ANALYSIS ON RIDESHARE DATA

1.PROBLEM DESCRIPTION

The problem described is to understand the consumer behavior and preferences in terms of the cost of ride-hailing services. The goal is to determine how much customers are willing to pay for these services and how it affects their choice of ride service provider. Despite the growing popularity of ride-hailing services like Uber and Lyft, the contribution of these services to the total vehicle miles traveled is relatively low, indicating significant room for attracting new customers and increasing their usage.

2.DATA DESCRIPTION

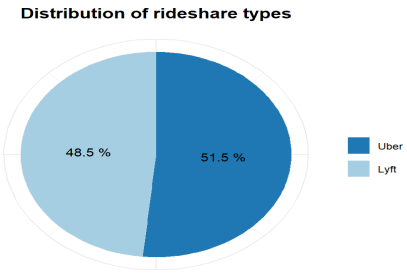
In this dataset we have 19160 records which represent the rideshare from Lyft and Uber in the Boston area. We also have weather data by location, day and time.

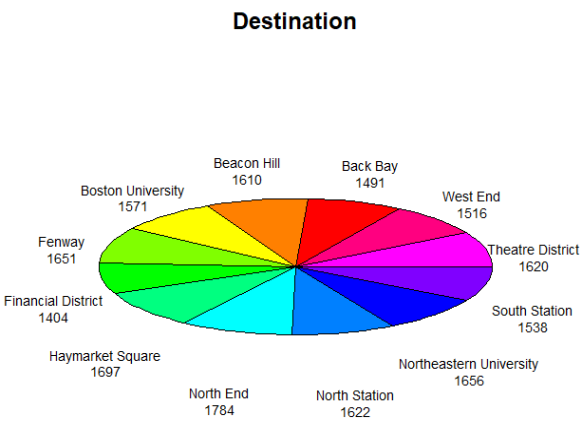
1. **Id** – This column gives a unique ID for each observation. We have 19160 unique ids
2. **Datetime** – In this column we have the date and column of the ride. The datatype of this column is date. We have 6031 unique datetime values. In our data set, we have data from 2020-08-01 04:57:00 UTC to 2020-12-28 19:53:00 UTC.
3. **Hour** – In this column we have hour which is extracted from the datetime column and its datatype is numerical.
4. **Day** – In this column we have the day of month which is extracted from the datetime column and its datatype is numerical. We have 15 unique values in the day column.
5. **Month –** In this column we have a month number which is extracted from datetime columns. We have data for the months of August to December.
6. **Weekday** – In this column we have the day of the week.
7. **Source** - In this column we have the location of pickup and it is of string data type. We have 12 unique sources in this column.
8. **Destinations** - In this column we have the destination of drop off and it is of string data type. We have 12 unique destinations in this column.
9. **Rideshare -** In this column we have the name of rideshare services - Lyft and Uber
10. **RideCategory** - In this column we have types of rideshare services of both- Lyft and Uber.
11. **Price** - In this column we have the total price of the ride. The price range is from 3.6$ to 61.2$.

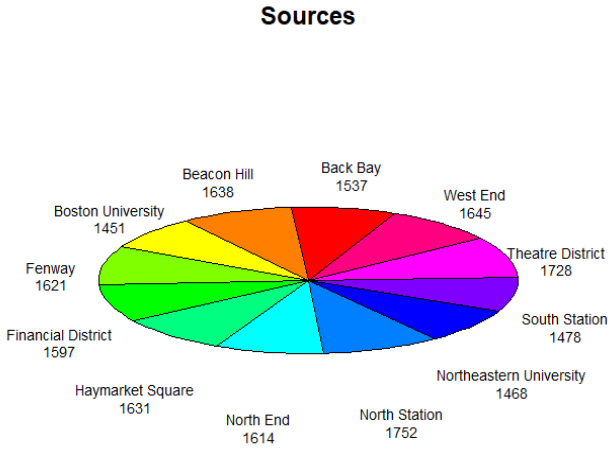
12. **Distance** - In this column we have the total distance of the ride. The smallest distance is 0.024 miles and maximum distance is 8.95 miles.
13. **SurgeMultiplier** - In this column we have a peak time price multiplier. We have 6 unique surge multipliers.
14. **Weather** - In this column we have the description of weather. e.g. cloudy, rainy. We have 7 unique weather descriptions.
15. **Temperature** - In this column we have the temperature during the ride.
16. **precipProbability** – In this column we have the likelihood of rain for a specific forecast period and location.
17. **Humidity** - In this column we have the humidity during the ride.
18. **Wind Speed** - In this column we have wind speed during the ride.
19. **Windgust** - In this column we have the increase in wind speed during the ride.
20. **Ozone** - In this column we have the air quality during the ride.

3.EXPLORATORY DATA ANALYSIS

3.1 First we checked if the data is balanced or not to make sure our findings are not biased. We can see that Uber and Lyft have almost equal numbers of rides. In the source column we have 12 different sources and from each source we have approximately 1500 rides.

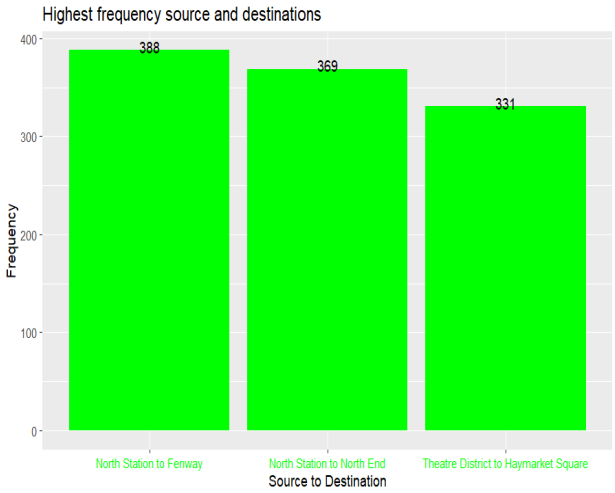


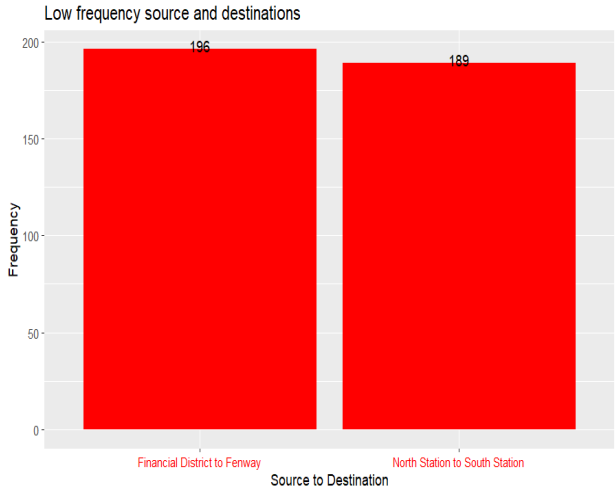




3.2 In the destination column, we have 12 unique destinations. The frequency of rides to these destinations ranges from 1404 to 1784.

3.3 Highest number of rides are from North Station to Fenway with 388 rides and lowest are from North Station to South Station with 189 rides.The highest distance is 8.952 miles from the Financial district to North Eastern University. And the lowest distance is 0.024 from South Station to Theatre District.

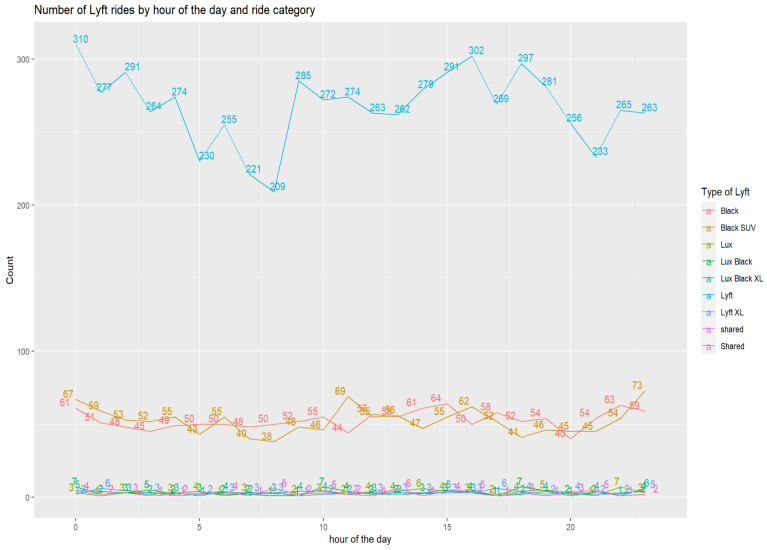




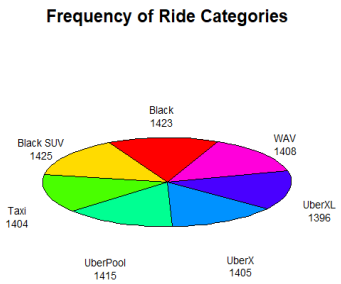
4. INSIGHTS

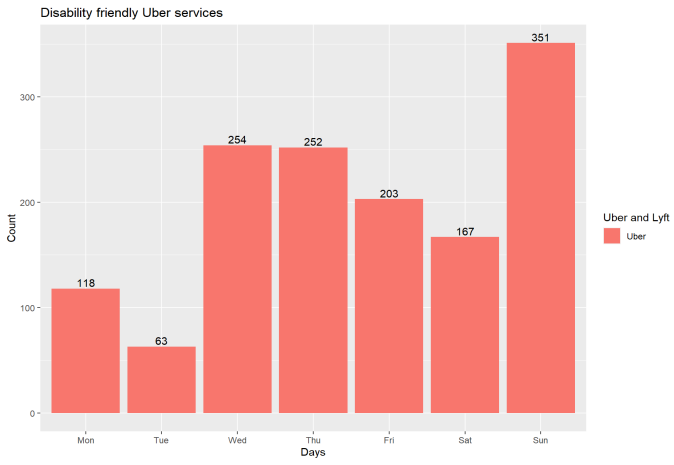
4.1 Insight 1 - Promoting higher segment cars and catering to different segments

The graph shows the frequency with which all the Lyft company cars are used during a 24-hour period. It can be seen that among the different segments of cars provided, Lyft is the most frequently used rideshare service during the hours of the day, as seen from the highest count of rides. Black and Black SUV are the next most popular options. Shared and Luxury segments are the least used. This trend however, is not observed with Uber, where all the segments are popular and highly used.



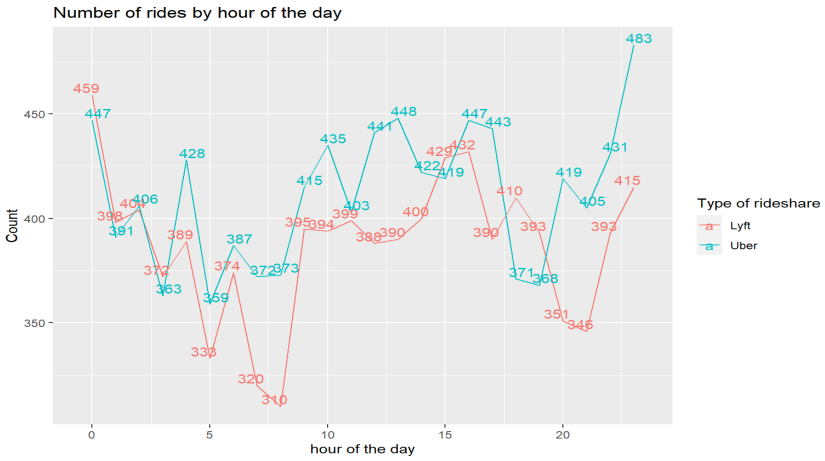
Despite this, both Uber and Lyft have similar average usage levels per hour,





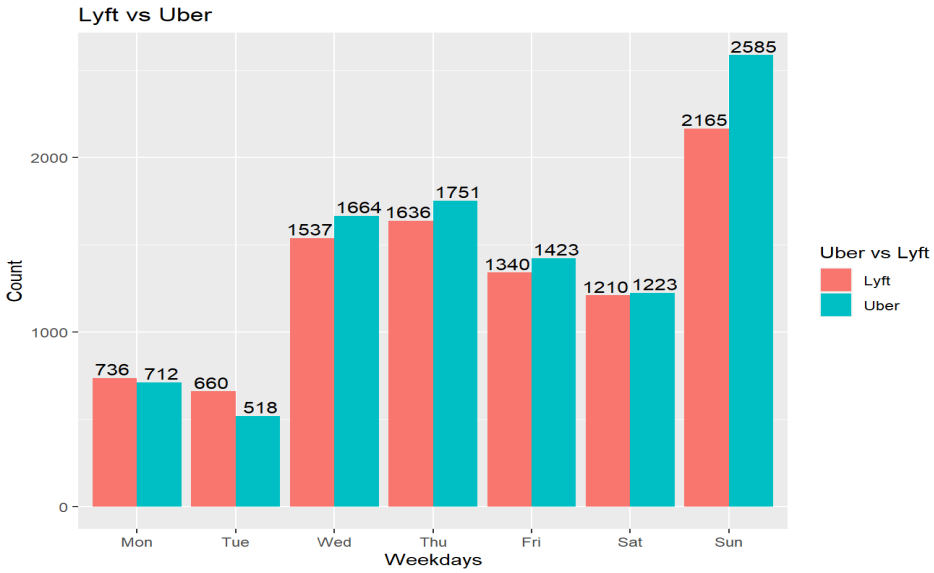
which suggests that there is potential for Lyft to increase their profit margins by engaging their larger, higher-priced cars. For Lyft, it highlights the opportunity to increase revenue by offering higher-priced car segments and, potentially, attract a different customer demographic. Lyft can increase its profitability by promoting its luxury segment. Additionally, the company can also explore new opportunities by entering the market for disability-friendly vehicles, as seen in the usage of such cars by Uber, as depicted in the graph. By expanding into this market, Lyft can diversify its offerings and attract a wider range of customers.

4.2 Insight 2 -Optimizing Customer Experience through Hourly and Weekly Demand Analysis



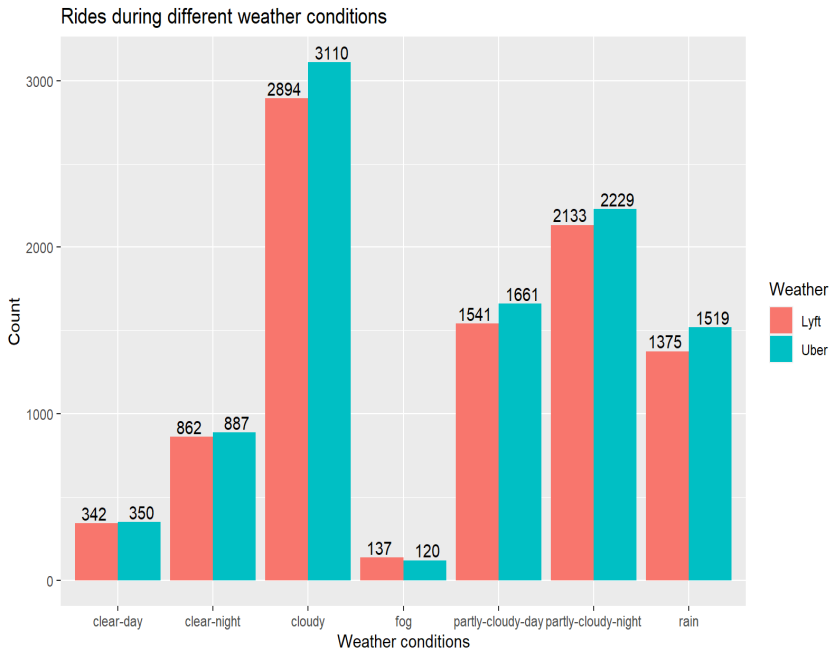
The graph on the left explains the usage of Uber and Lyft during the various hours of the day. Uber is noticeably used more than Lyft throughout. Usage peaks during the office hours and during the late evening hours. To increase profitability, both Uber and Lyft can leverage data analysis to identify the highest demand hours during the day and also on different days of the week. More vehicles and drivers can be deployed during peak hours to meet the increased demand and for more revenue generation. Additionally, they can also use dynamic pricing to boost their profits.

The companies can identify areas where demand is constantly high and increase their presence in such locations. Pricing can also be optimized to obtain more users and higher profits. Customer experience can also be improved through personalization. From the graph, we see that Tuesdays have the least number of bookings for both Uber and Lyft. Customers can be offered promotions and discounts on such days to attract customers to increase the same.



4.3 Insight 3 -Promoting actions with serge rate on different weather conditions

The graph above demonstrates that in different weather conditions, the demand orders number is different. On cloudy days, partly cloudy nights, and partly cloudy days are the top 3 situations in which the order is in high demand. On fog days, clear days and clear nights are the last 3 weather conditions in which the order is in very low demand.



The basic idea behind surge pricing is that during times of high demand, when there are more passengers requesting rides than there are drivers available, prices will temporarily increase. This incentivizes more drivers to come online and pick up passengers, which helps to restore balance to supply and demand.

```
table(data$weather, data$surgeMultiplier)
```

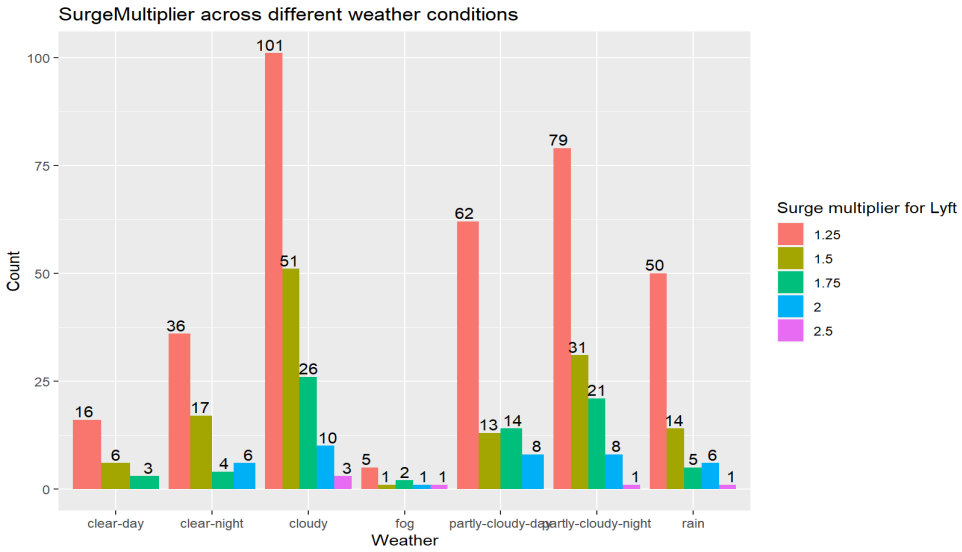
	1	1.25	1.5	1.75	2	2.5
clear-day	667	16	6	3	0	0
clear-night	1686	36	17	4	6	0
cloudy	5813	101	51	26	10	3
fog	247	5	1	2	1	1
partly-cloudy-day	3105	62	13	14	8	0
partly-cloudy-night	4222	79	31	21	8	1
rain	2818	50	14	5	6	1

```
#SurgeMultiplier
table(data$rideshare, data$surgeMultiplier)
```

	1	1.25	1.5	1.75	2	2.5
Lyft	8682	349	133	75	39	6
Uber	9876	0	0	0	0	0

```
#Only Lyft charges at a surge rate
```

We can find that only Lyft has the surge pricing model. They make 6 different kinds (1, 1.25, 1.5, 1.75, 2, 2.5)of surge rates to balance the supply and demand. We can find that in the highest demand weather conditions(cloudy), Lyft uses more surge rates, which increases their revenue.

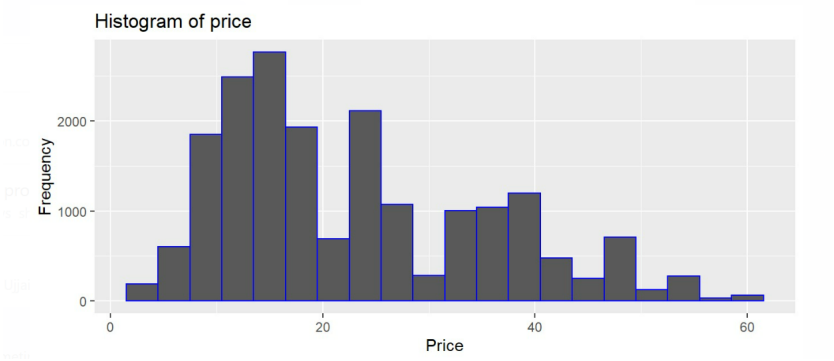


For Uber, Uber doesn’t have a surge rate on rides, so for Uber, they can import surge rates which can help to optimize the balance between affordability for passengers and profitability for the business.

For both Uber and Lyft, they need to use data and analytics to inform surge pricing strategy. Analyzing historical data to identify patterns in demand and supply, and use this information to make informed decisions about when and where to implement surge pricing is very important.

5.MODELING

5.1 - Model 1 with Log- Linear Regression

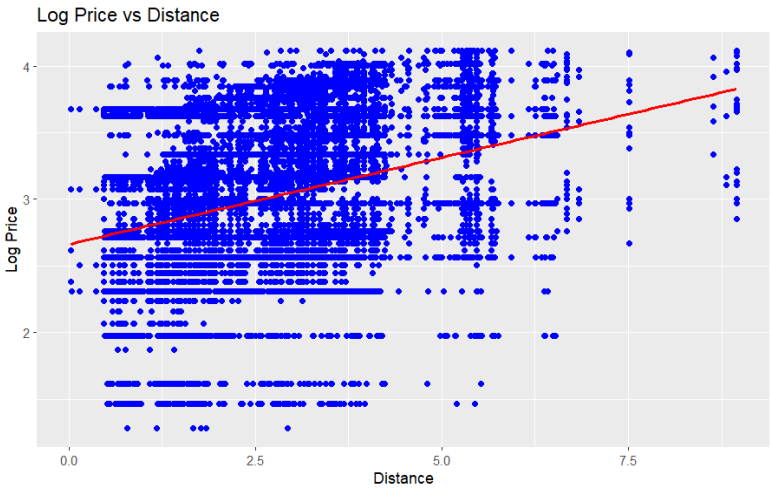


Usually, distance is the main factor governing price. So we begin by regressing the price on distance. On plotting a histogram of price, we observe that price is right skewed.

```
Call:
lm(formula = log(price) ~ distance, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.90945 -0.43460  0.03061  0.40320  1.24988

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.662117   0.008376  317.85   <2e-16 ***
distance    0.130431   0.002909   44.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5304 on 19158 degrees of freedom
Multiple R-squared:  0.09497,   Adjusted R-squared:  0.09493
F-statistic:  2010 on 1 and 19158 DF,  p-value: < 2.2e-16
```

Distance is statistically significant at 0.1% and we can interpret this table as when distance increases by 1 unit, price increases by 13%. But since our model explains very little variation in data, only 9.5 %, we will incorporate more variables to see if we get a better fit.

One more thing that can affect ride fare is the type of vehicle- XL cars have higher fare than normal cars. In our dataset, we have a column “rideCategory” which gives details on what type of car was booked. Since it has categorical values, we create dummy variables for it. This leads to creation of 14 more columns. So in our model we will consider only 13 columns.



With the addition of 13 columns, our model now explains 18% of variation in data. Variables like distance and ride categories of Black SUV, Lux, Lux_Black, Lux_Black_XL, Lyft, Lyft_XL and SHared are statistically significant at 0.1% while ride category Black is significant at 1%.

```
Call:
lm(formula = log(price) ~ distance + rideCategory_Black + rideCategory_Black_SUV +
    rideCategory_Lux + rideCategory_Lux_Black + rideCategory_Lux_Black_XL +
    rideCategory_Lyft + rideCategory_Lyft_XL + rideCategory_Shared +
    rideCategory_Taxi + rideCategory_UberPool + rideCategory_UberX,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.94141 -0.42962  0.00797  0.38039  2.06285

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.669379   0.011911 224.116  < 2e-16 ***
distance                  0.120091   0.002777  43.243  < 2e-16 ***
rideCategory_Black        0.041214   0.013611   3.028 0.002464 **
rideCategory_Black_SUV    0.048319   0.013626   3.546 0.000392 ***
rideCategory_Lux         -1.180233   0.061966 -19.046  < 2e-16 ***
rideCategory_Lux_Black   -1.225518   0.062852 -19.498  < 2e-16 ***
rideCategory_Lux_Black_XL -1.152568   0.061961 -18.601  < 2e-16 ***
rideCategory_Lyft         0.081029   0.011418   7.097 1.32e-12 ***
rideCategory_Lyft_XL     -1.173128   0.062870 -18.660  < 2e-16 ***
rideCategory_Shared      -1.220583   0.062395 -19.562  < 2e-16 ***
rideCategory_Taxi         0.003759   0.016492   0.228 0.819710
rideCategory_UberPool    -0.001155   0.016450  -0.070 0.944006
rideCategory_UberX        0.001524   0.016488   0.092 0.926376
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5044 on 19147 degrees of freedom
Multiple R-squared:  0.1818,    Adjusted R-squared:  0.1813
F-statistic: 354.6 on 12 and 19147 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = log(price) ~ distance + rideCategory_Black + rideCategory_Black_SUV +
    rideCategory_Lux + rideCategory_Lux_Black + rideCategory_Lux_Black_XL +
    rideCategory_Lyft + rideCategory_Lyft_XL + rideCategory_Shared +
    rideCategory_Taxi + rideCategory_UberPool + rideCategory_UberX +
    month_8 + month_9 + month_10 + month_11 + rideshare_Lyft,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.18089 -0.40192 -0.01185  0.33005  1.17361

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.059328   0.038067  54.097  < 2e-16 ***
distance                  0.088262   0.002505  35.238  < 2e-16 ***
rideCategory_Black        0.002346   0.013389   0.175    0.861
rideCategory_Black_SUV    0.008122   0.013386   0.607    0.544
rideCategory_Lux         -0.465955   0.057897  -8.048 8.91e-16 ***
rideCategory_Lux_Black   -0.491791   0.058707  -8.377  < 2e-16 ***
rideCategory_Lux_Black_XL -0.451694   0.057843  -7.809 6.06e-15 ***
rideCategory_Lyft         0.003136   0.015893   0.197    0.844
rideCategory_Lyft_XL     -0.460160   0.058653  -7.845 4.54e-15 ***
rideCategory_Shared      -0.501383   0.058269  -8.605  < 2e-16 ***
rideCategory_Taxi         0.003696   0.014639   0.252    0.801
rideCategory_UberPool     0.003241   0.014602   0.222    0.824
rideCategory_UberX        0.001165   0.014636   0.080    0.937
month_8                   0.014626   0.064651   0.226    0.821
month_9                  -0.017474   0.044985  -0.388    0.698
month_10                 -0.271148   0.039868  -6.801 1.07e-11 ***
month_11                  0.732325   0.037055  19.763  < 2e-16 ***
rideshare_Lyft            0.130011   0.012255  10.609  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4478 on 19142 degrees of freedom
Multiple R-squared:  0.3555,    Adjusted R-squared:  0.355
F-statistic: 621.2 on 17 and 19142 DF,  p-value: < 2.2e-16
```

Now, we add dummies of two more columns- month and rideShare to see how they contribute to model.

With the addition of more columns, model now explains 35.5% variation in data. Also we see that columns statistically significant at 0.1% are rideCatergories of Lux, Lux_Black, Lux_Black_XL, Lyft_XL, Shared, rideshare type Lyft and months of October and November.

To evaluate how good our model is, we make use of metrics like MAE, MSE, and MAPE. The values which we get are 0.37, 0.20 and 0.12 respectively. The AIC and BIC values are 23603 and 23753 respectively.

```
Call:
lm(formula = log(price) ~ distance + rideCategory_Lux + rideCategory_Lux_Black +
    rideCategory_Lux_Black_XL + rideCategory_Lyft_XL + rideCategory_Shared +
    month_10 + month_11 + rideshare_Lyft, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.18132 -0.40133 -0.01154  0.33053  1.17358

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.054067   0.020680  99.325  < 2e-16 ***
distance                   0.088229   0.002504  35.239  < 2e-16 ***
rideCategory_Lux          -0.469629   0.056111  -8.370  < 2e-16 ***
rideCategory_Lux_Black    -0.495457   0.056946  -8.700  < 2e-16 ***
rideCategory_Lux_Black_XL -0.455367   0.056056  -8.123 4.80e-16 ***
rideCategory_Lyft_XL      -0.463832   0.056890  -8.153 3.76e-16 ***
rideCategory_Shared       -0.505051   0.056495  -8.940  < 2e-16 ***
month_10                  -0.263166   0.024917 -10.562  < 2e-16 ***
month_11                   0.740364   0.020083  36.866  < 2e-16 ***
rideshare_Lyft             0.131012   0.006571  19.939  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4477 on 19150 degrees of freedom
Multiple R-squared:  0.3555,    Adjusted R-squared:  0.3552
F-statistic:  1174 on 9 and 19150 DF,  p-value: < 2.2e-16
```

If we remove all statistically insignificant variables, our R^2 value doesn't change much. But our new AIC and BIC values become 23588 and 23675 which is slightly lower than previous values.

5.2 Modeling - Model 2 with Predictive Logit

```
Call:
glm(formula = U_L ~ price + distance, family = binomial(link = "logit"),
    data = train_data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.367  -1.208   1.074   1.132   1.352

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.221819   0.044199   5.019 5.20e-07 ***
price       -0.011629   0.001465  -7.940 2.02e-15 ***
distance     0.045577   0.013827   3.296  0.00098 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18577  on 13411  degrees of freedom
Residual deviance: 18513  on 13409  degrees of freedom
AIC: 18519

Number of Fisher Scoring iterations: 3
```

Summary of a Logistic Regression:

In order to have a better analysis, the dataset is divided into two subsets, training (75%) and testing (25%).

This is the summary of a logistic regression model that predicts whether a ride will be an Uber (U) or a Lyft (L) based on two predictor variables, price and distance. The response variable is binary, with 1 representing an Uber ride and 0 representing a Lyft ride.

Based on the training set logit model, both distance and price are statistically significant.

The coefficients section provides the estimates of the model's parameters. Here are the key inferences:

The coefficient for the (Intercept) is 0.241035, which is the log odds of an Uber ride when both the price and distance are zero.

The coefficient for price is -0.012606, which means that for a one-unit increase in price, the log odds of an Uber ride decrease by 0.012606. This suggests that higher prices reduce the likelihood of people choosing an Uber ride.

The coefficient for distance is 0.044328, which means that for a one-unit increase in distance, the log odds of an Uber ride increase by 0.044328. This suggests that longer distances increase the likelihood of people choosing an Uber ride.

The significance codes indicate the level of significance of each predictor in the model. In this case, both price and distance have a p-value less than 0.05, indicating that they are both significant predictors of the response variable.

The deviance residuals section provides information about the goodness of fit of the model. The null deviance is the deviance of the model that only contains the intercept, while the residual deviance is the deviance of the model with the predictors. A smaller residual deviance indicates a better fit. The AIC (Akaike Information Criterion) is a measure of the model's goodness of fit that also takes into account the number of parameters in the model. A smaller AIC value indicates a better model.
In summary, this logistic regression model suggests that both price and distance are significant predictors of whether a ride will be an Uber or a Lyft. The model indicates that higher prices reduce the likelihood of people choosing an Uber ride, while longer distances increase the likelihood of people choosing an Uber ride.

This model is developed on a training dataset. After training the model, we tested this model on a testing dataset. The accuracy score is 54%. An accuracy of 54% means that the model correctly predicted the class labels for 54% of the samples in the test dataset. This metric is a good overall measure of performance. The Precision score is 30%. A precision of 30% means that 30% of the positive predictions made by the model were actually correct. Precision is a measure of the model's ability to avoid false positive predictions. A low precision score indicates that the model is making a large number of false positive predictions. The recall score is 55%. A recall of 55% means that the model correctly identified 55% of the positive cases in the test dataset. Recall is a measure of the model's ability to find all of the positive cases. A low recall score indicates that the model is missing a significant number of positive cases. The f1-score is 39%. An F1-score of 39% is the harmonic mean of precision and recall, and it provides a balance between the two metrics. The F1-score is often used as a single metric to summarize the performance of a binary classifier, as it provides a balance between precision and recall.

6.CONCLUSION

Based on these findings, the following recommendations can be made to the Lyft and Uber companies to improve their business revenue:

- According to the data, we can guide and deploy transportation capacity in advance according to different types of areas (such as schools and stations) according to the demand in different time periods. Improve the user experience and obtain higher profits by improving the page or redeploying specific high-end cars to areas with high demand for high-end cars. According to the weather forecast, combined with the demand for different types of cars in various regions in the past weather, it will be reasonably allocated in advance.
- To increase profitability, both Uber and Lyft can leverage data analysis to identify the highest demand hours during the day and also on different days of the week. By deploying more vehicles and drivers during peak hours, they can meet the increased demand and increase their revenue. Additionally, they can also use dynamic pricing to increase fares during peak hours, which can further boost their profits.
- To improve business revenue, both Lyft and Uber companies could focus on providing a better customer experience based on hourly and weekly demand analysis. For example, they can use this analysis to optimize vehicle availability and deployment, ensuring that there are enough vehicles available in areas where demand is high. They can also use this information to target promotions and discounts at specific times and days of the week when demand is lower.
- Moreover, understanding the customer preferences such as preferred mode of transportation, preferred destination, etc., can help both companies in optimizing their services and offering a better customer experience. For example, if customers prefer to use rideshare services for long-distance trips, Lyft and Uber can invest in improving the quality of their long-distance services, such as providing more comfortable vehicles, improving navigation systems, etc.
- Based on the data and insights, we can use regression analysis to model the relationship between the various factors affecting the cost of a ride, such as the weather, distance, time of day, location, etc. This model can be used to predict the cost of a ride and understand the impact of different factors on the cost. Additionally, this model can be used to optimize pricing strategies, such as dynamic pricing, to increase revenue.
- The rideshare market has significant room for growth, as the contribution of these services to the total vehicle miles traveled is relatively low. Both Uber and Lyft can increase their profitability by focusing on providing a better customer experience and

optimizing their services based on hourly and weekly demand analysis. By using data analysis and regression modeling, both companies can better understand the factors affecting the cost of a ride and use this information to make informed decisions about pricing, vehicle deployment, and customer engagement.