# *Optimizing Shared Micro Mobility Vehicle Practices: An Austin, Texas Case Study*


## *Abstract*

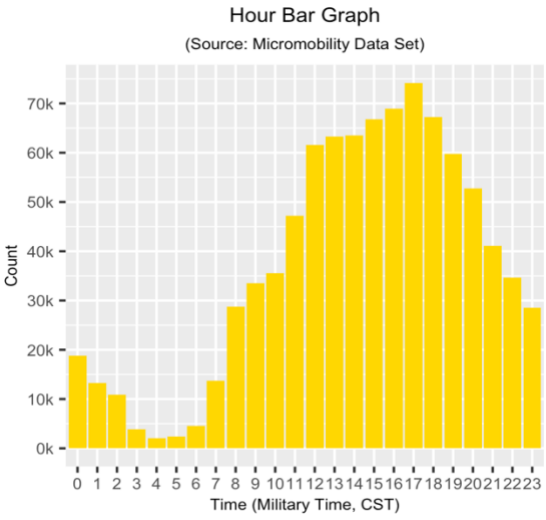
This report runs a multiple linear regression model using shared micro mobility data published by the City of Austin. This model will determine whether three categorical explanatory variables ("Hour," "Day of Week," and "Vehicle Type") and one numerical explanatory variable ("Trip Distance") are statistically significant to a numerical response variable ("Trip Duration"). Furthermore, we can derive an estimated regression line ($\hat{y} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon$). Knowing these aspects will enable companies to optimize vehicle distribution based on specific variables for the highest revenue–a result of maximized trip durations. In order to complete this experiment, we will look at partial F-tests and a T-test. After conducting partial F-tests with an $\alpha$ equal to 0.01, we conclude that all four variables are statistically significant in determining trip duration. From here, we get a final estimated regression line of $\hat{y} = 2.595$ + [(1.324 (if scooter)) + (-0.086 (if hour 1)) + (-0.13 (if hour 2)) + … + (-0.29 (if hour 23)) + (-1.4 (if day of week 1)) + … + (0.073 (if day of week 6))]×0.005minutes/meters $x$–see the column "estimate" in Fig. 5 for a complete list of adjustments.

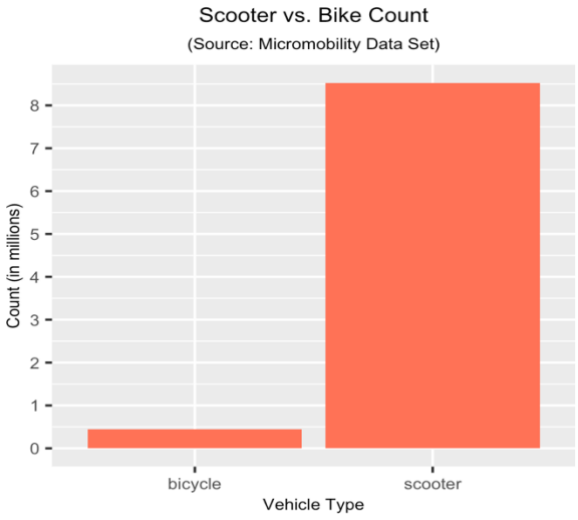
## *Background and Significance*

With “60 percent of all trips in the United States […] five miles or less,” shared micro mobility has seen exponential growth within America (Tonar). Furthermore, experts predict a market size of “200 to 300 billion […] by 2030” (Tonar). Companies–such as Lime, Uber, Bird, and more–provide for-charge scooters and bikes, so people do not have to use public transport or own a car. This introduction leads us to the question at hand: can we use four explanatory shared micro mobility variables to statistically estimate a response variable, or trip duration, in Austin, Texas? Trip duration is a critical metric as shared micro mobility companies charge per minute.

## *Methods*

The City of Austin provides shared micro mobility trip data, from 2018 to 2020, in an open data portal due to the “Shared Small Vehicle Mobility Systems operating rules” of Austin (Austin Transportation). Upon further examination of the rules document, this data is reported in a high-quality format. Moreover, it states that shared micro mobility companies “shall provide […] real-time and historical information for their entire fleet” through APIs (“Director Rules For Deployment And Operation Of Shared Small Mobility Systems”). An API is a an “application programming interface” that “let[s] your product or service communicate with other products and services” (“What is an API?”). In this instance, the API retrieves trip data without human contact. There are, however, still errors that occur and decrease the quality of the data set. For example, there can be human errors on the user front, due to people failing to quit the app, and technology errors, due to faulty gadgets. The extreme outliers in trip distance (original max trip distance: 2,147,483,647 meters) possibly show this error. To mitigate these errors, I calculated basic limitations to exclude extreme outliers, such as entries well beyond max speed, showing negative time, and more. These limitations left us with 8,966,884 entries for our multiple linear regression modeling–originally 9,031,956 entries. We will use six variables from the data set for our multiple linear regression: “Trip Duration,” “Trip Distance,” “Day of Week,” “Hour,” “Vehicle Type,” and “Census Tract Start.” The variable “Census Tract Start” will allow us to use the FIPS code of each trip to create a choropleth map of average trip duration. Knowing this information helps companies further optimize practices by signaling areas that achieve the highest average trip durations.



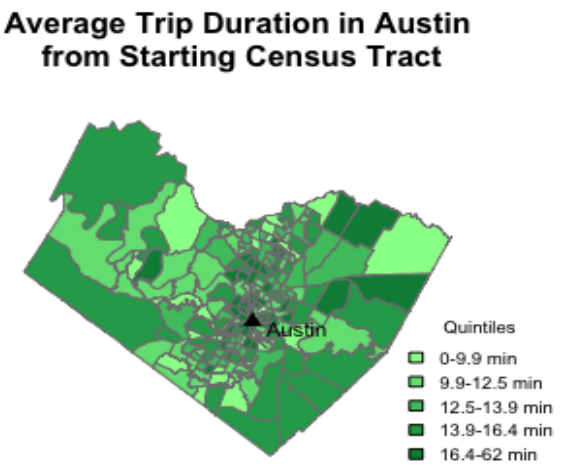
Before discussing the results, it is important to explore the data in question. One notable variable to discuss is the explanatory categorical variable "Vehicle Type," as seen in the bar graph labeled "Scooter vs. Bike Count." This graph shows the portion of reported trips on a bicycle versus scooter. It is clear, based on this visual, that the shared micro mobility type "scooter" comprises a majority of the data as 95.06% are scooter observations and 4.94% are bicycle observations. The other variable to spotlight, in particular, is the explanatory categorical variable "Hour," or hour of the day when the trip occurred, in the left graph labeled "Hour Bar Graph." When first looking at this bar chart, in a standard day format, it appears to have a bi-modal shape with more trips near bars 0 and 17. If, however, this time period is viewed starting at bar "4," there is approximately a symmetrical shape. One could argue this stance is a more correct way to view this data since shared micro mobility vehicles are charged in the early morning hours and people may be returning from a night out between bars 0-2. See Fig. 2, 3, and 4 in the "Appendix" for an exploratory analysis of the three other variables.

## *Multiple Regression Modeling Results*

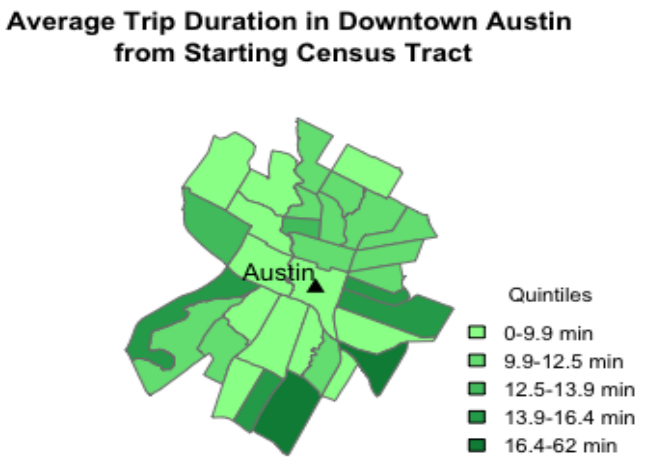
*Table of Partial F-Tests*

| | df | F-Test Statistic | Pr(>F) |
|---|---|---|---|
| Model without Day.of.Week | 6 | 12211 | < 2.2e-16 |
| Model without Hour | 23 | 2580.8 | < 2.2e-16 |
| Model without Trip Distance | 1 | 8880880 | < 2.2e-16 |
| Model without Vehicle.Type | 28 | 5512.7 | < 2.2e-16 |

Now, it is time to dive into the modeling and results. To determine the statistical significance of each variable in the model, we will look at the table labeled "Table of Partial F-Tests." Each row in this table highlights an ANOVA table between the full multiple regression model and a reduced model. If we set $\alpha$ equal to 0.01, we can see that in every scenario the missing variable was deemed statistically significant as the ANOVA tables showed the full models with a p-value of < 2.2e-16. Since our p-value is less than our $\alpha$, we reject the null hypothesis (($h_0$) $b_1$,$b_2$, $b_3$, $b_4$ = 0) and affirm that each variable is statistically significant. Looking at a T-test (Fig. 5), however, there are a few of slopes with a p-value greater than 0.01. These slopes include "Hour12", "Hour17", and "Hour18". Now that we determine "Hour," "Day of Week," "Vehicle Type," and "Trip Distance" are statistically significant for "Trip Duration," we derive the estimated regression line ($\hat{y} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon$). The estimated y-intercept of the regression line is 2.595 minutes. This number suggests that if a trip is 0 meters long, there will be a duration of 2.595 minutes. The estimated slope of the regression line is 0.005. This slope means that for every meter increase in the trip distance, there is is a 0.005 minute increase in the trip duration if all other variables are held constant. The slope, moreover, only has a practical meaning between 0 meters and 62,016 meters due to the range of this data. Beyond the y-intercept and slope, there are adjustments for each categorical variable. This conclusion leads us to a final estimated regression line of $\hat{y}$ = 2.595+ [(1.324 (if scooter)) + (-0.086 (if hour 1)) + (-0.13 (if hour 2)) + … + (-0.29 (if hour 23)) + (-1.4 (if day of week 1)) + … + (0.073 (if day of week 6))]×0.005minutes/meters $x$–see the column "estimate" in Fig. 5 for a complete list of adjustments.




(Source: Austin Micromobility Data Set)




(Source: Austin Micromobility Data Set)

In addition to the modeling, we can see more results by creating a choropleth map (labeled “Average Trip Duration in Austin from Starting Census Tract”) using the variable “Census Tract Start.” In an attempt to make this map more valuable, I calculated quintiles of average trip duration across the FIPS codes of Travis County. Upon first view, the average trip duration starting from downtown Austin appeared to be higher than the immediate surroundings. Until more complex maps in the future, we can manually zoom in on downtown Austin (as seen in the visual labeled “Average Trip Duration in downtown Austin from Starting Census Tract”) and see that assertion is incorrect.

## *Conclusion and Discussion*

Circling back to the proposed research question, shared micro mobility companies can look at Fig. 5 in the “Appendix” after we deemed these five variables statistically significant to generally optimize practices. Moreover, companies can see the greatest estimated regression line with its characteristics: deploying a scooter at hour 14:00 on a Saturday. Looking to the future, however, companies should filter data accurately to improve precision. Even as there are many entries, a handful of them brought down the precision of our analysis and estimated regression line. Moreover, this regression model failed five out of five assumption checks, as seen in Fig. 7 through 9 in the “Appendix.” Taking these steps will provide more valuable results in determining best practices. Another aspect for future study is to incorporate other variables such as month of the year to see the effects of festivals and seasons.

On the subject of the variable “Census Tract Start,” it will be critical to develop an interactive map in the future. An interactive map will allow individuals to seamlessly view the breakdown of Austin. On this topic, it is also important to investigate where Austin allows shared micro mobility vehicles to operate. These adjustments will improve the quality of the choropleth maps. Finally, it is valuable to create choropleth maps showing the two vehicle types separately in the future, when space permits.

Though the modeling in this case study provides a solid foundation, it should undergo further revisions in the future to increase and improve the precision of the results if companies are to use the modeling in optimizing practices to maximize revenue.

## *Works Cited*


Austin Transportation. “Shared Micromobility Vehicle Trips.” data.austintexas.gov, 03 Dec. 2018, https://data.austintexas.gov/Transportation-and-Mobility/Shared-Micromobility-Vehicle-Trips/7d8e-dm7r. Accessed 22 Apr. 2020.

“Director Rules For Deployment And Operation Of Shared Small Mobility Systems.” Austin Transportation Department, https://austintexas.gov/sites/default/files/files/Transportation/Dockless_Adopted_Tracked_Changes.pdf. Accessed 09 May 2020.

Tonar, Remington, and Ellis Talton. “Cities Need to Rethink Micromobility To Ensure It Works For All.” Forbes, 07 Jan. 2020, https://www.forbes.com/sites/ellistalton/2020/01/07/cities-need-to-rethink-micromobility-to-ensure-it-works-for-all/#49b44b312ebf. Accessed 23 Apr. 2020.

“What is an API?” Red Hat, https://www.redhat.com/en/topics/api/what-are-application-programming-interfaces. Accessed 09 May 2020.

## *Appendix*

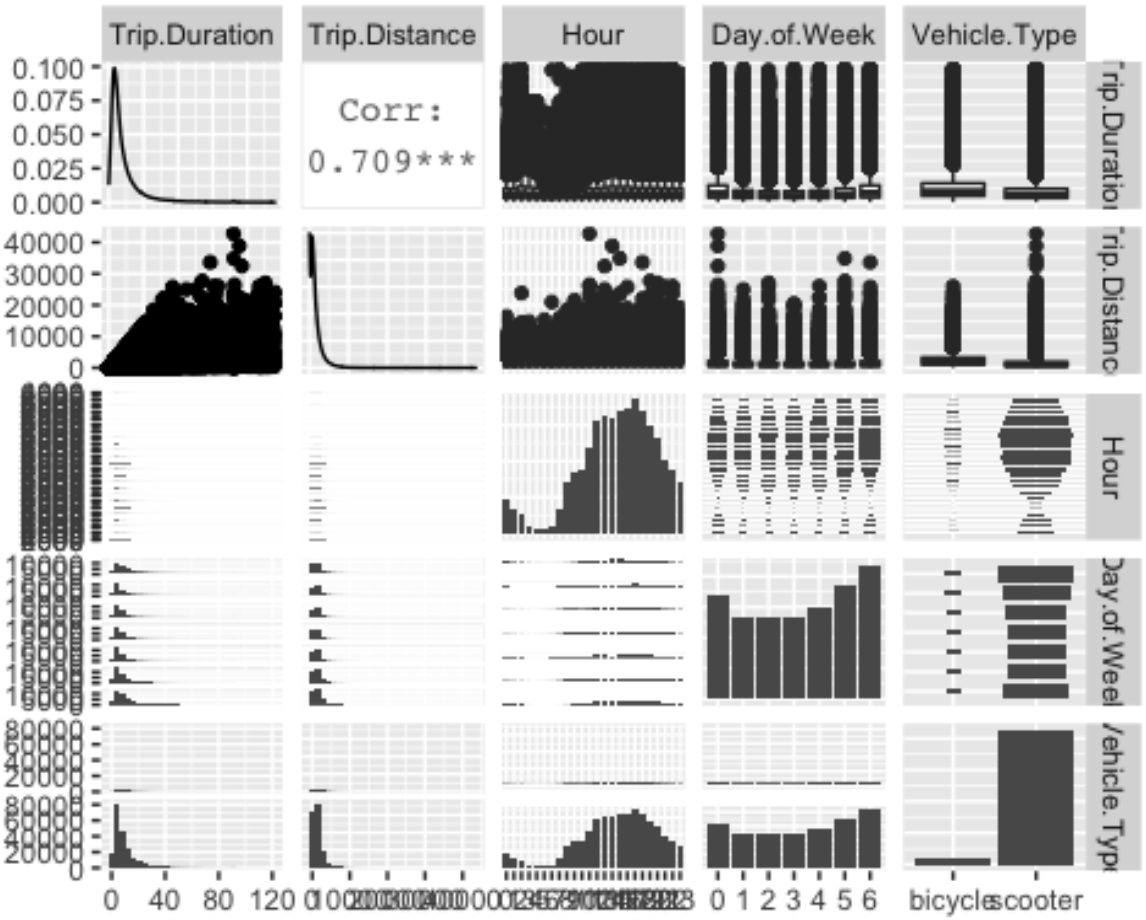



*Fig. 1: Associations Between All Variables*

In this matrix, we can see the associations between all variables using a random portion of data. Though the data was selected randomly, measures were taken to confirm not all entries were scooters in the categorical variable “Vehicle Type.” Specifically, for this analysis, we are looking at the first row and column. Though hard to read, this associations matrix does highlight trends and show visuals explored in other sections. One prominent example is the center cell [3,3] which shows approximately the graph labeled “Hour Bar Graph” in the “Method” section.

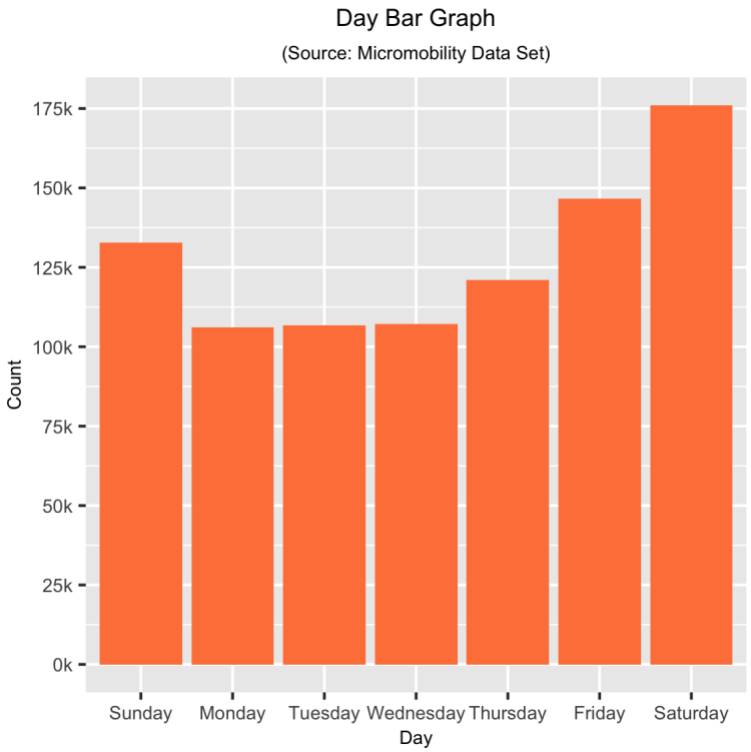



Fig. 2: Day of the Week Bar Graph

Here, we showcase the categorical variable “Day of Week” (in which the trip occurred). This bar graph shows the count for each day of the week. It does not show large differences in counts depending on the day. The bar graph first appears to be bi modal, with peak counts in bar 0 (Sunday) and bar 6 (Saturday). Viewed in a non-standard week format, this bar graph again would be approximately symmetrical. This makes sense as people tend to go out and be more active on weekends.

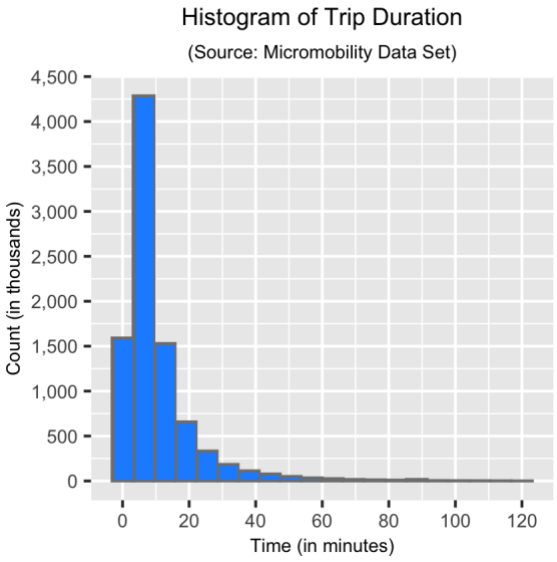



*Fig. 3: Trip Duration Histogram*

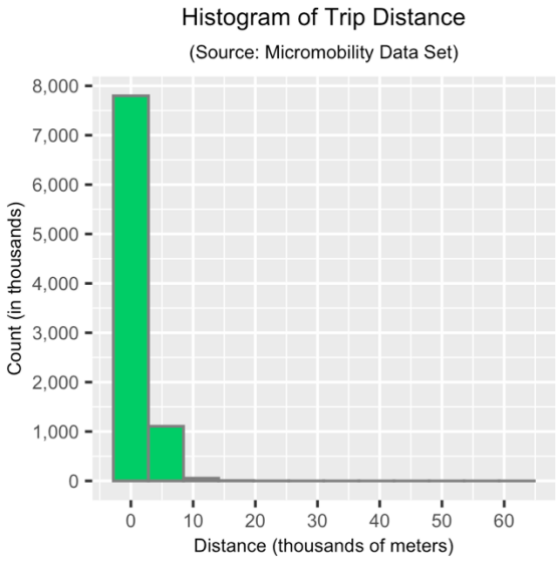



*Fig. 4: Trip Distance Histogram*

The histogram on the left displays the distribution of shared micro mobility "Trip Duration" in minutes, or our response variable. The values in this visual show a heavy skew right, as a majority of the trips (approximately $7{,}500{,}000$) are near the bins with values from 0 to 20 minutes. This makes sense logically as most people will not be using shared micro mobility vehicles for long periods of time.

The histogram on the right shows the distribution of shared micro mobility "Trip Distance" in meters. This histogram also has a heavy skew right as a majority of trip values are within the first two bin values–0 thousand meters and 5 thousand meters. There exist a few outliers, however, between 15,000 meters and 60,000 meters.

Table: Fig. 5: T-Test of the Multiple Linear Regression Model

| term | estimate | std.error | test stat. | p-value |
|---|---|---|---|---|
| (Intercept) | 2.5948327 | 0.0237 | 109.3580 | 0.000e+00 |
| Vehicle.Typescooter | 1.3242634 | 0.0129 | 102.5315 | 0.000e+00 |
| Hour1 | -0.0863274 | 0.0297 | -2.9030 | 3.696e-03 |
| Hour2 | -0.1337637 | 0.0315 | -4.2404 | 2.232e-05 |
| Hour3 | -0.3659918 | 0.0464 | -7.8857 | 3.128e-15 |
| Hour4 | -1.4911564 | 0.0606 | -24.6166 | 8.482e-134 |
| Hour5 | -2.3107133 | 0.0566 | -40.8254 | 0.000e+00 |
| Hour6 | -2.4847393 | 0.0435 | -57.0916 | 0.000e+00 |
| Hour7 | -2.3021072 | 0.0295 | -77.9468 | 0.000e+00 |
| Hour8 | -2.1995392 | 0.0247 | -89.1064 | 0.000e+00 |
| Hour9 | -1.5587256 | 0.0240 | -65.0475 | 0.000e+00 |
| Hour10 | -0.7641493 | 0.0237 | -32.3037 | 6.388e-229 |
| Hour11 | -0.1737307 | 0.0226 | -7.6828 | 1.557e-14 |
| Hour12 | 0.0224841 | 0.0219 | 1.0288 | 3.036e-01 |
| Hour13 | 0.3675315 | 0.0218 | 16.8770 | 6.655e-64 |
| Hour14 | 0.5720483 | 0.0218 | 26.2836 | 2.996e-152 |
| Hour15 | 0.4865170 | 0.0216 | 22.4777 | 6.912e-112 |
| Hour16 | 0.3372966 | 0.0216 | 15.6328 | 4.362e-55 |
| Hour17 | -0.0200855 | 0.0214 | -0.9374 | 3.485e-01 |
| Hour18 | -0.0076962 | 0.0216 | -0.3556 | 7.222e-01 |
| Hour19 | 0.1468271 | 0.0219 | 6.6937 | 2.177e-11 |
| Hour20 | 0.2600759 | 0.0223 | 11.6709 | 1.796e-31 |
| Hour21 | 0.0656250 | 0.0231 | 2.8418 | 4.485e-03 |
| Hour22 | -0.1576159 | 0.0238 | -6.6314 | 3.324e-11 |
| Hour23 | -0.2921768 | 0.0246 | -11.8571 | 1.979e-32 |
| Day.of.Week1 | -1.4248600 | 0.0109 | -131.1940 | 0.000e+00 |
| Day.of.Week2 | -1.7596987 | 0.0109 | -161.9328 | 0.000e+00 |
| Day.of.Week3 | -1.8069260 | 0.0109 | -166.5356 | 0.000e+00 |
| Day.of.Week4 | -1.5603426 | 0.0105 | -148.6184 | 0.000e+00 |
| Day.of.Week5 | -1.0114887 | 0.0100 | -101.2875 | 0.000e+00 |
| Day.of.Week6 | 0.0730971 | 0.0095 | 7.6579 | 1.890e-14 |
| Trip.Distance | 0.0052256 | 0.0000 | 2980.0805 | 0.000e+00 |

This figure, which shows our T-test, displays the values for each of the variables that comprise the estimated regression line as discussed earlier.

Another key aspect to highlight in this model is that the coefficient of determination, or $r^2$, is 0.503. This number displays the variation in the response variable, "Trip Duration," explained by the other variables. In effect, 0.503 of variability in duration of a trip can be understood by knowing "Trip Distance," "Vehicle Type," "Hour," and "Day of Week."

Diving into the characteristics more, the RMSE (root mean square error) for the multiple linear regression model is 8.28 minutes. Essentially, this number estimates the standard

deviation, or the error, to be 8.28 minutes. Considering that 75% of the data is within 12.25, this RMSE is very high.

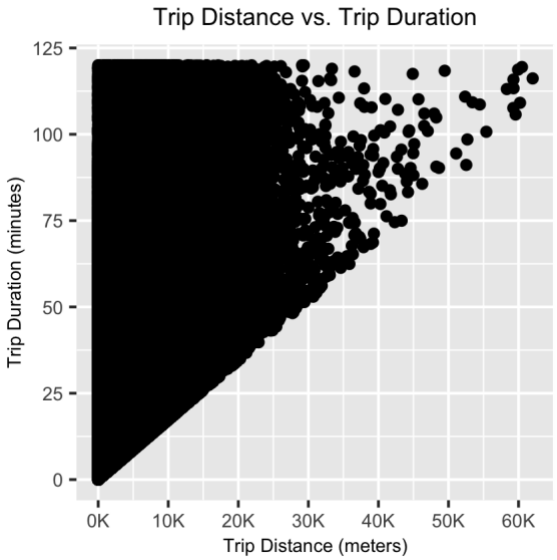



Fig. 6: Scatter Plot: Trip Distance vs. Trip Duration

## *Assumptions Checking*

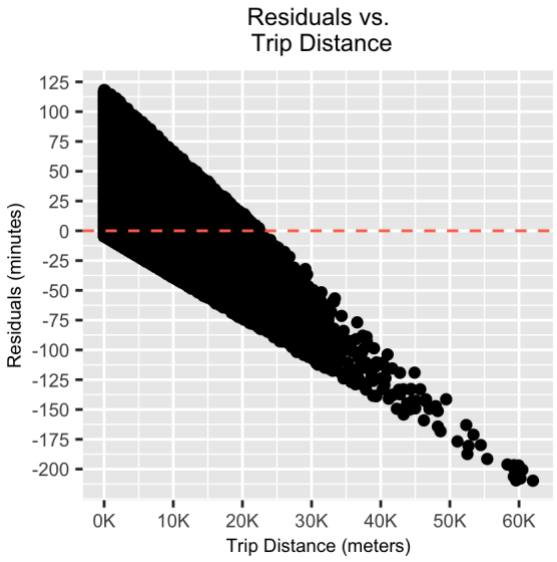



*Fig. 7: Assumption Check #1*

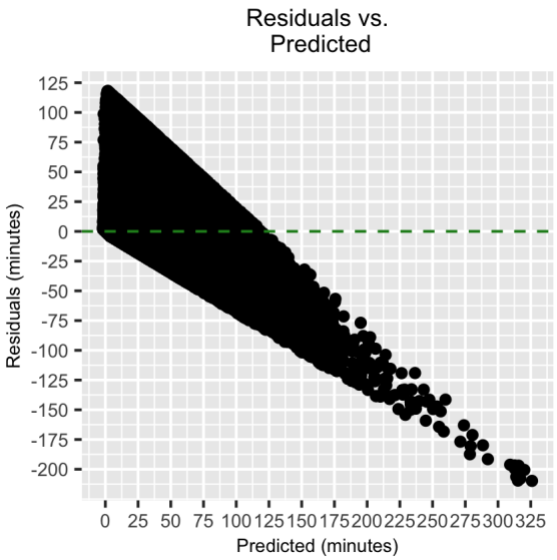



Fig. 8: Assumption Check #2

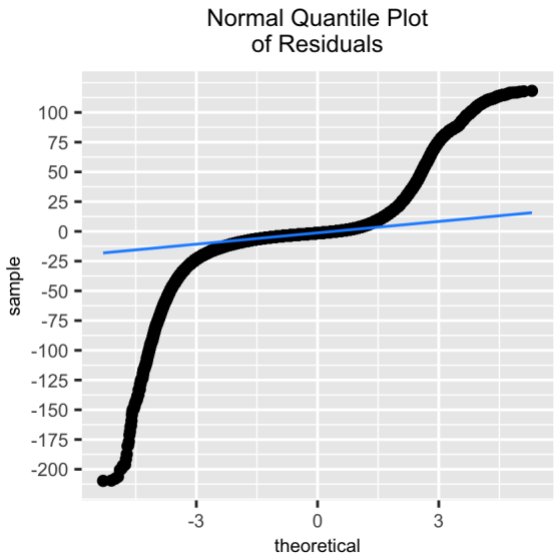



Fig. 9: Assumption Check #3

This multiple linear regression model fails the first assumption that the $x$ variable, or "Trip Duration," is measured without error. In the very beginning, I filtered out data deemed logically incorrect which could lead to this error. This multiple regression test fails the next assumption that there is constant variance. This assumption is violated because the data residuals are not evenly spread across the line on the residuals versus "Trip Duration" plot above (Fig. 7), as they form a more conic shape. Furthermore, "Residuals vs. Predicted" (Fig. 8) does not show constant variance as well. Similarly, this multiple regression test also fails the error normality assumption as it does not follow the normal quantile plot well due to the heavy tails (Fig. 9). Another assumption is that this data is a simple random sample. This data does not represent a simple random sample as I used every single data point until that time and filtered data. The last assumption, that the errors are independent of each other, is also violated by this data set. This data set violates the independent error assumption because if one reporting chip was faulty in a specific shared micro mobility vehicle, then it would report error in other trips as well if used again.