Multiple Linear Regression - Prediction of Traffic Volume

STAT 4355 with Dr. Shin, Spring 2022
Jay Glessner, Alex May, Muhammad Zubair, Nick Azar

Background

With the average American worker spending nearly an hour a day in traffic (United States Census Bureau), the prediction and understanding of traffic flows has a direct impact on the daily lives of millions of workers across the country.

A variety of models have been successfully employed in predicting traffic conditions in different cities, which is a critical aspect of advanced traffic management systems (Sohaee). Utilizing traffic prediction is a more cost-effective and modern approach to resolving congestion in major cities, compared to traditional infrastructure-based approaches.

Using data collected from 2012-2018 by the Minnesota Department of Transportation near Minneapolis-St. Paul, we aim to predict the total traffic volume based on environmental and social conditions by creating a multiple linear regression model. By identifying the critical regressors and revising towards the most useful model, traffic volume prediction models could be modified and used nationally to advise commuters and travelers of when high traffic volumes are predicted.

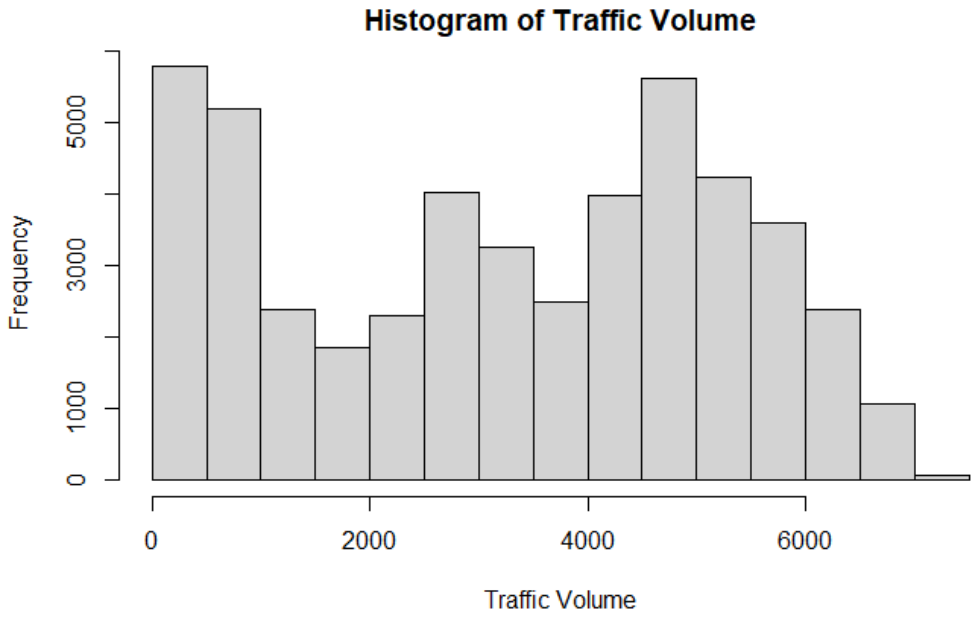
Additionally, we utilized the MN DOT analysis of the same data to confirm our understanding of the traffic and weather patterns as well as draw additional conclusions and expansions (Minnesota).

Data Description

The dataset was found on the UCI Machine Learning Repository and consists of 48,204 occurrences across 8 attributes. Each instance is a snapshot of the weather conditions during that hour along with the traffic volume on I-94W near Minneapolis-St. Paul. Data was recorded regularly over the space of 6 years.

Response

The response variable is the total traffic volume measured in cars per hour, ranging continuously from a minimum of 0 cars to a maximum of 7,280



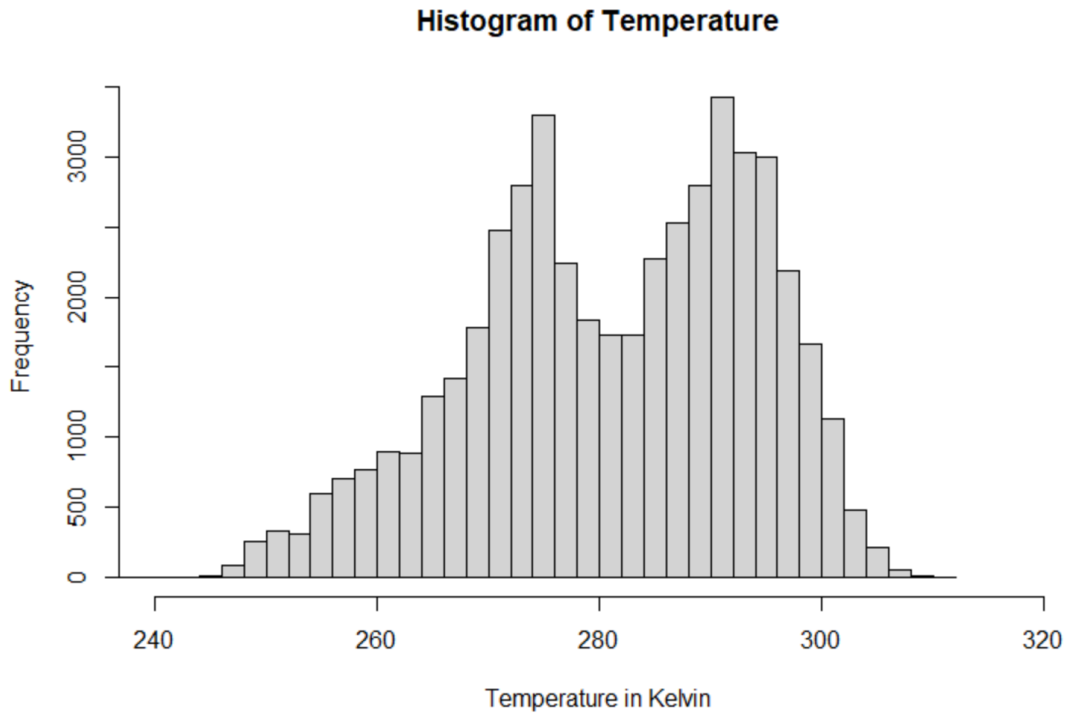
Regressors

The dataset initially used 8 regressors:

1) **Date-Time** was recorded several times a day starting October 2012 and continuing to 2018

Year:	2012	2013	2014	2015	2016	2017	2018
Observations:	2559	8573	4839	4373	9306	10605	7949

2) **Temperature** is the ambient air temperature at the time of recording (range 243.4 Kelvin to 310.1K). The raw data did include 10 observations with a reported temperature of 0 degrees Kelvin. These entries were treated as erroneous and not included in temperature-related analysis.



3) **Holiday** indicates what holiday, if any, was observed that day. The holidays included were the federal holidays of Columbus Day, Veterans Day, Thanksgiving, Christmas, New Year's Day, Washington's Birthday (President's Day), Memorial Day, Independence Day, Labor Day, Martin Luther King Jr Day, and the state recognition of the Minnesota State Fair

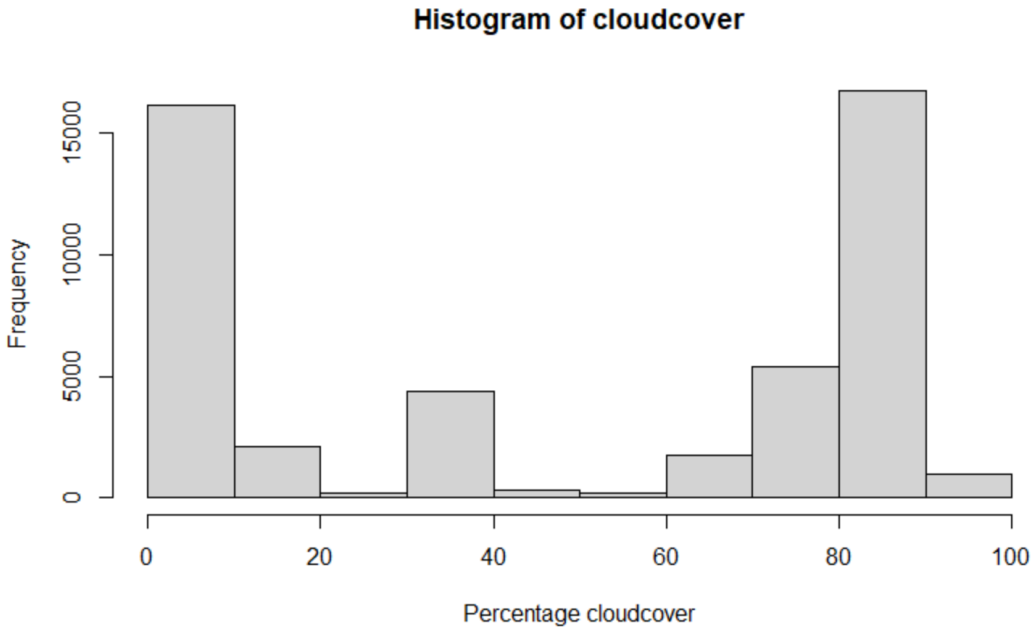
4) **Rain** is the amount of rain that occurred in the hour (range 0mm to 9831mm)
Rain contains a significant outlier, with an average value near 0 and a maximum value near 9.8m of rain, which occurred in July 2016 during a flash flood.

	Min	Median	Average	Max
Rain	0.0 mm	0.0 mm	0.334 mm	9831.3 mm

5) **Snow** is the amount of snow that occurred in the hour (range 0mm to 0.5mm)

	Min	Median	Average	Max
Snow	0.0 mm	0.0 mm	0.0002 mm	0.51 mm

6) **Clouds** is the percentage of cloud cover in the sky (range 0% to 100%)



7) **Weather Main** is a short description of the dominant weather at that time (eg, ‘cloudy’, ‘rainy’). There were a total of 11 factors for weather main.

Most frequently recorded weather:

Weather:	Clouds	Clear	Mist	Rain	Snow	Drizzle	Haze	Thunder
Count:	15164	13391	5950	5672	2876	1821	1360	1034
% of instances:	31.5%	27.8%	12.3%	11.8%	5.6%	3.8%	2.8%	2.1%

8) **Weather Description** is a longer description of the dominant weather (eg, ‘thunderstorm heavy rain’). There were a total of 38 recorded factors.

Data Cleaning

Code used in data cleaning is located in Appendix A

Date & Time

Our first data cleaning task was changing the units and format for Date-Time and Temperature regressors. For date-time, the original format was YYYY-MM-DD HH:MM, which we converted into separate columns for date (MM-DD) and time (HH:MM)

Holiday

The original holiday attribute was a character listing the holiday observed, but only for the first observation of that day. The holiday attribute was modified to be a logical value (true or false) and, using the date, expanded to apply to each observation on that day. This increased the occurrence of holidays to a total of 2.88% of all observations.

	Count of Non-Holiday	Count of Holiday
Before Cleaning	48143	61 *(0.13%)*
After Cleaning	46818	1386 *(2.88%)*

Weather Max Occurance

We also decided to measure which type of weather was the most recorded across all observations for a given day - the overview of the weather for that day. This information was added as a new column of the data frame.

Traffic Volume Intervals

For easier comparison and categorization, we created traffic volume intervals of 250 cars per hour in addition to retaining the original precise traffic volume.

Temperature

For the temperature attribute, we converted Kelvin to Celsius, which made the range (-30°, 37°C). It is noteworthy that data included 10 observations reporting an impossible temperature of 0 degrees Kelvin - these observations had the temperature marked as NA due to erroneous reporting.

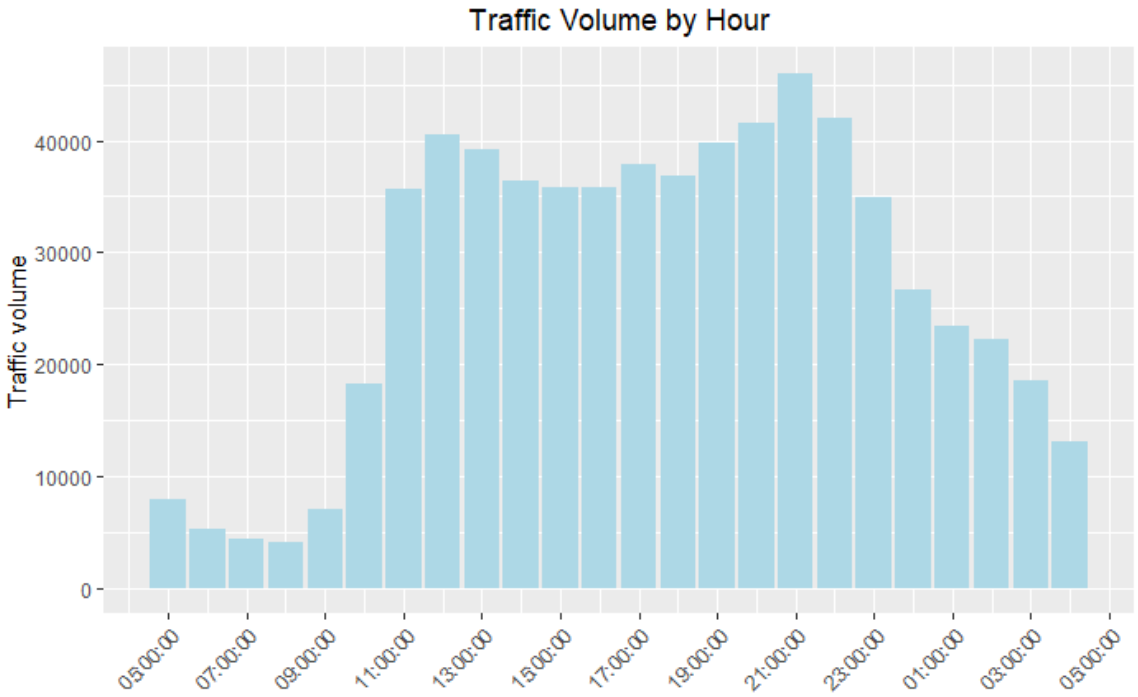
Cleaned Data Overview

```
holiday          temp            rain_1h           snow_1h           clouds_all
None:46818   Min.   :-29.7600   Min.   :  0.000   Min.   :0.0000000   Min.   :  0.00
Yes : 1386   1st Qu.: -0.9675   1st Qu.:  0.000   1st Qu.:0.0000000   1st Qu.:  1.00
             Median :  9.3100   Median :  0.000   Median :0.0000000   Median : 64.00
             Mean   :  8.1142   Mean   :  0.334   Mean   :0.0002224   Mean   : 49.36
             3rd Qu.: 18.6600   3rd Qu.:  0.000   3rd Qu.:0.0000000   3rd Qu.: 90.00
             Max.   : 36.9200   Max.   :9831.300   Max.   :0.5100000   Max.   :100.00
             NA's   :10
weather_main       weather_description   date_time                   traffic_volume
Length:48204       Length:48204        Min.   :2012-10-02 09:00:00   Min.   :   0
Class :character   Class :character    1st Qu.:2014-02-06 11:45:00   1st Qu.:1193
Mode  :character   Mode  :character    Median :2016-06-11 03:30:00   Median :3380
                                       Mean   :2016-01-05 10:46:16   Mean   :3260
                                       3rd Qu.:2017-08-11 06:00:00   3rd Qu.:4933
                                       Max.   :2018-09-30 23:00:00   Max.   :7280

     date                time                Date                Time              TimeSec
Min.   :2012-10-02   Length:48204       Min.   :2012-10-02   Length:48204       Min.   :    0
1st Qu.:2014-02-06   Class :character   1st Qu.:2014-02-06   Class :character   1st Qu.:21600
Median :2016-06-11   Mode  :character   Median :2016-06-11   Mode  :character   Median :39600
Mean   :2016-01-04                      Mean   :2016-01-04                      Mean   :41188
3rd Qu.:2017-08-11                      3rd Qu.:2017-08-11                      3rd Qu.:61200
Max.   :2018-09-30                      Max.   :2018-09-30                      Max.   :82800

       max.occurence   traffic_interval_of_250
sky is clear   :15713   2      : 5419
mist           : 9893   19     : 2853
overcast clouds: 5348   20     : 2758
broken clouds  : 3901   4      : 2669
light rain     : 3522   3      : 2536
light snow     : 2219   18     : 2405
(Other)        : 7608   (Other):29564
```

1 - Preliminary Analysis & Data Visualization

[Appendix B]

Figure 1.1

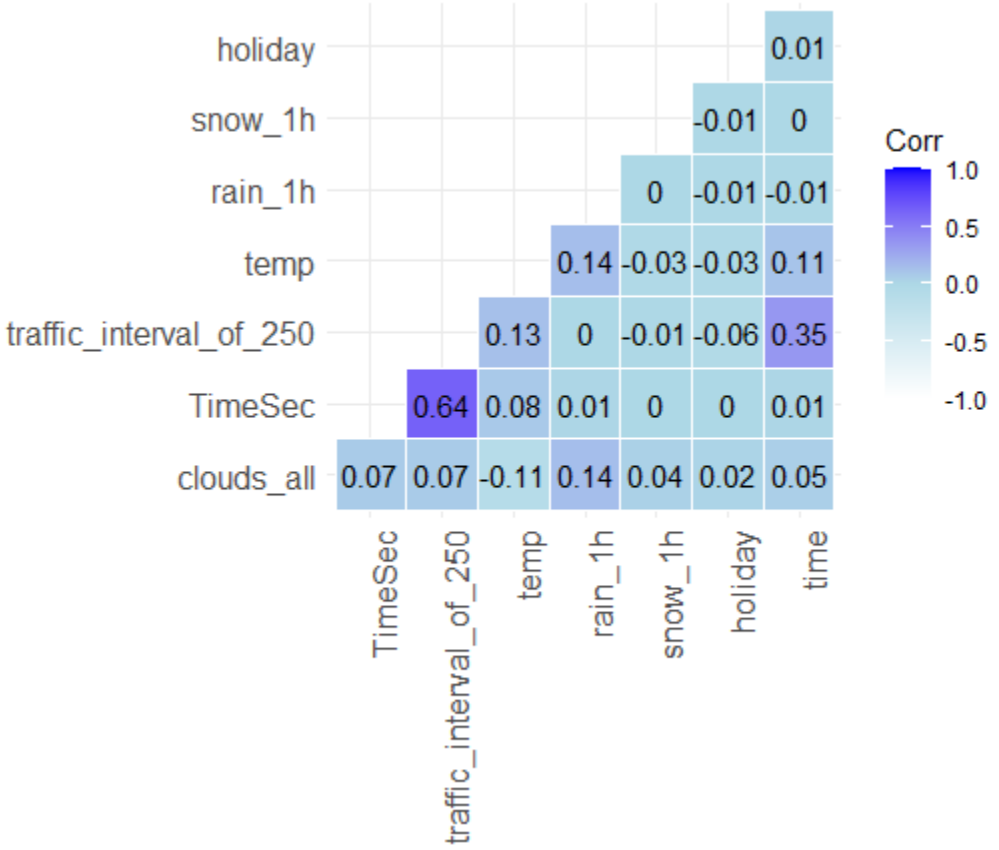


We start by visualizing the distribution of traffic volume over the course of the day.

As could be expected, the highest traffic volumes typically occur around 11:00 and 21:00, with the lowest traffic volumes occurring in the early morning around 7:00 prior to morning rush hour.

Additionally, most hours do not show a strong difference from adjacent hours (ie, 17:00 and 18:00 show similar traffic volume), and the trend is more general to the course of the day rather than particular hours. We use this observation in creating the regression model, choosing to represent time in 4 intervals rather than individual hours.

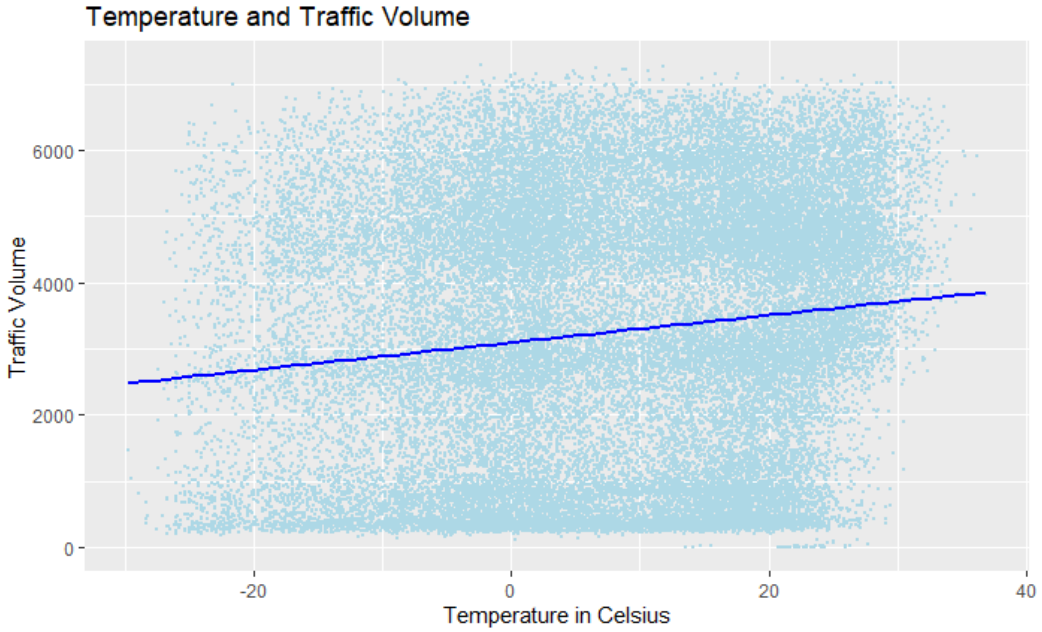
Figure 1.2



Through the correlation matrix, it's clear that traffic volume has the highest correlation with time, followed by temperature, and low correlation with the other environmental conditions.

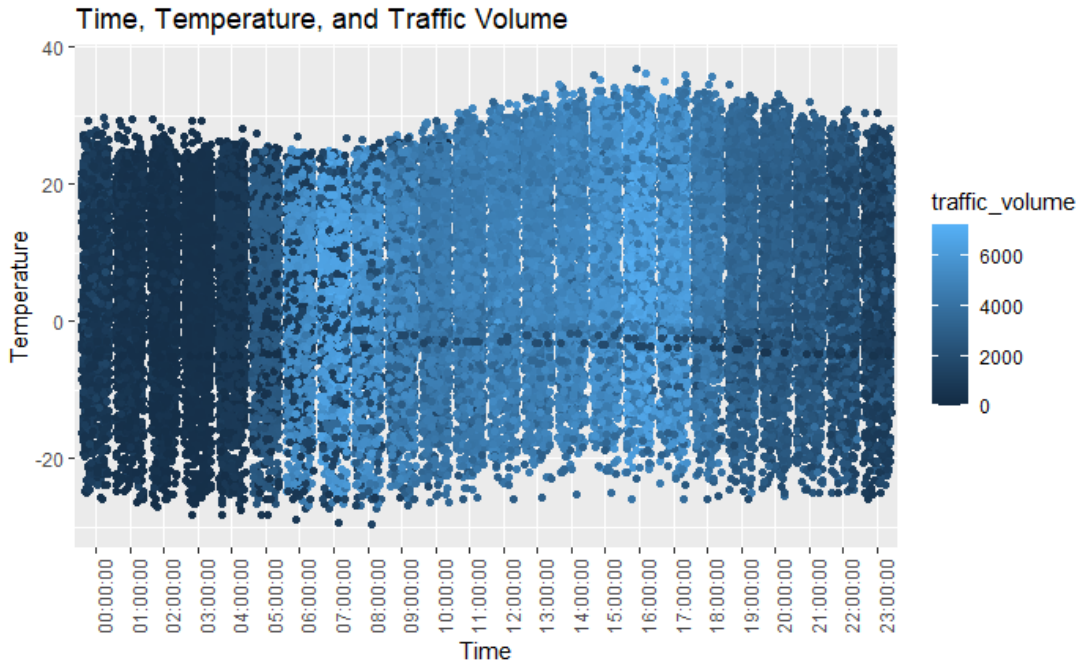
It's noteworthy that some environmental conditions display a strong association with one another - in particular, hourly rainfall has a strong correlation with both temperature and cloud cover. Additionally, temperature shows a correlation with time of day, most likely because temperatures typically fall overnight and rise during the day.

Figure 1.3A



Modeling the individual relationship between temperature and traffic volume, we see a general positive association. Although traffic volumes are widely distributed, there is an association and we see that temperature is correlated with traffic volume to some degree.

Figure 1.3B



We can see a combined relationship between time, temperature and traffic volume. Clearly temperature and time of day are associated, as temperatures trend lowest just before daybreak and peak in the mid-afternoon. Traffic volume tends to be highest in the morning and late afternoon, as was shown by Figure 1.1.

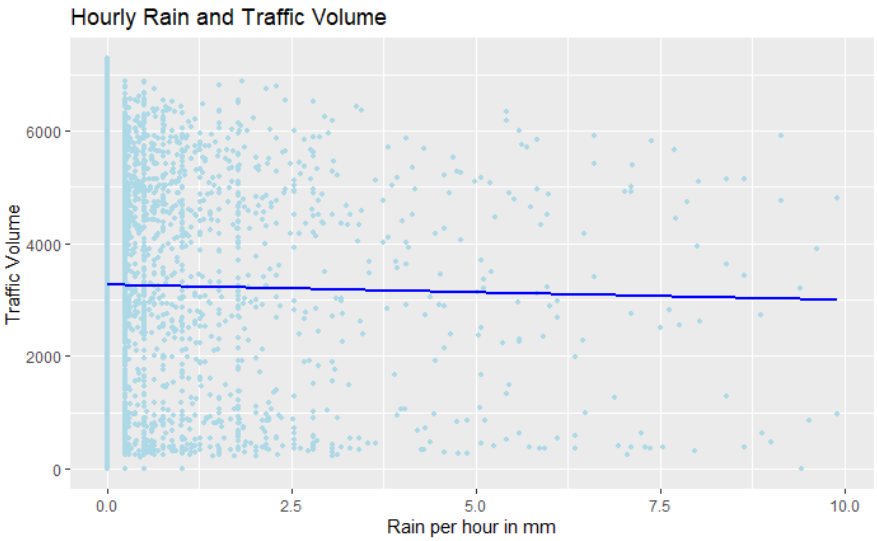


Figure 1.4
Hourly rainfall, on the other hand, has little to no association with traffic volume.

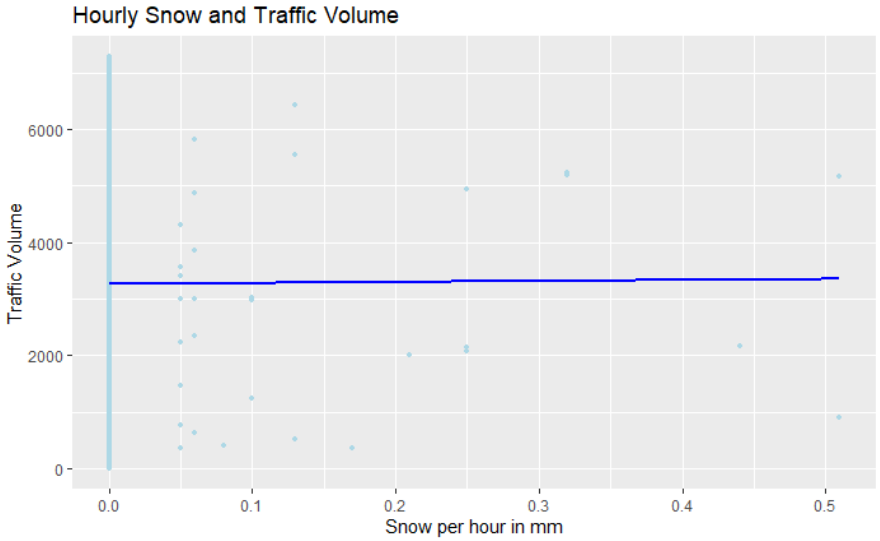


Figure 1.5
Similarly, we see hourly snowfall does not appear to impact the traffic volume. The results for both rainfall and snow are sensible; Minnesota receives both frequently, and has the infrastructure for the city freeways to continue operating. Drivers are experienced in both weather conditions and unlikely to change their routines or avoid driving due to normal volumes of rain or snow.

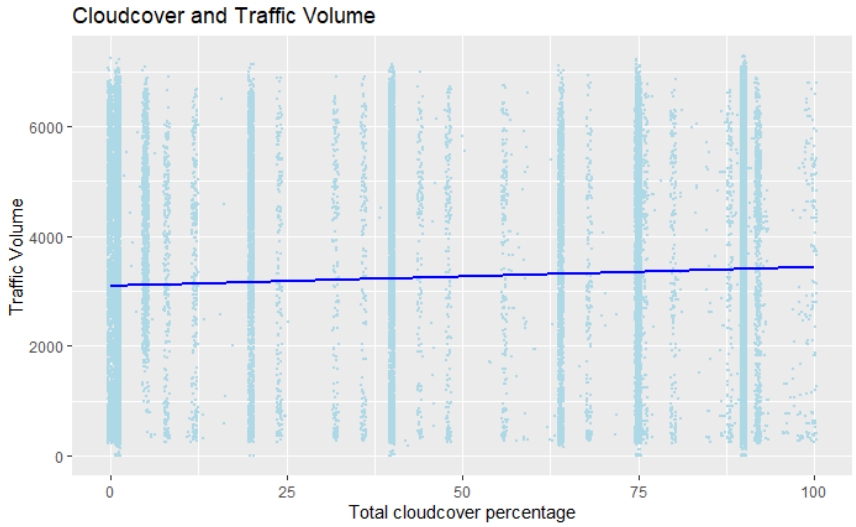
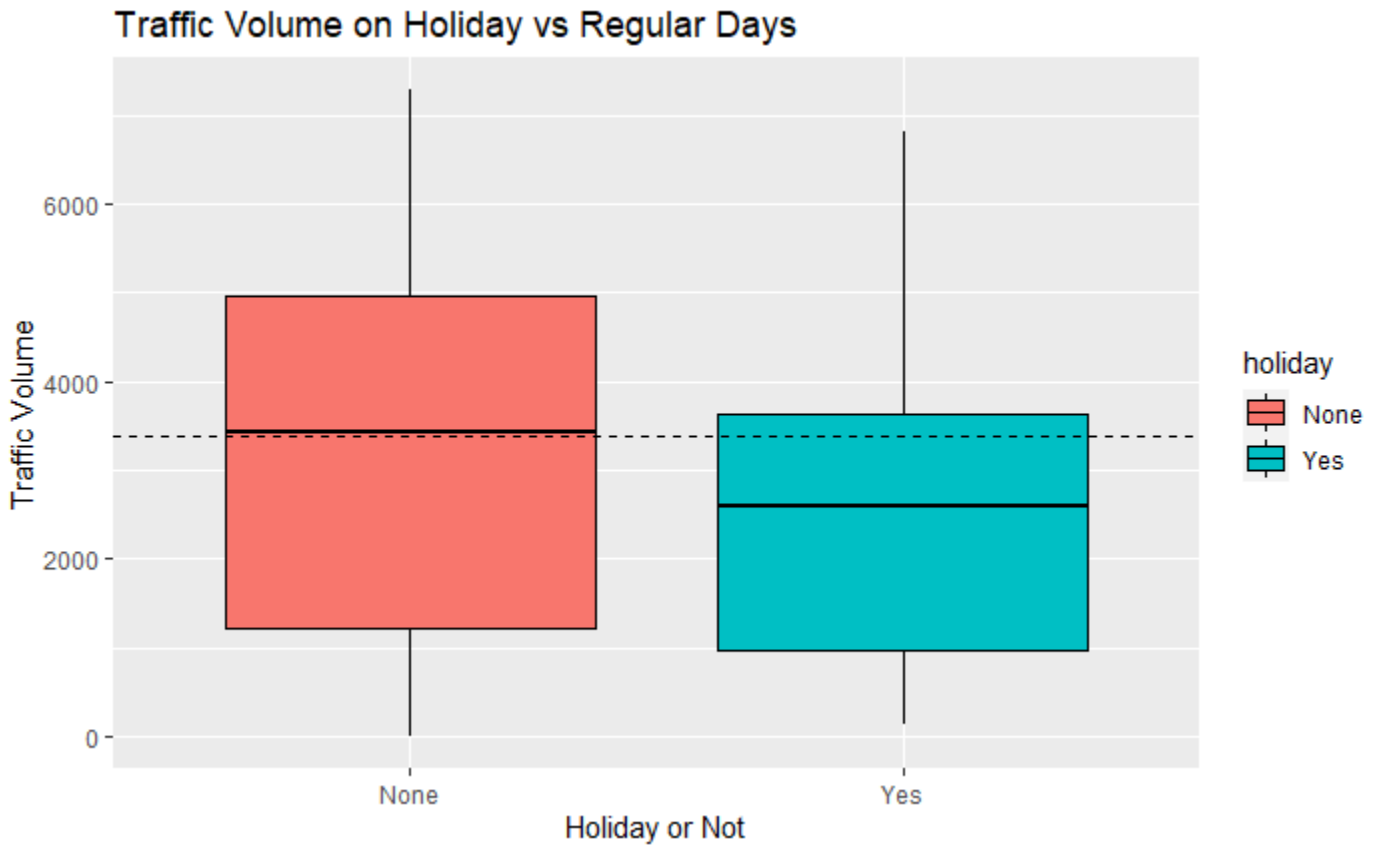


Figure 1.6
Along with rain and snow, cloudcover predictably does not have much influence over traffic. We see striations in the horizontal axis, representing that certain cloud cover percentages are more likely to be recorded than others.

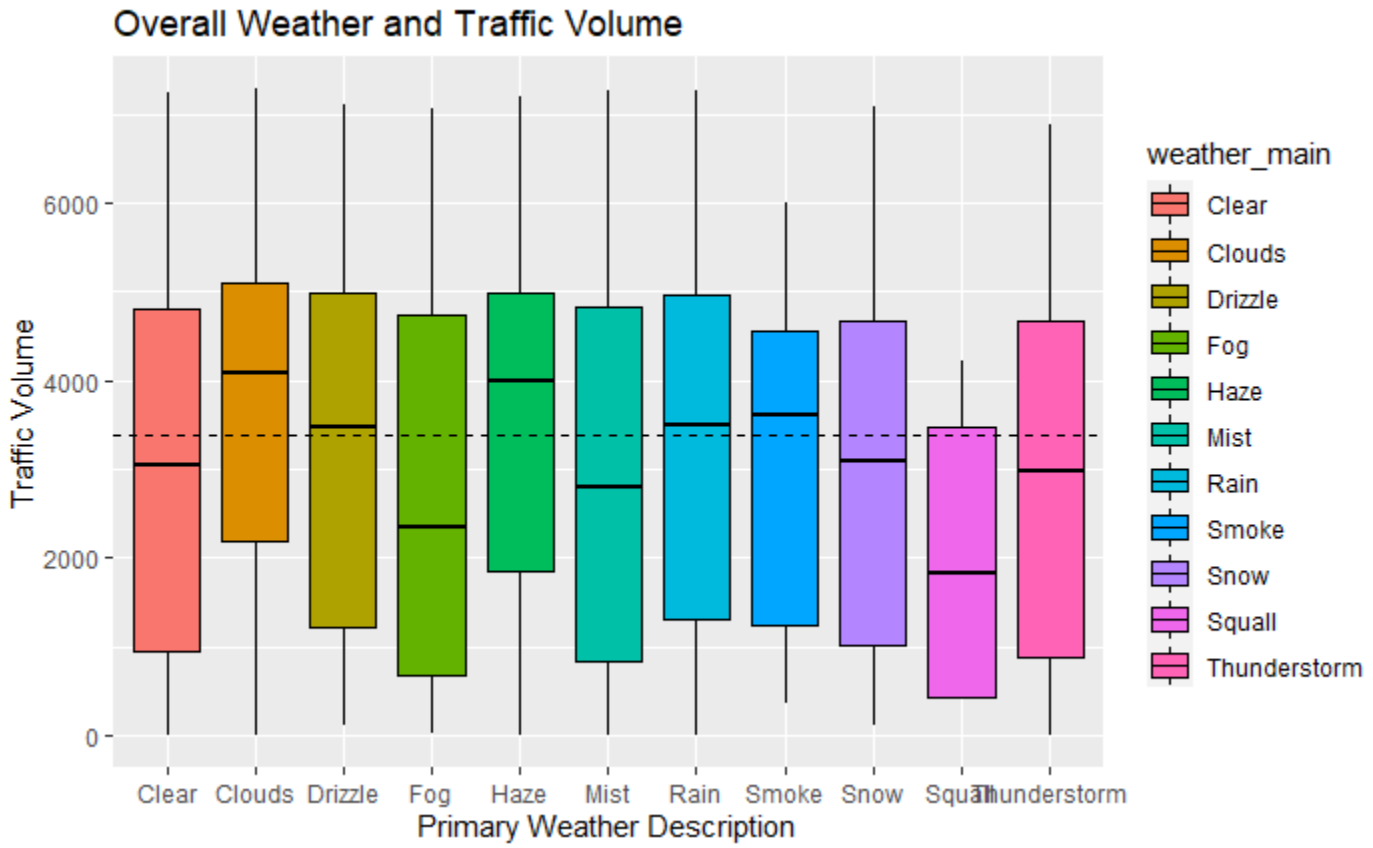
Figure 1.7



We can also represent the impact of holidays on traffic volume. Representing the overall median traffic volume of 3380 cars per hour with the dashed line, holidays clearly have lower traffic volume in general compared to non-holidays. Non-holidays have a median very near the overall median, while holidays have a significantly lower median. This is likely attributable to holidays reducing the number of people commuting to work on that day, which is a significant proportion of freeway drivers on regular days.

	Overall	Non-Holiday	Holiday
Observations	48204	46818	1386
Median Traffic Volume	3380	3435	2587
Difference in Median		+ 1.6%	- 23.5%

Figure 1.8



Finally, we can observe the relationship between broader weather patterns and the traffic volume. The dashed line again represents the overall median traffic volume across all observations (3380 cars per hour).

Most weather types have a median similar to the overall median, but there are some exceptions. In particular, we see ‘squall’, ‘fog’, and ‘mist’ are associated with lower traffic volumes, while clouds and haze have the highest median volume:

	Overall	Squall	Fog	Mist	Clouds	Haze
Observations	48204	4	912	5950	15164	1360
Median Traffic	3380	1818	2339	2800	4072	3987
Change in Median		- 46.2%	- 30.8%	- 17.2%	+ 20.5%	+ 17.9%

From the table above, we see that while squall in particular has a lower median traffic volume, it is also the least common primary weather description with only 4 observations. Fog and mist both have more obserservations, and are closer to the median.

Cloudy days have a notable higher median, as well as a significant proportion of the overall observations. Haze has a moderately higher median, but again has a lower number of observations.

2 - Primary Model

Appendix C

From the exploratory analysis of the *time* regressor, we divided time into 5 intervals:

	Time interval	Occurrences
night	(19:00 - 01:00]	19042
early	(01:00 - 07:00]	12364
midday	(07:00 - 13:00]	11988
afternoon	(13:00 - 19:00]	9810

Our primary model resulted in an adjusted R^2 value of 0.4929:

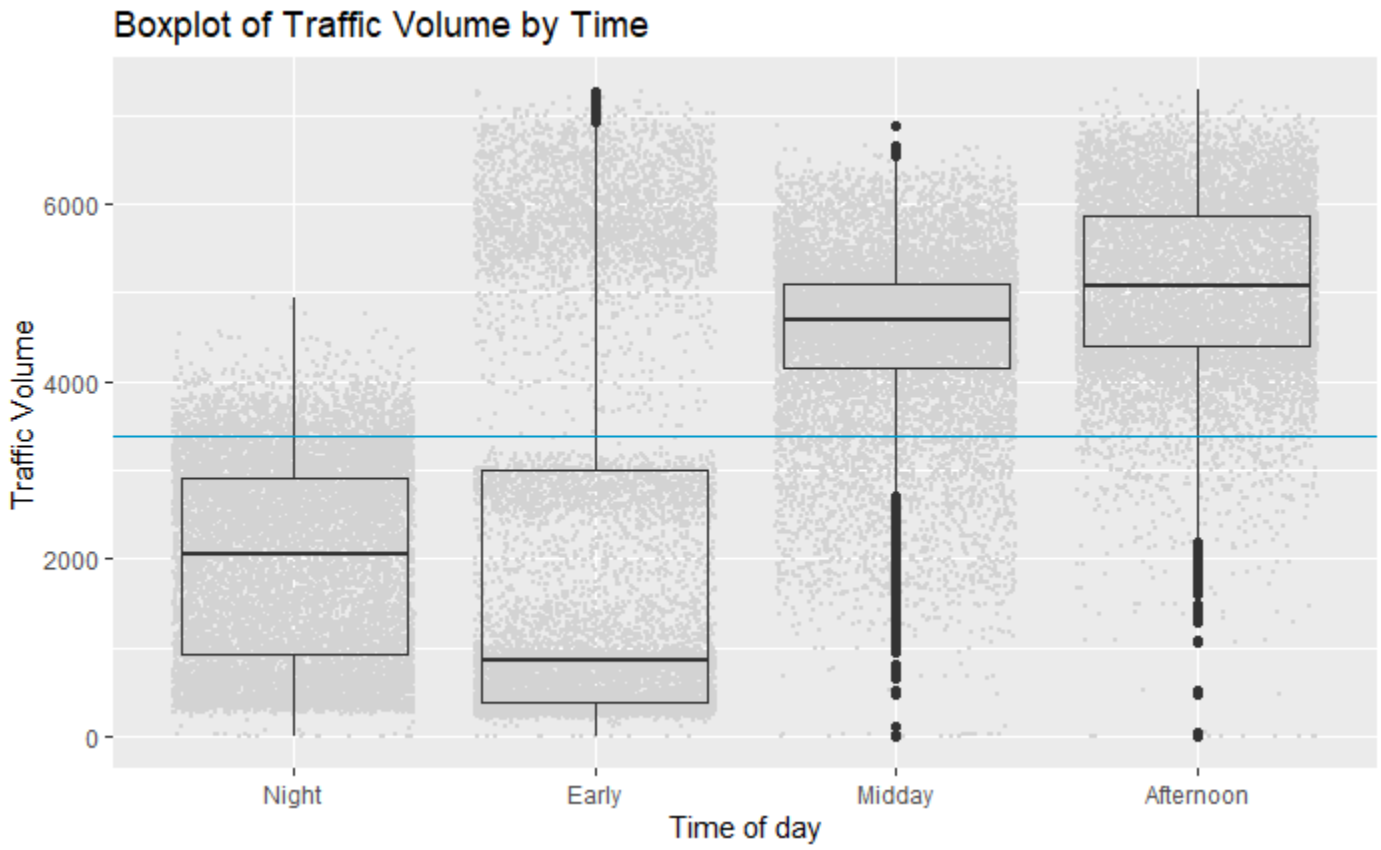
Figure 2.1

```
Residuals:
    Min      1Q  Median      3Q     Max
-5299.4 -1137.9   -41.5   775.9  5445.2

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                1864.05474   16.20845 115.005  < 2e-16 ***
holidayYes                 -778.02044   38.61823 -20.146  < 2e-16 ***
temp                          8.81455    0.55334  15.930  < 2e-16 ***
rain_1h                       0.01797    0.14391   0.125 0.900632
snow_1h                    -128.52159  790.40343  -0.163 0.870832
clouds_all                    0.27126    0.26580   1.021 0.307482
weather_mainClouds          147.99454   23.55636   6.283 3.36e-10 ***
weather_mainDrizzle          60.07370   41.45164   1.449 0.147275
weather_mainFog            -110.59292   50.40213  -2.194 0.028225 *
weather_mainHaze             -7.57680   43.00955  -0.176 0.860165
weather_mainMist            -40.02225   27.97956  -1.430 0.152606
weather_mainRain              3.11098   29.72235   0.105 0.916640
weather_mainSmoke          -406.97085  316.99288  -1.284 0.199201
weather_mainSnow           -103.55496   36.09861  -2.869 0.004124 **
weather_mainSquall         -846.24973  707.88583  -1.195 0.231913
weather_mainThunderstorm   -184.40244   49.76558  -3.705 0.000211 ***
timeIntervalEarly           169.36426   17.65533   9.593  < 2e-16 ***
timeIntervalMidday         2543.02919   17.65118 144.071  < 2e-16 ***
timeIntervalAfternoon      3076.05957   18.82075 163.440  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1415 on 48175 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.493,     Adjusted R-squared:  0.4929
F-statistic:  2603 on 18 and 48175 DF,  p-value: < 2.2e-16
```

From the primary model we see that several observations in Section 1 seem to be correct. Holidays, temperature, and time appear to be highly significant, while hourly rain, clouds, and snow are insignificant. Some of the weather descriptions highlighted under Figure 1.8 (clouds and fog) are indeed of interest, and we add snow to the list of potentially influential weather events. Squall and haze do not seem to be of value in explaining variation in traffic volume.

We can see the impact of the time intervals on the traffic volume by modeling it with a box plot

Figure 2.2



Here, we again represent the overall median traffic volume with the horizontal line at 3380 cars per hour. We see that early morning, morning, afternoon, and evening all have traffic consistently above the overall median, while nighttime traffic is significantly lower with a median of only 799 cars per hour.

Interestingly, the early morning interval has a much wider distribution with dense clustering above and below the median. This is likely related to the hours covered by early morning, as traffic is seen to reach its lowest point during this time and then rise rapidly with morning commuters.

	Overall	Early	Night	Midday	Afternoon
Median Volume	3380	835	2040	4681	5073
Change		-75.3%	-39.6%	+38.5%	+50%

3 - Reduced Model

Appendix D

From the data exploration in Section 1 and the primary model in Section 2, we created a reduced model of the regressors most likely to be significant. The adjusted R^2 value of the reduced model is the same: 0.4929

Figure 3.1

```
Residuals:
    Min      1Q  Median      3Q     Max
-5304.8 -1138.3   -41.8   775.1  5447.3

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1865.3313    16.1572 115.449  < 2e-16 ***
holidayYes                 -777.6861    38.6156 -20.139  < 2e-16 ***
temp                          8.7240     0.5459  15.980  < 2e-16 ***
weather_mainClouds          164.7372    16.9055   9.745  < 2e-16 ***
weather_mainDrizzle          82.1425    35.3636   2.323 0.020194 *
weather_mainFog             -96.9672    48.5768  -1.996 0.045922 *
weather_mainHaze              7.1920    40.5039   0.178 0.859066
weather_mainMist            -22.6766    22.1557  -1.024 0.306072
weather_mainRain             22.8123    22.6075   1.009 0.312952
weather_mainSmoke          -392.1395   316.6530  -1.238 0.215578
weather_mainSnow            -83.3839    30.0676  -2.773 0.005553 **
weather_mainSquall         -824.7105   707.5581  -1.166 0.243793
weather_mainThunderstorm   -165.1901    46.0687  -3.586 0.000336 ***
timeIntervalEarly           168.1786    17.6118   9.549  < 2e-16 ***
timeIntervalMidday         2543.4375    17.6464 144.134  < 2e-16 ***
timeIntervalAfternoon      3076.8484    18.8046 163.622  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1415 on 48178 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.493,     Adjusted R-squared:  0.4929
F-statistic:  3124 on 15 and 48178 DF,  p-value: < 2.2e-16
```

To compare the primary and reduced models, we completed an ANOVA analysis table.

Figure 3.2

Model 1: traffic_volume ~ holiday + temp + weather_main + timeInterval
Model 2: traffic_volume ~ holiday + temp + rain_1h + snow_1h + clouds_all + weather_main + date + timeInterval

	Res.df	RSS	Df	Sum of Sq	F	Pr(>F)
1	48178	9.6439e+10				
2	48175	9.6437e+10	3	2159846	0.3596	**0.7822**

From the p-value of 0.782, we fail to reject the null hypothesis and conclude that the dropped regressors of date, hourly rain, snow, and cloud cover are likely to be insignificant, and so we adopt the reduced model for the rest of the analysis.

To confirm that no further model reduction is required, we also assessed the variance inflation factors for the reduced model.

Figure 3.3

Regressor	VIF		
holiday	1.0028	weather - clouds	1.4838
temperature	1.1591	weather - drizzle	1.0947
Early	1.4237	weather - fog	1.0548
Midday	1.4009	weather - haze	1.0832
Afternoon	1.3802	weather - mist	1.2790
		weather - rain	1.2778
		weather - smoke	1.0014
		weather - snow	1.2214
		weather - squall	1.003
		weather - thunderstorm	1.0728

From the variance inflation factor table, we see all values are below 2, showing no evidence of multicollinearity. We conclude that the reduced model does not require further variable selection and continue analysis using the model outlined in Figure 3.1.

4 - Residual Analysis Pre-Transformation

Appendix E

Continuing analysis, we explore the residual values to identify outlying residuals.

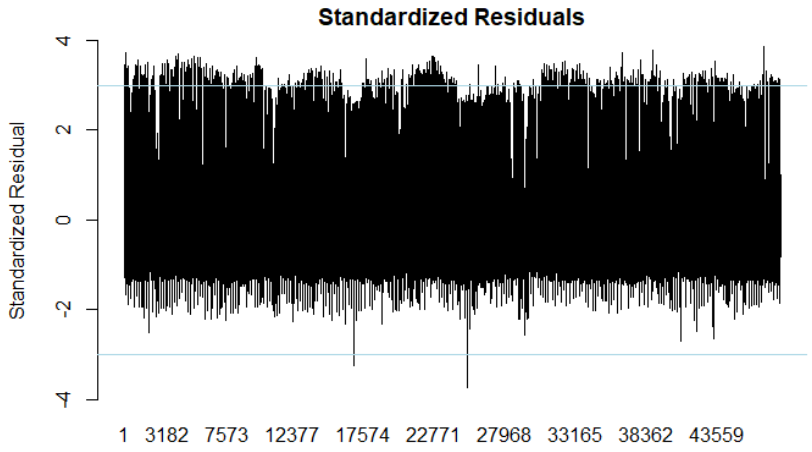


Figure 4.1
The standardized residuals identify 667 residuals above 3 and 31 below -3

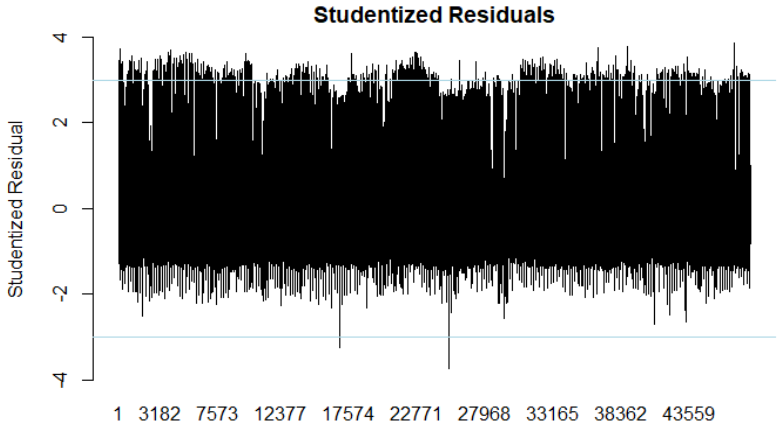


Figure 4.2
The studentized residuals identify 669 residuals above 3 and 31 below -3

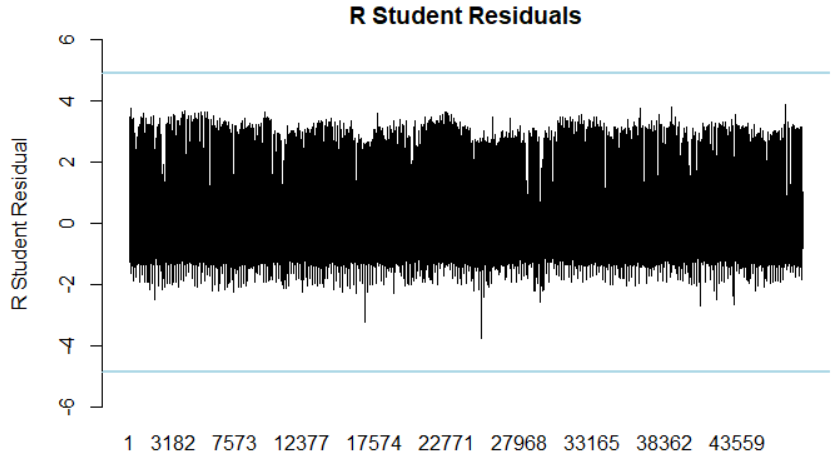
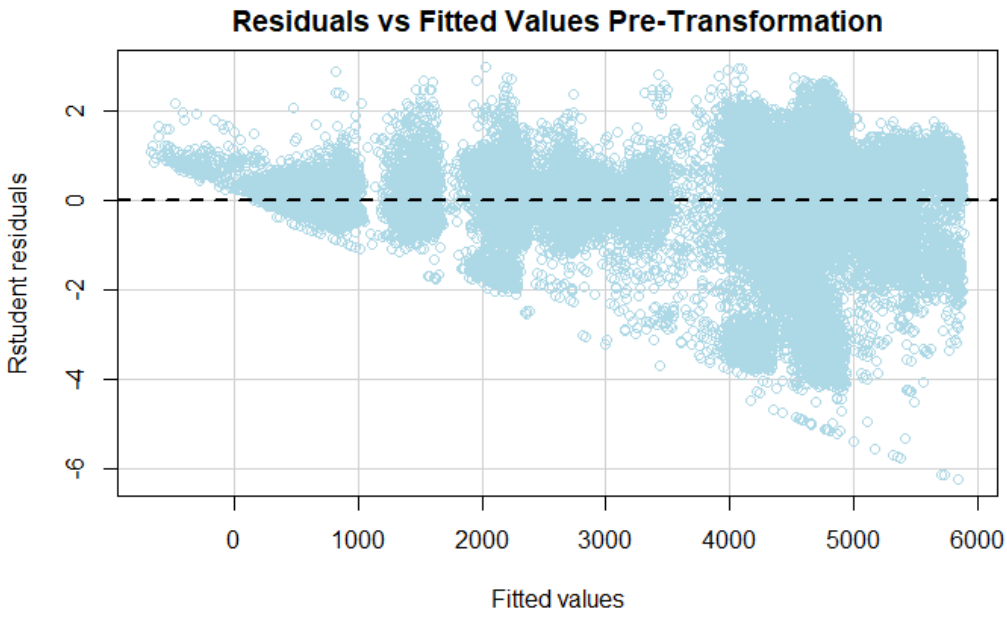


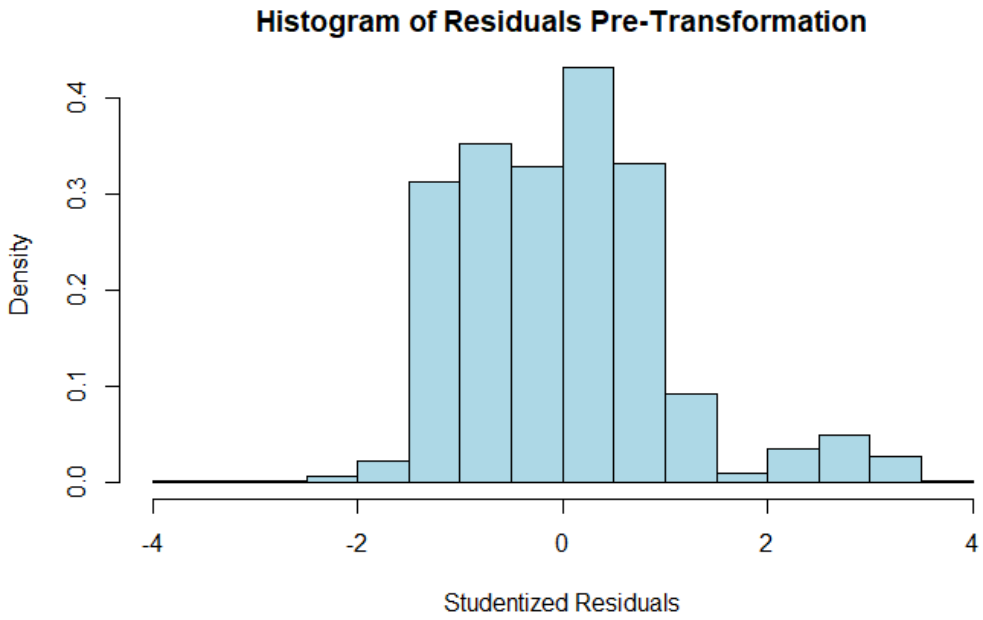
Figure 4.3
Finally, the R student residuals indicate no observations outside the acceptance range, calculated as -4.8 to 4.8

Figure 4.4



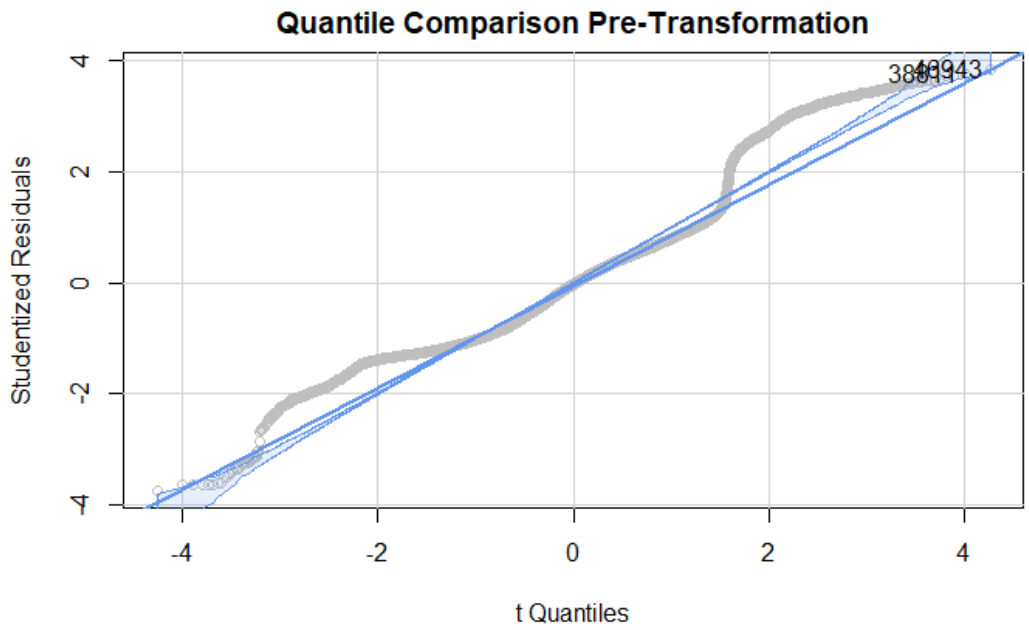
Our plot of residual vs fitted values shows some interesting features. The data displays a fan or cone shape, violating the constant variance assumption and indicating that a transformation is needed. There are more residuals clustered significantly below the 0 line than above

Figure 4.5



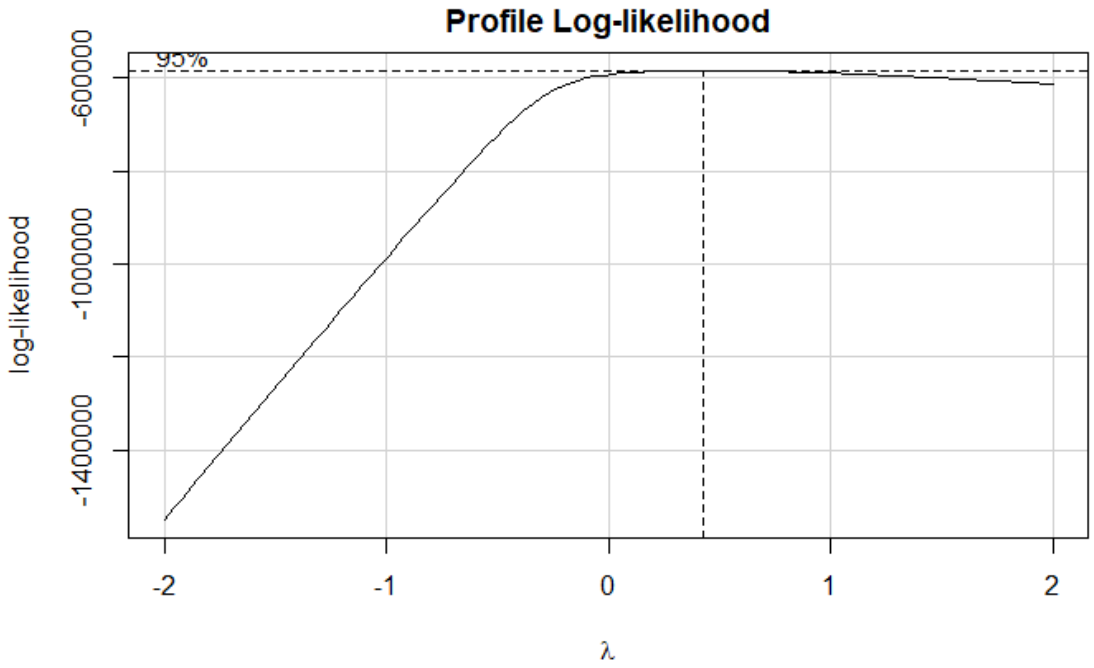
The histogram of residuals displays relative normality, roughly centered around zero. However, we see a somewhat asymmetric distribution on either side of the peak, and the peak itself is flatter than ideal.

Figure 4.6



The quantile comparison plot also has some interesting features indicating a transformation will be necessary. We see violations of the expected linearity throughout the plot, increasing away from the center of the plot

Figure 4.7



The box cox plot results in an ideal value of λ = 0.424 . This value, near 0.5, is affirmed by the fan shape of the residual plot in Figure 4.4, which suggests a square root transformation on the response.

5 - Transformation

Appendix F

From the results of Section 4, especially Figures 4.4 and 4.7, we decided to try using the transformation

$$y^* = y^{0.5} = \sqrt{y}$$

on the response variable for traffic volume.

Figure 5.1

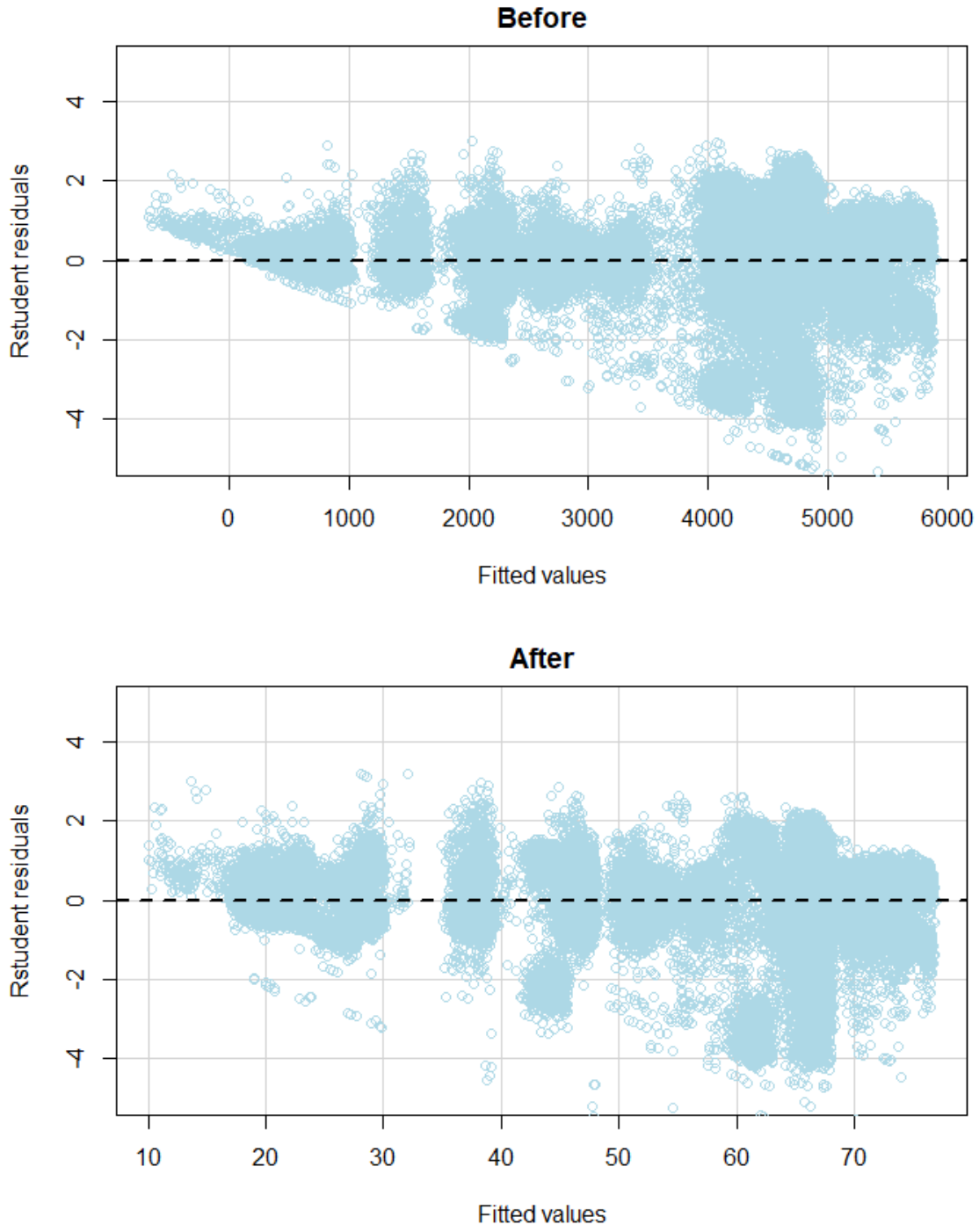
```
Residuals:
    Min      1Q  Median      3Q     Max
-71.727 -10.008   0.464   8.067  49.408

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              41.119769   0.166974 246.264  < 2e-16 ***
holidayYes               -6.770944   0.399068 -16.967  < 2e-16 ***
temp                      0.087271   0.005642  15.468  < 2e-16 ***
weather_mainClouds        1.698043   0.174708   9.719  < 2e-16 ***
weather_mainDrizzle       0.808524   0.365461   2.212 0.026948 *
weather_mainFog          -0.905419   0.502012  -1.804 0.071303 .
weather_mainHaze          0.297492   0.418583   0.711 0.477265
weather_mainMist         -0.285649   0.228966  -1.248 0.212198
weather_mainRain          0.242086   0.233635   1.036 0.300128
weather_mainSmoke        -3.211571   3.272414  -0.981 0.326397
weather_mainSnow         -0.570321   0.310730  -1.835 0.066449 .
weather_mainSquall       -8.243312   7.312177  -1.127 0.259604
weather_mainThunderstorm -1.579741   0.476092  -3.318 0.000907 ***
timeIntervalEarly        -2.218613   0.182007 -12.190  < 2e-16 ***
timeIntervalMidday       24.464043   0.182364 134.149  < 2e-16 ***
timeIntervalAfternoon    28.311810   0.194334 145.686  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.62 on 48178 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.4782,    Adjusted R-squared:  0.478
F-statistic:  2943 on 15 and 48178 DF,  p-value: < 2.2e-16
```

After applying the square root transformation, we see that many of the significant variables from before have remained as such, while some were made less significant which will allow for a more accurate analysis. In accordance with the more potent regressors being identified, the residual standard error has dropped drastically, which is further incorporated down below when analyzing residuals.

6 - Residual Analysis Post-Transformation

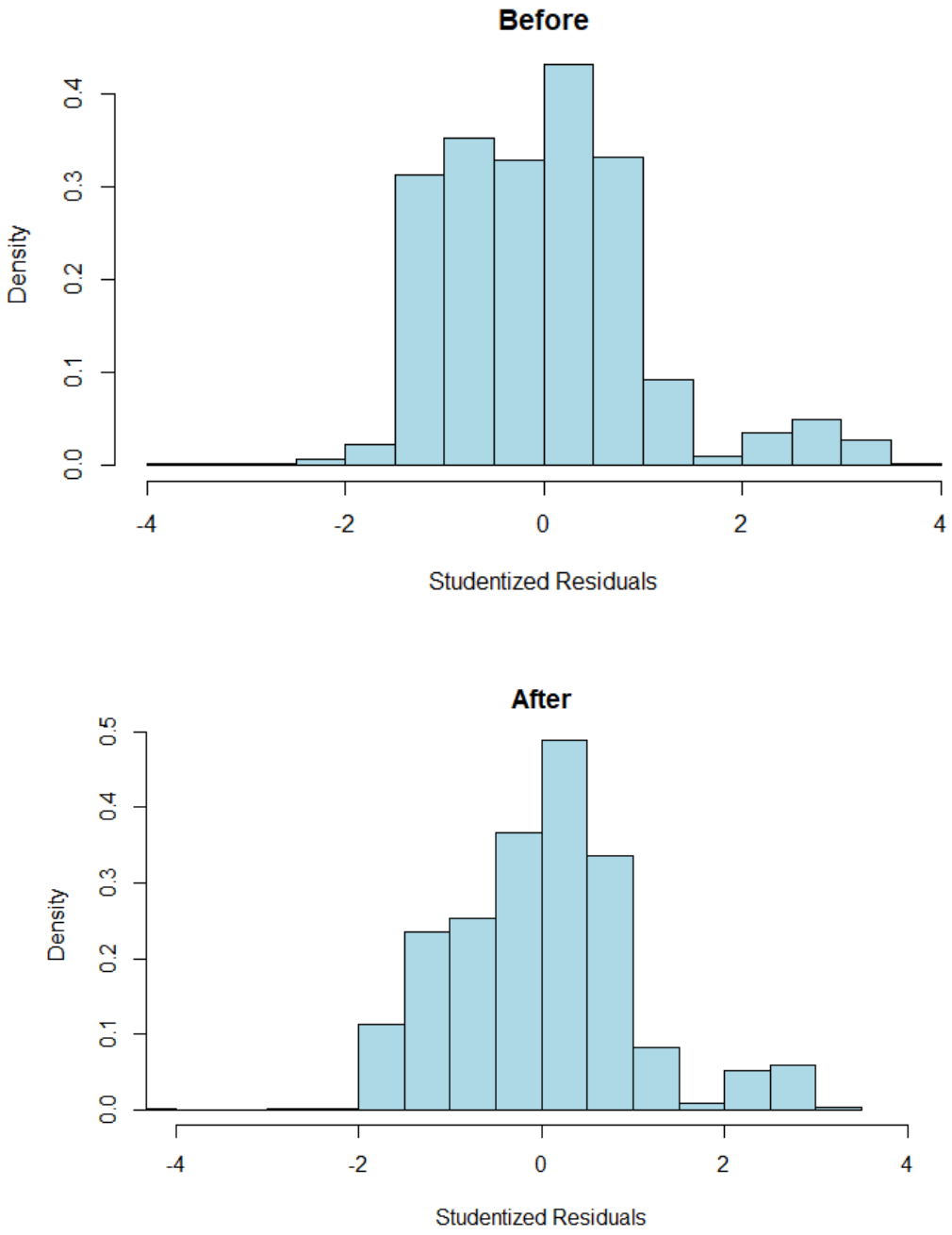
[Appendix G]

Figure 6.1



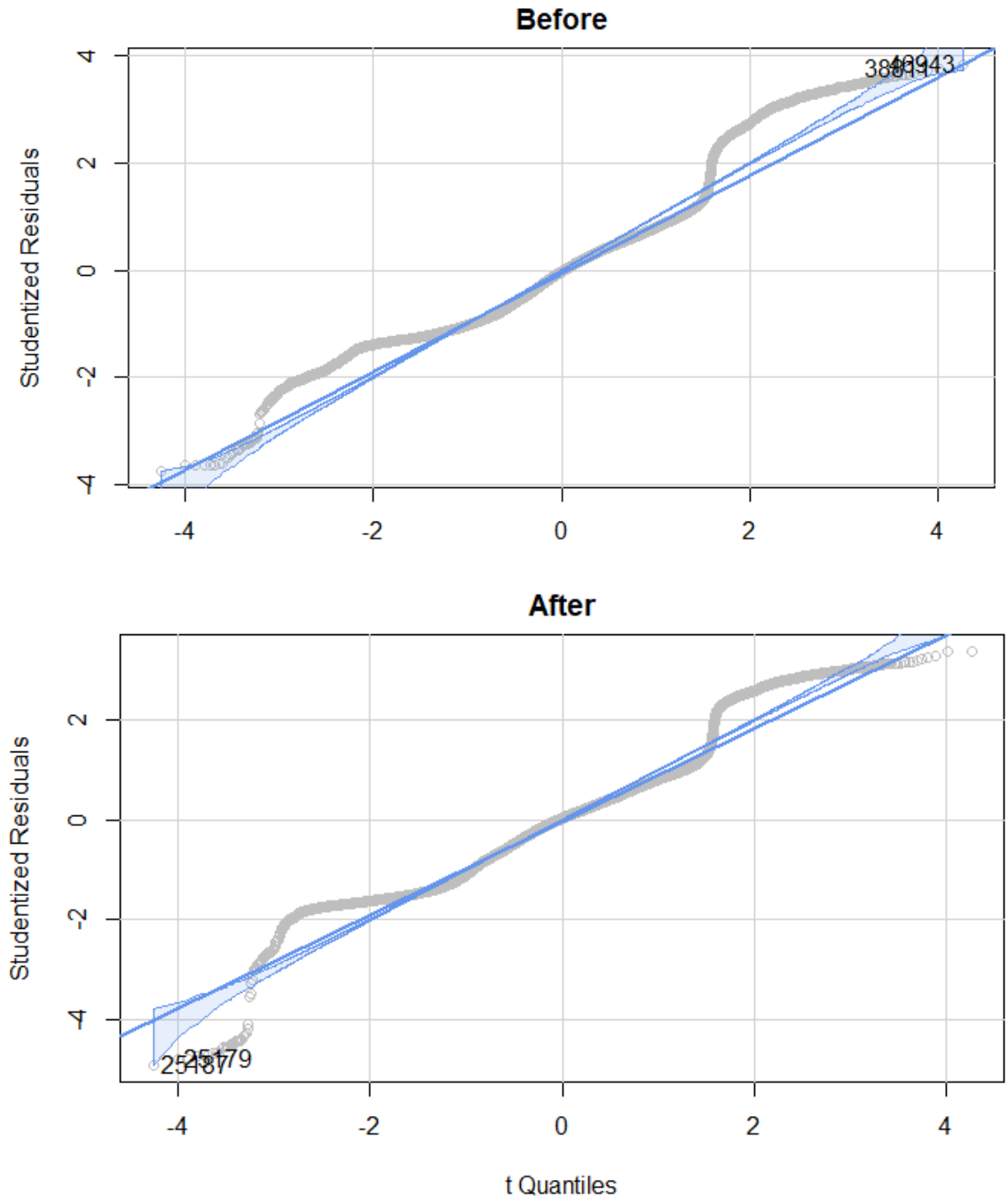
In comparing the plot of residuals vs fitted values before and after transformation, we see a moderate improvement in conformity to the constant variance assumption. The residuals are more evenly distributed around zero, with a less pronounced fan shape, particularly along the bottom. However, the transformation failed to alleviate all nonconformity in the residuals.

Figure 6.2



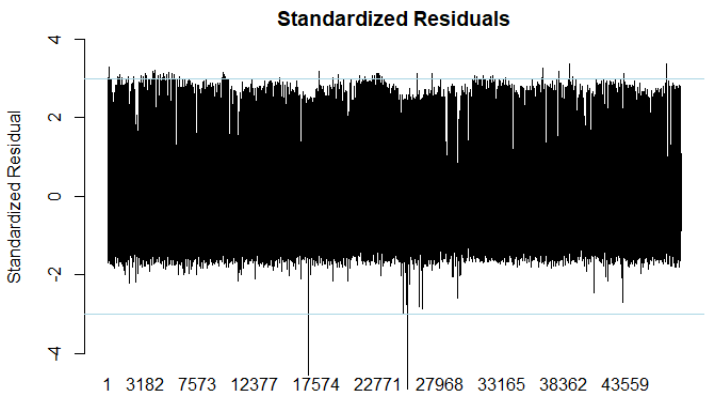
Similarly, we see a mild improvement in the normality of the residuals. Residuals are more centered around 0, with an even peak. Additionally, the distribution is more symmetric around the peak, which is more prominent than before compared to the adjacent residuals.

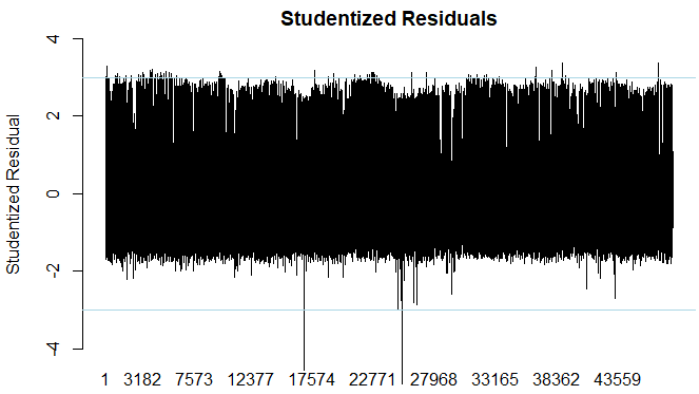
Figure 6.3

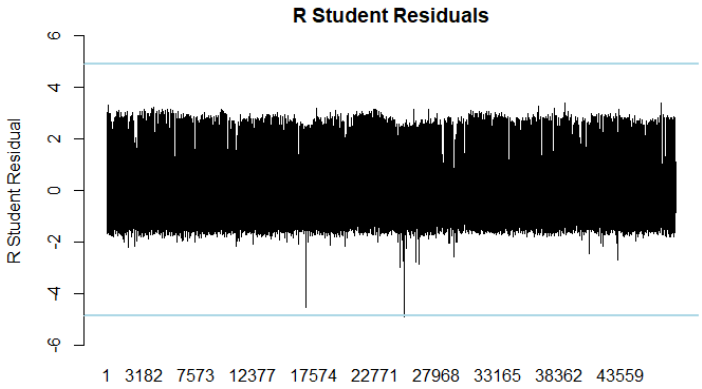


Comparing the QQ plots provides little insight into the efficacy of the transformation. This is most likely related to the data itself being difficult to perform linear regression on, as we see from other measures that the transformation was effective in improving the residuals.

Figure 6.4 - 6.6







We see that the number of residuals across all three methods was dramatically affected by the transformation. We saw a dramatic decrease in the total number of outliers which indicates that our transformation was justified now that only the most significant of variables are above and below the defined cutoff regions.

We can summarize the change in the residuals in the following table:

Figure 6.7

	Standardized	Studentized	R student	Total
Acceptance Range	- 3 to 3	- 3 to 3	- 4.8 to 4.8	
Pre-Transformation	667 above 31 below	669 above 31 below	0 above 0 below	700 *(1.45%)*
Post-Transformation	71 above 32 below	71 above 32 below	0 above 1 below	103 *(0.21%)*

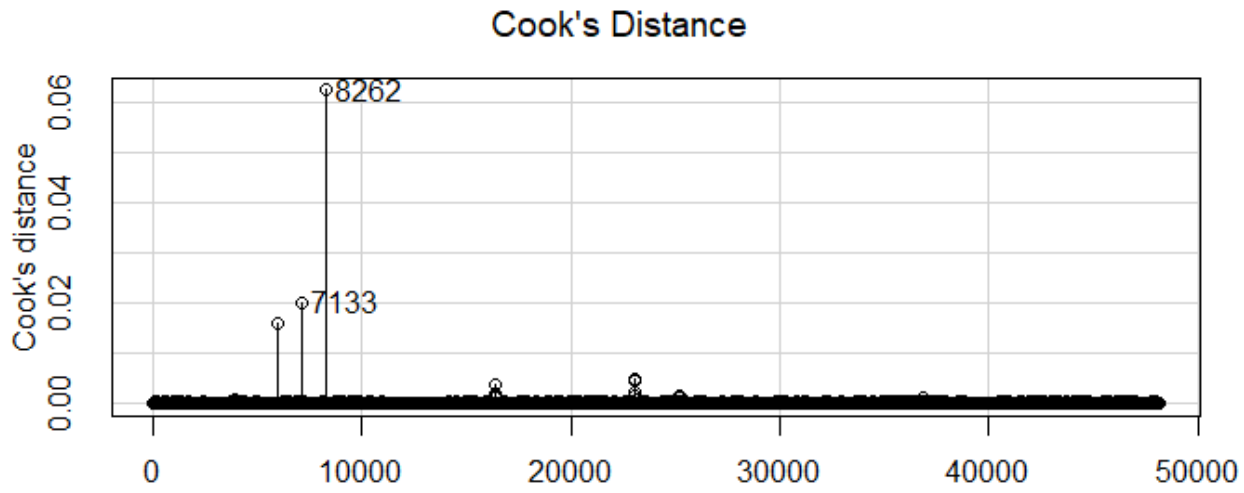
From the graphs, it is also clear that the majority of identified residuals are only marginally outside of the acceptance range.

7 - Influence Analysis

Appendix H

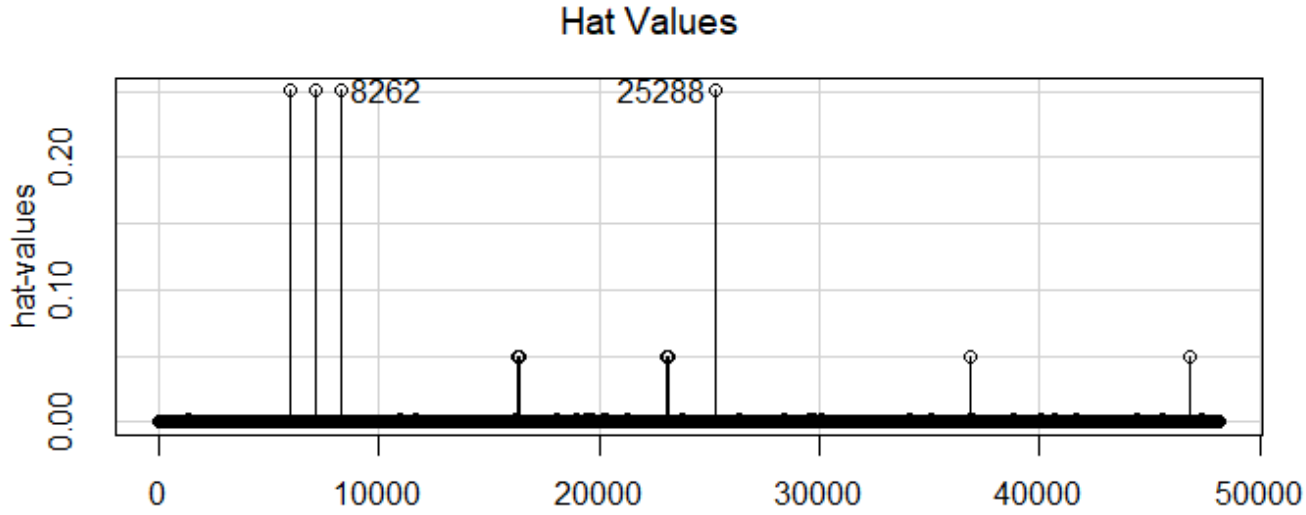
We conclude the model-building process by applying measures of influence to identify potential outliers or observations wielding unreasonable influence over the model.

Figure 7.1



All the cook's distance values are significantly less than 1, indicating that no points are exerting an outsized influence on the model parameters.

Figure 7.2



We use the hat value cutoff of *2p / n = 30 / 48204 = 0.0006*. We find the majority of points are therefore potentially influential in the x-space, likely due to having such a large data set.

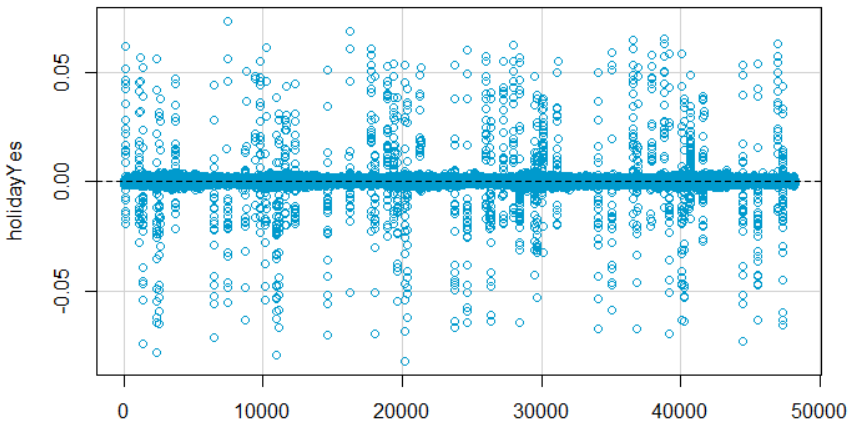


Figure 7.3 - Holiday

The DFBETAs plot for the holiday regressor shows strong clustering around zero, with an even distribution above and below, indicating that the regression coefficient for holiday is not being strongly impacted by any individual observation

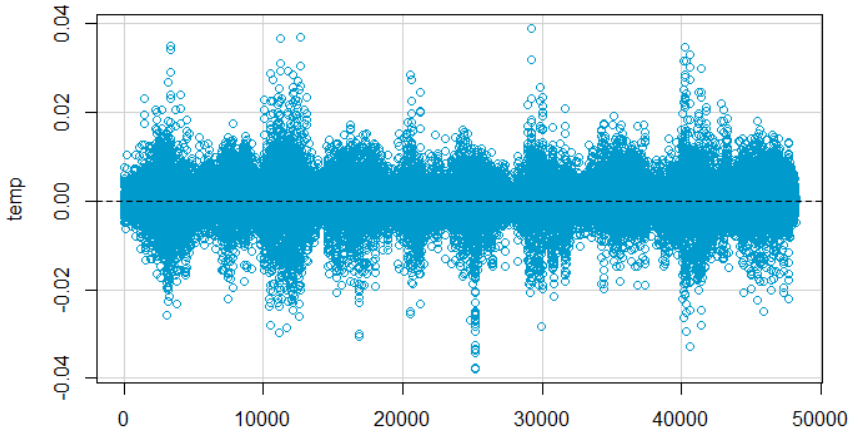
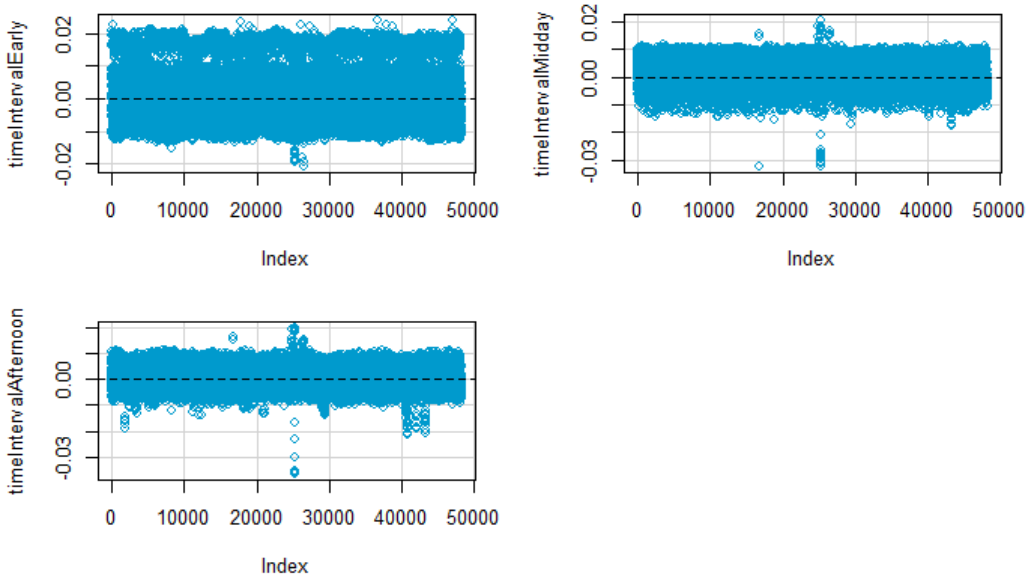


Figure 7.4 - Temperature

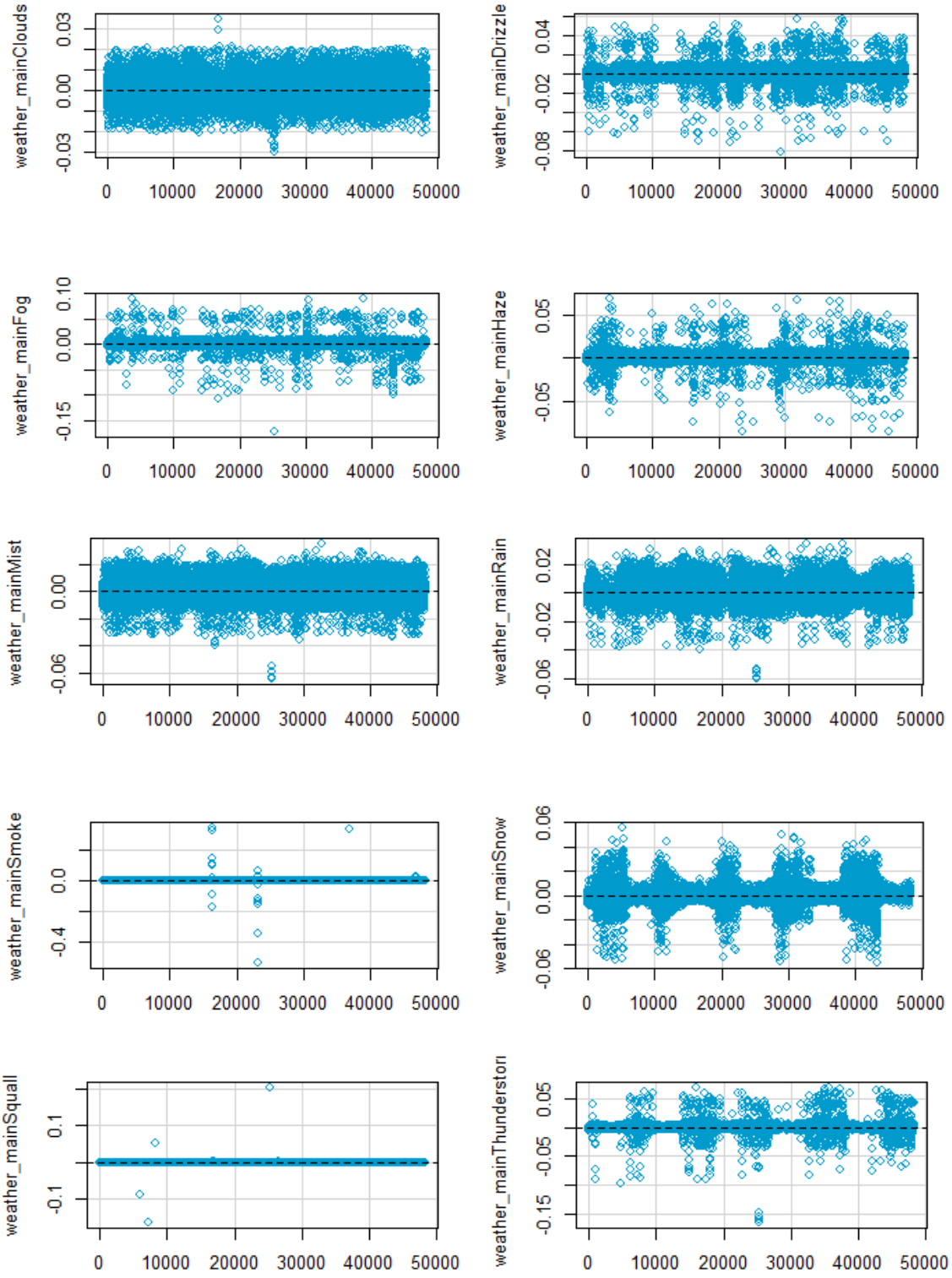
We see that the temperature DFBETAs plot is similarly compact and symmetrically clustered around zero, again indicating that the model coefficient for temperature is not overly influenced by particular points

Figure 7.5 - Time Intervals



Across the time intervals, we see very clean symmetric clustering around zero with very few values straying from this layout. A handful of points stand out from the crowd, but these remain sufficiently small to be considered not overly influential.

Figure 7.6 - Weather



Finally, we see the plots for weather are similarly satisfying, with all values well within the range to indicate no events having an outsized effect on the regression coefficients for weather.

Interestingly, we can observe the frequency of different weather events throughout the year. The plots for squall and smoke both indicate their low number of observations, while we see clouds are frequently represented throughout the data over time. Snow and thunder both display a seasonal pattern, being closely clustered around zero during some intervals of the data but having an increase in occurrences resulting in a wider spread during other intervals.

Conclusions

Our completed model is:

```
Residuals:
    Min      1Q  Median      3Q     Max
-71.727 -10.008   0.464   8.067  49.408

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              41.119769   0.166974 246.264  < 2e-16 ***
holidayYes               -6.770944   0.399068 -16.967  < 2e-16 ***
temp                      0.087271   0.005642  15.468  < 2e-16 ***
weather_mainClouds        1.698043   0.174708   9.719  < 2e-16 ***
weather_mainDrizzle       0.808524   0.365461   2.212 0.026948 *
weather_mainFog          -0.905419   0.502012  -1.804 0.071303 .
weather_mainHaze          0.297492   0.418583   0.711 0.477265
weather_mainMist         -0.285649   0.228966  -1.248 0.212198
weather_mainRain          0.242086   0.233635   1.036 0.300128
weather_mainSmoke        -3.211571   3.272414  -0.981 0.326397
weather_mainSnow         -0.570321   0.310730  -1.835 0.066449 .
weather_mainSquall       -8.243312   7.312177  -1.127 0.259604
weather_mainThunderstorm -1.579741   0.476092  -3.318 0.000907 ***
timeIntervalEarly        -2.218613   0.182007 -12.190  < 2e-16 ***
timeIntervalMidday       24.464043   0.182364 134.149  < 2e-16 ***
timeIntervalAfternoon    28.311810   0.194334 145.686  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.62 on 48178 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.4782,    Adjusted R-squared:  0.478
F-statistic:  2943 on 15 and 48178 DF,  p-value: < 2.2e-16
```

We see from the completed model that the time of day, temperature, clouds, thunderstorms, and whether the day is a holiday are highly significant in explaining variability in the traffic volume. Moreover, it’s clear that traffic volume will usually only be affected by uncommon events rather than something that may happen every day of a given week, such as cloud coverage. Naturally there exists a practical limit to our data as we are only capable of observing the impact of time and natural causes, so a day to day event such as construction or car accidents are out of consideration, thus at times there may be unlikely quantities which would appear as outliers but we have no means of identifying their cause.

Reflection

Throughout the project process, we consistently reached roadblocks in data transformation and our final model still does not perfectly satisfy several normality and linearity assumptions. We attempted several types of transformation and ultimately chose the square root on the response as it produced the most positive change in the residuals.

Additionally, our team could have benefited from meeting more often and having more focused work time, potentially meeting in-person more frequently instead of virtually. At times, we had difficulty with organization as several members had R code that they were working on at the same time, leading to inconsistencies that unnecessarily slowed down the process of creating the presentation and report.

However, by working on the report we were able to resolve many of these errors and achieve a cleaner analysis overall, although we did have to make changes from the presentation (such as adjusting what hours are in which time intervals). In conclusion, we reached a satisfactory model that we believe accurately reflects the data, which unfortunately was not ideal for linear modeling from the beginning. All group members learned significantly from the process of working with more difficult data as opposed to the pre-cleaned and prepared data used in homework and examples.

For future analysis of traffic volume, we would want to include information on car accidents, nearby construction, and other information that may dramatically change the traffic on a given road. For instance, construction in one area may not change the total traffic volume in the region, but could influence if drivers choose a specific road over another. Additional areas for expansion would be to create a comparable model for other cities and compare the most significant regressors to see if traffic patterns may be more variable in other areas - for instance thunder in California, where it is infrequent, could be more impactful than in a state where lightning is a common occurrence.

Sources

Minnesota Department of Transportation. “Monthly Report, Station No. 301.” January 2018. Accessed March 2022.

Sohaee, Nassim, et al. “Short-term traffic volume prediction using neural networks.” https://illidanlab.github.io/big_traffic/2018/papers/sohaee2018shortterm.pdf. Accessed 18 April 2022.

United States Census Bureau. “Census Bureau Estimates Show Average One-Way Travel Time to Work Rises to All-Time High.” *Census Bureau*, 18 March 2021, https://www.census.gov/newsroom/press-releases/2021/one-way-travel-time-to-work-rises.html. Accessed 18 April 2022.

Appendix

Member Roles

Alex May: Presentation and guidance of interpretation
Jay Glessner: Report and development of analysis approach
Muhammad Zubair: Majority of early coding and data cleaning
Nick Azar: Attended most meetings and gave a portion of the presentation

A - Data Cleaning

```
# Libraries
library(tidyr)
library(readr)
library(dplyr)
library(stringr)
library(ggplot2)
library(ggcorrplot)
library(ggdark)
library(scales)
library(car)
library(MASS)
library(DAAG)
library(lubridate)

# Data Cleaning
# Read in the data frame
df <- read_csv("Metro_Interstate_Traffic_Volume.csv.gz")

## Data Cleaning - Date & Time
# Separating the date_time into date and time column
df$date <- sapply(strsplit(as.character(df$date_time)," "), "[", 1)
df$date <- as.Date(df$date, format="%Y-%m-%d")
df$time <- sapply(strsplit(as.character(df$date_time)," "), "[", 2)

df$Date <- as.Date(df$date_time)
df$Time <- format(as.POSIXct(df$date_time), format = "%H:%M:%S")

df$TimeSec <- as.numeric(as.POSIXct(df$Time, format = "%H:%M:%S")) %% 86400

summary(df$temp)
df$temp[which(df$temp < 5)] <- NA
summary(df$temp)

hist(df$traffic_volume, main = "Histogram of Traffic Volume", xlab = "Traffic Volume")

## Data Cleaning - Holiday
# Mark holiday as a logical value
df$holiday <- ifelse(df$holiday == "None", "None", "Yes")
```

```
# Table of holidays before cleaning
table(df$holiday)

# Pull the dates of each holiday
df_holiday <- subset(df, holiday=='Yes')
df_holiday <- subset(df, select = c(holiday, date))

# Marking entire day as holiday, instead of only first hour
# Holidays in 2013
df$holiday[format(df$date, '%Y-%m-%d') %in% '2012-10-08'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2012-11-12'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2012-11-22'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2012-12-25'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2013-01-01'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2013-02-18'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2013-05-27'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2013-07-04'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2013-09-02'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2013-10-14'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2013-11-11'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2013-11-28'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2013-12-25'] <- "Yes"
# Holidays in 2014
df$holiday[format(df$date, '%Y-%m-%d') %in% '2014-01-01'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2014-01-20'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2014-02-17'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2014-05-26'] <- "Yes"
# Holidays in 2015
df$holiday[format(df$date, '%Y-%m-%d') %in% '2015-07-03'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2015-08-27'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2015-09-07'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2015-10-12'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2015-11-11'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2015-11-26'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2015-12-25'] <- "Yes"
# Holidays in 2016
df$holiday[format(df$date, '%Y-%m-%d') %in% '2016-01-01'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2016-02-15'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2016-05-30'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2016-07-04'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2016-08-25'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2016-09-05'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2016-10-10'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2016-11-11'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2016-11-24'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2016-12-26'] <- "Yes"
# Holidays in 2017
df$holiday[format(df$date, '%Y-%m-%d') %in% '2017-01-02'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2017-01-16'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2017-02-20'] <- "Yes"
```

```
df$holiday[format(df$date, '%Y-%m-%d') %in% '2017-05-29'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2017-07-04'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2017-08-24'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2017-09-04'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2017-10-09'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2017-11-10'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2017-11-23'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2017-12-25'] <- "Yes
# Holidays in 2018
df$holiday[format(df$date, '%Y-%m-%d') %in% '2018-01-01'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2018-01-15'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2018-02-19'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2018-05-28'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2018-07-04'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2018-08-23'] <- "Yes"
df$holiday[format(df$date, '%Y-%m-%d') %in% '2018-09-03'] <- "Yes"

# Checking to see if values were added
table(df$holiday)

# Making a copy of df for future data cleaning purposes
df_holiday_cleaned <- data.frame(df)

## Data Cleaning - Weather frequency
# Identify which types of weather occur the most
df_max_occur <- df_holiday_cleaned %>%
       gather("key", "value", weather_description) %>%
       group_by(date) %>%
       dplyr::summarise(max.occurence = names(which.max(table(value))))

head(df_max_occur)

# Merging the df of max occurrences with the original df
df_max_occur$date <- as.Date(df_max_occur$date, format="%Y-%m-%d")
df_clean <- full_join(df_holiday_cleaned, df_max_occur,
                    by = c("date" = "date"))

## Data Cleaning - Traffic volume intervals
# Create intervals of traffic volume for easier comparison
df_clean$traffic_interval_of_250 <- floor(df_clean$traffic_volume / 250)+1
df_clean$traffic_interval_of_250 <- as.factor(df_clean$traffic_interval_of_250)
```

Data Cleaning - Cleaned data frame

```
df_clean$max.occurence <- as.factor(df_clean$max.occurence)
df_clean$holiday <- as.factor(df_clean$holiday)
df_clean$temp <- df_clean$temp - 273.15

summary(df_clean)
```

B - Data Visualization

```
# Data Visualization
# Create a data frame for data visualizations
df_visualization <- subset(df_clean,
                           select = -c(max.occurence, weather_description, date_time))
df_visualization$holiday <- ifelse(df_visualization$holiday == "Yes", 1, 0)
df_visualization$weather_main <- as.factor(df_visualization$weather_main)

str(df_visualization)

## Visualization - Time series
df_visualization$time <- as.POSIXct(df_visualization$time, format="%H:%M:%S")
df_visualization$traffic_interval_of_250 <- as.integer(df_visualization$traffic_interval_of_250)

## Visualization - Correlation matrix
df_cor_mat <- subset(df_visualization, !is.na(temp),
             select = -c(date, traffic_volume))

df_cor_mat$rain_1h <- ifelse(df_cor_mat$rain_1h == 0, 0, 1)
df_cor_mat$snow_1h <- ifelse(df_cor_mat$snow_1h == 0, 0, 1)
df_cor_mat$time <- sapply(strsplit(as.character(df_cor_mat$time)," "), "[", 2)
df_cor_mat$time <- sapply(strsplit(as.character(df_cor_mat$time),":"), "[", 1)
df_cor_mat$time <- as.numeric(df_cor_mat$time)

df_cor_mat <- df_cor_mat[sapply(df_cor_mat,is.numeric)]
corr <- cor(df_cor_mat)
ggcorrplot(corr, lab = TRUE, hc.order = TRUE, type = "lower",
       outline.color = "white",
       colors = c("white", "lightblue", "blue")) +
theme(axis.text.x = element_text(angle = 90, hjust=1))

## Visualization - Histogram of Response
ggplot(df_visualization, aes(x=time, y=traffic_interval_of_250))+
  geom_bar(size=1.5, stat = "identity", fill="LIGHTBLUE") +
  labs(x=NULL, y="Traffic volume", title = "Traffic Volume by Hour") +
  scale_x_datetime(breaks = date_breaks("2 hour"),
                   labels=date_format("%H:%M:%S"))+
  theme(axis.text.x = element_text(angle = 45, hjust=1)) +
  theme(plot.title = element_text(hjust = 0.5))

## Visualization - Temperature
ggplot(subset(df_clean, temp > -50), aes(x = temp, y = traffic_volume))+
  geom_point(color = "lightblue", cex = 0.5)+
  geom_smooth(method = lm, formula = y~ x, se = F, color = "blue")+
  labs(x = "Temperature in Celsius", y = "Traffic Volume", title = "Temperature and Traffic
Volume")


ggplot(subset(df_clean, temp > -50), aes(y = temp, x = time, color = traffic_volume))+
  geom_jitter()+
  labs(x = "Time", y = "Temperature", title = "Time, Temperature, and Traffic Volume")+
```

```
  theme(axis.text.x = element_text(angle = 90))

## Visualization - Hourly Rain
ggplot(subset(df_clean, rain_1h < 10), aes(x = rain_1h, y = traffic_volume))+
  geom_point(color = "lightblue", cex = 1)+
  geom_smooth(method = lm, formula = y ~ x, se = F, color = "blue")+
  labs(x = "Rain per hour in mm", y = "Traffic Volume", title = "Hourly Rain and Traffic Volume")

## Visualization - Hourly Snow
ggplot(df_clean, aes(x = snow_1h, y = traffic_volume))+
  geom_point(color = "lightblue", cex = 1)+
  geom_smooth(method = lm, formula = y ~ x, se = F, color = "blue")+
  labs(x = "Snow per hour in mm", y = "Traffic Volume", title = "Hourly Snow and Traffic Volume")

## Visualization - Cloudcover
ggplot(df_clean, aes(x = clouds_all, y = traffic_volume))+
  geom_jitter(cex = 0.5, color = "lightblue")+
  geom_smooth(method = lm, formula = y ~ x, se = F, color = "blue")+
  labs(x = "Total cloudcover percentage", y = "Traffic Volume", title = "Cloudcover and Traffic
Volume")

## Visualization - Holiday
ggplot(df_clean, aes(x = holiday, y = traffic_volume, fill = holiday))+
  geom_boxplot(color = "black")+
  labs(x = "Holiday or Not", y = "Traffic Volume", title = "Traffic Volume on Holiday vs Regular
Days")+
  geom_hline(yintercept = 3380, linetype = "dashed", color = "black")

## Visualization - Weather Main
ggplot(df_clean, aes(x = weather_main, y = traffic_volume, fill = weather_main))+
  geom_boxplot(color = "black")+
  labs(x = "Primary Weather Description", y = "Traffic Volume",
       title = "Overall Weather and Traffic Volume")+
  geom_hline(yintercept = 3380, linetype = "dashed", color = "black")
```

C - Primary Model

```
# Primary Model

## Primary - Time intervals

df_time_factor <- df_visualization

df_time_factor$groups <- cut(df_time_factor$TimeSec,
                             breaks = c(-1, 21600, 43200, 64800, Inf),
                             labels = c("Night", "Early", "Midday", "Afternoon"))

table(df_time_factor$groups)

df_model <- subset(df_clean,
                   select = -c(date_time, weather_description, date,
Date,TimeSec, max.occurence, traffic_interval_of_250))

df_model$timeInterval <- factor(df_time_factor$groups)

summary(df_model)

## Primary - Linear model with all regressors
fit1 <- lm(data = df_model, traffic_volume~. -time - Time)
summary(fit1)
```

D - Reduced Model

```
# Reduced model
df_reduced1 <- subset(df_model, select = -c( rain_1h, snow_1h, clouds_all))
fit2 <- lm(data = df_reduced1,  traffic_volume ~. -time - Time)
fit2b <- lm(data = df_reduced1, traffic_volume ~. -timeInterval)
summary(fit2)

## Reduced - Time Intervals Box
ggplot(data = df_reduced1, aes(x = timeInterval, y = traffic_volume))+
  geom_jitter(cex = 0.2, color = "lightgray")+
  geom_boxplot(fill = NA)+
  geom_hline(yintercept = 3380, color = "deepskyblue3")+
  labs(x = "Time of day", y = "Traffic Volume", title = "Boxplot of Traffic Volume by Time")
median(df_reduced1$traffic_volume[which(df_reduced1$timeInterval == "Early"
      & !is.na(df_reduced1$timeInterval))])
median(df_reduced1$traffic_volume[which(df_reduced1$timeInterval == "Midday")])
median(df_reduced1$traffic_volume[which(df_reduced1$timeInterval == "Afternoon")])
median(df_reduced1$traffic_volume[which(df_reduced1$timeInterval == "Night")])

## Reduced - ANOVA table comparison
anova(fit2, fit1)

## Reduced - VIF table
vif(fit2)
```

E - Analysis Pre-Transformation

```
#Standardized residuals
length(which(stdres(fit2) > 3))
length(which(stdres(fit2) < -3))
barplot(height = stdres(fit2),
       main = "Standardized Residuals", xlab = "index", ylab = "Standardized Residual",
        ylim = c(-4, 4))
abline(h = 3, col = "lightblue")
abline(h = -3, col = "lightblue")

#Studentized residuals
length(which(studres(fit2) > 3))
length(which(studres(fit2) < -3))
barplot(height = studres(fit2),
       main = "Studentized Residuals", xlab = "index",
       ylab = "Studentized Residual", ylim = c(-4, 4))
abline(h = 3, col = "lightblue")
abline(h = -3, col = "lightblue")

#R student residuals
cor.qt <- qt(0.05/(2 * 48204), 48176, lower.tail = F)

length(which(rstudent(fit2) > cor.qt))
length(which(rstudent(fit2) < -cor.qt))
barplot(height = rstudent(fit2),
       main = "R Student Residuals", xlab = "index", ylab = "R Student Residual", ylim = c(-6, 6))
abline(h=cor.qt, col = "lightblue", lwd=2)
abline(h=-cor.qt, col = "lightblue", lwd=2)

## Residuals - Residuals vs fitted, histogram, QQ
residualPlot(fit2b, type = "rstudent", quadratic=F, col = "lightblue", main = "Residuals vs Fitted

hist(studres(fit2), breaks = 25, freq=F,
       col = "lightblue",
       main = "Histogram of Residuals Pre-Transformation",
       xlab = "Studentized Residuals")

qqPlot(fit2, col = "grey", col.lines = "cornflowerblue",
       ylab = "Studentized Residuals",
       main = "Quantile Comparison Pre-Transformation")

## Residuals - Summary
df_ResPre <- df_reduced1[which(stdres(fit2) > 3 | stdres(fit2) < -3 | studres(fit2) > 3 |
studres(fit2) < -3 | rstudent(fit2) > cor.qt | rstudent(fit2) < -cor.qt), ]

## Residuals - box cox
df_positive <- df_reduced1
df_positive$traffic_volume <- (df_positive$traffic_volume + 0.01)
fit3 <- lm(data = df_positive, traffic_volume~. )
boxcox.traffic <- boxCox(fit3)
boxcox.traffic$x[which.max(boxcox.traffic$y)]
```

F - Transformation

```
# Square root transformation on the response
df_transformed <- data.frame(traffic_volume = (df_reduced1[,4])^(0.5),
                             holiday = df_reduced1[,1],
                             temp = df_reduced1[,2],
                             weather_main = df_reduced1[,3],
                             time = df_reduced1[,5],
                             Time = df_reduced1[,6],
                             timeInterval = df_reduced1[,7])
summary(df_transformed)

fit4 <- lm(data = df_transformed, traffic_volume~. -time -Time)
fit4b <- lm(data = df_transformed, traffic_volume~. -timeInterval)

summary(fit4)
```

G - Analysis Post-Transformation

```
## New Residuals - Histogram comparison
hist(studres(fit2), breaks = 25, freq=F,
       col = "lightblue",
       main = "Before",
       xlab = "Studentized Residuals",
       xlim = c(-4, 4))
hist(studres(fit4), breaks = 25, freq=F,
       col = "lightblue",
       main = "After",
       xlab = "Studentized Residuals",
       xlim = c(-4, 4))

## New Residuals - QQ plot comparison
qqPlot(fit2, col = "grey", col.lines = "cornflowerblue",
       ylab = "Studentized Residuals",
       main = "Before",
       ylim = c(-4, 4))
qqPlot(fit4, col = "grey", col.lines = "cornflowerblue",
       ylab = "Studentized Residuals",
       main = "After",
       ylim = c(-4, 4))

## New Residuals - Standardized
barplot(height = stdres(fit4),
       main = "Standardized Residuals", xlab = "index",
       ylab = "Standardized Residual", ylim = c(-4, 4))
abline(h = 3, col = "lightblue")
abline(h = -3, col = "lightblue")

#Studentized residuals
barplot(height = studres(fit4),
       main = "Studentized Residuals", xlab = "index",
       ylab = "Studentized Residual", ylim = c(-4, 4))
```

```
abline(h = 3, col = "lightblue")
abline(h = -3, col = "lightblue")

#R student residuals
cor.qt <- qt(0.05/(2 * 48204), 48176, lower.tail = F)

barplot(height = rstudent(fit4),
       main = "R Student Residuals", xlab = "index",
       ylab = "R Student Residual", ylim = c(-6, 6))
abline(h=cor.qt, col = "lightblue", lwd=2)
abline(h=-cor.qt, col = "lightblue", lwd=2)

# Standardized
length(which(stdres(fit4) > 3))
length(which(stdres(fit4) < -3))

length(which(stdres(fit4) > 3.2))
length(which(stdres(fit4) < -3.2))

# Studentized
length(which(studres(fit4) > 3))
length(which(studres(fit4) < -3))

length(which(studres(fit4) > 3.2))
length(which(studres(fit4) < -3.2))

# R student
length(which(rstudent(fit4) > cor.qt))
length(which(rstudent(fit4) < -cor.qt))

# Total residuals
df_ResPost <- df_transformed[which(stdres(fit4) > 3 | stdres(fit4) < -3 | studres(fit4) > 3 |
studres(fit4) < -3 | rstudent(fit4) > cor.qt | rstudent(fit4) < -cor.qt), ]
```

H - Influence Analysis

```
## Influence - Measures
inf <- influence.measures(fit4)

# summary(inf)
influenceIndexPlot(fit4,vars=c("Cook"),
                   main = "Cook's Distance")
influenceIndexPlot(fit4,vars=c("Hat"),
                   main = "Hat Values")
## Influence - DFBETAS
dfbetasPlots(fit4, intercept = F,
             terms = "holiday",
             main = "Holiday",
             col = "deepskyblue3")

dfbetasPlots(fit4, intercept = F,
             terms = "temp",
             col = "deepskyblue3")

dfbetasPlots(fit4, intercept = F,
             terms = "timeInterval",
             col = "deepskyblue3")

dfbetasPlots(fit4, intercept = F,
             terms = "weather_main",
             col = "deepskyblue3",
             layout = c(2, 3),
             ask = F,
             xlab = NA)
```