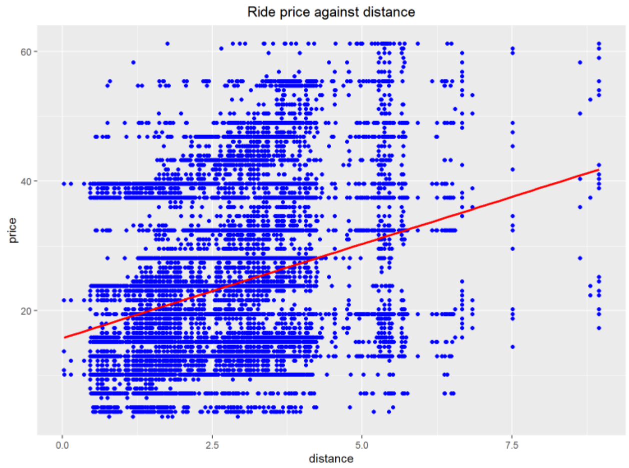
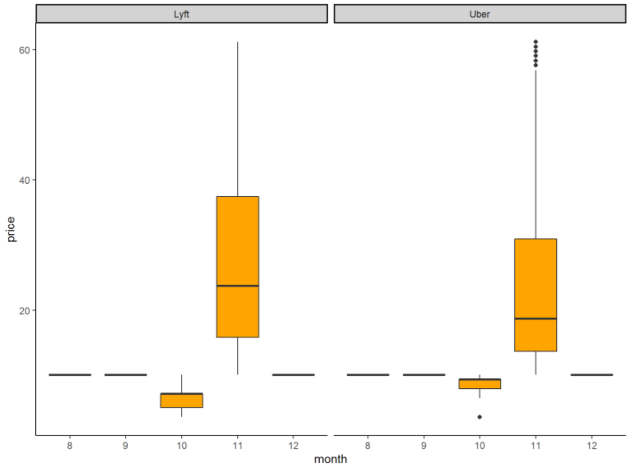
Group 8 - Assignment 2

Group Members: Yihsuan Tseng, Chia-Ling Ni, Mihir Sandeep Kungulwar, Tejaswini Edupuganti

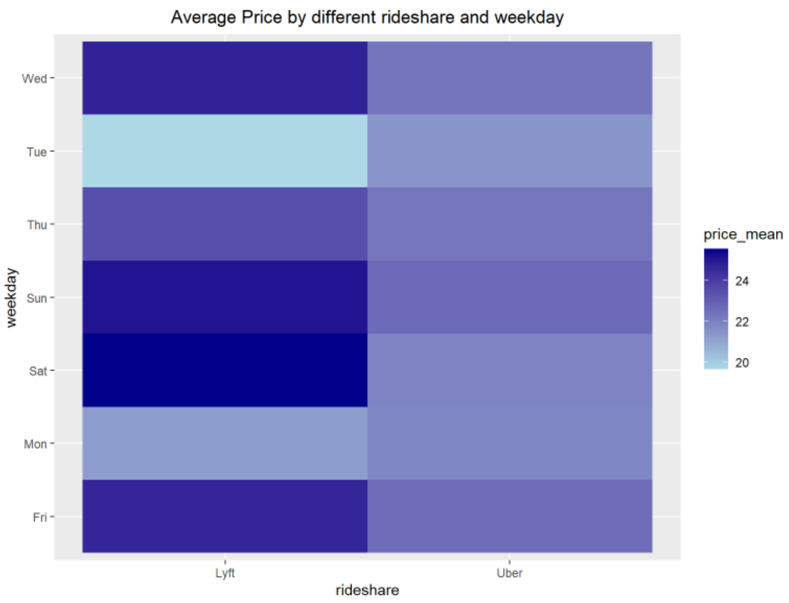
What are some insights gained from exploratory data analysis? (Specify three with visuals)



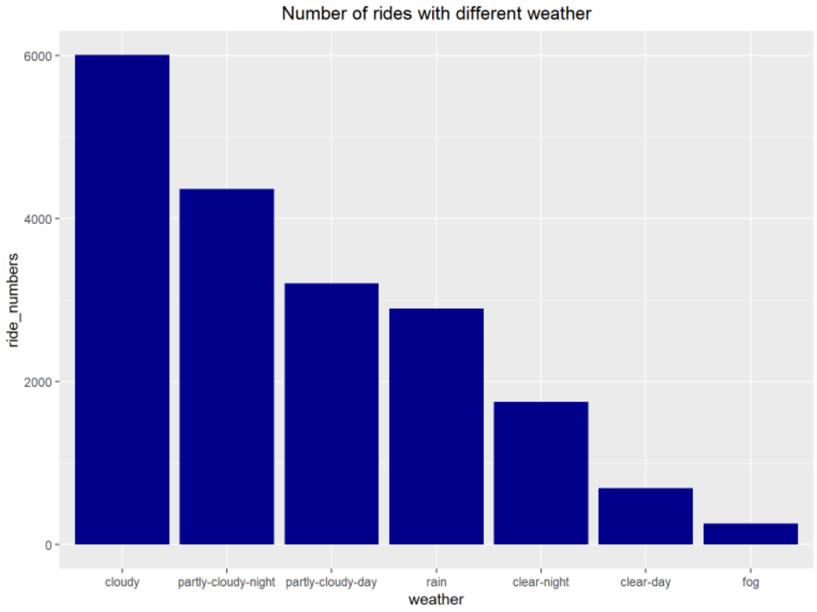
- Insight from plot 1: Most of the distance condition per ride is under 5.00. We can also see that distance and price have a positive relationship.



- Insight from plot 2: The price varies a lot, and the average price is higher in November compared to August, September, October and December with both Lyft and Uber. And the average price from Lyft is higher than Uber in November, however, we can also see that the average price in October from Lyft is lower than Uber.



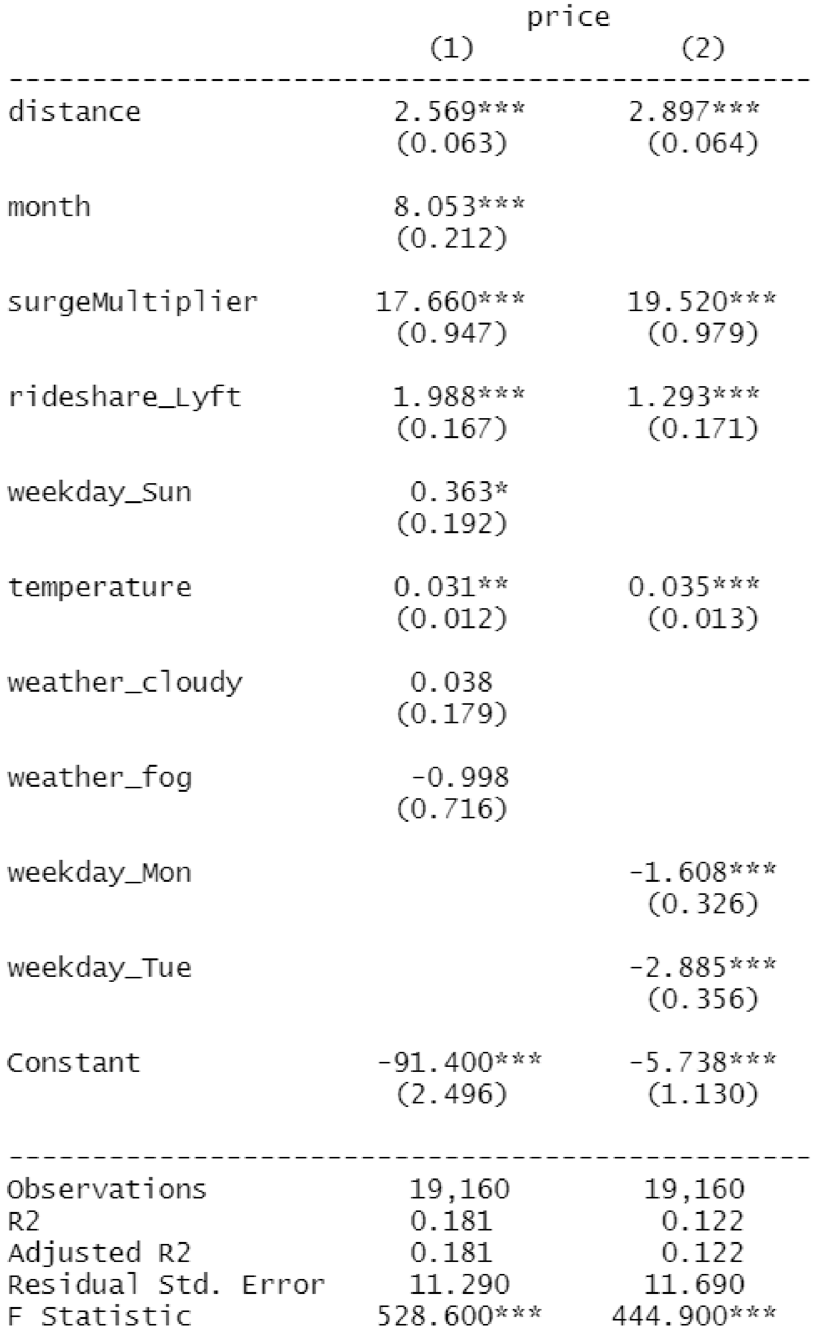
- Insight from plot 3: Under different weekdays, the average price range is larger and fluctuates with Lyft, while the average price range is similar to Uber.



- Insight from plot 4: People prefer to take rides on cloudy days, and they tend to not take rides on foggy days or clear-days.

Describe what techniques you used and why you felt they were appropriate.	We used techniques like dummy encoding, multivariate linear regression and correlation. The reason for using these techniques is because of the kind of data that was under inspection. Having a few categorical variables in the form of text was holding some very important value that we would have missed out on if we did not encode it in the binary form which is much better for the linear model to consume. Also, a linear model was an obvious choice considering we are dealing with a continuous target variable. Correlation was also necessary to be known since we do not want to supply unimportant or redundant information to the model as it risks multicollinearity. We also used visual techniques to represent the insights as it is a better way to abstract the complexity of the process.

Describe insights gained from your analysis with a discussion of what the results mean and how you assessed performance of the models.



```
                              price
                      (1)              (2)
----------------------------------------------
distance            2.569***        2.897***
                   (0.063)          (0.064)

month               8.053***
                   (0.212)

surgeMultiplier    17.660***       19.520***
                   (0.947)          (0.979)

rideshare_Lyft      1.988***        1.293***
                   (0.167)          (0.171)

weekday_Sun          0.363*
                   (0.192)

temperature         0.031**         0.035***
                   (0.012)          (0.013)

weather_cloudy       0.038
                   (0.179)

weather_fog         -0.998
                   (0.716)

weekday_Mon                        -1.608***
                                    (0.326)

weekday_Tue                        -2.885***
                                    (0.356)

Constant           -91.400***      -5.738***
                   (2.496)          (1.130)

----------------------------------------------
Observations         19,160          19,160
R2                   0.181           0.122
Adjusted R2          0.181           0.122
Residual Std. Error  11.290          11.690
F Statistic        528.600***      444.900***
```

R-Squared of the regression models:
For the first linear regression model, R-squared=0.181, which means the descriptors explain 18% of the variation in price. For the second linear regression model, R-squared=0.122, it shows that the model explains 12.2% of the variation in price. It demonstrates that "Month" might be an important feature for the model to make a prediction.

RMSE (root mean squared error) of the training and testing sets:
Using multi-linear regression model 1 to do the training and

testing with set.seed(0), we could get RMSE 11.31 on the training set, and RMSE 11.22 on the testing set.

To compare, multi-linear regression model 2 gets RMSE 11.7, and RMSE 11.62 on the testing set. Based on the result above, the first model has minimized the RMSE.

MAE (mean absolute error) of the regression models:
On regression model 1: 9.279
On regression model 2: 9.638

To summarize and assess the quality of the prediction model, we use MAE(mean absolute error) to test both the regression. The MAE for regression model1 is 0.36, smaller than regression model 2.

Conclusion
Based on R-Squared, RMSE , and MAE scores, **multi-linear regression model1** is more suitable to make a prediction for the dataset.