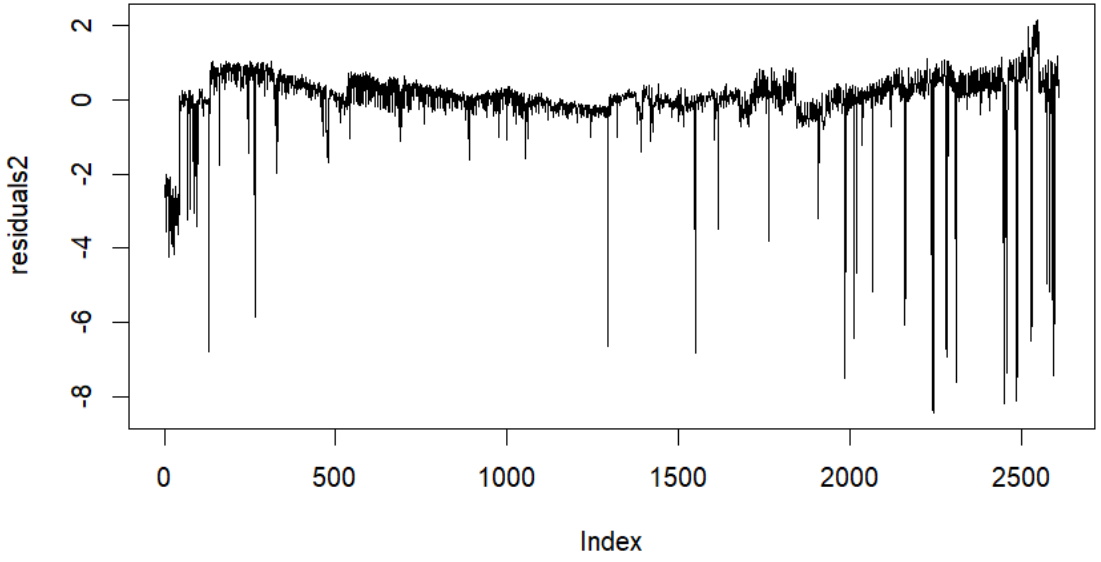
- These 4 circled sections of the plot seem to be different time series, so we created different indicators for them
- The first dip is presumable from COVID, although it's about 3 months after lockdowns started

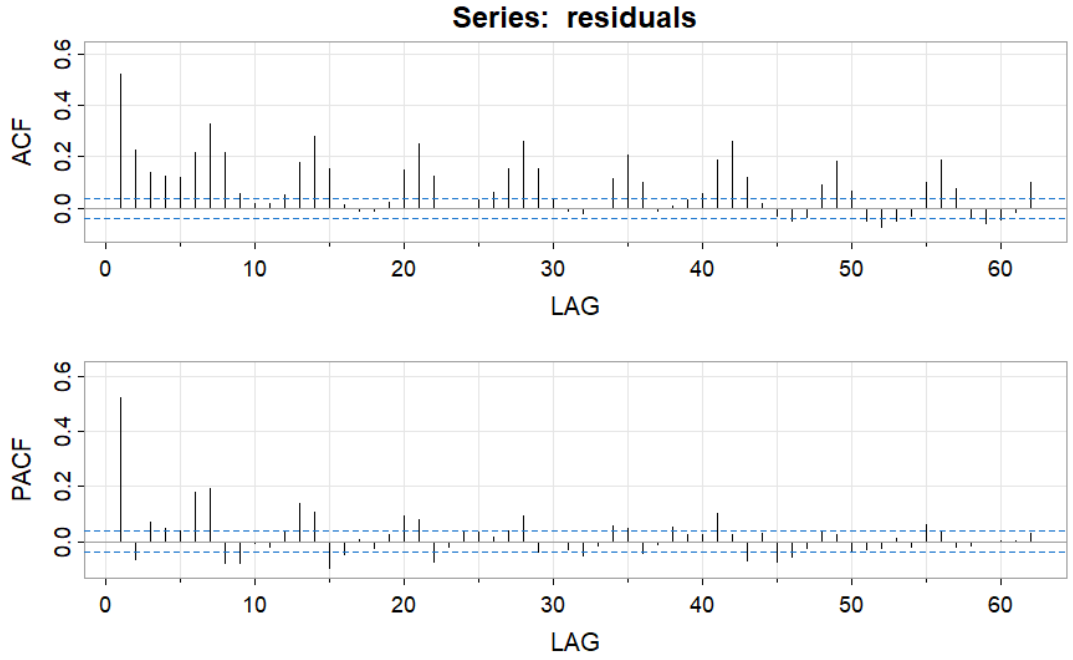
Regression Model

```
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
```

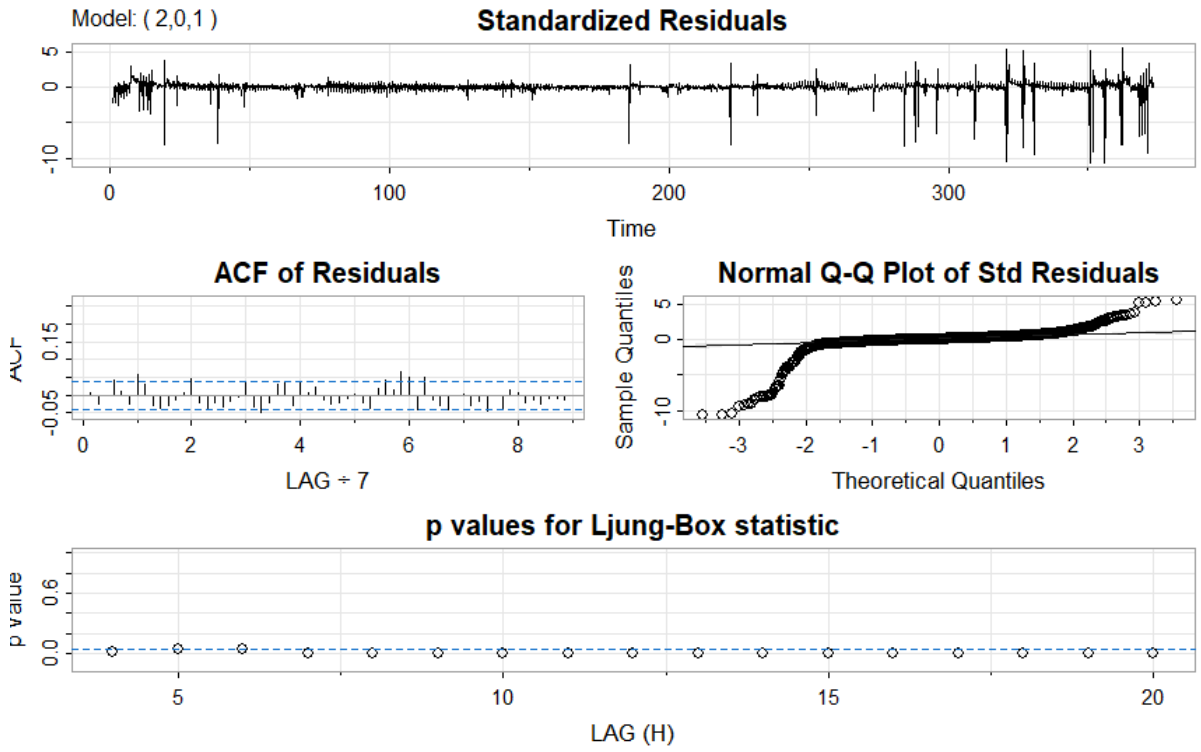
- We were interested in the following variables:
  - Time (centered)
  - Time squared (centered)
  - Indicators for day of the week (ind1 through ind6)
  - Whether the data was from post-June 2020
  - Whether the data was from post-September 2020
  - Whether the data was from post-January 2023

Regression Model



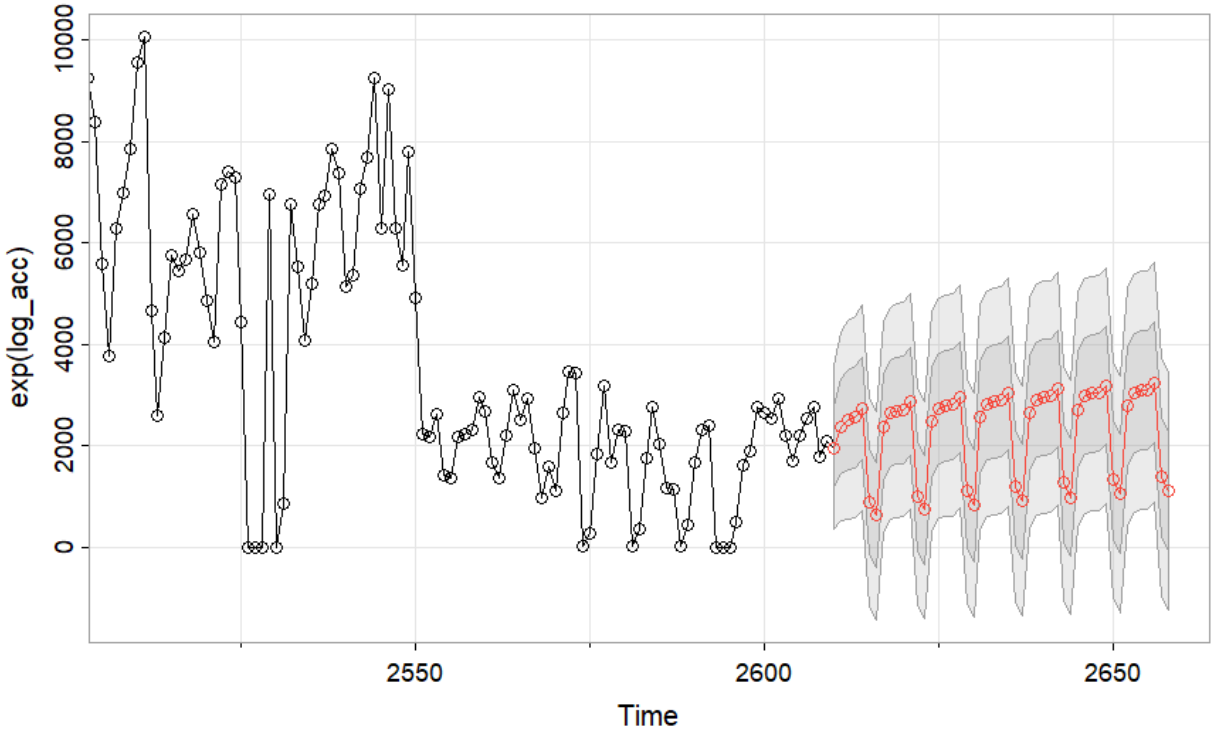


Regression Model



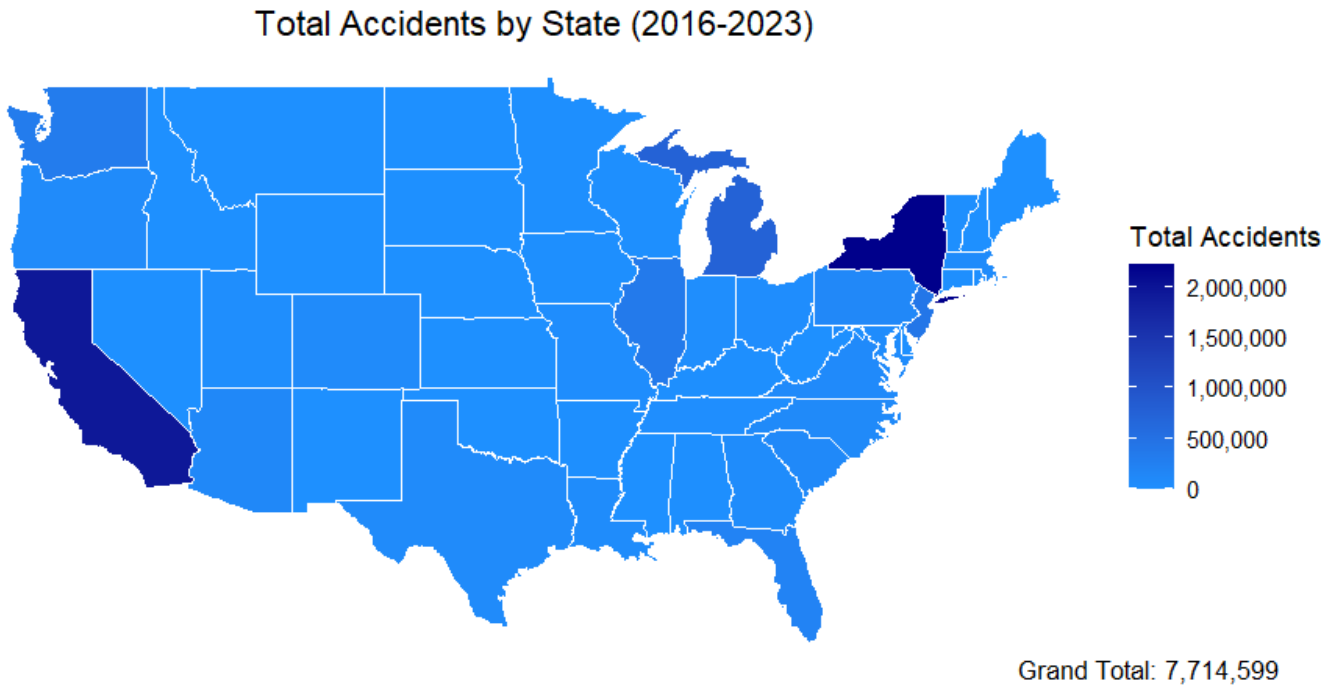

```
Coefficients:
           Estimate     SE   t.value p.value
ar1          1.4460 0.0195   74.2879  0.0000
ar2         -0.4470 0.0196  -22.8538  0.0000
ma1         -0.9603 0.0064 -149.1637  0.0000
intercept    7.0655 0.3186   22.1796  0.0000
tc           0.0008 0.0003    2.1947  0.0283
ind1_trim    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
```

Regression Model: Forecasting

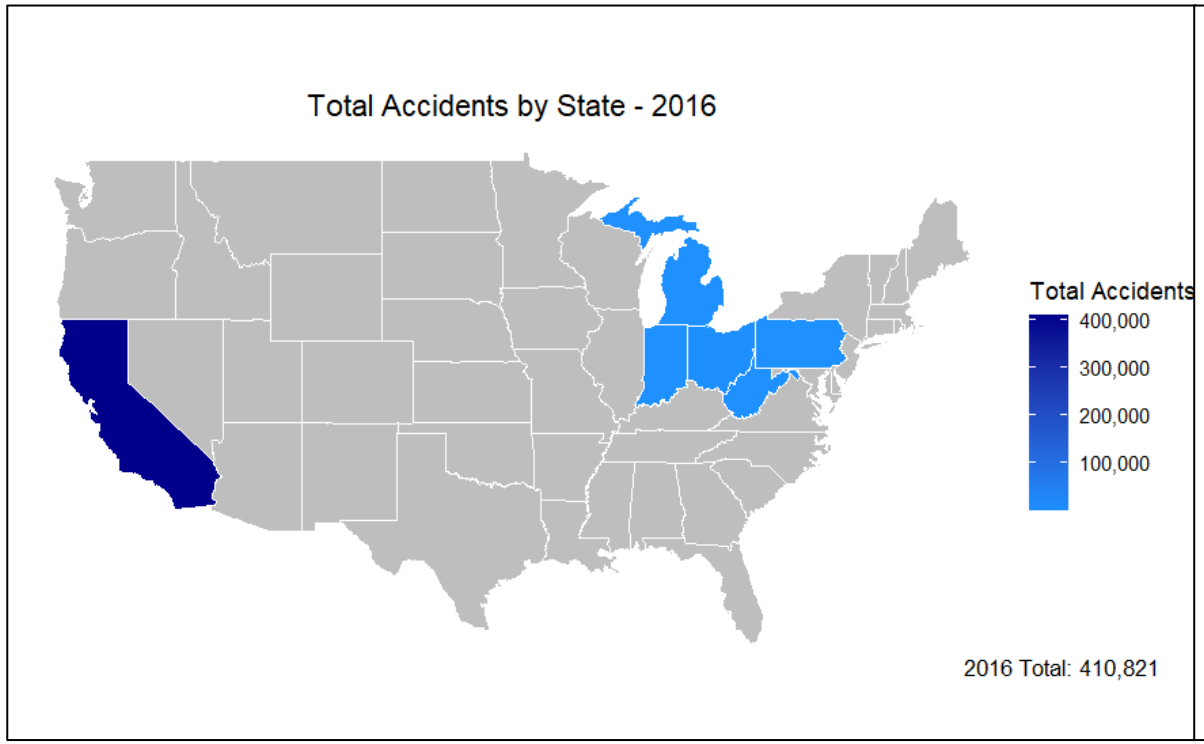


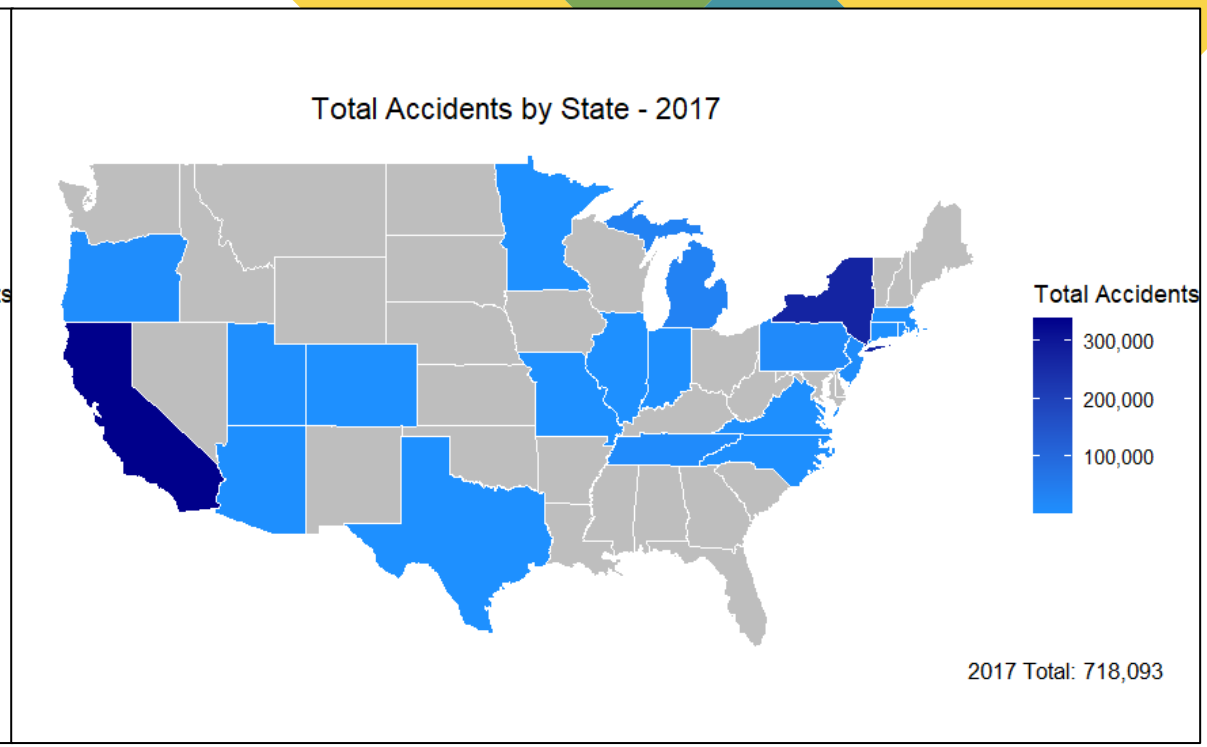

```
$pred
Time Series:
Start = 2610
End = 2658
Frequency = 1
 [1] 1962.7334 2379.3365 2506.6529 2565.6712 2737.2217  896.9726  625.3482 2367.7406
2637.7854 2690.6136
[11] 2711.3604 2862.8458 1011.6904  733.7736 2472.2041 2739.4703 2790.1267 2809.0189
2958.8210 1106.0802
[21]  826.6388 2563.5868 2829.4029 2878.6367 2896.1309 3044.5582 1190.4648  909.6926
2645.3308 2909.8579
[31] 2957.8230 2974.0685 3121.2669 1265.9639  984.0012 2718.4676 2981.8415 3028.6715
3043.7999 3189.8987
[41] 1333.5135 1050.4856 2783.9037 3046.2458 3092.0602 3106.1891 3251.3041 1393.9507
1109.9698
```

State Accident Trends

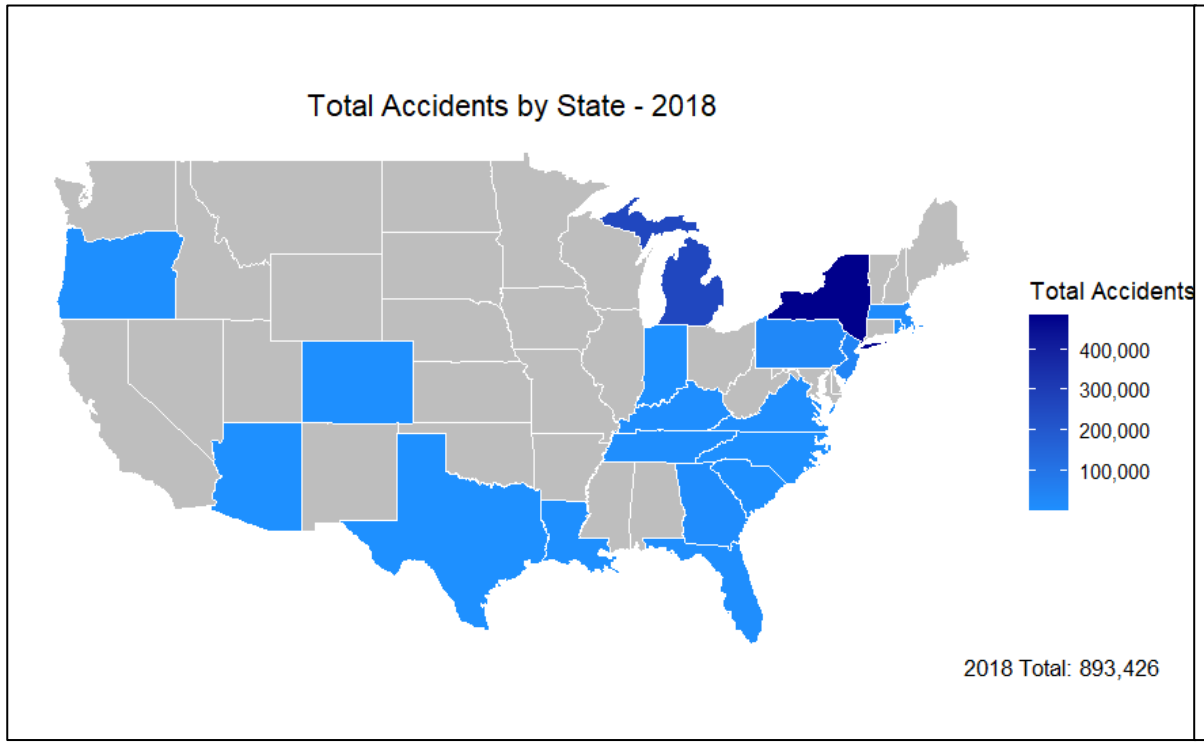


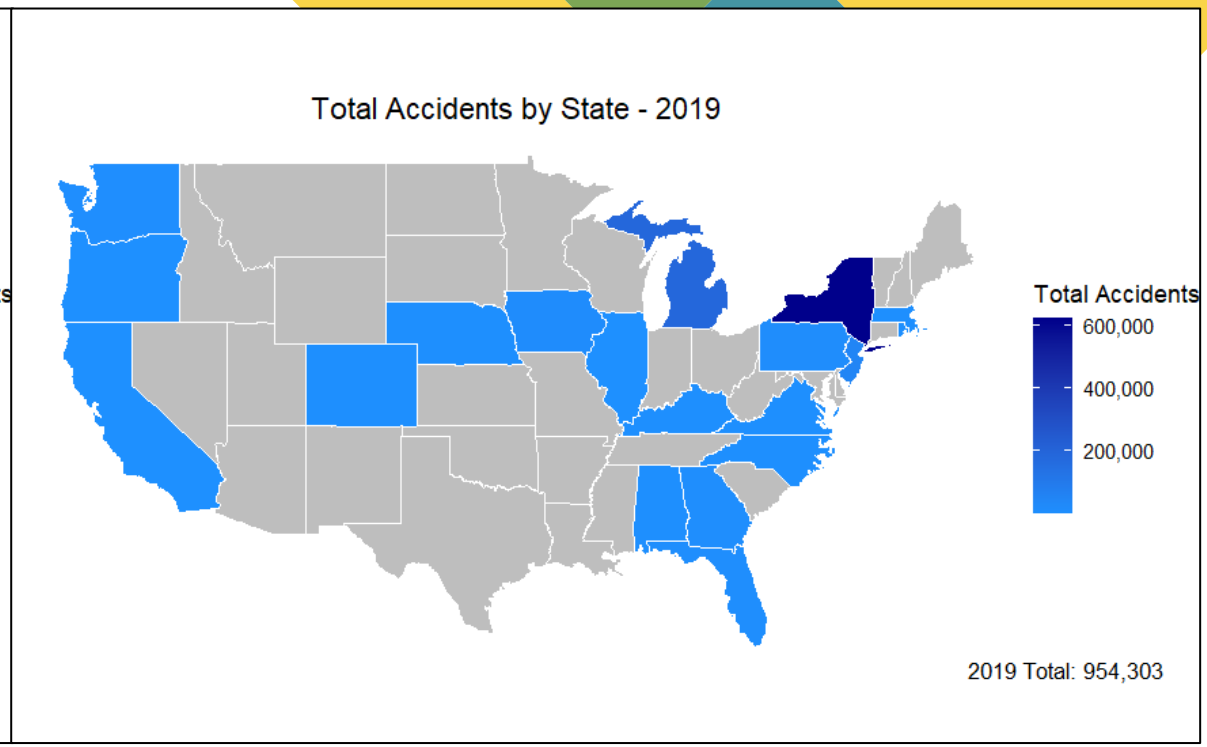
State Accident Trends



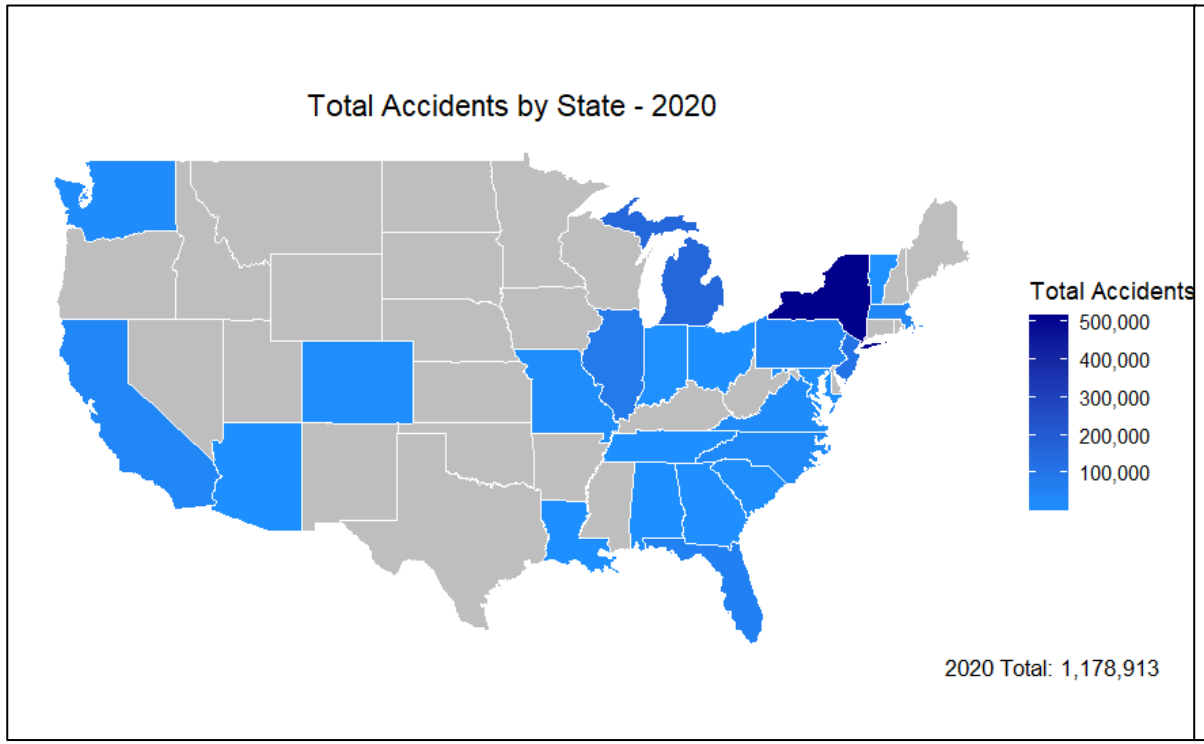


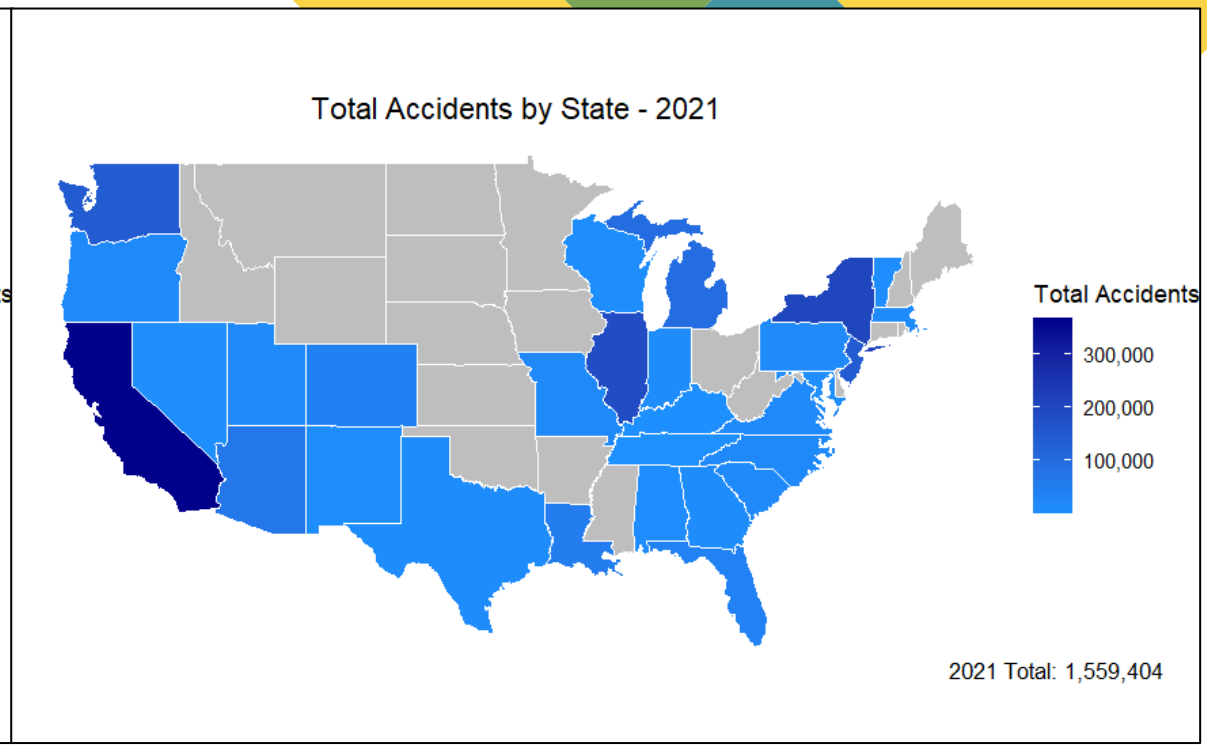
State Accident Trends



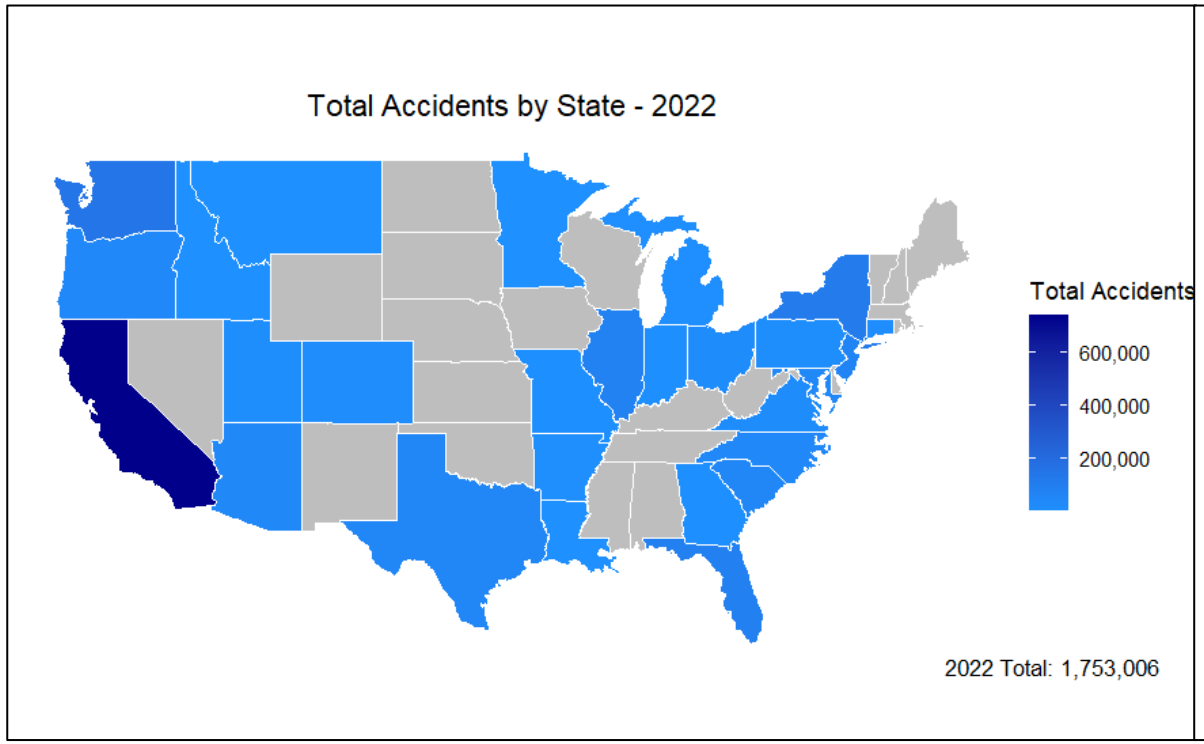


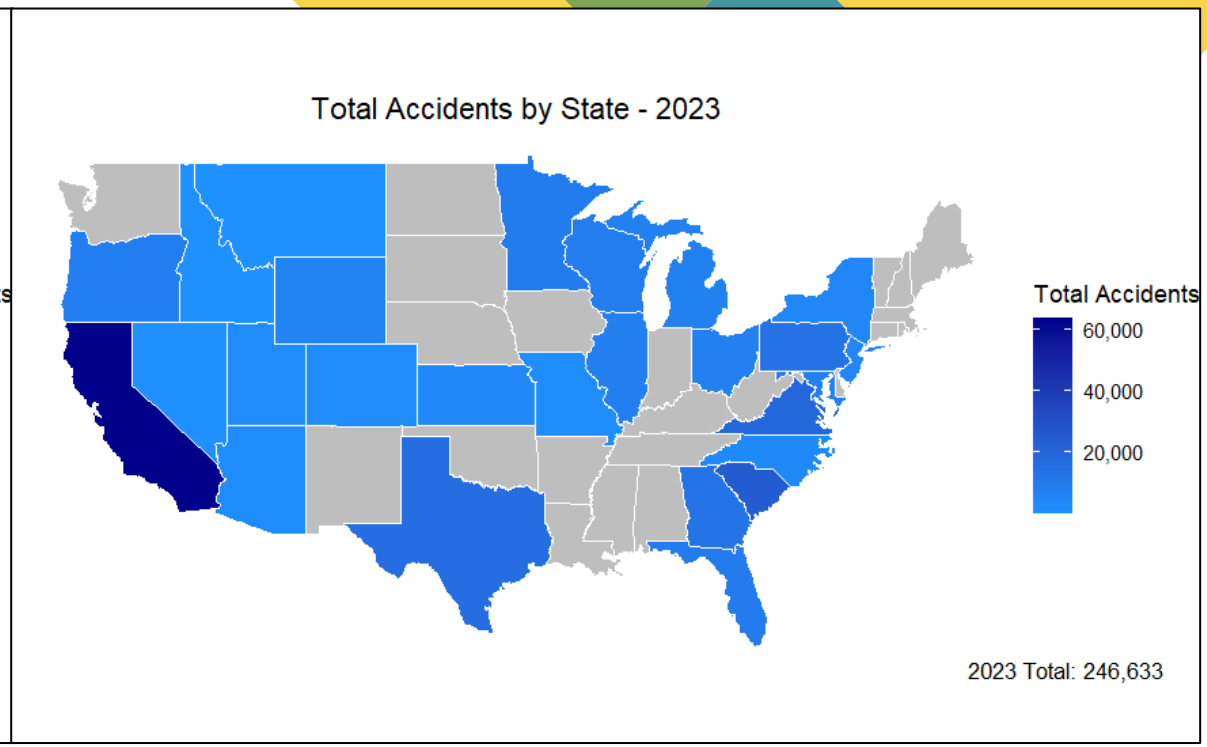
State Accident Trends





State Accident Trends





Limitations and Further Considerations

- Data collection had many dates with either no accidents or 1 accident (likely a placeholder) recorded
- Many states had years where no accidents were recorded
  - In fact, Delaware, Maine, Mississippi, New Hampshire, North Dakota, Oklahoma, and South Dakota all have 0 reported total accidents
- The many jumps in accidents resulted in very concerning Normal Q-Q plots
- If more time, run SARIMA model only on states where there is data for each year
- Explore more of the other predictors (like temperature, weather conditions, time zone, etc.) and their relationships with number of daily accidents, as well as interactions between predictors

Conclusion

- The SARIMA (5,0,4,1,0,1) model with s=7 and no indicators was the best at predicting total daily accidents
- Weekdays tend to have more accidents than the weekend
- As time goes on and driving becomes more popular, more accidents will tend to occur
- 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

Thank you

Any questions?