Introduction & Overview

New York City is known to have massive traffic with a large potential for accidents to happen each day. We have obtained a dataset over the course of 2 years, or 731 days, containing the number of collisions on each day. The data spans from January 1st of 2016 to December 31st of 2017. The data also includes variables windspeed (AWND), precipitation (PRCP), minimum temperature (TMIN), and maximum temperature (TMAX) in celsius. Using this data examined the potential presence of cyclical trends throughout the days of the year to determine if collisions are more prone to happen on specific days. We compared this to a deterministic time series plot for months of the year and days of the week in an effort to decide which model would be most appropriate for forecasting daily collisions in New York. We used a hold-out sample of 131 observations and a training sample of 600 observations. The data shows signs of seasonal behavior with drastic increases and decreases at different points in the time series plot at figure 1. The periodogram in figure 2 has period 6.98 as having the strongest amplitude which indicates weekly seasonality in the data. Looking more in depth at the seasons throughout the data, we developed box-plots to identify potential trends within the days of the week and another for months of the year. The daily box-plot in figure 3 shows an obvious trend of the weekends having much lower collision averages than the weekdays. Day 1 is Friday. Sunday has the least collision average which is to be expected. The monthly box-plot in figure 4 shows some signs of a trend with the months at the end of the year having higher collision averages and the month of June having a much higher average than the rest of the months. Month 1 is January.

Figure 1

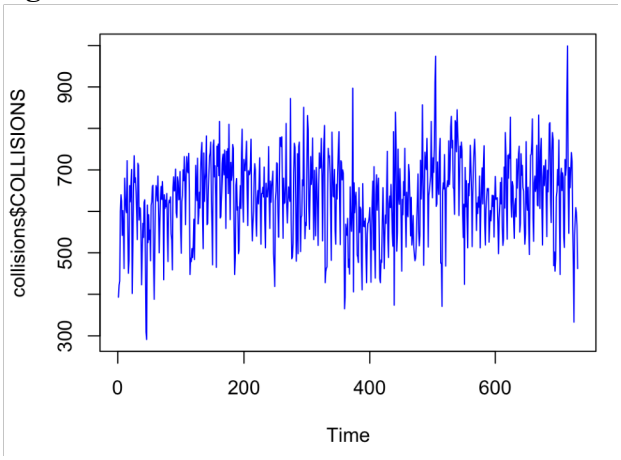
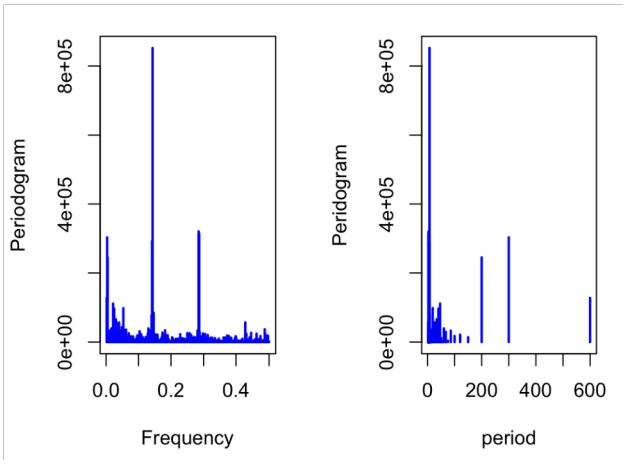


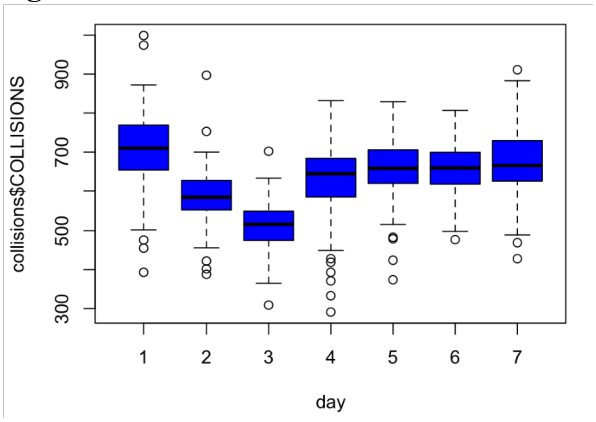
Figure 2

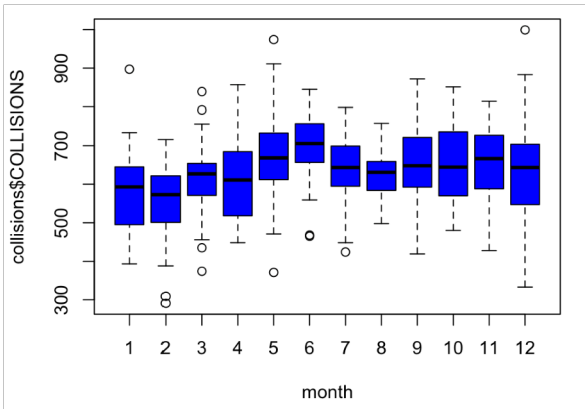


	harmonics	period	frequency	amplitude
[1,]	86	6.976744	0.143333333	852150.75
[2,]	171	3.508772	0.285000000	320903.19
[3,]	172	3.488372	0.286666667	315208.30
[4,]	2	300.000000	0.003333333	303350.86
[5,]	85	7.058824	0.141666667	292366.08
[6,]	3	200.000000	0.005000000	245108.83
[7,]	1	600.000000	0.001666667	128294.86

Figure 3

Figure 4





Univariate Time-Series Models

2.1

The first model we examined was a seasonal dummies model using daily and monthly seasonality. When evaluating the dummies model figure 5, we found Saturday and Sunday to experience the largest decrease in collisions when compared to Friday with the most significance at a 95% confidence level. This seems appropriate after examining the daily box-plot in figure 3 which had much lower averages for the weekends than the weekdays. All of the days exhibit a

significant p-value except for Thursday at this level, indicating the days play a large role on the number of collisions that occur. Looking at the months within the model, the most significant month is July which also correlates with our box plot from figure 5 where July had the highest average daily collisions. Most of the months are significant with p-values less than 0.05 except for February, March, April, and December. This means that the number of daily collisions are mainly affected in the middle of the year towards the end in New York City. Our

model has a R^2 of about 0.5 and the f-statistic shows the overall model as being significant in explaining the variation in collision occurrences. The low R^2means the model may not capture all of the variability affecting the daily collisions, but some of it is explained here.

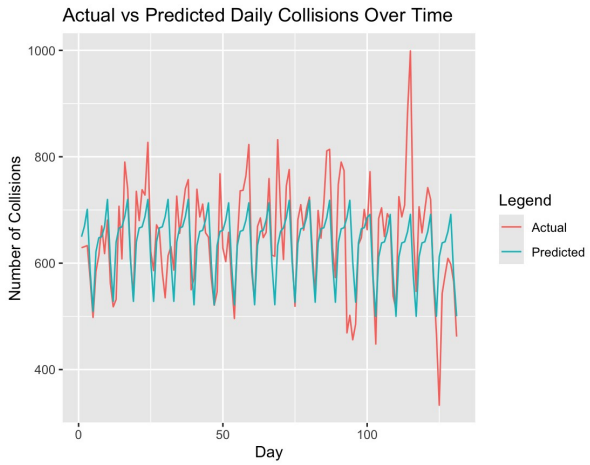
After training our model we tested it on our holdout sample using mean absolute percent error (MAPE). Our model received a MAPE value of 0.0906 which means our model's predictions are 9.06% off from the actual values. This shows our model handles new data very well and can predict the number of daily collisions very accurately. This is also demonstrated in a plot of our actual values in the holdout sample and the predicted values in figure 6, where we can see out predicted line follow the actual time series very closely with little variation.

Figure 5: Daily Dummies Model

```
Call:
lm(formula = COLLISIONS ~ train_time + nday + nmonth, data = train_collisions)

Residuals:
    Min      1Q  Median      3Q     Max
-290.69  -34.46    4.11   42.63  358.67

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.515e+02  1.209e+01  53.886  < 2e-16 ***
train_time   6.561e-03  1.796e-02   0.365 0.715025
nday2       -1.156e+02  1.092e+01 -10.585  < 2e-16 ***
nday3       -1.915e+02  1.092e+01 -17.528  < 2e-16 ***
nday4       -7.985e+01  1.093e+01  -7.308 8.99e-13 ***
nday5       -5.355e+01  1.093e+01  -4.901 1.24e-06 ***
nday6       -5.148e+01  1.096e+01  -4.699 3.27e-06 ***
nday7       -3.235e+01  1.095e+01  -2.954 0.003267 **
nmonth2     -1.970e+01  1.315e+01  -1.498 0.134548
nmonth3      3.268e+01  1.291e+01   2.531 0.011636 *
nmonth4      3.020e+01  1.307e+01   2.311 0.021187 *
nmonth5      8.667e+01  1.304e+01   6.645 6.97e-11 ***
nmonth6      1.192e+02  1.325e+01   8.994  < 2e-16 ***
nmonth7      6.289e+01  1.327e+01   4.741 2.68e-06 ***
nmonth8      4.966e+01  1.378e+01   3.603 0.000341 ***
nmonth9      6.809e+01  1.597e+01   4.265 2.33e-05 ***
nmonth10     6.148e+01  1.583e+01   3.883 0.000115 ***
nmonth11     6.605e+01  1.608e+01   4.108 4.56e-05 ***
nmonth12     3.937e+01  1.599e+01   2.462 0.014087 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71.59 on 581 degrees of freedom
Multiple R-squared:  0.4983,    Adjusted R-squared:  0.4827
F-statistic: 32.05 on 18 and 581 DF,  p-value: < 2.2e-16
```

```
> dummies_MAPE
[1] 0.09063112
```

Figure 6



2.2

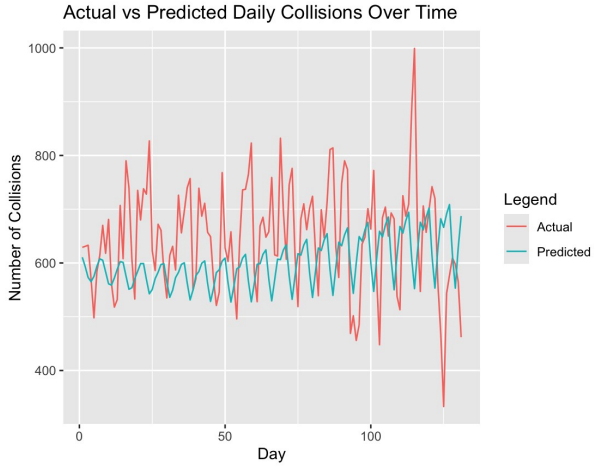
Additionally, we created a cyclical model measuring the number of collisions using time and 7 paired sine and cosine terms as predictors to capture the cyclical patterns. We chose 7 pairs because period 7 had the most prevalent amplitude in figure 2, indicating a potential cyclical effect. The sine and cosine terms allow the model to account for periodic fluctuations in collision occurrences. The model shows significant coefficients at all sine and cosine pairs at the 0.05 significance level, excluding sine 3, cosine 4, and both sine and cosine 8. Our trend variable also didn't seem to be significant. Cosine 2 had the most significance indicating period 2 significantly impacts collision occurrences. The model exhibits a slightly lower R^2 than our seasonal dummies model at 0.4254 meaning the cyclical patterns only explain 42.54% of the model. Although, the model has a significant F-statistic, so there is some presence of a cyclical trend in the model. After using the model to predict values within the holdout sample, the model has a MAPE value of 0.1463. This means that on average the model's prediction only deviates about 14.63% of the values. This is more than our seasonal dummies model but the cyclical model is still able to predict a large amount of the daily collisions. When we plotted the actual versus predicted values in figure 8 we saw that the cyclical model stays within the bounds of the actual values and follows the actual time series increases too.

Figure 7

```
> cyclical_MAPE
[1] 0.1463397
```

```
Call:
lm(formula = COLLISIONS ~ train_time + cos2 + sin2 + cos3 + sin3 +
    cos4 + sin4 + cos5 + sin5 + cos6 + sin6 + cos7 + sin7 + cos8 +
    sin8, data = train_collisions)

Residuals:
    Min      1Q  Median      3Q     Max
-305.55  -40.10    7.39   47.25  352.25

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 620.89733   13.40730  46.310  < 2e-16 ***
train_time    0.01916    0.04339   0.442 0.658969
cos2         43.87714    4.41204   9.945  < 2e-16 ***
sin2         30.22918    4.41274   6.850 1.87e-11 ***
cos3         29.41497    4.41204   6.667 6.06e-11 ***
sin3         14.34791    4.41196   3.252 0.001212 **
cos4        -28.58493    4.41204  -6.479 1.97e-10 ***
sin4        -15.24067    4.41196  -3.454 0.000592 ***
cos5         -2.78180    4.41204  -0.631 0.528613
sin5        -35.47890    6.05249  -5.862 7.66e-09 ***
cos6        -22.72271    4.41204  -5.150 3.56e-07 ***
sin6        -21.44753    4.41276  -4.860 1.51e-06 ***
cos7        -20.74957    4.41204  -4.703 3.20e-06 ***
sin7        -22.15391    5.20518  -4.256 2.42e-05 ***
cos8         -4.85611    4.41204  -1.101 0.271502
sin8         12.48067    9.38838   1.329 0.184244
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.42 on 584 degrees of freedom
Multiple R-squared:  0.4254,    Adjusted R-squared:  0.4107
F-statistic: 28.83 on 15 and 584 DF,  p-value: < 2.2e-16
```

Figure 8



2.3 ACF Residuals

Lastly, we looked at the autocorrelation functions (ACF) for both our dummies and cyclical models. The residuals of both models are very similar with significant lags in the same places and following similar patterns where the lags fall and rise. Both residuals show the lags being chopped off after lag 2. The cyclical residuals seem to be slightly higher at most lags,

presenting possibly stronger autocorrelation in the cyclical residuals. Overall the time series cannot be dismissed as white noise because there are prevalent lags outside of the 2 standard error bounds.

Figure 9

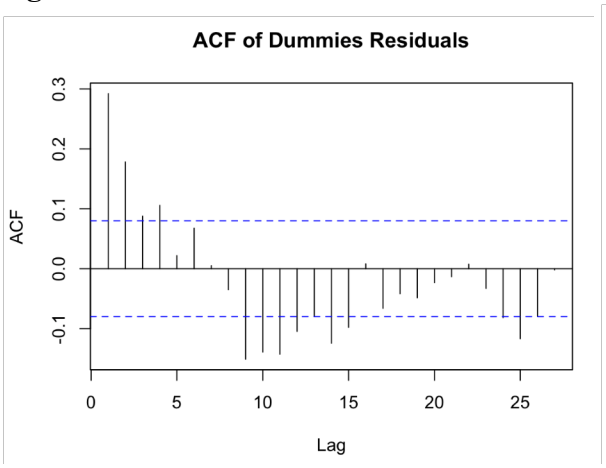
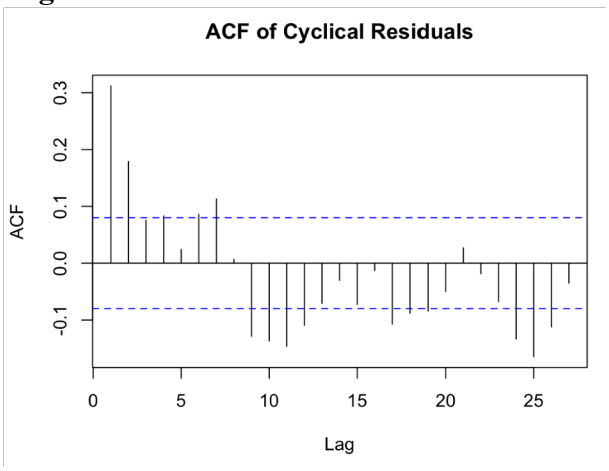


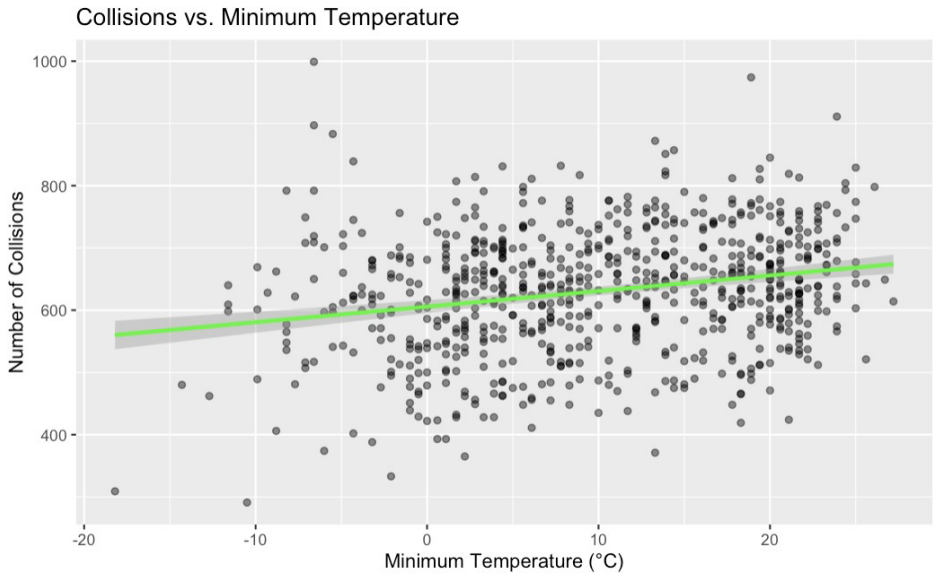
Figure 10



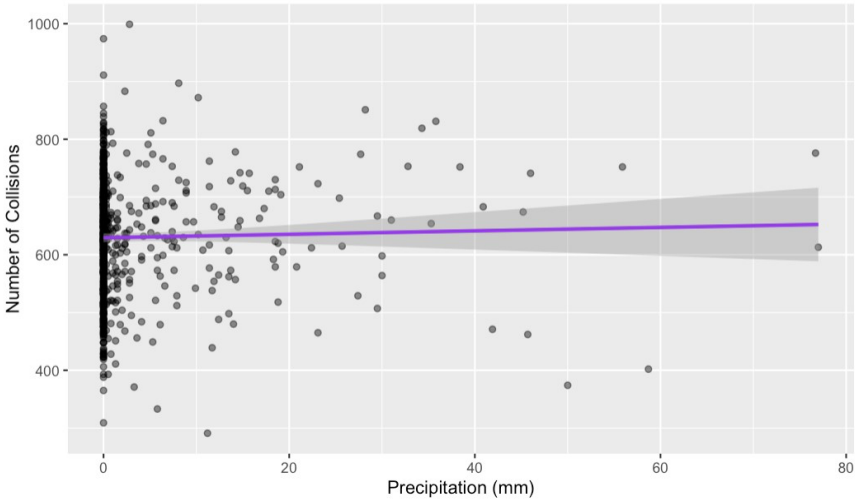
Time Series Regression Models

3.1 Discussion of independent variables. Correlation analysis and scatter plots.

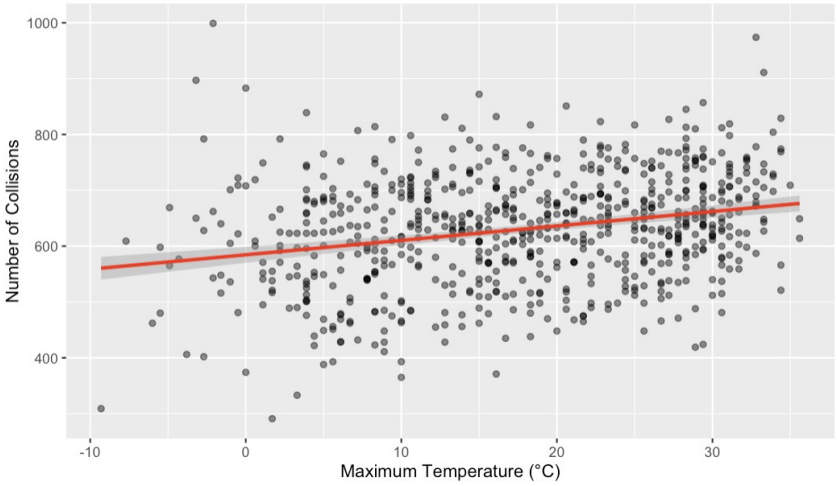
Scatter plots of each factor involved in collisions. PRCP, MIN TEMP. , Max Temp, and Wind Speed. Plots include a simple linear regression line to help visualize any trends:



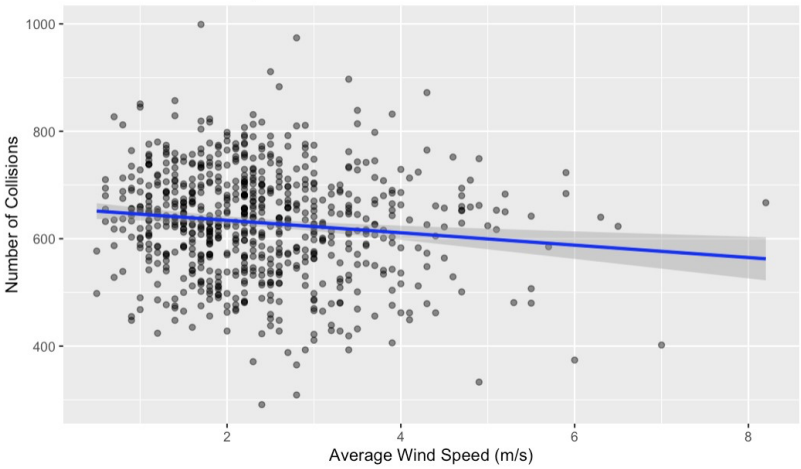
Collisions vs. Precipitation



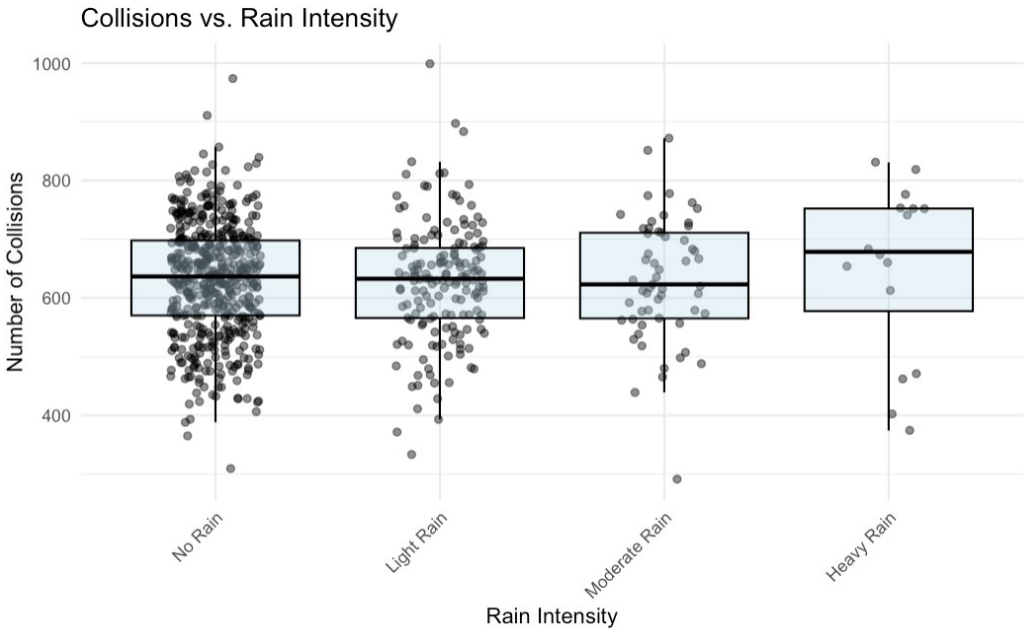
Collisions vs. Maximum Temperature



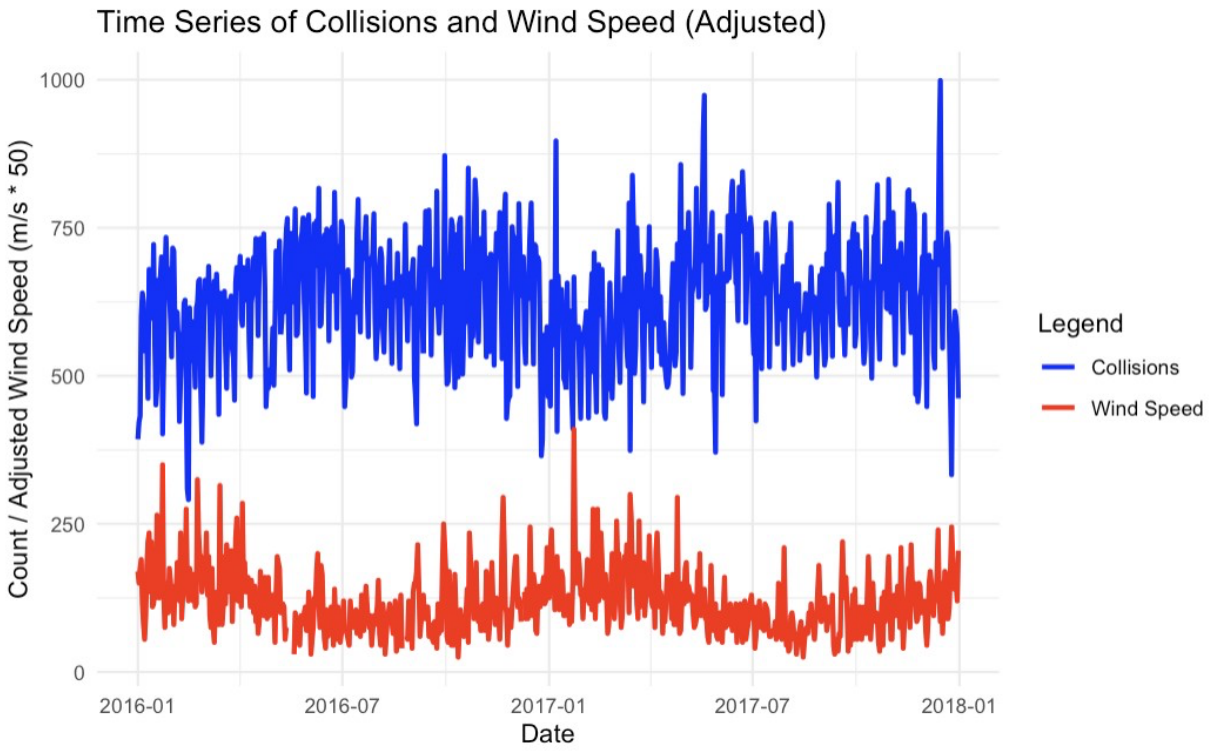
Collisions vs. Wind Speed



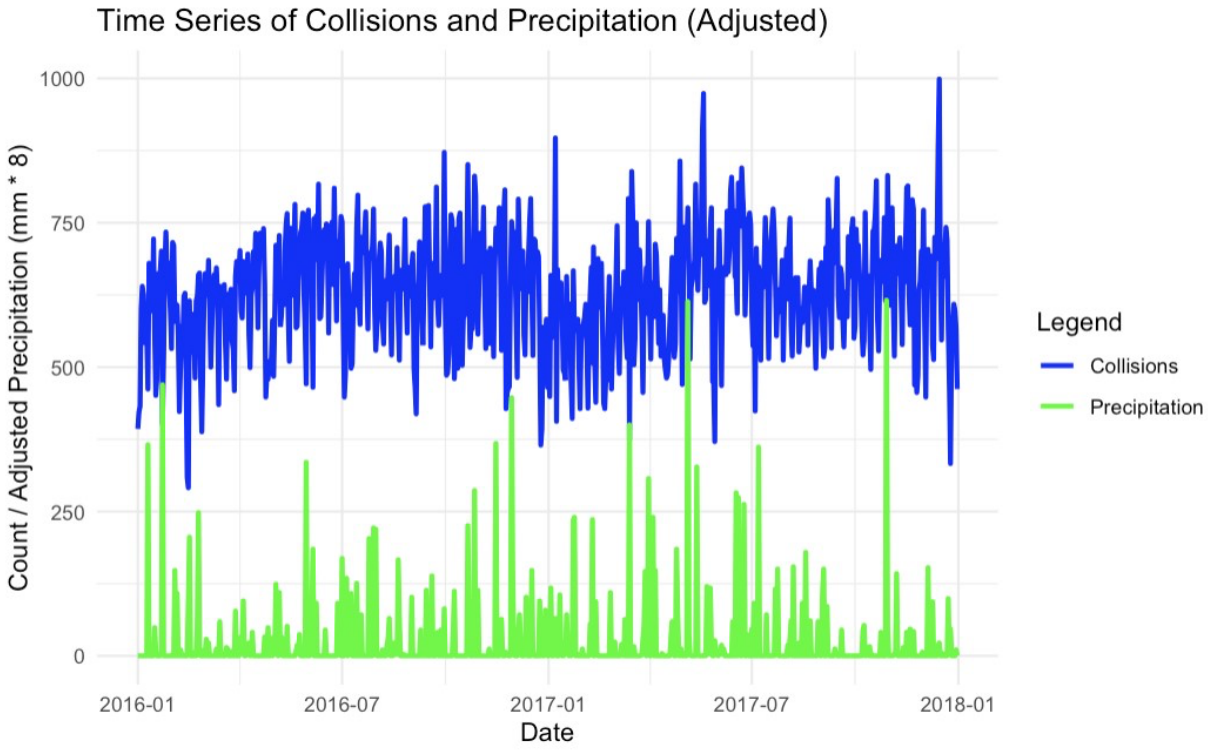
Now let's look at PRCP as a categorical variable… with PRCP = 0 being ‘No Rain’, PRCP 1-10 being ‘Light Rain’, PRCP 11-13 being ‘Moderate Rain’, and 31-100 being ‘Heavy Rain’. Below are four box plots to help visualize this.



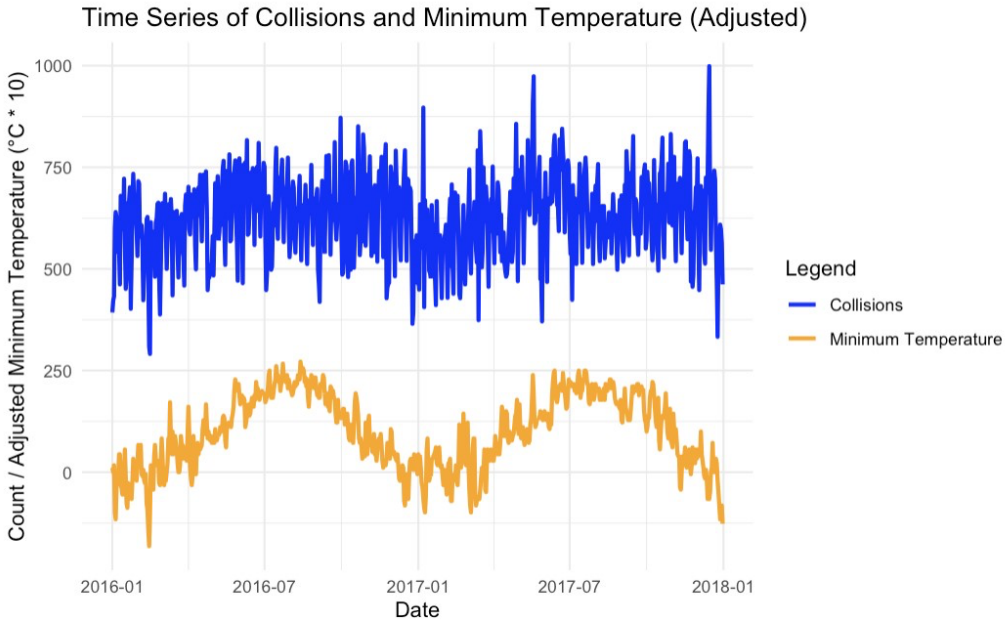
Now let's look at some overlay time series plots. The plot below shows collisions and wind speed over time. Please note that wind speed was multiplied by 50 for better visualization.



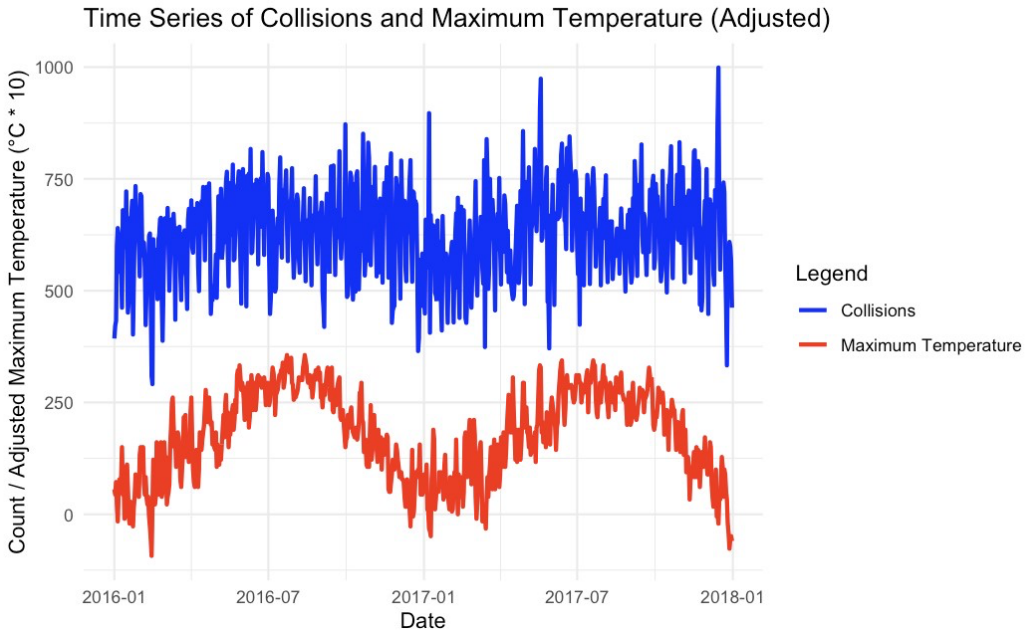
The next plot below shows collisions and PRCP levels over time. Please note that PRCP was multiplied by 8 for better visualization.



The next plot below shows collisions and TMIN over time. Please note that TMIN was multiplied by 10 for better visualization.



The next plot below shows collisions and TMAX over time. Please note that TMAX was multiplied by 10 for better visualization.



3.2 Comparison of "candidate" models in terms of fit and hold-out sample.

We will compare the performance of some possible combinations of linear regression models and how they perform on the testing data.

These are the following 'candidate models':

Models with all variables
model_all <- lm(COLLISIONS ~ AWND + TMAX + TMIN + PRCP)

Models with three variables
model_AWND_TMAX_TMIN <- lm(COLLISIONS ~ AWND + TMAX + TMIN)
model_AWND_TMAX_PRCP <- lm(COLLISIONS ~ AWND + TMAX + PRCP)
model_AWND_TMIN_PRCP <- lm(COLLISIONS ~ AWND + TMIN + PRCP)
model_TMAX_TMIN_PRCP <- lm(COLLISIONS ~ TMAX + TMIN + PRCP)

Models with two variables
model_AWND_TMAX <- lm(COLLISIONS ~ AWND + TMAX)
model_AWND_TMIN <- lm(COLLISIONS ~ AWND + TMIN)
model_AWND_PRCP <- lm(COLLISIONS ~ AWND + PRCP)

model_TMAX_TMIN <- lm(COLLISIONS ~ TMAX + TMIN)
model_TMAX_PRCP <- lm(COLLISIONS ~ TMAX + PRCP) model_TMIN_PRCP
<- lm(COLLISIONS ~ TMIN + PRCP)

Models with single variables
model_AWND <- lm(COLLISIONS ~ AWND)
model_TMAX <- lm(COLLISIONS ~ TMAX)
model_TMIN <- lm(COLLISIONS ~ TMIN)
model_PRCP <- lm(COLLISIONS ~ PRCP)

After testing all these models on the testing sample of 131 observations. After running these models TMAX single variable model had the lowest MSE (mean squared error):

Model	MSE
$mse_TMAX	9471.756
$mse_AWND	10010.05

These results show us that the best-performing LM model is a single variable model of TMAX and the worst-performing model is also a single variable model using AWND. Overall it also shows us that using all the variables in our model may just overfit it and not be the best in predicting future collisions.

Next, let's look at an ARIMA model fitted to the training data to make predictions on the testing dataset. These predictions are based on the TMAX values from the testing data.

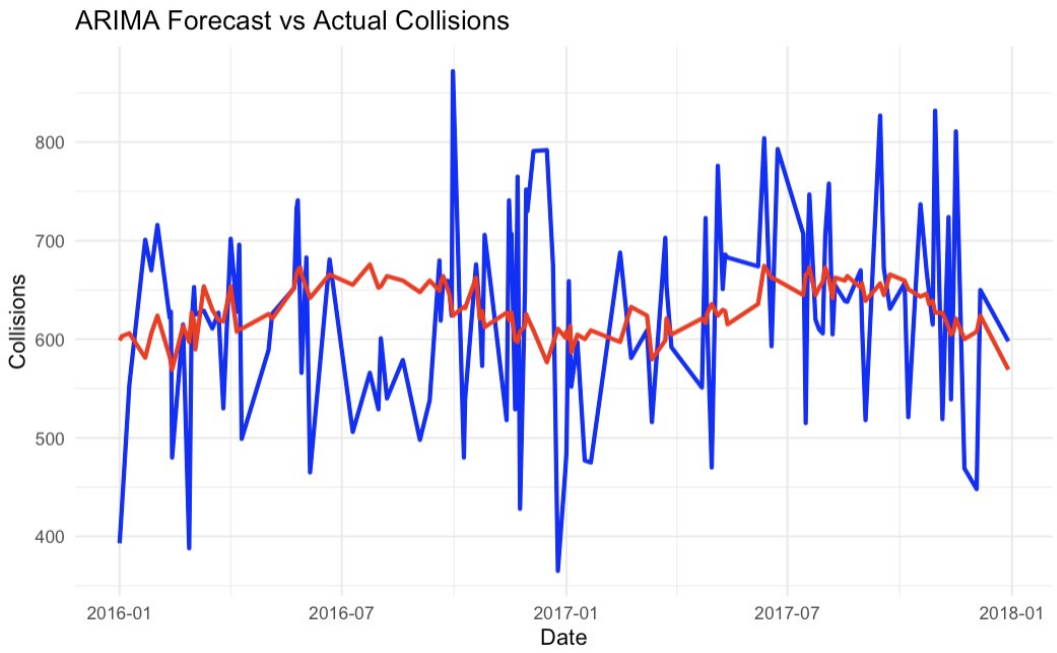
```
Arima(training_data$COLLISIONS, order = c(0,0,0),seasonal =
        list(order = c(0,0,0), period = 7), xreg =
        training_data$TMAX)
```

Model	MSE
ARIMA(0,0,0)	9263.52735773

This model shows a lower MSE then that of LM model.

3.3 Looking at residuals of the model(s).

Below is the actual crash data plotted against the ARIMA model predictions.




```
Regression with ARIMA(0,0,0) errors

Coefficients:
      intercept    xreg
       583.9222  2.6741
s.e.     8.0668  0.3943

sigma^2 = 9295:  log likelihood = -3591.52
AIC=7189.03   AICc=7189.07   BIC=7202.22

Training set error measures:
                      ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set -5.65592e-13 96.24722 75.85789 -2.575699 12.78271 0.6810791 -0.03040976
```

The coefficient for TMAX is 2.6741 with a standard error of 0.3943. This indicates that for each unit increase in TMAX, collisions increase by approximately 2.6741, again estimated with good precision.

Overall, the model appears to do a reasonably good job predicting collisions based on TMAX. The errors (ME, RMSE, MAE) are relatively low, and the low ACF1 value suggests that the model's residuals do not show significant autocorrelation

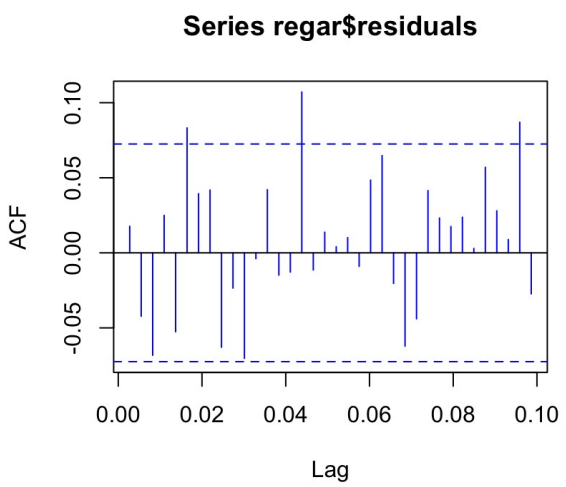
4.1???

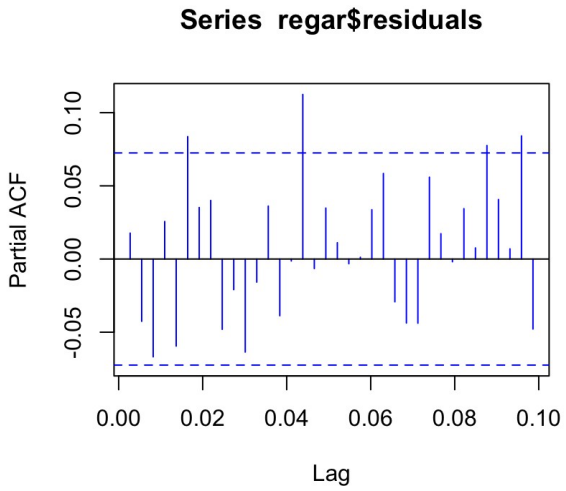
4.2 Analysis and modeling of regression model residuals

```
Series: collisions_ts
ARIMA(1,0,1)(0,1,1)[7]

Coefficients:
         ar1      ma1     sma1
      0.8101  -0.4630  -0.9813
s.e.  0.0472   0.0725   0.0223

sigma^2 = 4937:  log likelihood = -4115.85
AIC=8239.71   AICc=8239.77   BIC=8258.05

Training set error measures:
                    ME    RMSE      MAE         MPE     MAPE      MASE        ACF1
Training set 6.097002 69.7821 50.92034 -0.3566747 8.477883 0.5248928 0.01769581
```

According to the absolute value of t-statistic 0.8101/0.0472, -0.4630/0.0725, and -0.9813/0.0223 are all greater than 1.96. We can conclude that both Φ1, θ1 and seasonal θ1coefficients are different than 0. The positive AR(1) coefficient (0.8101) indicates a strong positive relationship between the current value and the previous value in the series. The negative MA(1) coefficient (-0.4630) suggests that past errors negatively impact the current value. The highly negative seasonal MA(1) coefficient (-0.9813) indicates a strong seasonal pattern with a weekly periodicity (m=7). The low ACF1 value (0.01769581) suggests that there is minimal autocorrelation in the residuals, indicating that the model has adequately captured the temporal dependencies in the data.

4.3 ARIMA models for the variable of interest





According to the graph, we can see that all the ACF are within the range of 2 s.e. We can conclude that this series is stationary and it is white noise.

```
> Box.test(regar$residuals, lag = 36, type = "Ljung-Box")

        Box-Ljung test

data:  regar$residuals
X-squared = 54.284, df = 36, p-value = 0.02582
```

Ljung-Box Test: The Ljung-Box test results in a p-value that is higher than 0.05. As a result, we cannot reject the null hypothesis. As a result, we can conclude that the series is stationary and it is white noise.

Conclusion

The ARIMA(1,0,1)(0,1,1)[7] model applied to the New York City daily traffic collisions data demonstrates a good fit with significant coefficients and low residual autocorrelation. The model captures both short-term dependencies and seasonal patterns, particularly weekly seasonality, as indicated by the significant seasonal moving average coefficient. The training set error metrics suggest that the model's predictions are reasonably accurate, with an RMSE of 69.7821 and a MAPE of 8.477883. These results indicate that the model's predictions deviate by about 8.48% on average from the actual values, which is acceptable for practical purposes.

Further analysis shows that the model effectively captures the cyclical nature of the data, with a strong positive AR(1) coefficient suggesting that past values positively influence future values, and a significant seasonal MA(1) coefficient indicating a strong weekly pattern. The minimal autocorrelation in the residuals, as evidenced by a low ACF1 value, further confirms the model's adequacy in capturing the underlying temporal dependencies.

Overall, the ARIMA(1,0,1)(0,1,1)[7] model is a robust choice for forecasting daily traffic collisions in New York City. Its effectiveness can be enhanced by continually validating its predictions against new data and exploring potential improvements through alternative modeling approaches or incorporating additional variables.