**Week 10- Final Project Part 3**

**Zachary DeNoto**

**Uber and Lyft Prices**

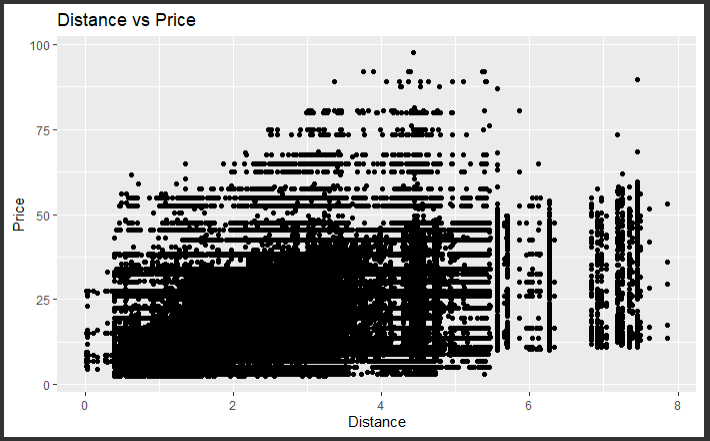
**Summarize the problem statement you addressed.**

In the past few years, ride share companies such as Uber and Lyft have become very popular as an alternative to the long used taxis. My interest lies in the difference between the two. As someone who has used both, I have always wondered if there was a large difference between the two. When I was a college student there were only two deciding factors for my friends and I to determine which rideshare app to use: price and how quickly it can pick us up. I still use the apps to get around if I am in a city without a vehicle or if I plan to go out with friends and we do not want a designated driver. In a little over a week I am heading to a trip in Europe and may even use the services over there. As with many elements in life, price is one of the largest determining factors. This led me to my interest in seeing which one is more expensive and what the determining factors are for price.

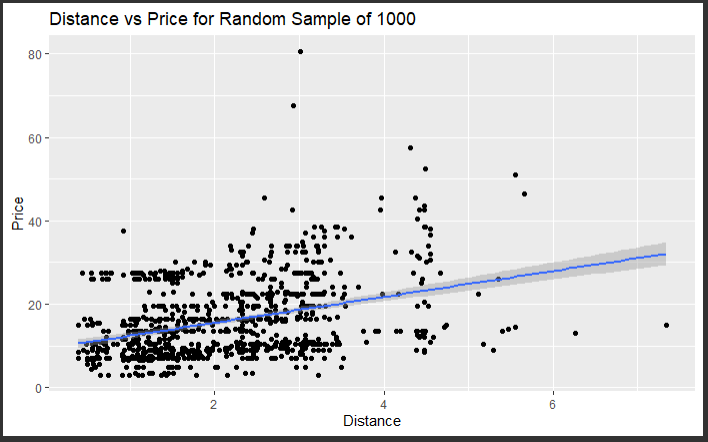
**Summarize how you addressed this problem statement (the data used and the methodology employed).**

I found a dataset on Kaggle.com which contained data from around 700,000 Uber and Lyft rides in the Boston area from November to December 2018. The data was very well split up with about half of the data being from Uber rides and the other half being from Lyft rides. I looked at the data and noticed that it could be cleansed as the date variable was in epoch timestamp. In excel I converted the variable and broke it into 3 different variables, one for Date, one for Day of the week, and one for time of the day, which was broken into 24 timeslots (0-23). I removed the source, destination, product id, and id variables as I did not need them. For the variable name, which is the type of Uber or Lyft ride, there is a ride type called Taxi where you can order a taxi using Uber. The variable had no prices whenever the name variable was Taxi, so I removed all Taxi name variables in the dataset. The final step that I needed to do was to convert the date variable into a date data type.

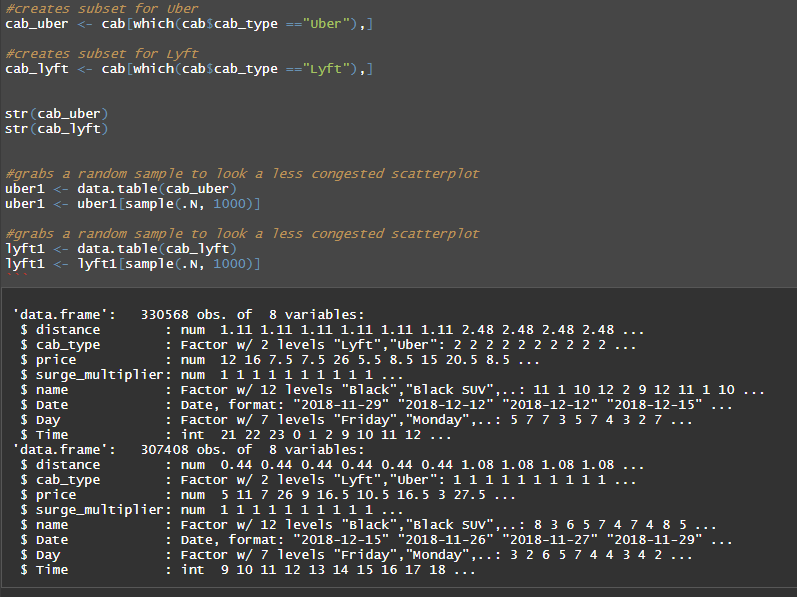
Once the data was cleansed, exploration of the data could occur. At first I ran a scatterplot to compare distance against price to see if there was a clear pattern.

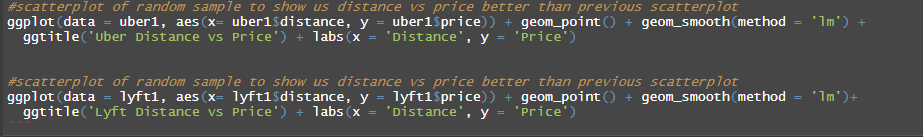


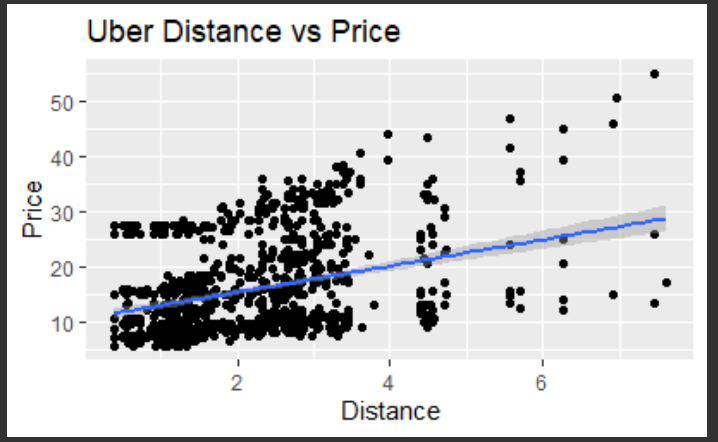
The problem with the scatterplot was that there were too many data points to clearly see if there was a pattern or trend. So I then took a small, random sample of 1000 points to see the data better.

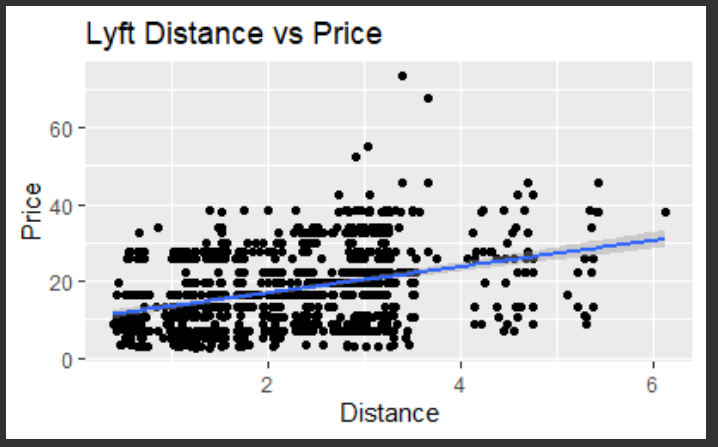


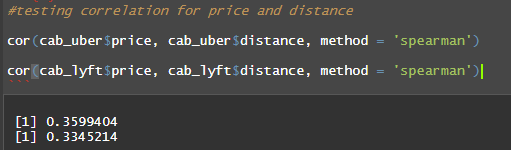
The next thing I did was to split up the data into two subsets, one for Uber and one for Lyft. I then ran many different tests including scatterplots, correlation tests, linear regression, and created new subsets based on day of the week, weekdays, weekends, day time, and night time for both Uber and Lyft, as seen below.

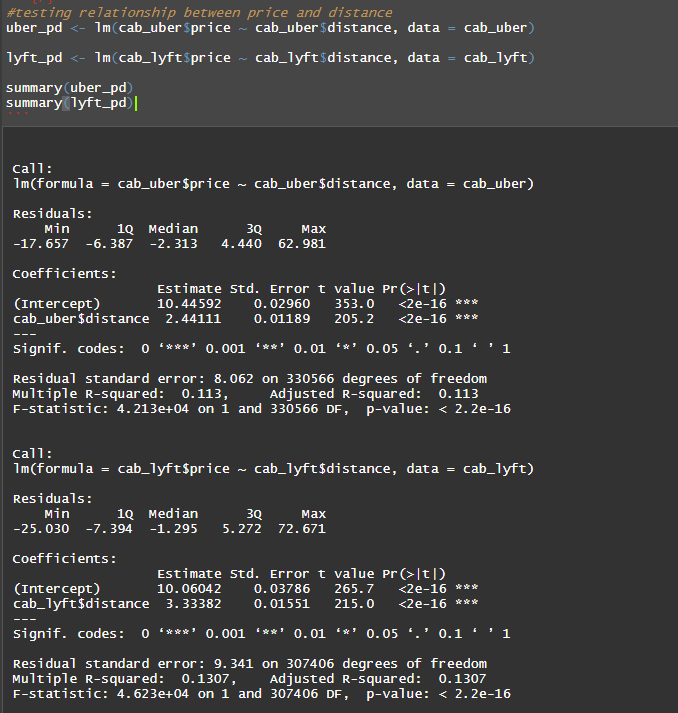


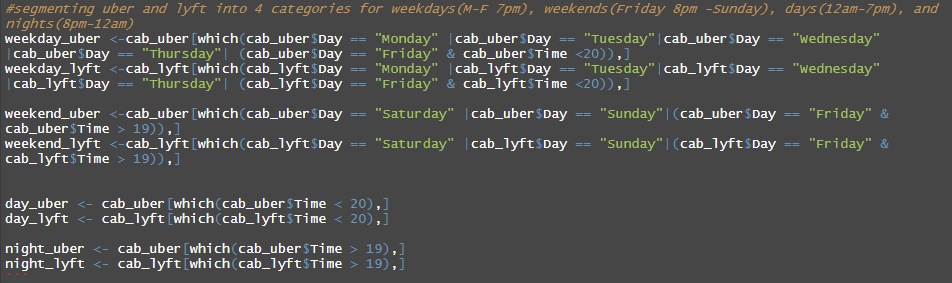


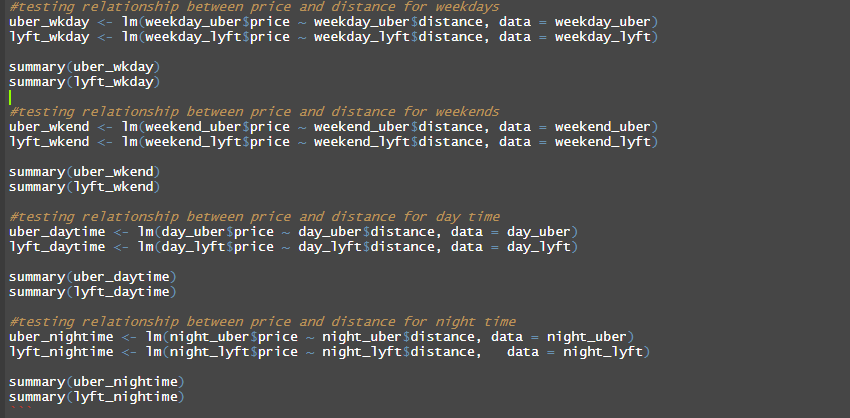


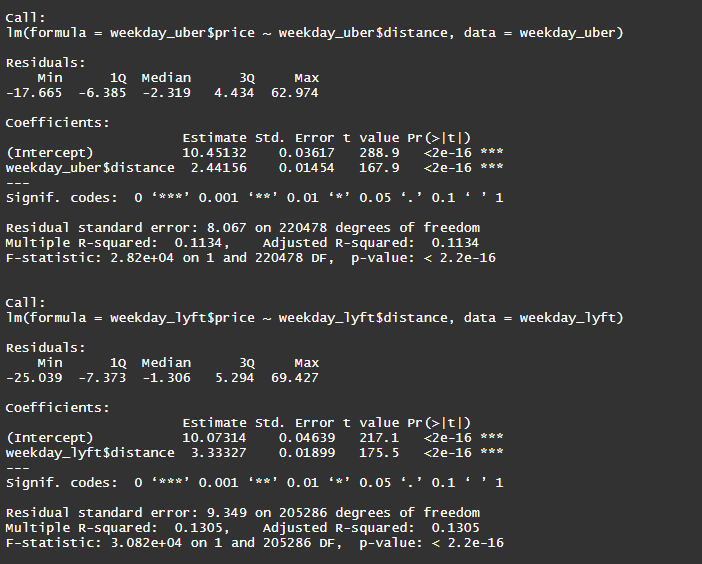


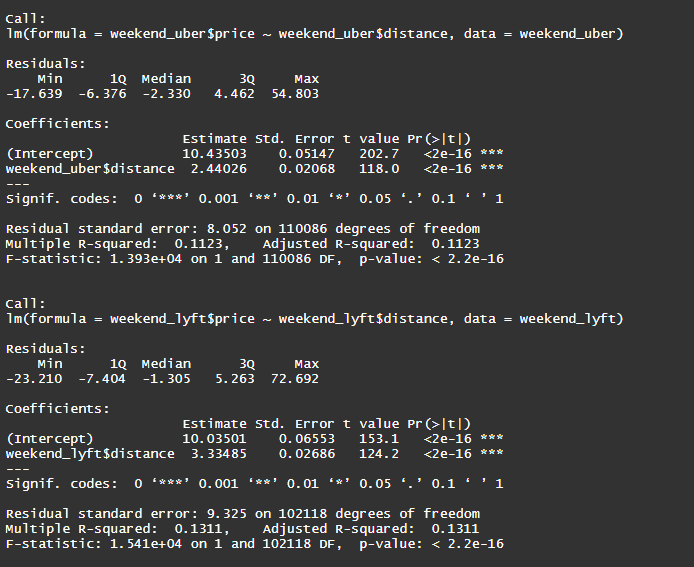


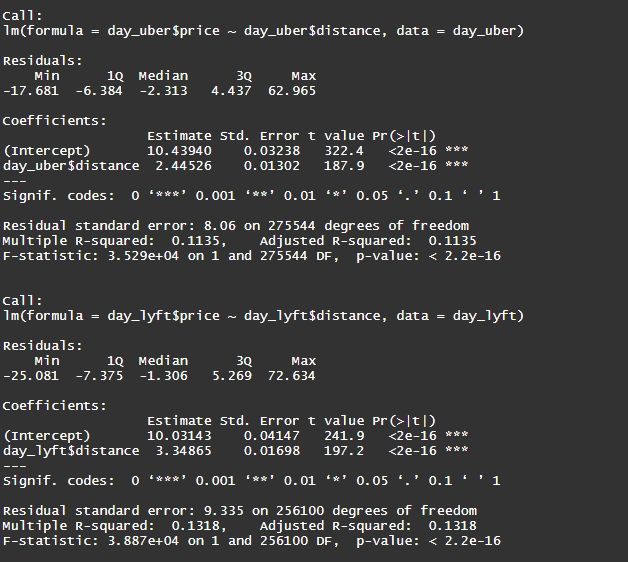


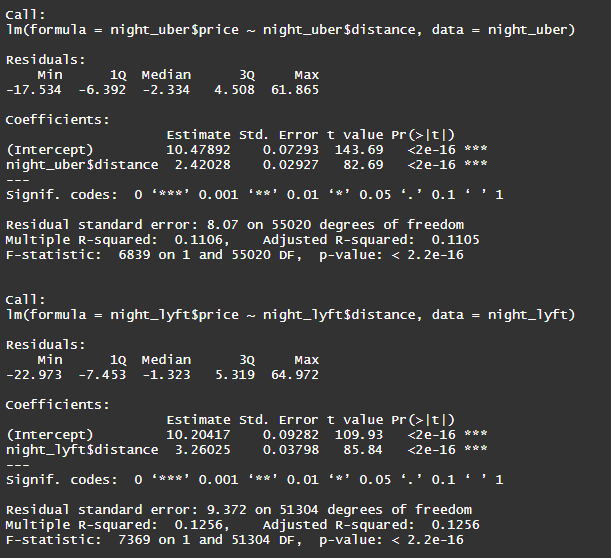


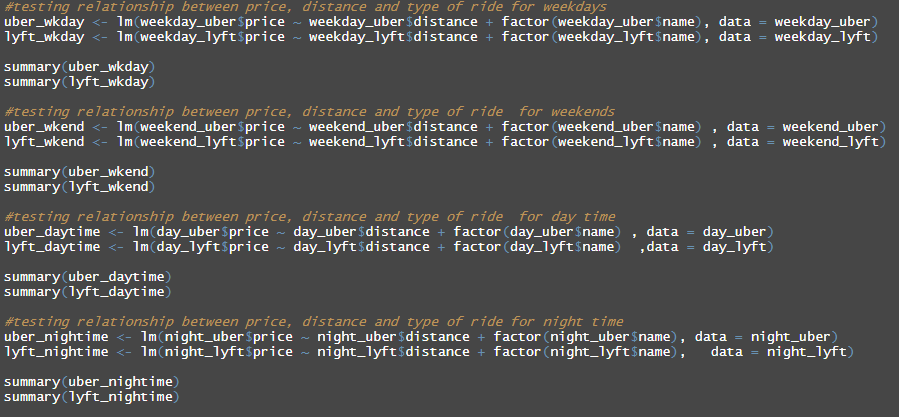


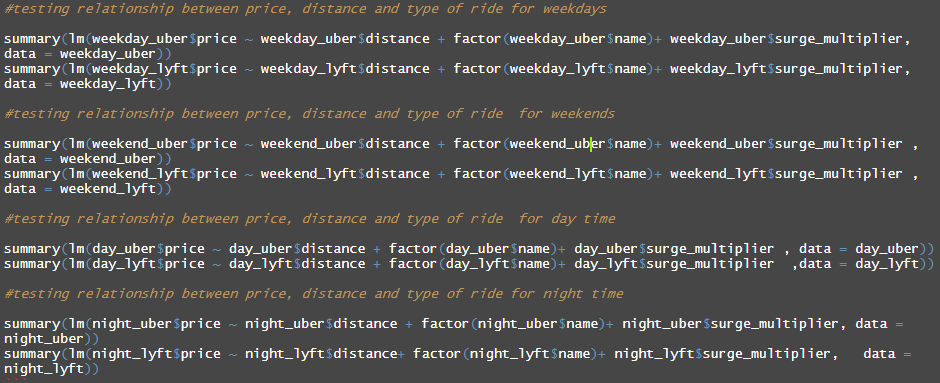


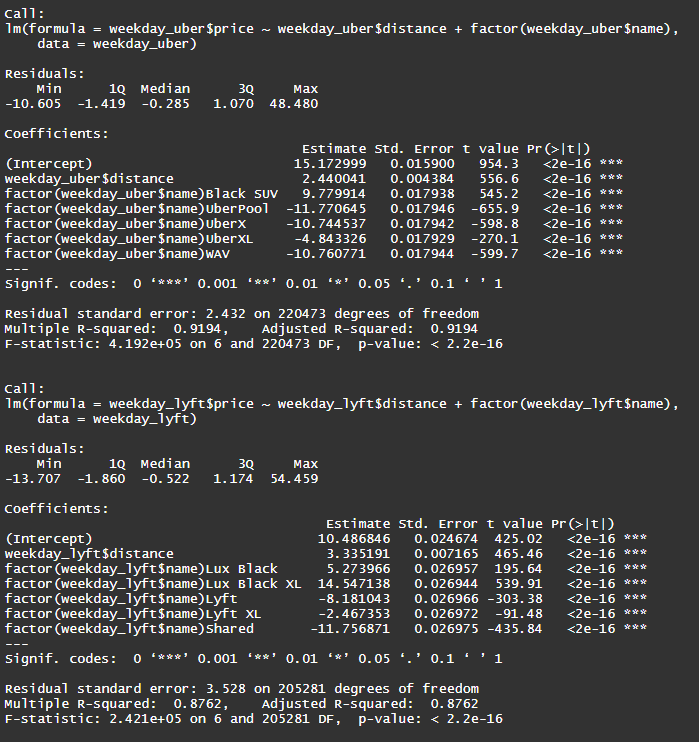


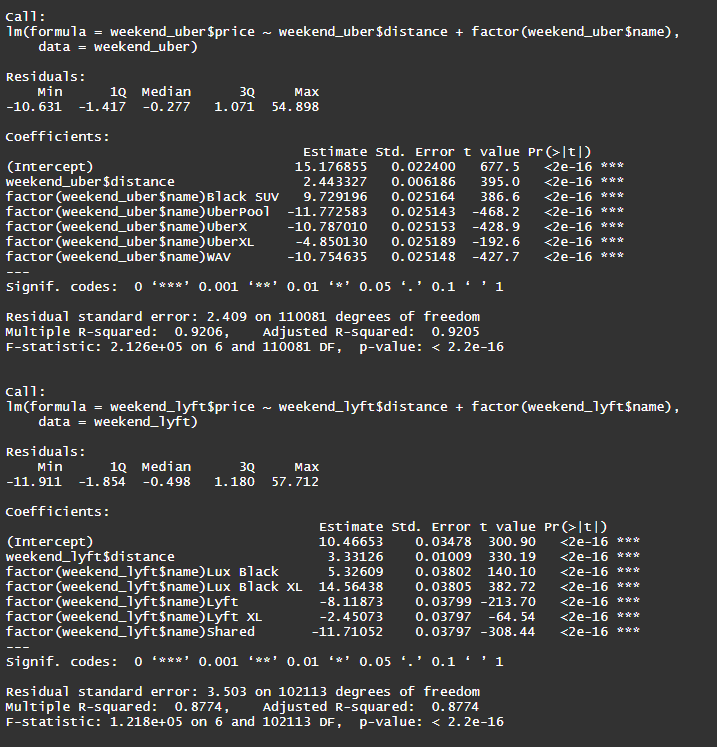


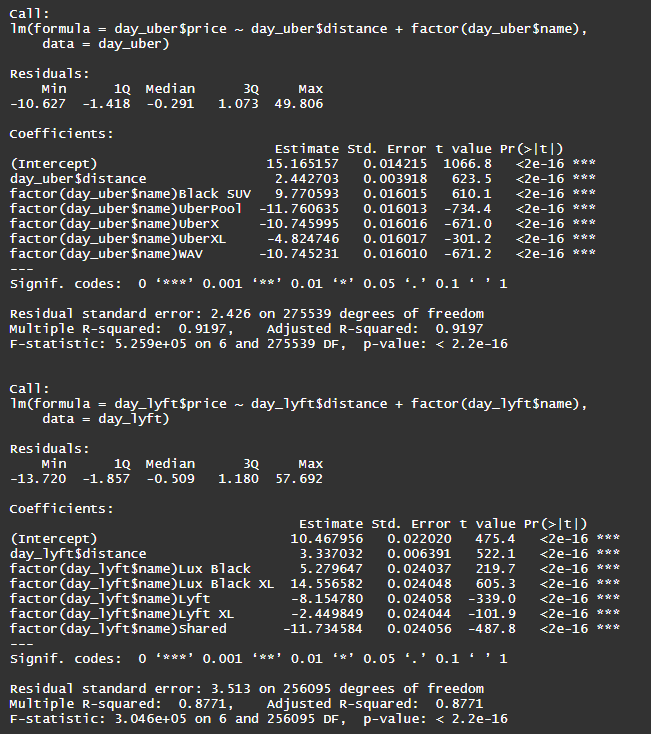


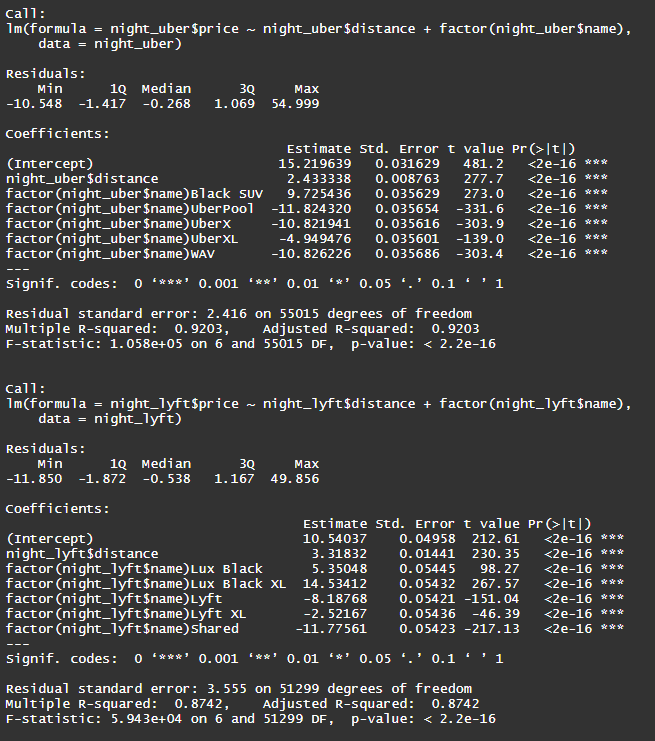


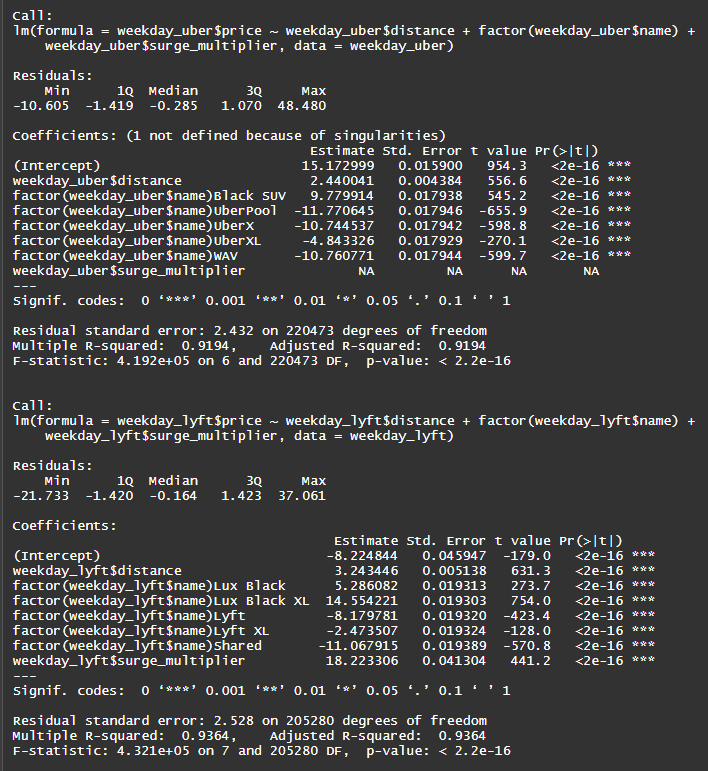


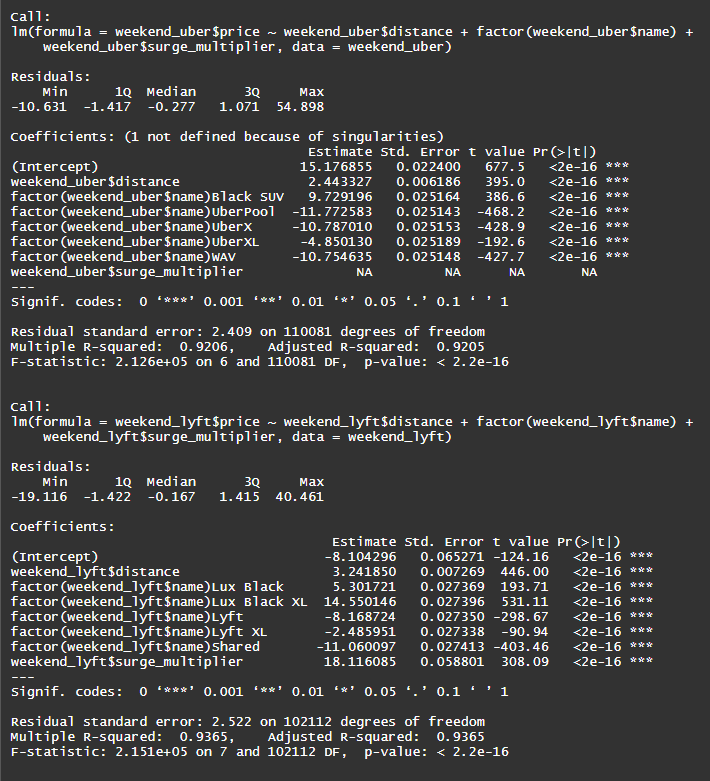


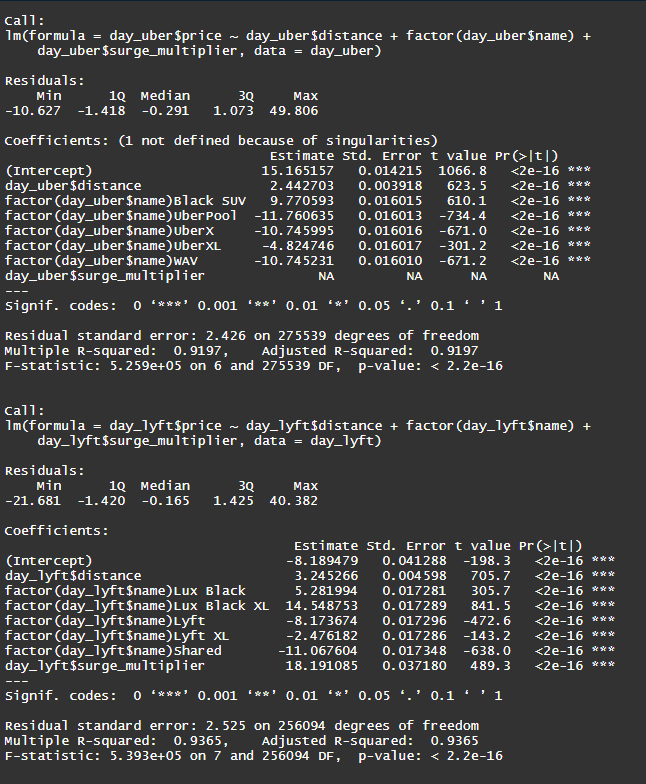


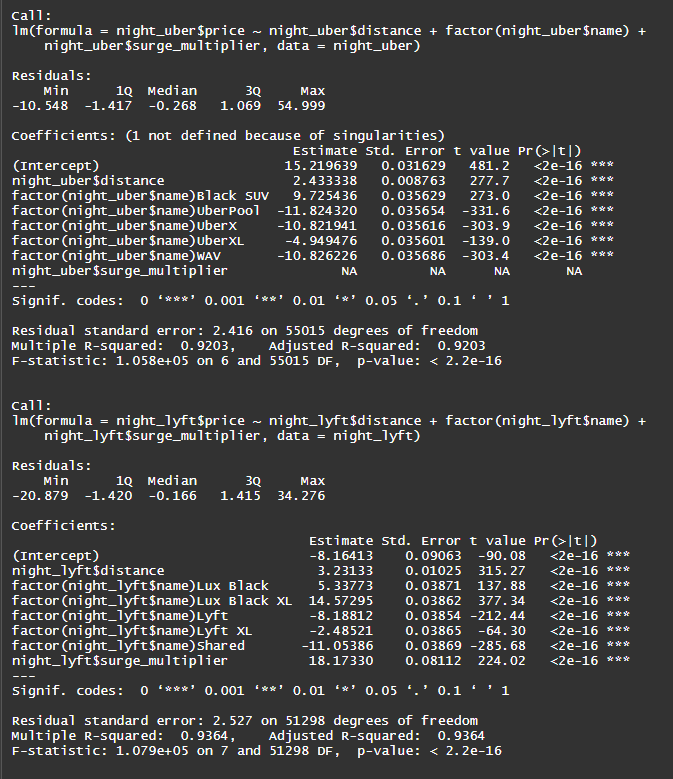


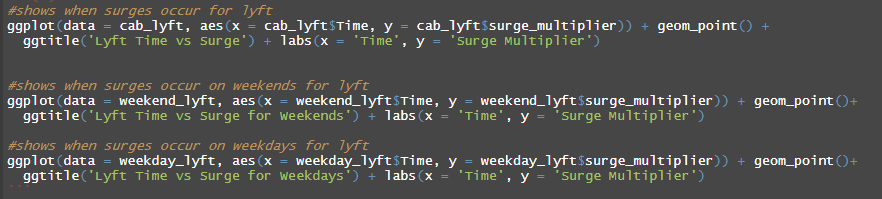


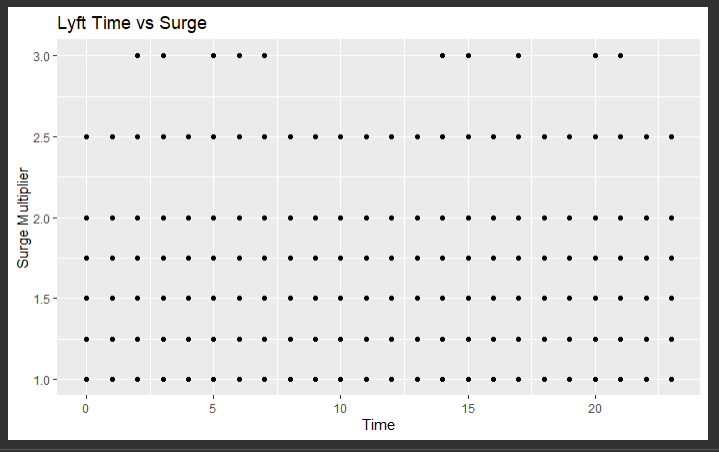


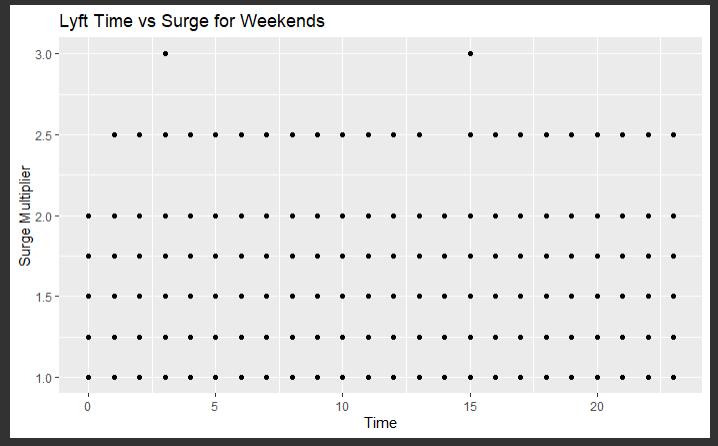


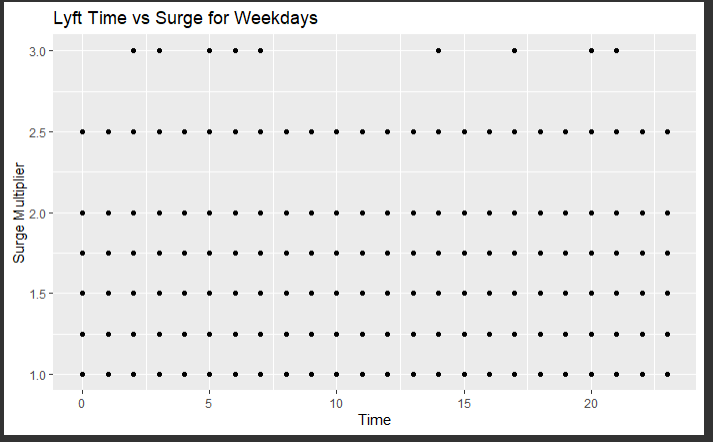


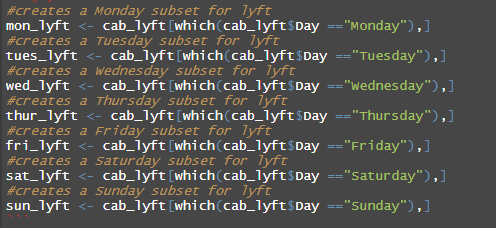


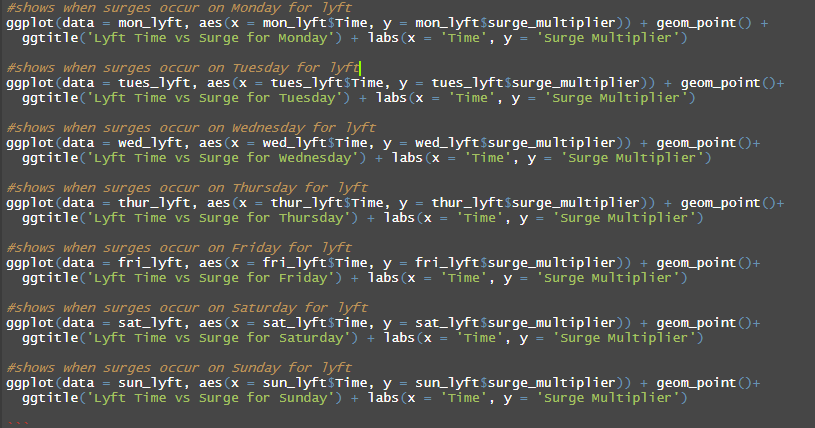


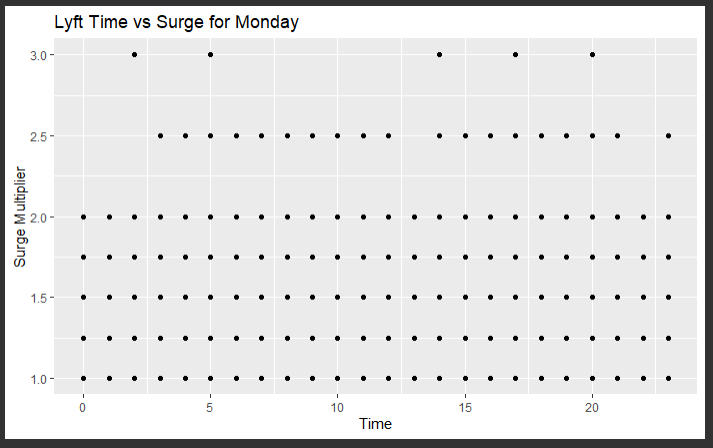


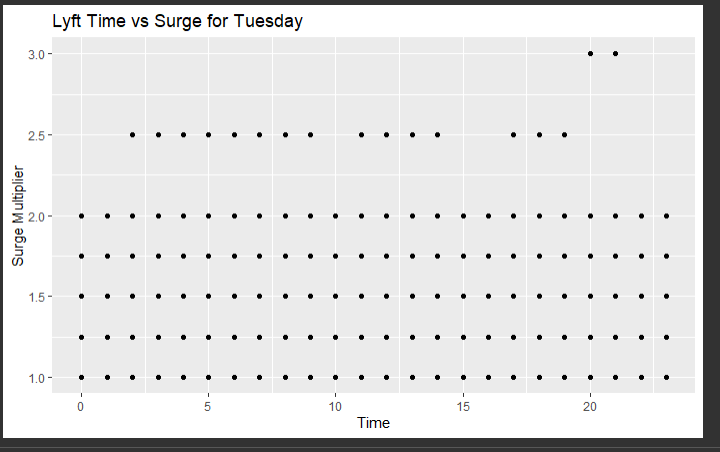


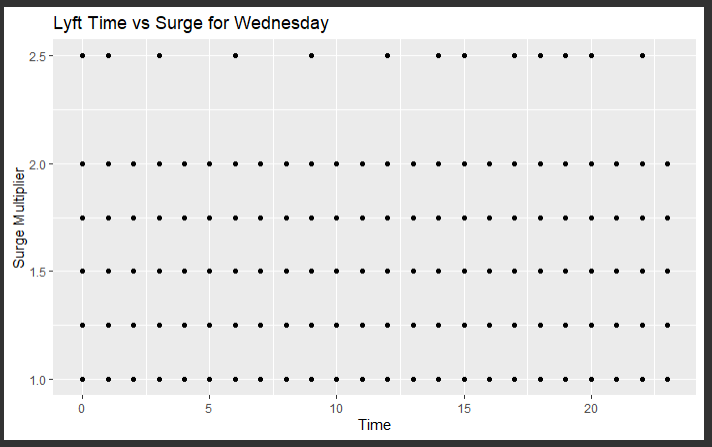


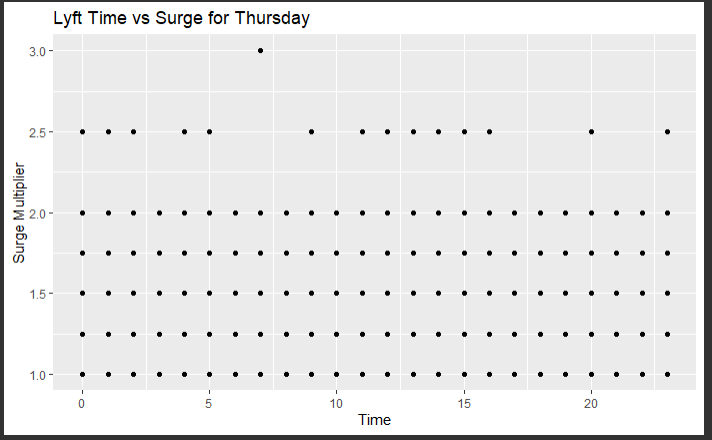


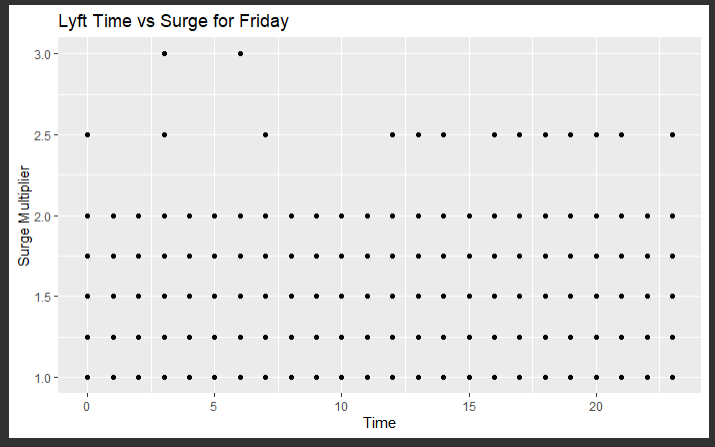


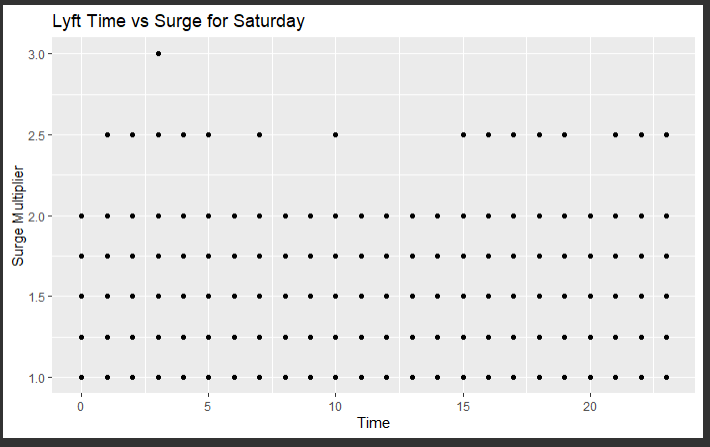


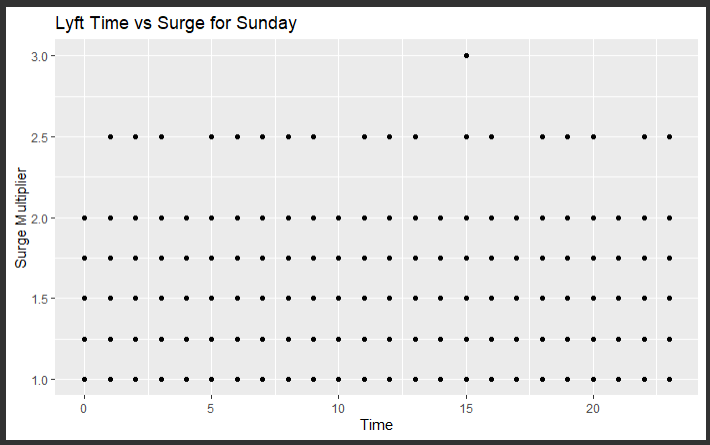


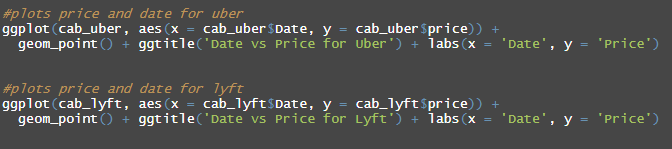


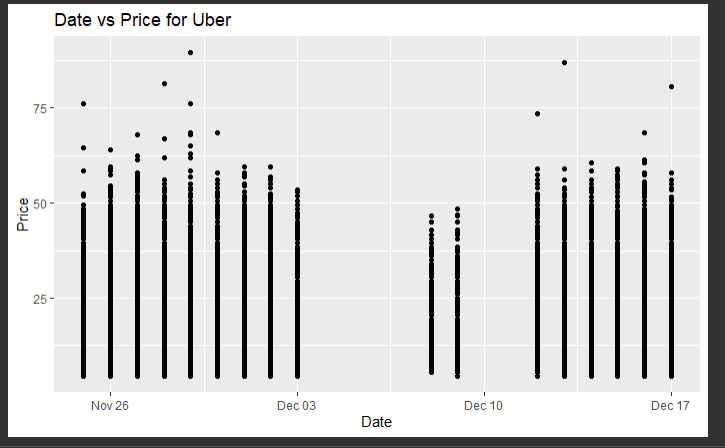


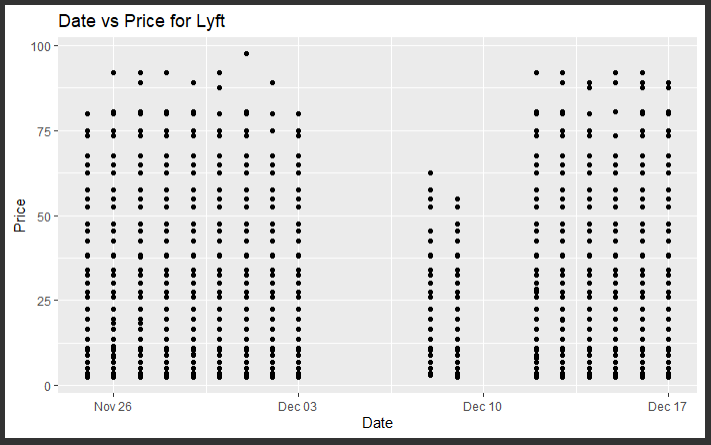


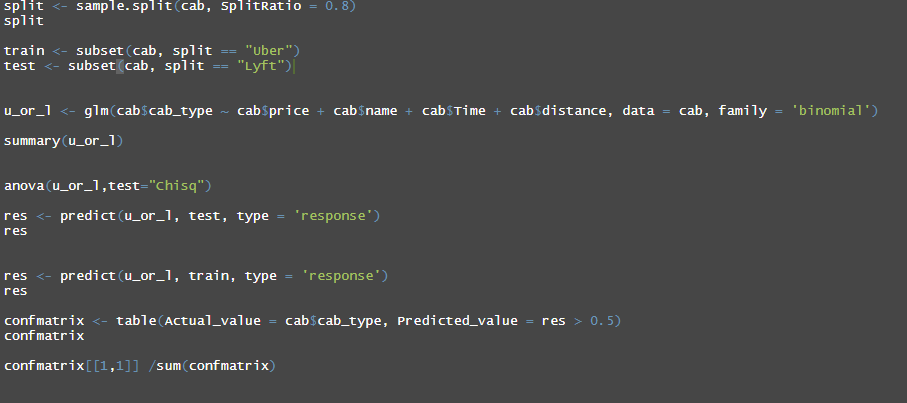


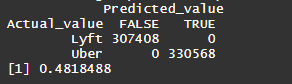












This provided a lot of insights which can be found in the following section. Lastly, I tried to use both k-nearest neighbor and the predict function. However, when I tried to use k-nearest neighbor it crashed the program every time due to the amount of data.

**Summarize the interesting insights that your analysis provided.**

* Uber’s prices are more concentrated than Lyft. Uber’s prices are more similar for distances as compared to Lyfts.
* There is a low correlation between distance and price for Uber and Lyft. However, the correlation is slightly higher for Uber.
* Lyft is more affected by time and day of the week for surge prices.
* Surge prices can be at any time, but certain days and times have a better chance to not have the highest surge multiplier (please see next section).
* The prediction model was only 48% accurate with the dataset.
* When using linear regression, Lyft distance affects price more than Uber.
* When using multi linear regression and factoring the variable name (cab type), Uber’s prices were affected more by the variables as compared to Lyft.

**Summarize the implications to the consumer (target audience) of your analysis.**

Though I was not fully able to determine if Uber or Lyft is more expensive than the other, there are several important outcomes that came from this analysis. The first is that Uber’s prices are more clustered than Lyft’s. If you were to get a price for a ride and decide to wait, there is a higher chance the price would vary more with Lyft than with Uber. There is always a chance there will be a surge depending on the number of people who want rides and the amount of drivers, but with Lyft there is a higher chance to get a 3x surge at 3pm and 3am. Wednesday is the best day for a Lyft ride as the likelihood of getting a 3x surge is very minimal. In this dataset there were no 3x surges on Wednesdays. The highest surges occurred at 2am, 3am, 5am, 6am, 7am, 2pm, 3pm, 5pm, 8pm, and 9pm. This means that using Lyft between 8am-1pm and after 9pm have the best chance for not getting as high of a surge.

**Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.**

Throughout the project I came across several limitations which were not anticipated. The first issue which came up was discovering that the surge factor for Uber were all 1’s. This is due to the fact that Lyft discloses the surge multiplier whereas Uber included it in the price behind the scenes. If you go through the dataset, the surge multiplier for all the Uber rides are 1 where Lyft surge multiplier ranges from 1-3. This also causes our data to not be fully accurate as the Uber prices potentially could be higher than the Lyft prices due to the built in surges. The other limitation was the lack of variables and data. There are many factors which determine the price of each rideshare company that is secret to each of them. If I had other data like price of alternative ride share apps at the same time, number of ride requests at that time, and number of cars available at that time, it could have helped to determine the pricing difference.

I am not that familiar with the Boston area that the dataset is from, but perhaps someone with knowledge of the Boston area would be able to split the data better by location. The original dataset had a variable for source and destination, but since I am not familiar with Boston, I removed the variables. The dataset is only for November and December of 2018 in Boston. It does not contain information on other timeframes or locations. The results from this dataset could be vastly different if the timeframe and location were different. If someone had data for a large timeframe and all locations he could get much better results, especially if the Uber surge multiplier was there.