Traffic crashes and injuries forecast

Time Series Final Project

Team 5

05/22/2024

Wednesday

Aayush Verma

Ashmita Mukherjee

Mitali Dighe

Jahanvi Reddy

Agenda

- Problem Statement
- Business Objective
- Data Overview
- Exploratory Data Analysis
- Experimental Analysis
- Modeling Crashes and Injuries
- Model Selection
- Results
- Future Scope

Problem Statement

- Traffic collisions in urban areas pose significant challenges to public safety and place a heavy burden on emergency response systems
- To enhance public safety and optimize resource allocation, this project aims to analyze historical traffic collision data and the injuries caused due to these crashes

Business Objective

- The primary goal of this project is to develop forecasting models that can accurately forecast the frequency of traffic collisions and accidents
- We also want to predict injuries along with the traffic crashes
- These predictions will assist law enforcement agencies and healthcare facilities in anticipating periods of increased workload, thereby enabling better preparedness and response strategies

Data Overview

The dataset is recent and has crash and injuries data with timestamps from 1st January 2021 to 30th April 2024

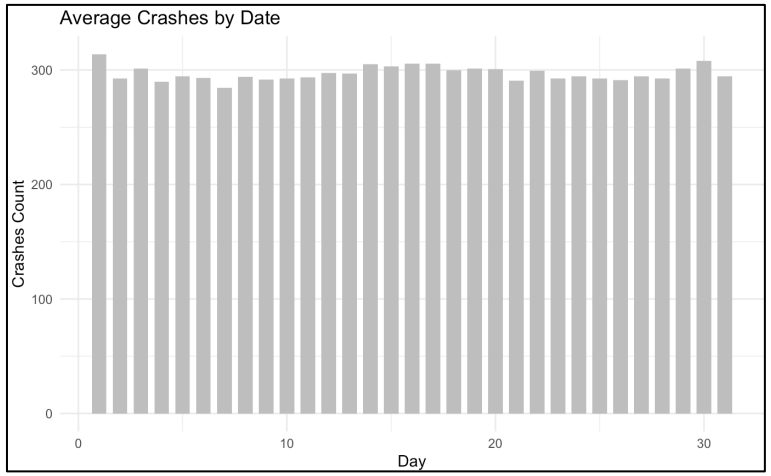
We downsampled the data and analyzed by creating the time series in 3 sampling rates: Monthly, Weekly and Daily

Daily sampling is the most beneficial way to go according to our use case of aiding law enforcement and medical professionals

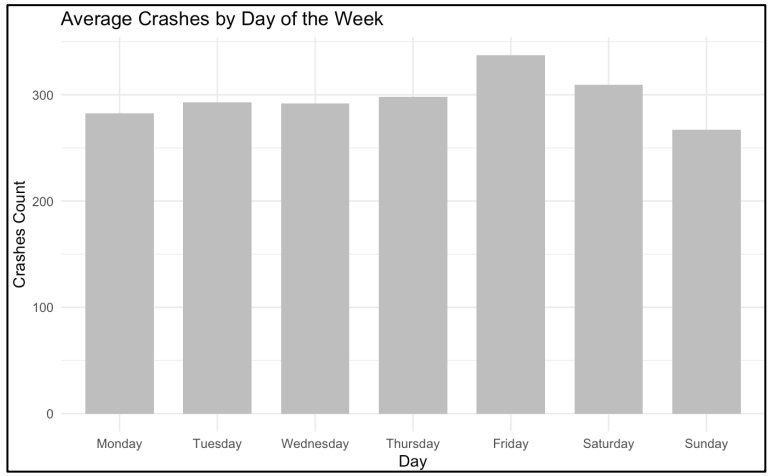
Exploratory Data Analysis

Day

- On an average, 285 to 315 traffic crashes are observed
- Higher number of crashes are observed on 30th and 1st of the months
- This could be because more people go out at the beginning or end of the month



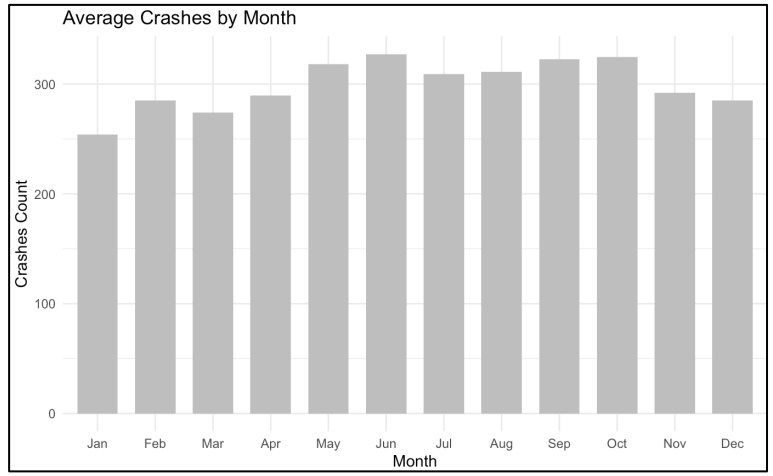
Week



- On an average, 267 to 337 traffic crashes are observed
- Highest traffic crashes are observed on Fridays and Saturdays
- This could be because more people go out on weekends (Friday and Saturday nights)

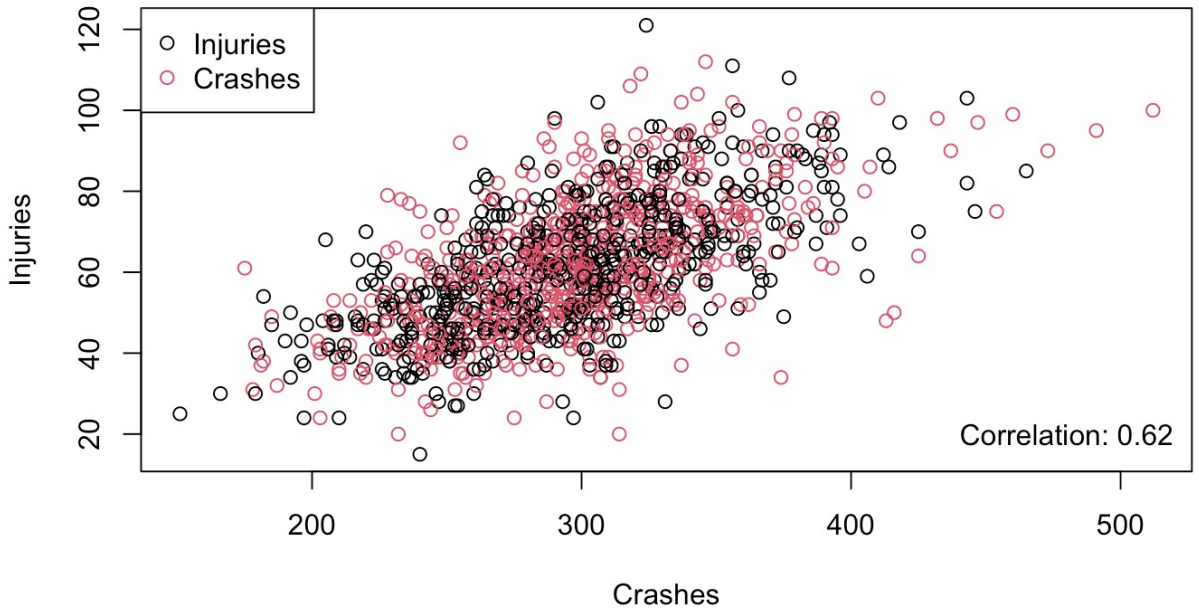
Month

- On an average, 254 to 327 traffic crashes are observed
- Highest traffic crashes are observed in the month of June
- This could be because more people go out in summers



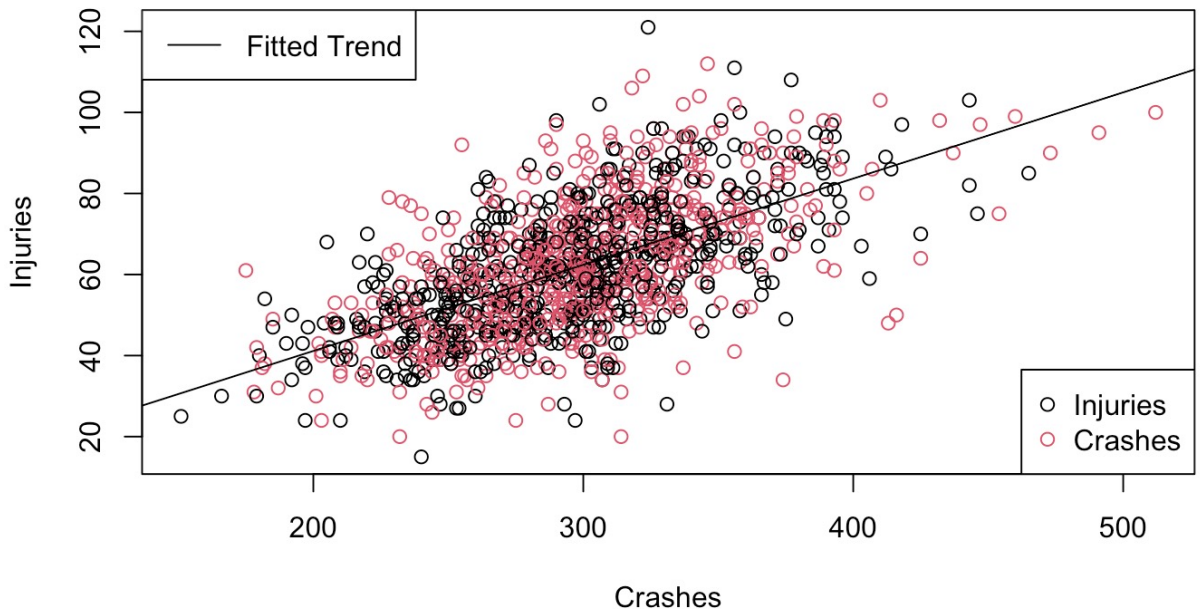
Exploratory Data Analysis

Crashes vs Injuries - Correlation



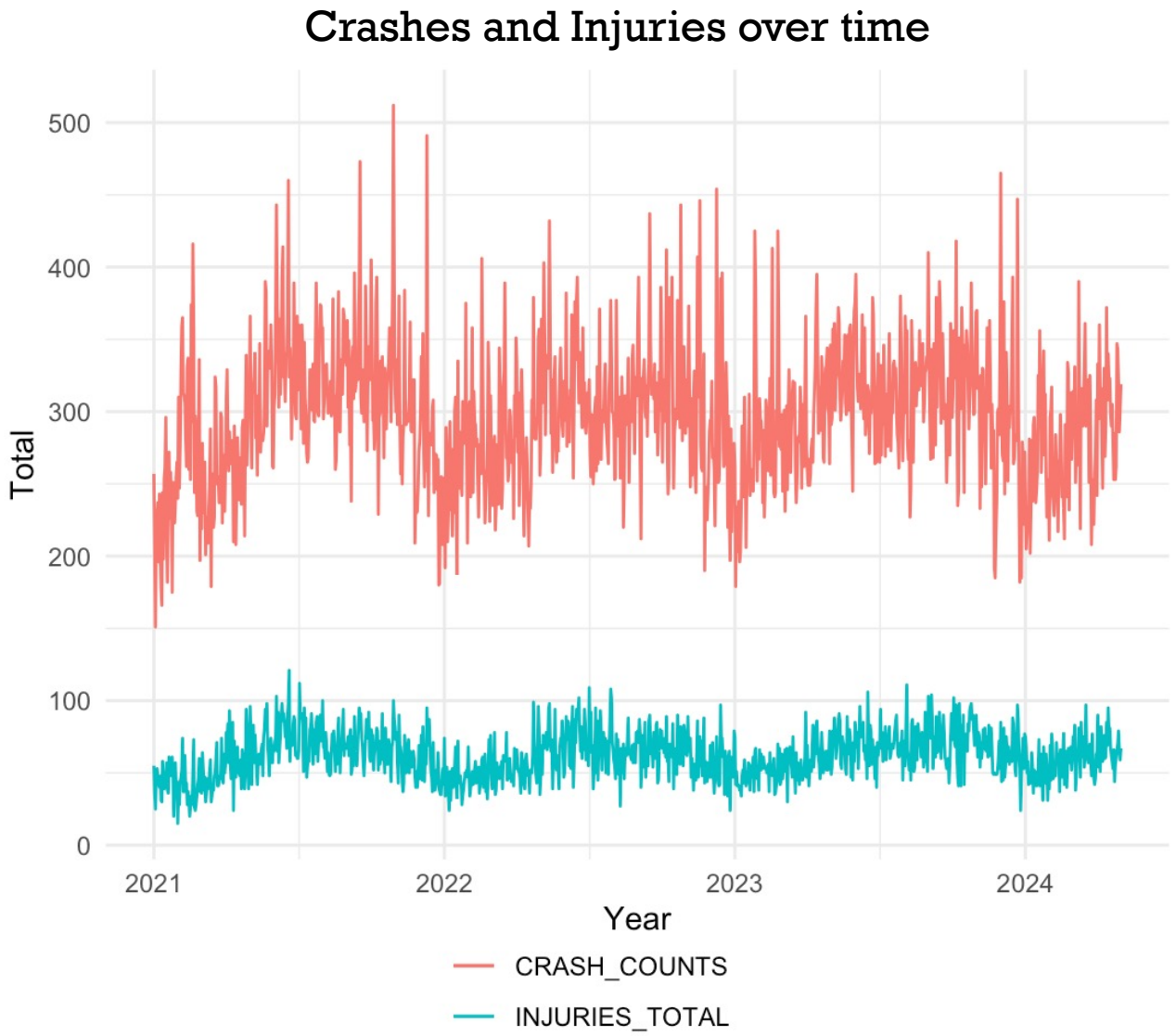
- Injuries have a positive correlation with crashes
- We can see the correlation between crashes and injuries is ~0.62

Crashes vs Injuries – Fitted Trend



- A simple linear regression model to fit a trend
- Observe the data along with the fitted trend

Exploratory Data Analysis



There is no specific trend. However, we can see seasonality in both crashes and injuries

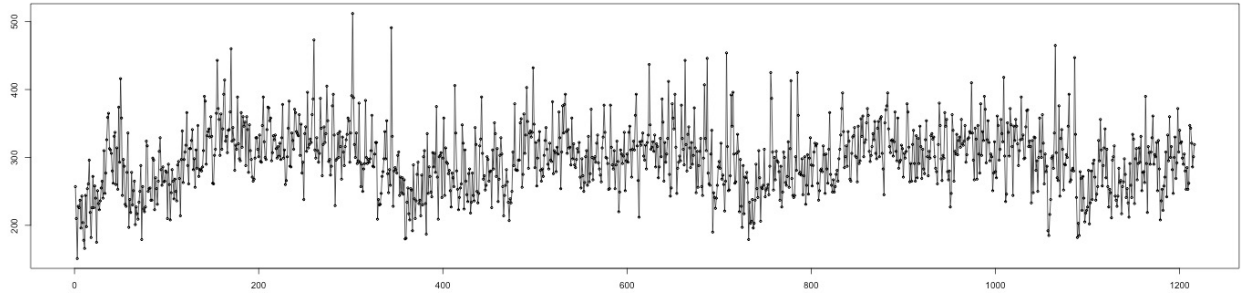
Time Period: January 21 to April 2024. We can see volatility and fluctuations in data

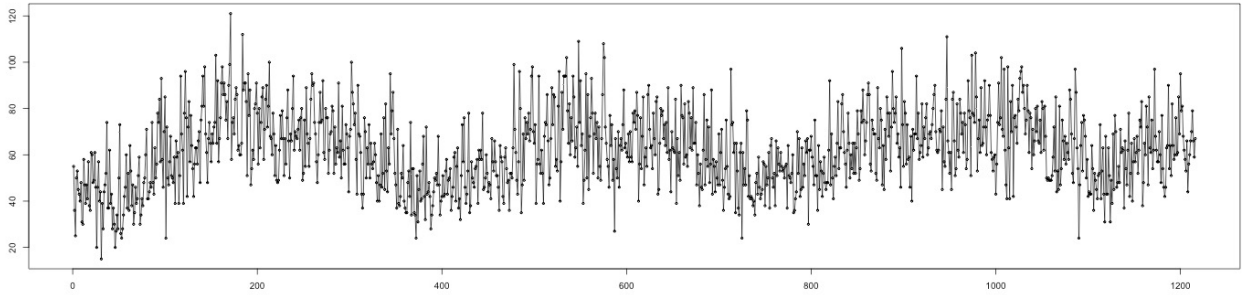
Within the year, we can see there could be a possibility of multiple seasonal components

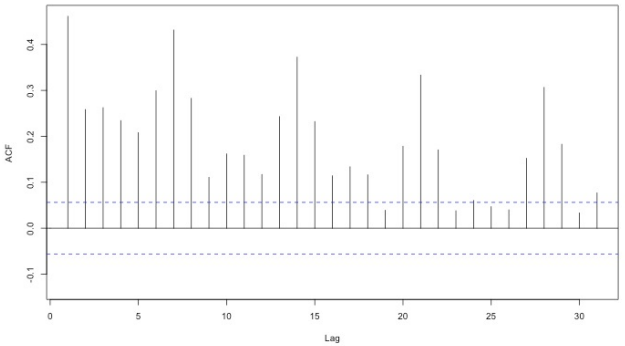
Periodograms can help us identify the potential multiple seasonal components

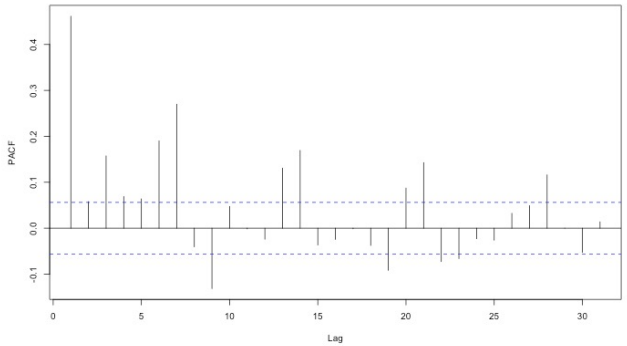
Experimental Analysis

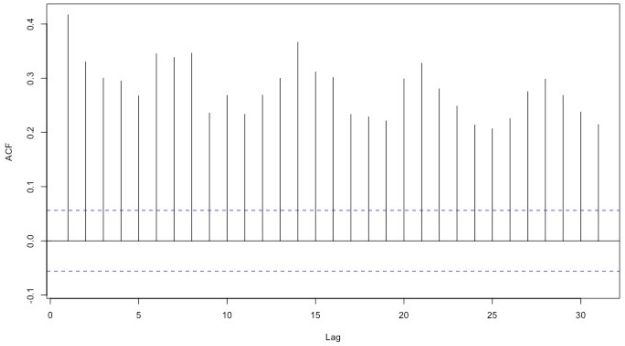
To analyze trends and patterns, we converted the crashes and injuries data into time series objects. The below ACF and PACF plots are observed.

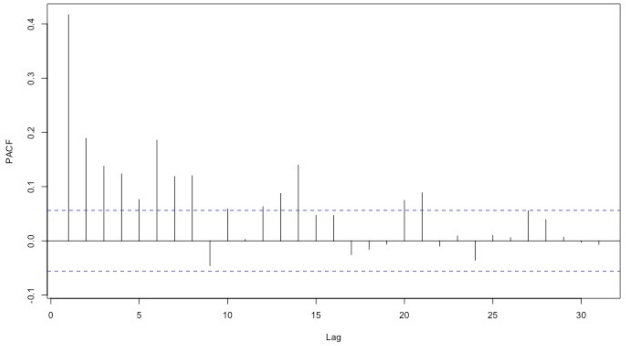








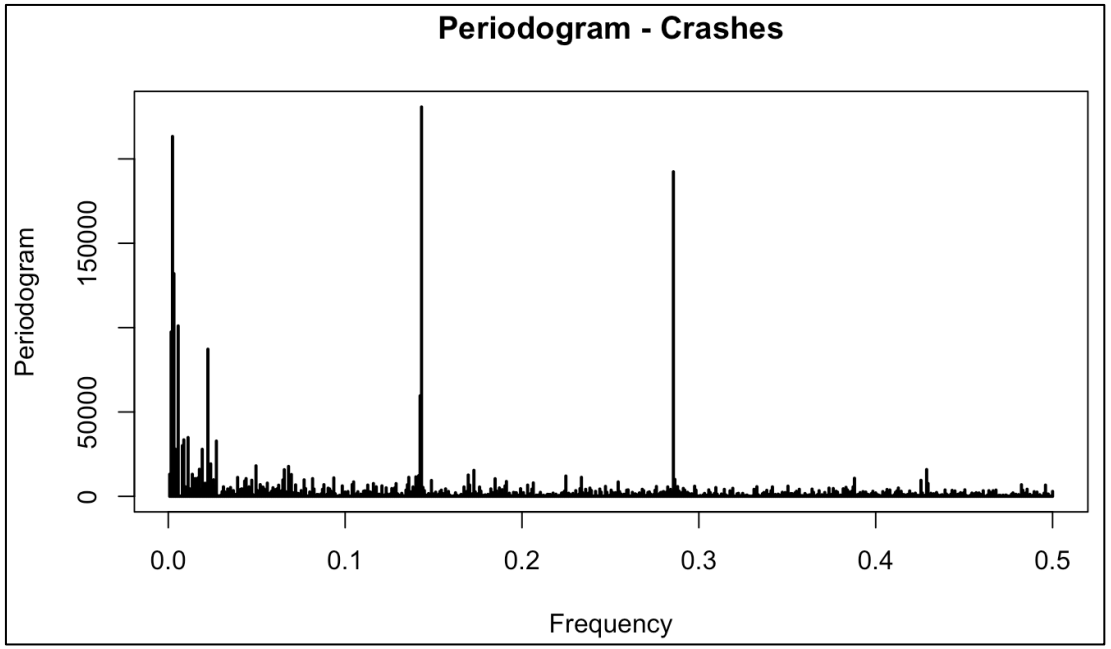




For both crashes and injuries, the time series are clearly not stationary with multiple significant lags. It would require us to do multiple seasonal differencing to achieve stationarity. The ACF plot is dragging and the PACF also shows multiple significant lags.

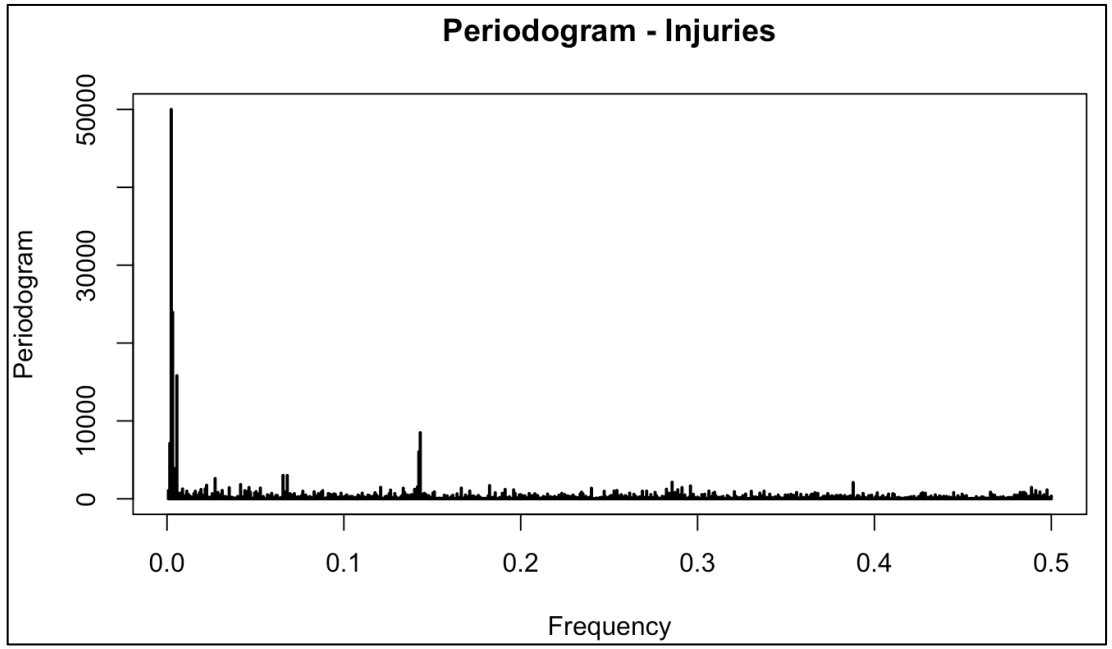
Experimental Analysis

By conducting an exhaustive spectral analysis of the injuries and crashes data, multiple seasonal components have been observed in both the time series.



4 Seasonal components:

5, 7, 183, 365 (Days)

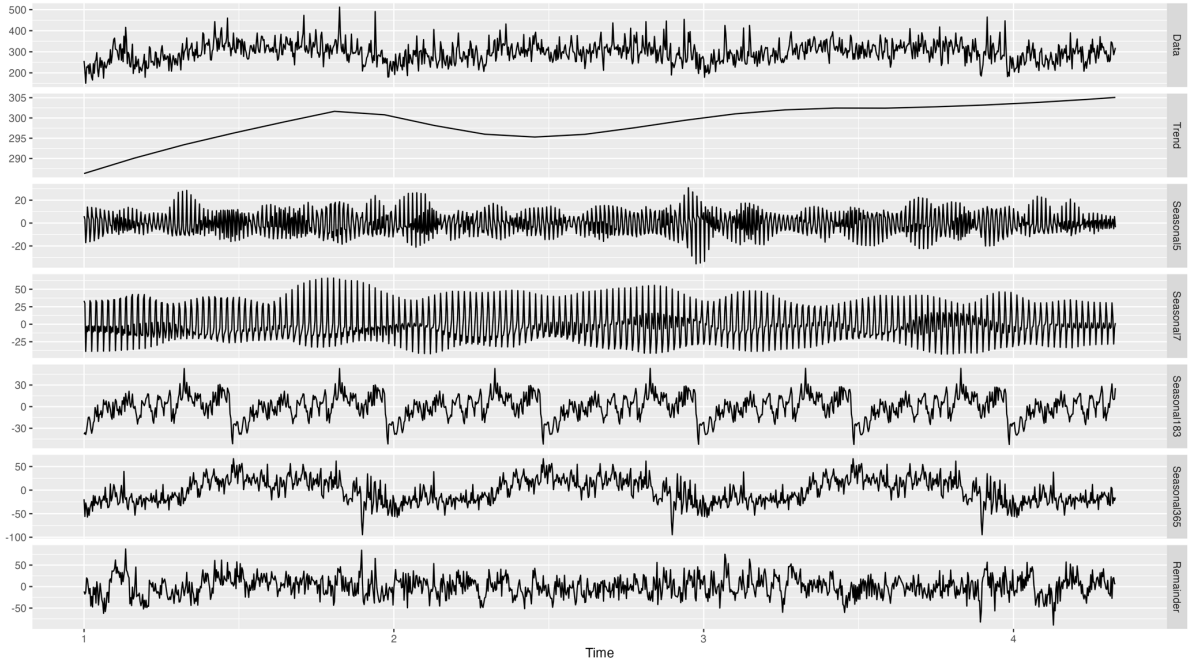


2 Seasonal components:

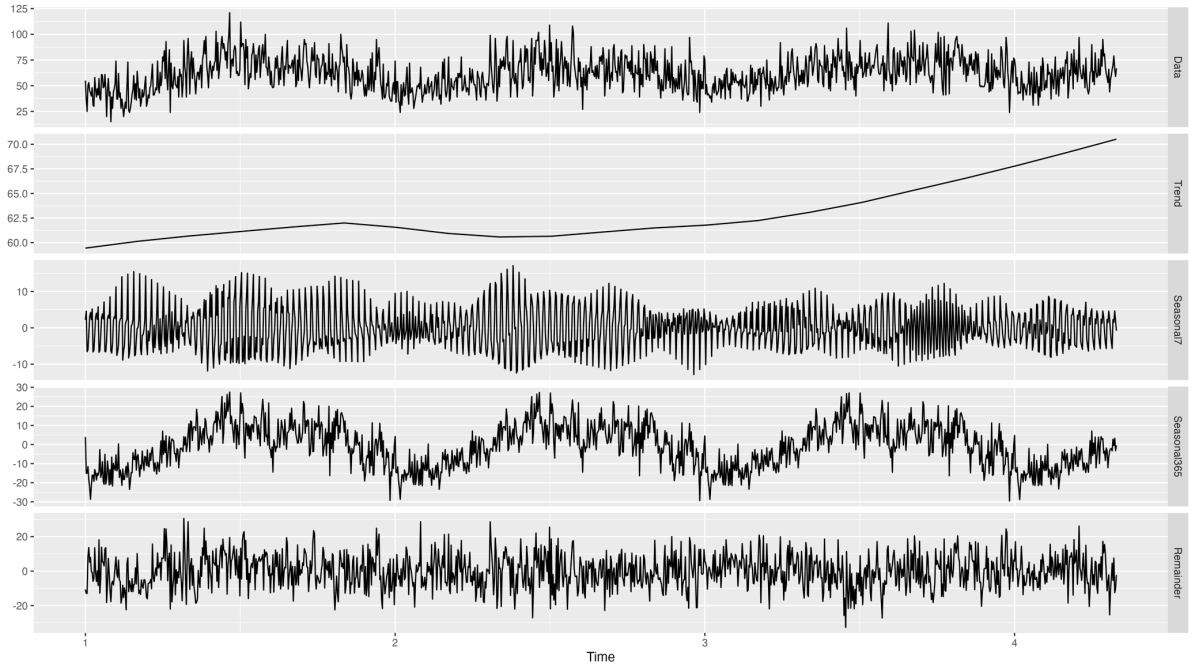
7, 365 (Days)

Experimental Analysis

To understand the underlying patterns in our data, we performed a multi-seasonal time series decomposition using the mstl() function for both injuries and crashes into seasonal, trend and remainder components.



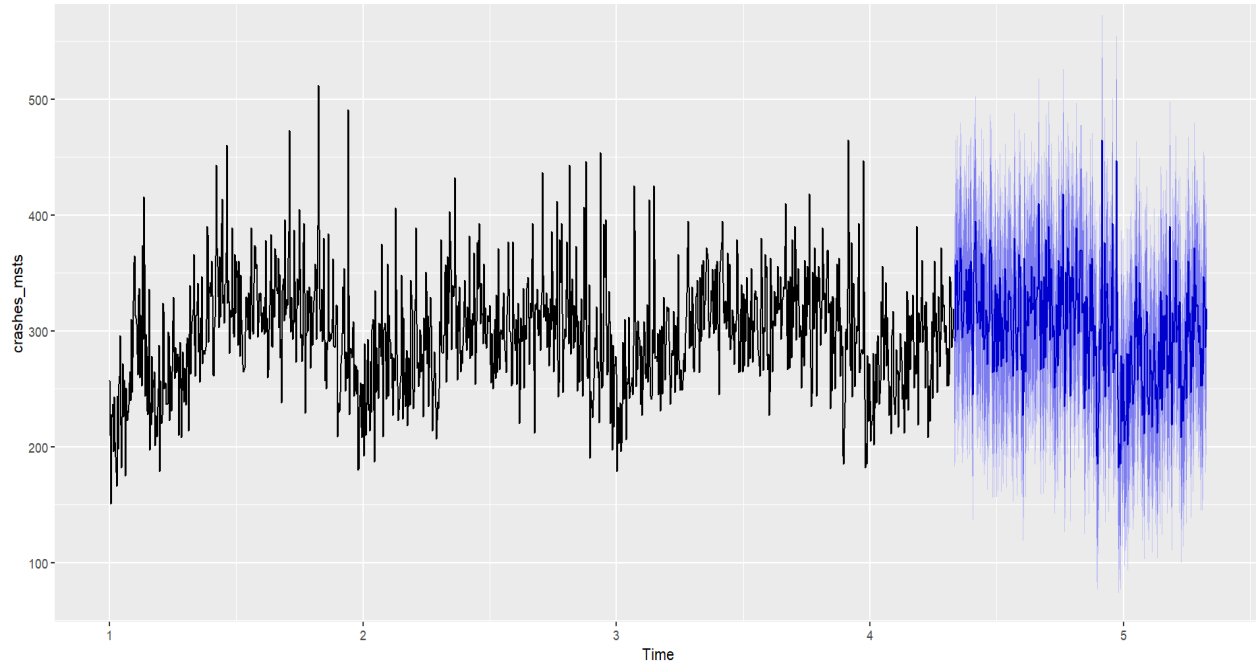
CRASHES: Identified multiple seasonality, including 5-day, weekly, biannual and yearly patterns

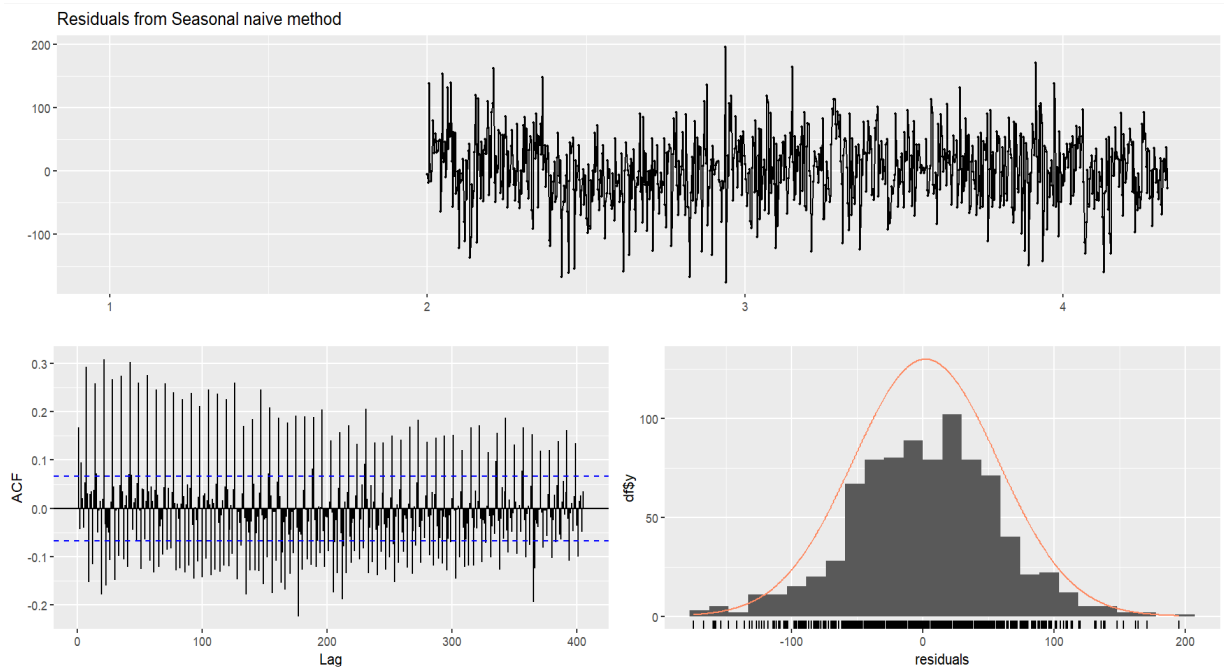


INJURIES: Identified weekly and yearly seasonal patterns.

Modeling - Seasonal Naive

A seasonal naive model has been developed as a benchmark model to evaluate the trends for a 365-day forecast.



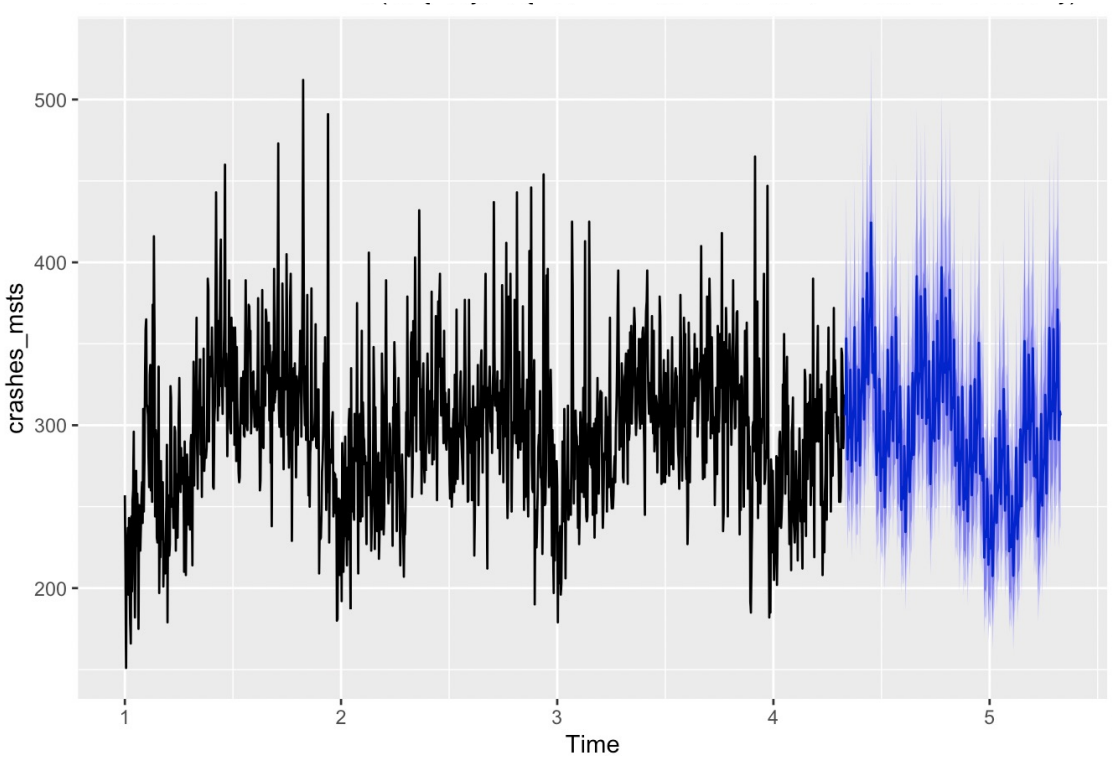


The forecasts are very jittery and do not seem to capture the variation in the data. They do capture some of the seasonality in the data though, letting it act as a valid benchmark.

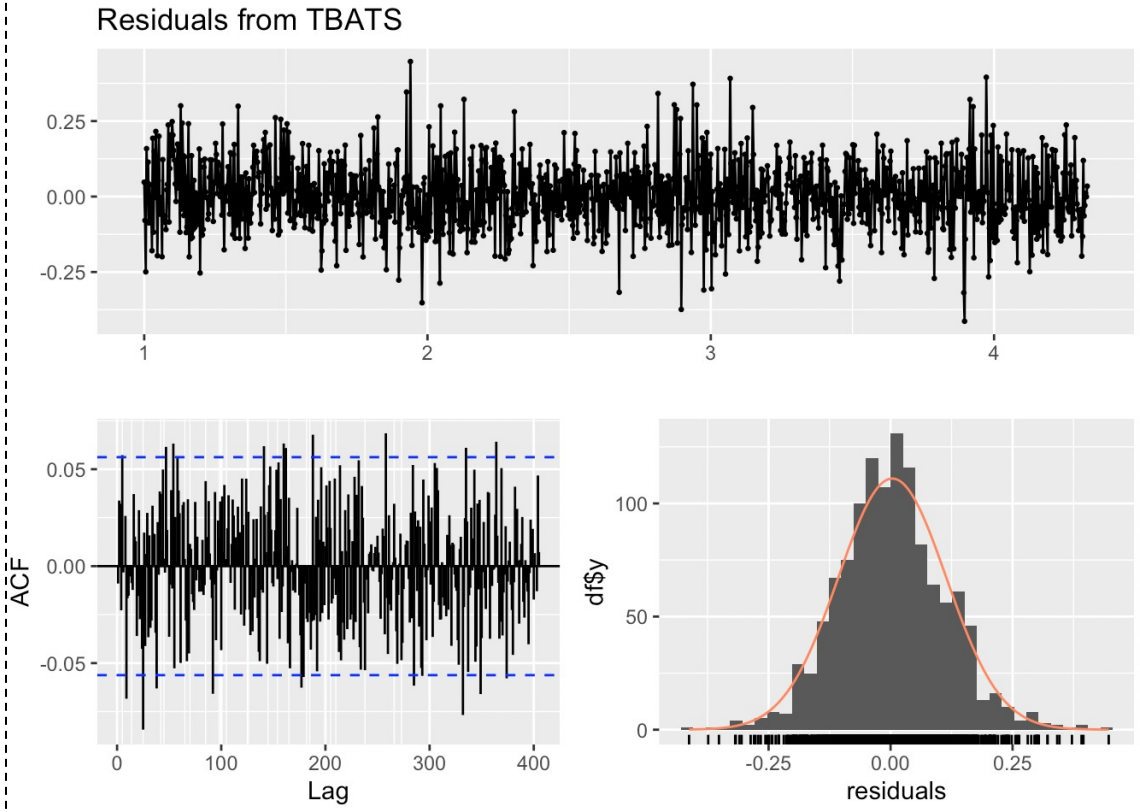
The model shows moderate errors and despite a low ACF, it fails the Ljung-Box test (p-value < 2.2e-16), suggesting unreliable predictions. A high positive RMSE of 55.22 shows a scope of improvement with model refinement.

Modeling - TBATS

We fit a TBATS model considering all the seasonal patterns discussed and forecast for a 1-year period. The fitted model is TBATS (0, {1,0}, - , {<5,2>,<7,3>,<183,5>,<365,3>})



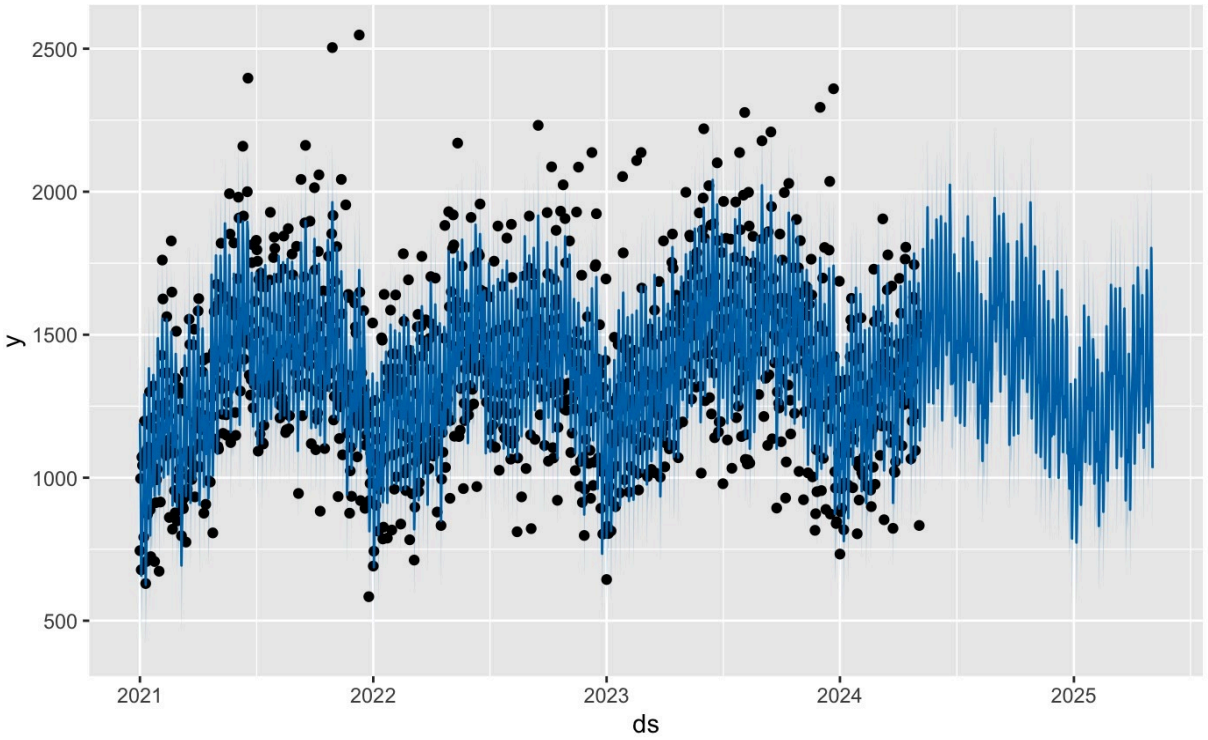
The forecast looks like it is able to capture the seasonality in the data.



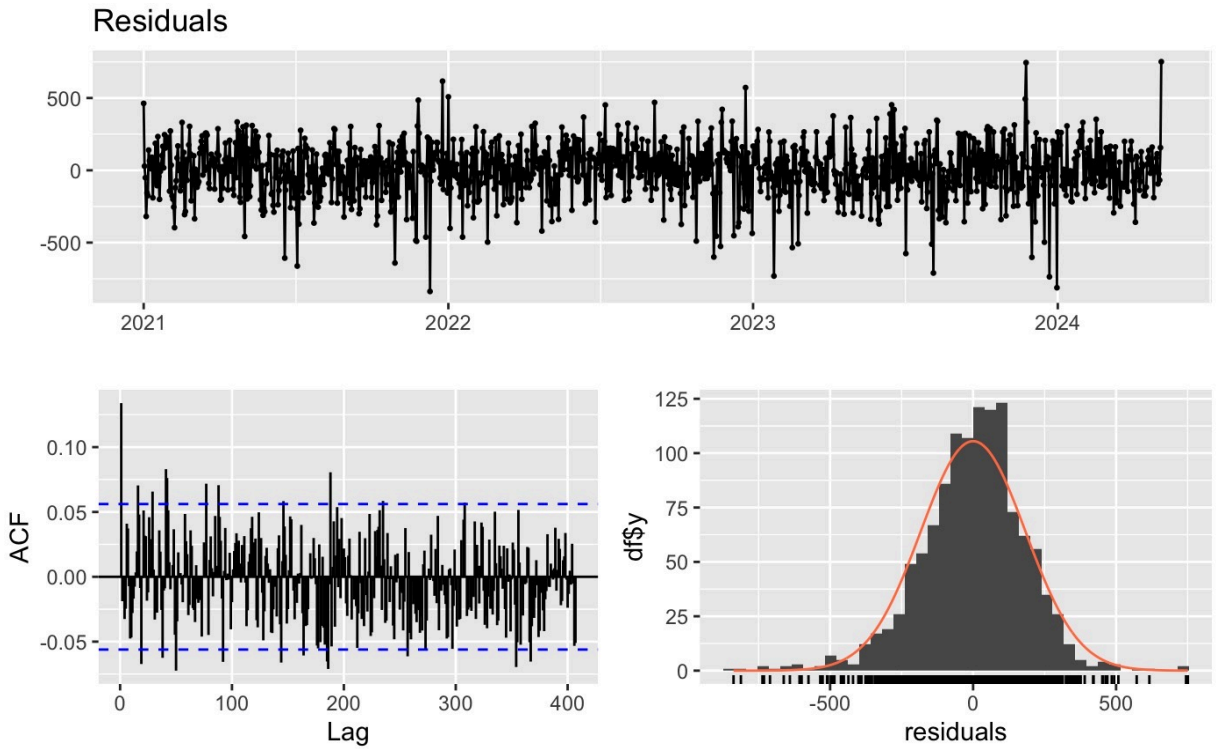
The residuals are normally distributed, but the model fails the Ljung box test with a p-value of 0.001. The ACF plot also suggests that there are some significant lags and some seasonality which is not captured by the model fit.

Modeling - Prophet

We Used Facebook's Prophet library to fit a model with 5,7,183 and 365 periodicity seasonality's to predict the number of crashes.



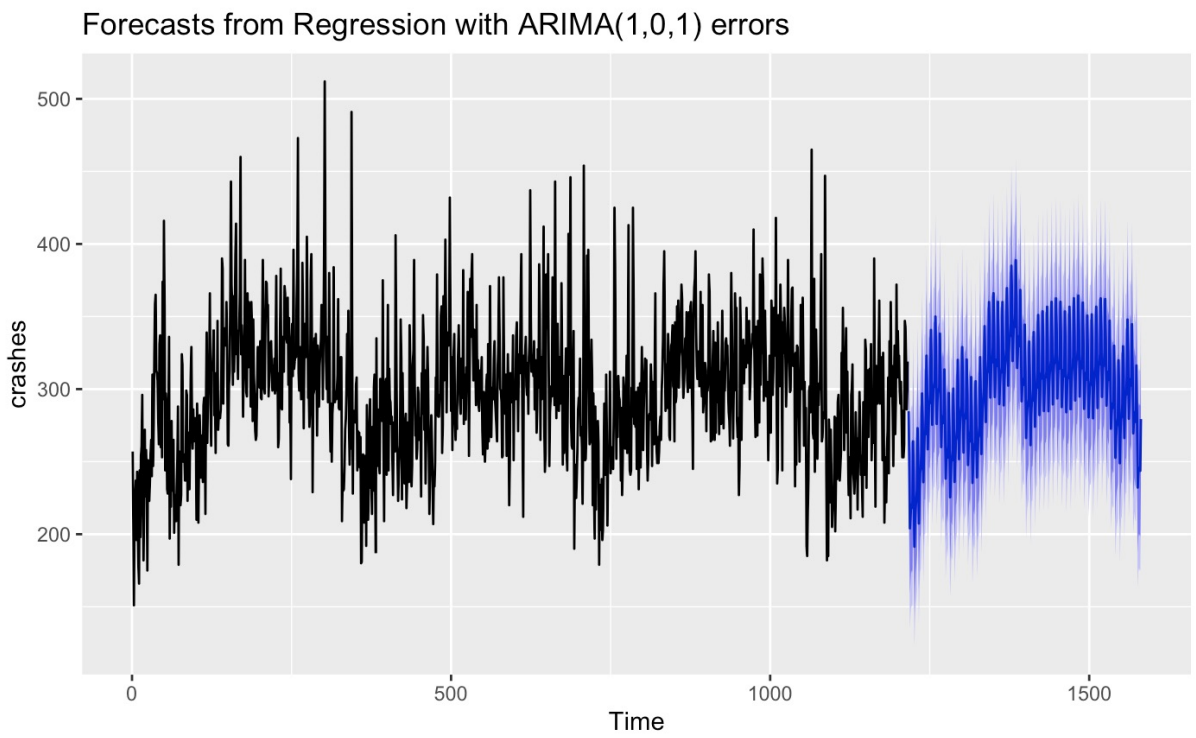
Forecasts from the model look reasonable in their predictions of the future, following trends that we can see in the training set. There are some outliers in the data, which contribute to our High RMSE value as seen in the scatter.



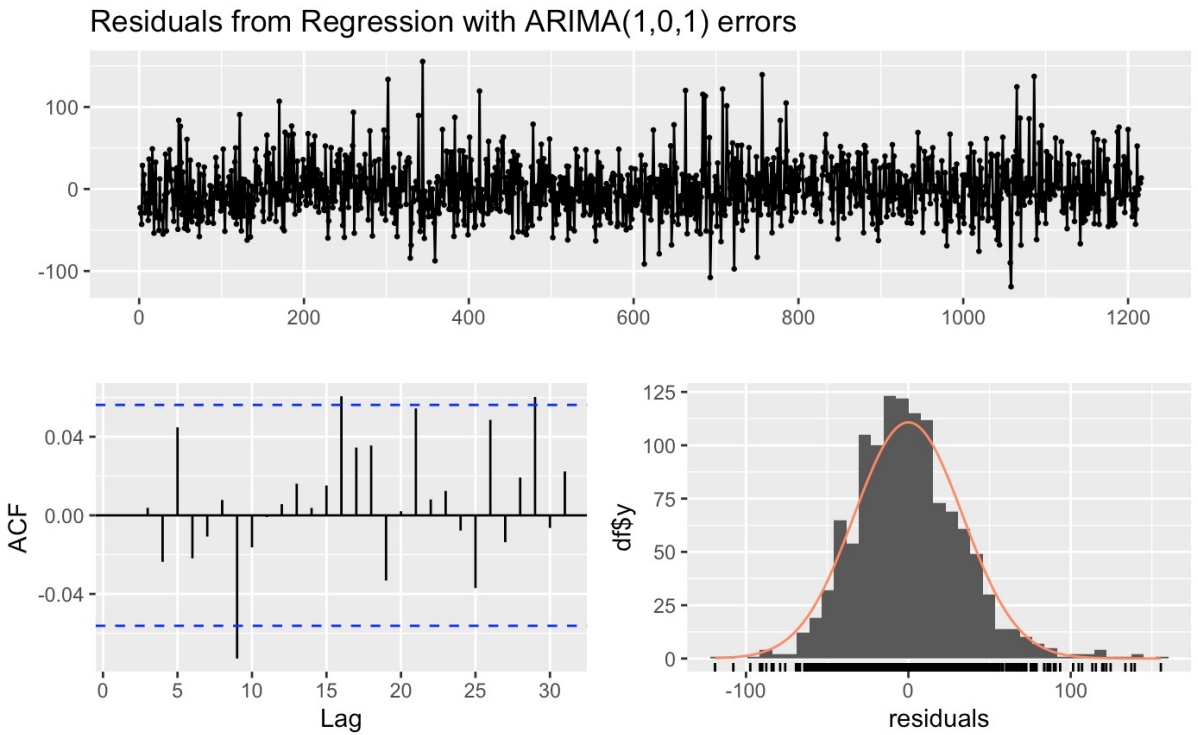
Given that the model does seem reasonable, The residuals still do not pass the Ljung-Box test with a p-value of 2e-3. This model has not explained all the variance in the data even if the RMSE and MAPE value is one of the lowest.

Modeling - ARIMAX

ARIMAX Model: We use Fourier terms and fit a model using the auto.arima() function. It fits a Regression with ARIMA(1,0,1) errors



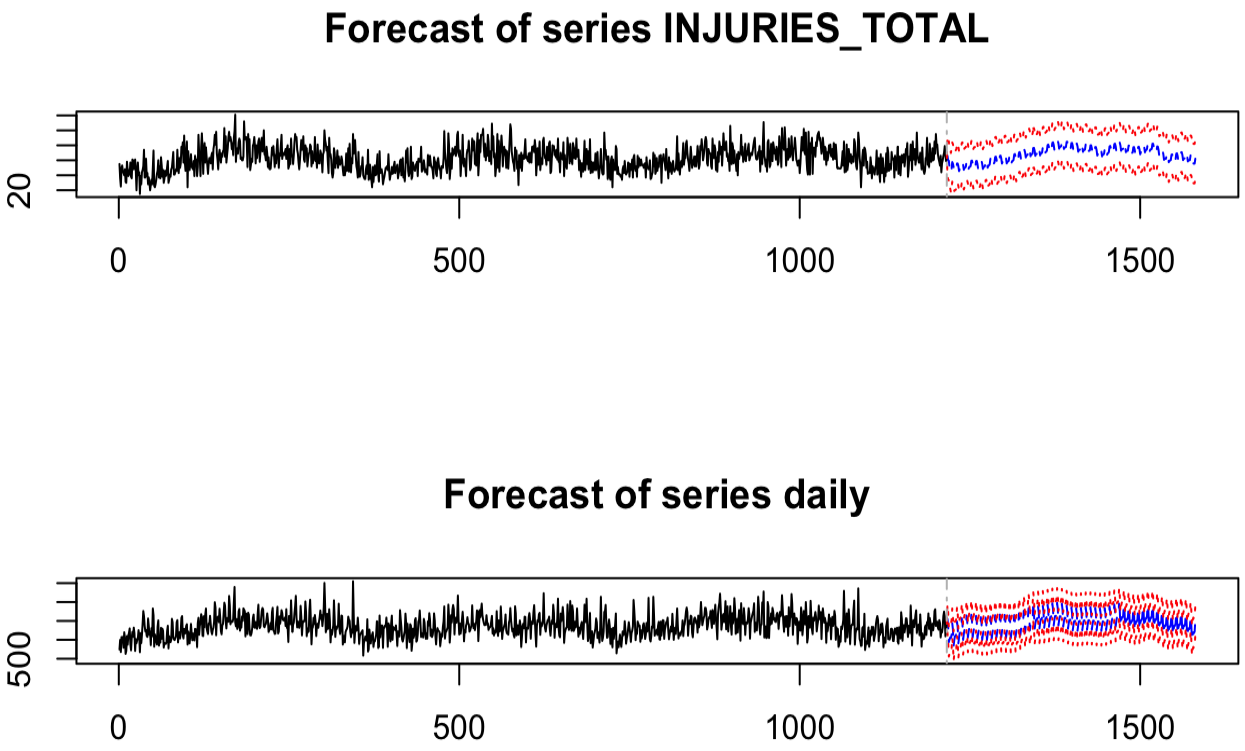
Forecasts from the model look reasonable, and the model is able to capture the underlying patterns and seasonality in the crashes data.



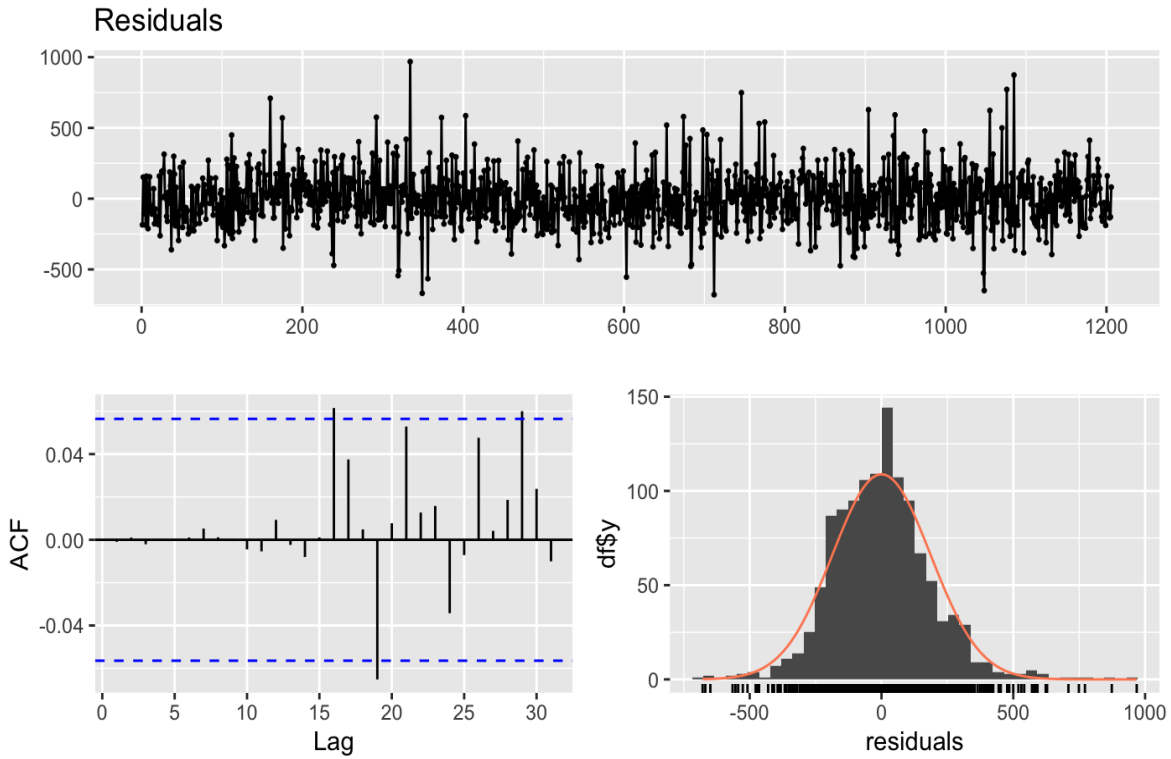
The model residuals pass the Ljung-Box test with a p-value of 0.21 and are fairly normally distributed. The ACF plot also does not show any significant pattern of correlation between the lags. This model is our best model.

Modeling Injuries - VAR

We Used a VectorAutoRegressive model to predict the Number of Injuries using the number of crashes and added seasonality's using Fourier terms.



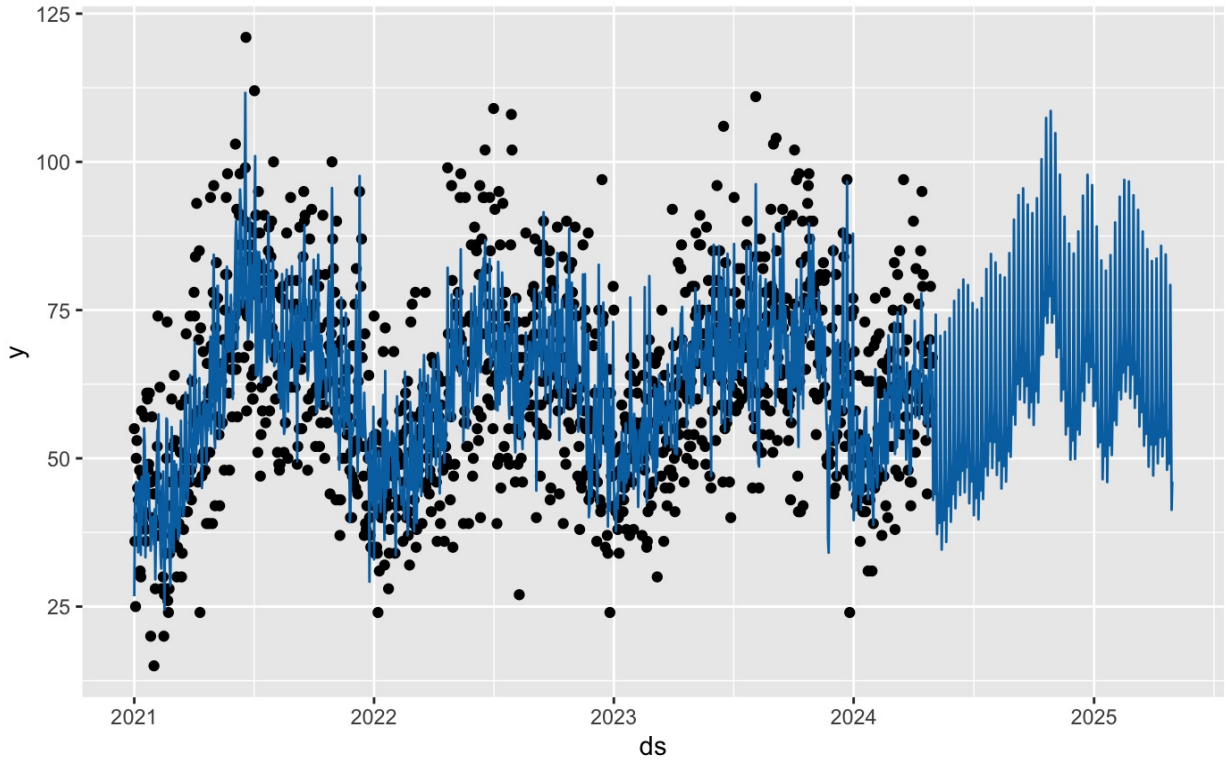
The VAR model performs quite well on the predictions with a tight confidence interval and mirroring the observed seasonality.



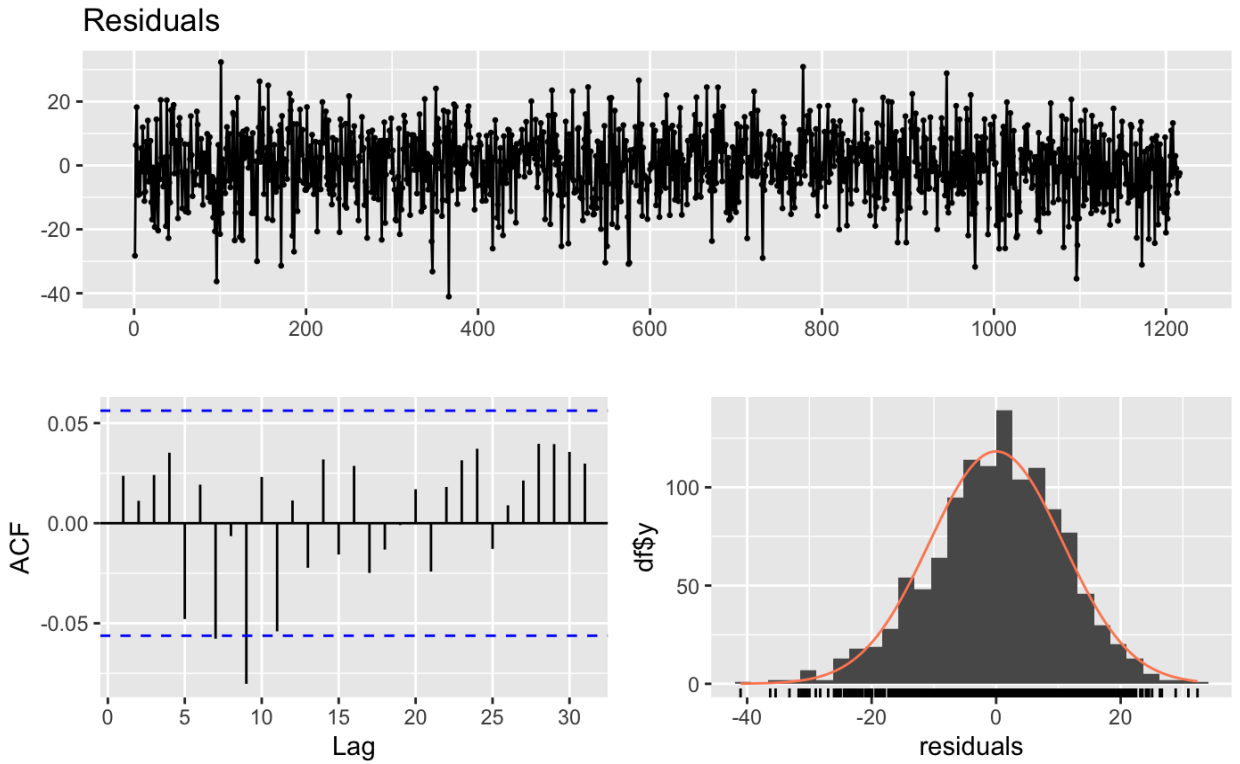
The Residuals of this model look quite normally distributed, low autocorrelation is confirmed by the Ljung Box test with p-value of 0.14. This model performs quite well with an RMSE of 12.7 and MAPE of 17.96 but still falls behind other models

Modeling Injuries - Prophet

We Used Facebook's Prophet library to fit a model with 7 and 365 periodicity seasonality's. We additionally added the Number of crashes as an external regressor to help predict the number of injuries



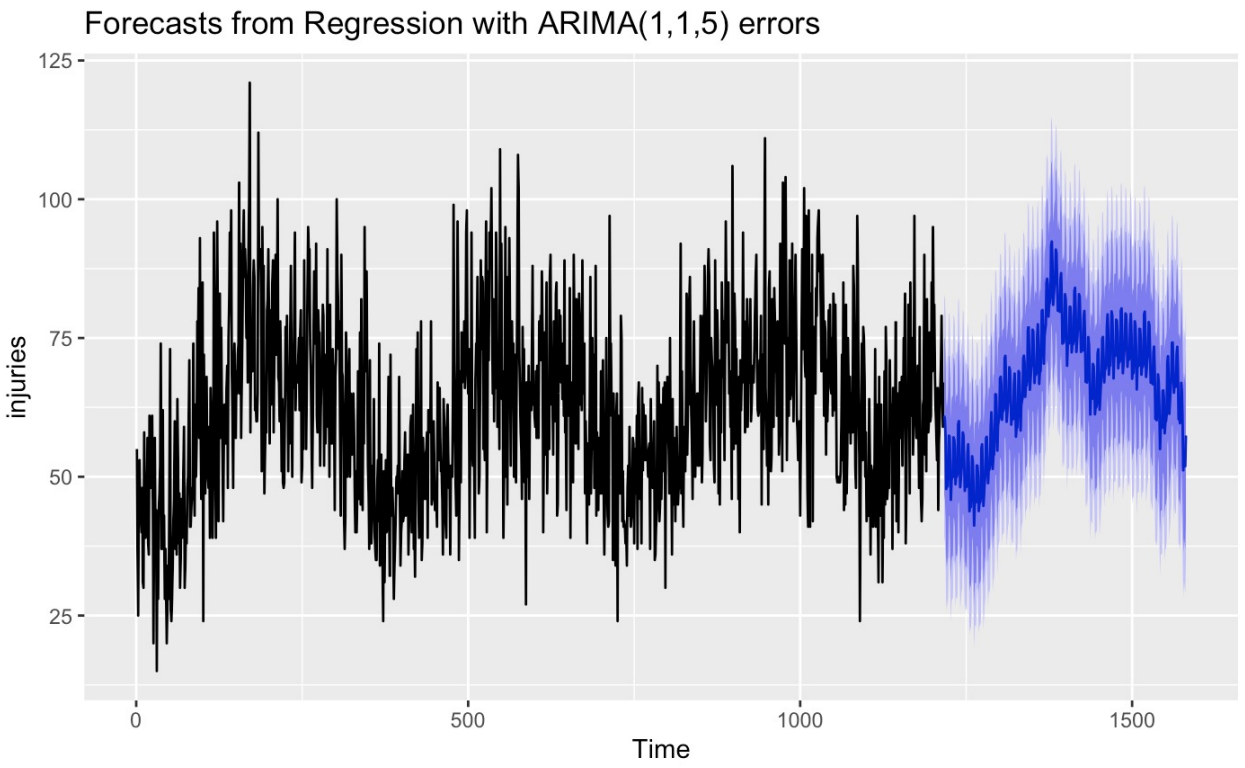
Forecasts, from the model are really variant as we can see from the fast changes in seasonality. They do not seem to be a good representation of reality.

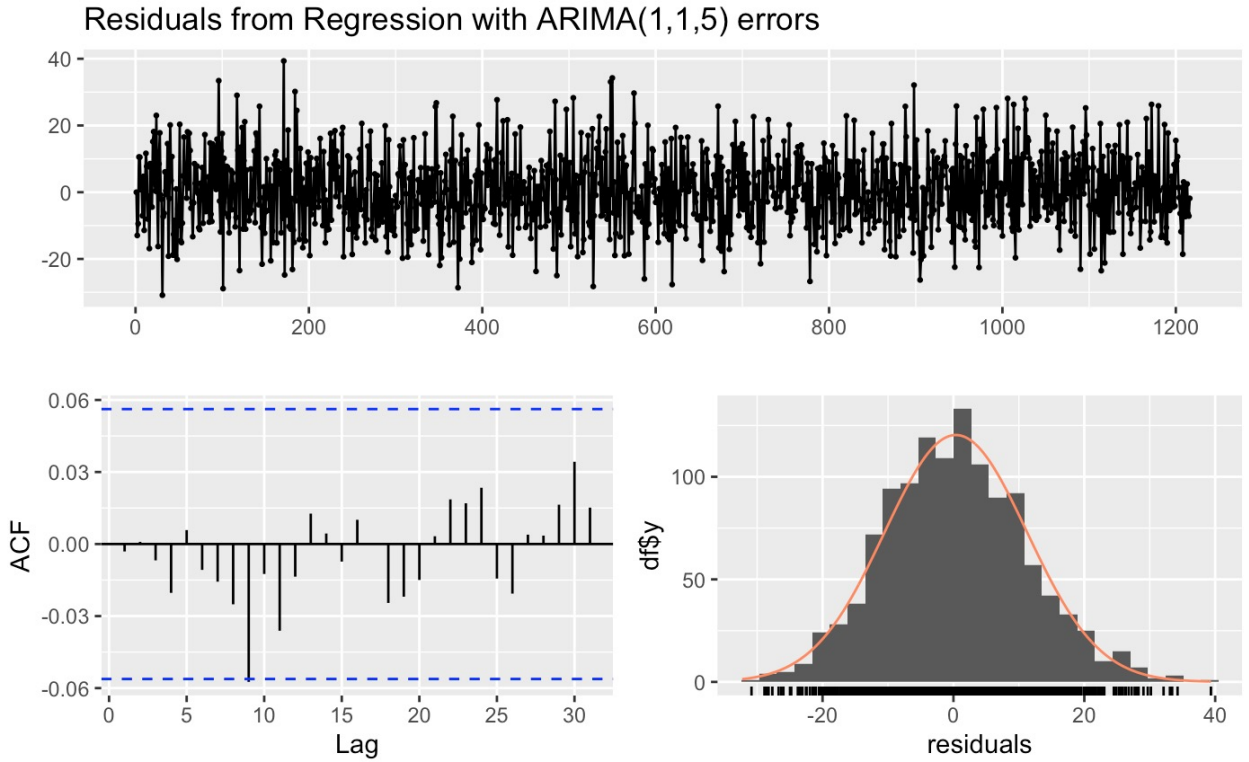


The model does not pass the Ljung Box test with a p-value of 0.04 even after having a low RMSE and MAPE values of 10.74 and 14.93 respectively. Echoing our previous assumptions. Prophet is Overfitting to the Series.

Modeling Injuries - ARIMAX

ARIMAX Model: We use Fourier terms for injuries and combine with crashes to fit a model using the auto.arima() function. It fits a Regression with ARIMA(1,1,5) errors





We used the predicted values for crashes from ARIMAX model (for crashes) to feed as one of the regressors for this multivariate model. The model is able to predict injuries based on crashes and handle multiple seasonality in data.

The model residuals pass the Ljung-Box test with a p-value of 0.194 and are normally distributed. The ACF plot does not show any significant lags. This model is also our best model with and RMSE of 10.88 and a MAPE of 15.22

Model Evaluation and Selection

CRASHES

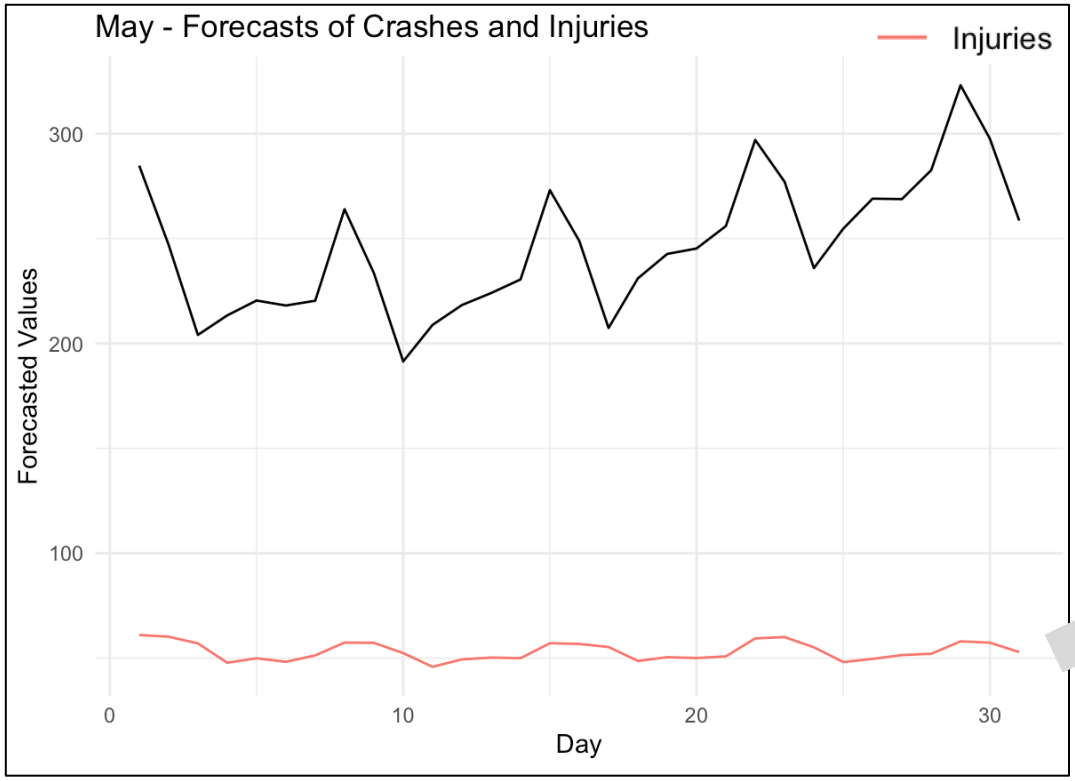
Model	RMSE	MAE	MAPE	Residuals Ljung-Box Test	P-value
SNAIVE	55.2	43.2	14.6	Fail	2.2e-16
TBATS	33.2	25.3	8.6	Fail	1e-3
PROPHET	33.1	25.2	8.6	Fail	2e-3
ARIMAX	33.3	25.6	8.7	Pass	0.21
STLM	34.2	26.1	8.7	Fail	2.2e-16

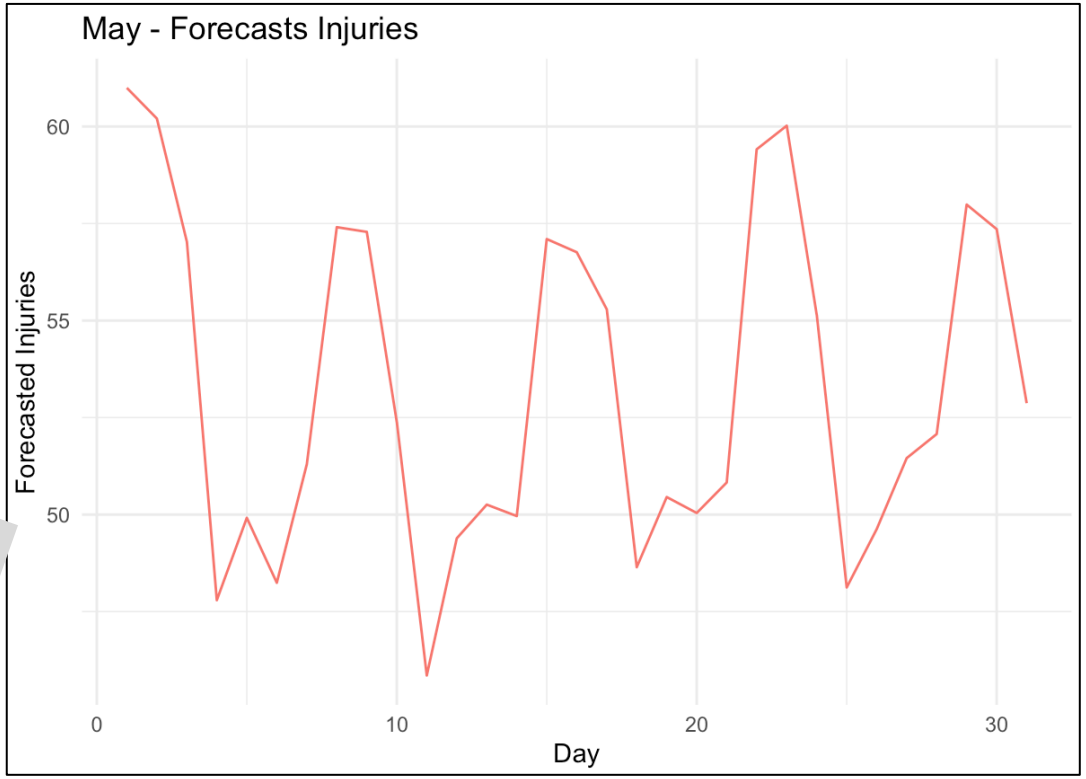
INJURIES

Model	RMSE	MAE	MAPE	Residuals Ljung-Box Test	P-value
PROPHET	10.7	8.4	14.9	Fail	4e-2
ARIMAX	10.8	8.6	15.2	Pass	0.19
VAR	12.7	10.2	17.9	Pass	0.14

Results - May 2024 Forecasts

We use our best models [ARIMAX models for crashes and injuries] to generate forecasts for May 2024. These results can enable the law enforcement agencies and healthcare facilities to manage workload.





Forecasted crashes for May indicate the following:

- Most crashes in the last week of May
- More crashes on dates 1st, 8th, 15th, 22nd, 23rd
- This can be useful for law enforcement agencies

Forecasted injuries for May indicate the following:

- Most injuries in the third week of May
- More injuries on dates 1st, 2nd, 8th, 9th, 15th, 16th
- This can be useful to guide the healthcare facilities

Future Scope

Granularity

Further down sampling the data to Hourly data will enable us uncover more complex patterns

Modeling

Experimental analysis of data and modeling to give hourly predictions for crashes and injuries

Improvements

Use cross-validation techniques for model improvement. Incorporate more data (for eg. Weather, Traffic data)

Recommendations

Customised for law enforcement agencies and healthcare facilities by the hourly forecasts

More sophisticated models will allow policymakers to simulate the effects of various traffic laws and infrastructure changes before implementation. Forecasted data can be increasingly used to design targeted public safety campaigns, focusing on times and locations where data predicts high risk.

Thank You!

Appendix

- Data Preparation/Aggregation
- Differencing for time series stationarity
- Modeling crashes – STLM()
- Residuals Linear model – TSLM()

Data Preparation

Raw Data

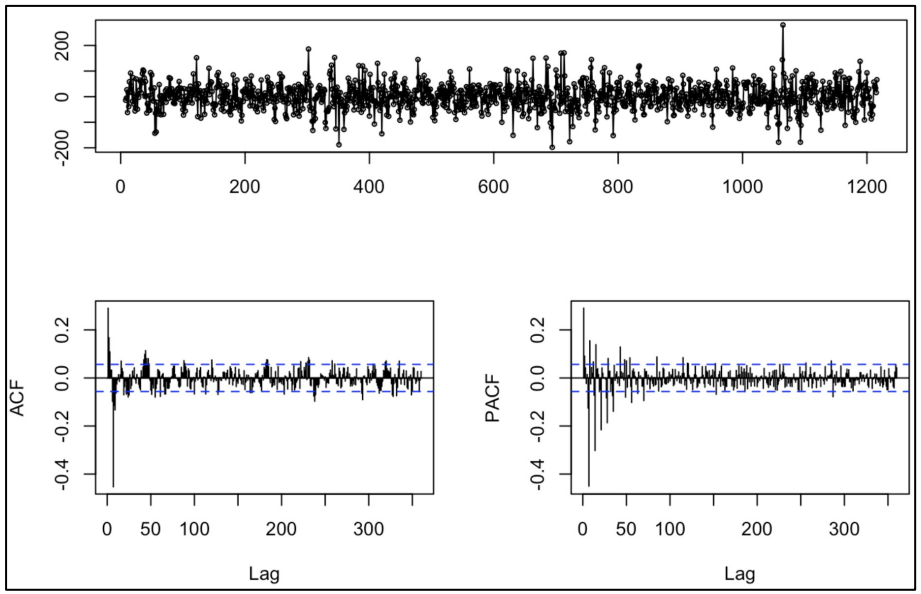
	CRASH_RECORD_ID	CRASH_DATE	INJURIES_TOTAL
0	6afead8030a2926d5d6d1aac2cef8edd466528d67f2958...	05/05/2024 01:30:00 AM	0.0
1	b2bba1a01839245d118b2e25f598f63b8c2a7b4bf42841...	05/05/2024 12:12:00 AM	0.0
2	5ed4cc024c8aaebaa1fde70838725f79d11509da30176b...	05/05/2024 12:01:00 AM	0.0
3	807ccb04c67193ae1244daf6e70bbda37f6ec905bc1424...	05/04/2024 11:06:00 PM	0.0
4	fb16b82c76af22bede438ab73d73ffdcc8b23ed799fee9...	05/04/2024 11:02:00 PM	0.0

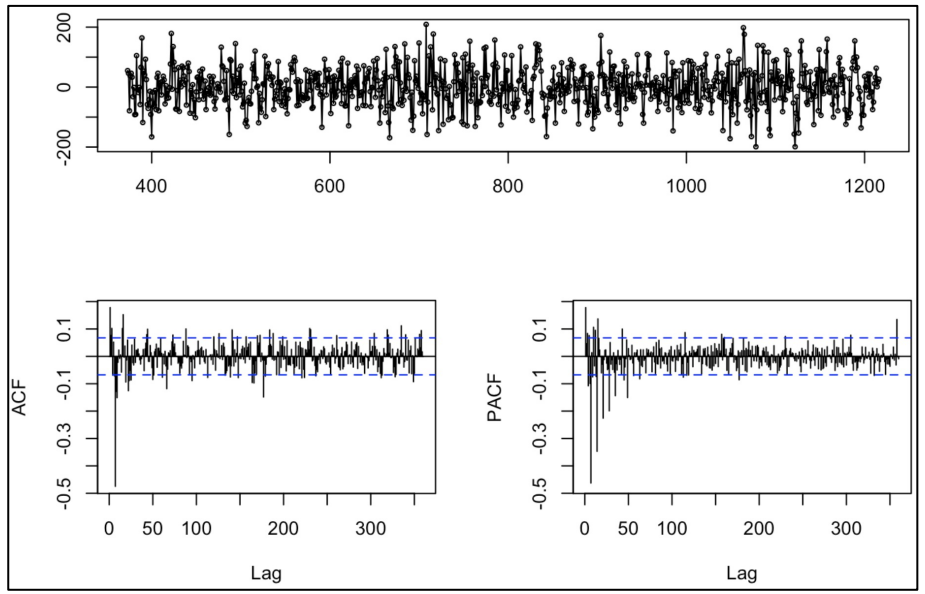
Final Data

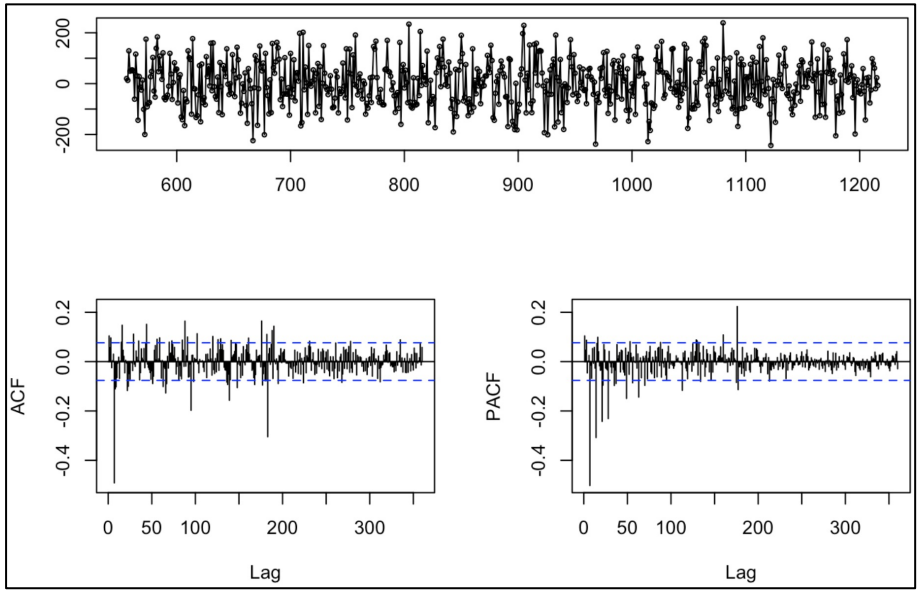
CRASH_DATE	INJURIES_TOTAL	CRASH_COUNTS
2021-01-01	55	257
2021-01-02	36	210
2021-01-03	25	151
2021-01-04	50	228
2021-01-05	53	226

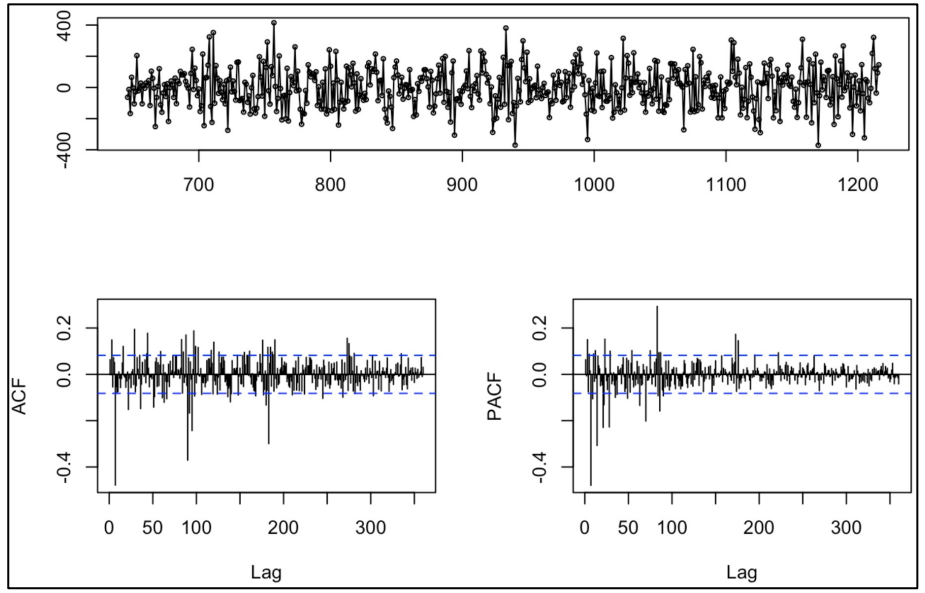
- The raw data has a unique row for each crash and a timestamp associated with the crash.
- Each row also has the total number of injuries associated with that crash.
- We extracted date from timestamp and down sampled the data to day level data. Aggregated the data to calculate crashes and injuries as follows:
  - Crashes = Count(Rows)
  - Injuries = Sum(Injuries)

Data Differencing





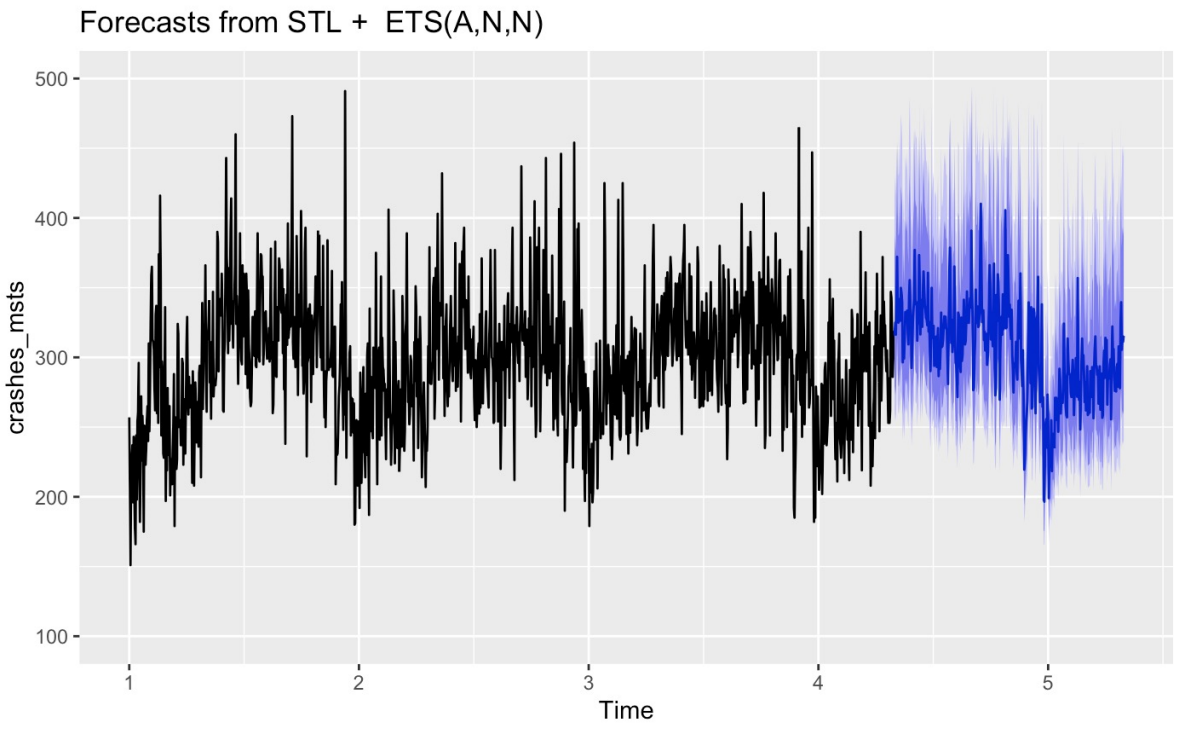


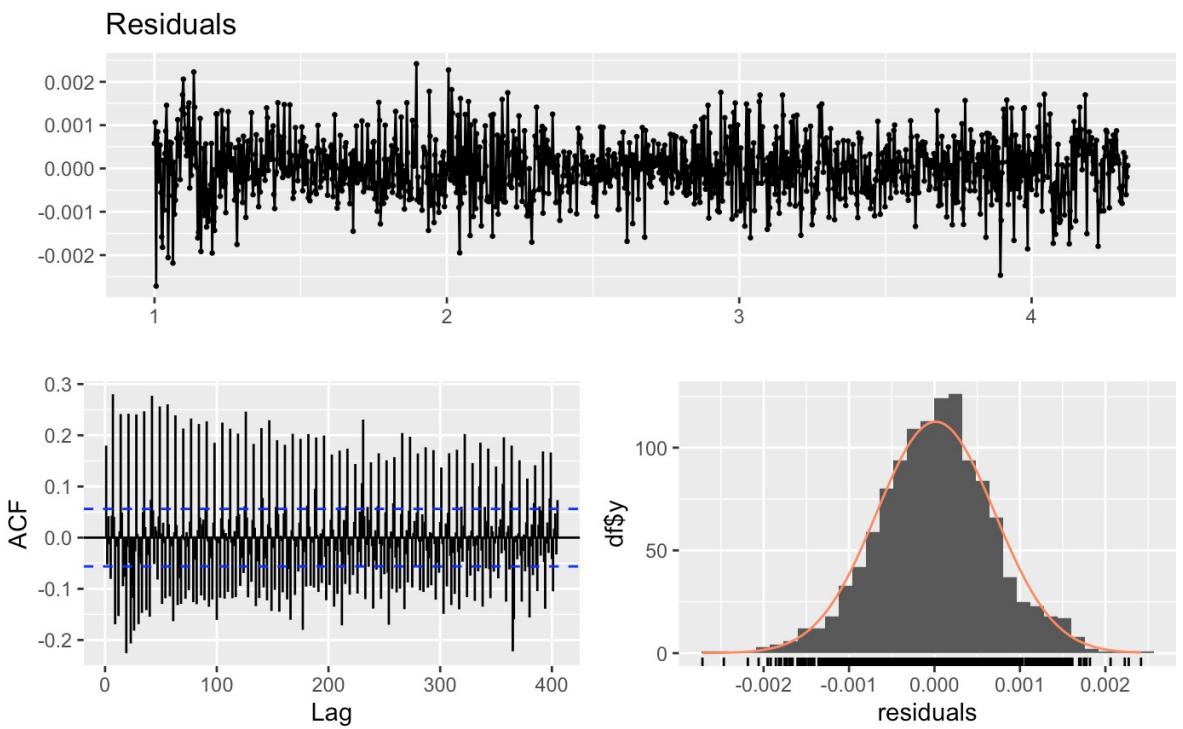


- After 4 rounds of seasonal differencing of the data, we were able to achieve near stationarity.
- Passes Ljung-Box test with a p-value of 0.1289
- Passes Augmented Dickey-Fuller test with a p-value of 0.01
- Passes KPSS test for level and trend stationarity with p-values 0.1 each

Modeling - STLM

STLM is used to fit a model to the crashes data with multiple seasonal periods 5,7,183 and 365. It fits an ETS(A,N,N) model with additive errors. The model is not able to identify and account for the seasonality.

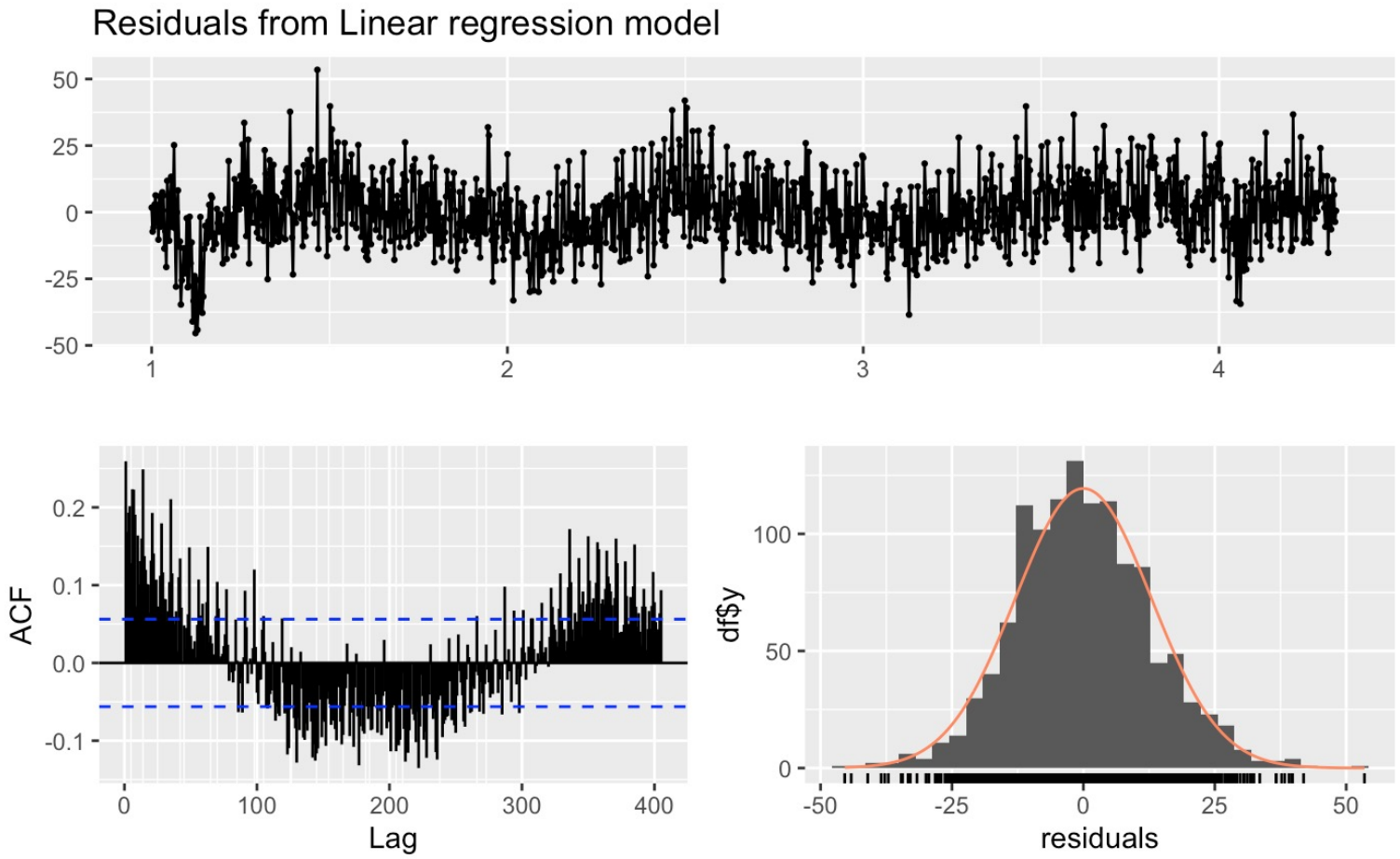




Even though the forecasts look reasonable and in line with the overall pattern of the data, the residuals are highly autocorrelated hence not a reliable model.

The model residuals fail the Ljung-Box test (p-value < 2.2e-16), suggesting high autocorrelation between residuals with multiple significant lag. We can see seasonality in residuals.

Residuals of Linear model




```
## Call:
## tslm(formula = injuries_msts ~ crashes_msts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.405  -9.009  -0.607   8.048  53.467
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.413382   2.336094  -0.605    0.545
## crashes_msts  0.212796   0.007766  27.401   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.97 on 1214 degrees of freedom
## Multiple R-squared:  0.3821, Adjusted R-squared:  0.3816
## F-statistic: 750.8 on 1 and 1214 DF,  p-value: < 2.2e-16
```

The model has R squared of 0.38, which means that crashes time series is able to capture 38% variation in the injuries time series. However, the residuals show seasonality and clearly indicate that the model has not been able to capture the underlying patterns in the data.