Olivia_Lee_CS555_Term_Project

#SECTION 1: DATA PREPARATION AND DATA CLEANING

#Part 1. Loading packages used in project

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.4.0      ✔ purrr   0.3.5
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.5.0
## ✔ readr   2.1.3      ✔ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

#Part 2. Reading CSV file

This data set from Kaggle contains data from Uber and Lyft rides within Boston from November 2018 to December 2018.

```
#data = read.csv(file = 'rideshare_kaggle.csv') #Reading csv file downloaded from Kaggle
#data %>% head()

#I commented these lines out as only the final cleaned data set will be included in submission and used for analy
sis.
```

#Part 3. Data Cleaning

a. Filtering two conditions:

1. I filtered cab_type to be the most commonly used Uber rides, which is UberX, to eliminate more expensive options like UberBlack or UberSUV. I have also done the same with Lyft rides and picking Lyft's equivalent of UberX, which is just Lyft. This is to have a fair playing field for the price as they are all the same type of ride.
2. Using drop_na() to drop any NA values in the data set.

b. Selecting columns

I have chosen the following columns to use in my data set: i. name ii. price ii. distance

c. Additional columns

I have added the following columns to the data set: i. log_price: This is the log10 transformation of price. I did this to normalise the price distribution as the price variable is greatly right skewed. ii. far_distance: This is a binary variable that indicates whether this ride was a relatively long distance. To indicate that the ride was long, it takes the value 1 if the distance was more than 5 and 0 otherwise.

Hence, the following columns are used in the final data set: i. name ii. price ii. distance iv. log_price v. far_distance

```
# data_cleaned = data %>% filter(name == 'UberX' | name == "Lyft") %>% select(name, price, distance) %>% mutate(l
og_price = log10(price), far_distance = as.integer(distance > 5)) %>% drop_na() #Filtering data
#
# #Taking 250 samples of each cab type so we can have an equal amount of data for Uber and Lyft for analysis. I a
m taking 250 samples each because there is a 500 sample limit for this assignment.
# sample_uber = data_cleaned %>% filter(name == "UberX") %>% sample_n(250, replace = FALSE)
# sample_lyft = data_cleaned %>% filter(name == "Lyft") %>% sample_n(250, replace = FALSE)
#
# #Combining the two samples together in one data frame from analysis.
# uber_lyft_sample = rbind(sample_uber, sample_lyft)
#
#write.csv(sample_uber, "sample_uber.csv") #Saving sample as csv so it does not randomise the sample again
#write.csv(sample_lyft, "sample_lyft.csv") #Saving sample as csv so it does not randomise the sample again
# write.csv(uber_lyft_sample, "uber_lyft.csv") #Saving final data set to csv file. This will be the final cleaned
data set that I will use for analysis and that will be included in my submission.

#I commented the lines above so it does not take another random sample and change data set used for analysis.

sample_uber = read.csv(file = 'sample_uber.csv')
sample_lyft = read.csv(file = 'sample_lyft.csv')
df = read.csv(file = 'uber_lyft.csv') #Reading csv file
df %>% head()
```

```
##   X  name price distance log_price far_distance
## 1 1 UberX   9.5     2.04 0.9777236            0
## 2 2 UberX   7.0     0.39 0.8450980            0
## 3 3 UberX  12.5     3.07 1.0969100            0
## 4 4 UberX  16.0     6.91 1.2041200            1
## 5 5 UberX   8.5     2.17 0.9294189            0
## 6 6 UberX   8.0     1.22 0.9030900            0
```

c. Identifying outliers

Using the IQR method, I checked for outliers in the numerical columns in the data set which are price and distance. Firstly, I created a function to identify outliers for a given column in the dataframe to avoid repetitiveness and keep consistency.

```
#Function to identidy outliers using the IQR method
outliers = function(df, col_name){ #the input will be the dataframe and the column that we want to find the outli
ers of
  Q1 = quantile(col_name, 0.25) #1st Quartile of Data
  Q3 = quantile(col_name, 0.75) #3rd Quartile of Data
  IQR = Q3 - Q1 #Interquartile Range
  min_outliers = Q1 - (1.5*IQR) #anything below the 'minimum' is an outlier
  max_outliers = Q3 + (1.5*IQR) #anything above the 'maximum' is an outlier
  outliers = df[col_name > max_outliers | col_name < min_outliers, ]  #Filtering data to get the outliers in the
data
  return(outliers) #Returns the dataframe with outliers of the given column
}
```

Then, I used this function for the numerical variables to identify outliers.

i. Price

There are maximum outliers but no minimum outliers. After analyzing the outliers, I have decided to keep all the data regardless of outliers, as I do not think that they are mistakes in the data and should be included in analysis.

```
outliers(df, df$price)
```

```
##       X  name price distance log_price far_distance
## 43   43 UberX  17.0     4.72  1.230449            0
## 95   95 UberX  18.5     4.49  1.267172            0
## 188 188 UberX  24.0     5.70  1.380211            1
## 212 212 UberX  18.0     1.80  1.255273            0
## 215 215 UberX  17.0     2.32  1.230449            0
## 227 227 UberX  17.5     3.25  1.243038            0
## 245 245 UberX  17.5     2.62  1.243038            0
## 281 281  Lyft  16.5     5.41  1.217484            1
## 405 405  Lyft  22.5     3.93  1.352183            0
## 408 408  Lyft  22.5     2.90  1.352183            0
## 411 411  Lyft  22.5     3.23  1.352183            0
## 428 428  Lyft  16.5     3.37  1.217484            0
## 430 430  Lyft  19.5     4.68  1.290035            0
## 431 431  Lyft  16.5     3.94  1.217484            0
## 458 458  Lyft  19.5     2.95  1.290035            0
```

ii. Distance

There are maximum outliers but no minimum outliers. After analyzing the outliers, I have decided to keep all the data regardless of outliers, as I do not think that they are mistakes in the data and should be included in analysis.

```
outliers(df, df$distance)
```

```
##       X  name price distance log_price far_distance
## 4     4 UberX  16.0     6.91  1.204120            1
## 109 109 UberX  14.0     5.56  1.146128            1
## 124 124 UberX  16.0     5.56  1.204120            1
## 188 188 UberX  24.0     5.70  1.380211            1
## 281 281  Lyft  16.5     5.41  1.217484            1
## 293 293  Lyft  13.5     5.43  1.130334            1
```

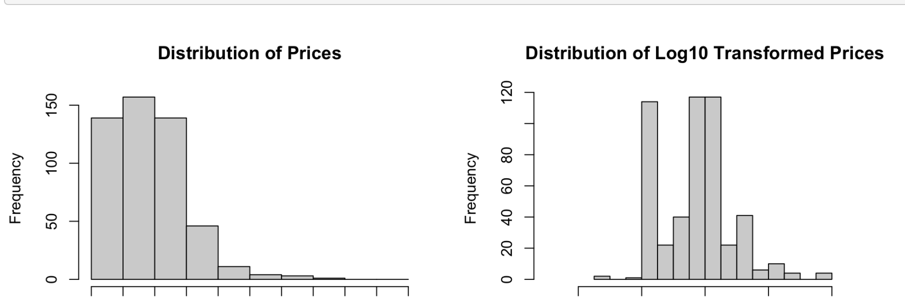
#SECTION 2: DATA VISUALIZATION

#Part 1. Distribution of data

From the histogram and boxplot below, we can observe that the distribution of prices is right-skewed. After the log10 transformation, prices are fairly normally distributed, with a slight skew to the left. We will use the normalised log10 transformation of price in our hypothesis testing.

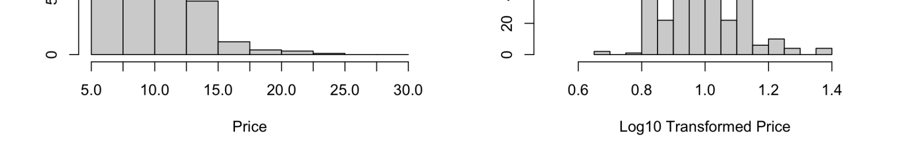
a. Histogram of Price

```
par(mfrow=c(1,2))
hist1 = hist(df$price, main = "Distribution of Prices", xlab = "Price", ylab = "Frequency", breaks = seq(5, 30,
2.5), xlim = c(5, 30), xaxp=c(5, 30, 10))
hist2 = hist(df$log_price, main = "Distribution of Log10 Transformed Prices", xlab = "Log10 Transformed Price", y
lab = "Frequency", xlim = c(0.5, 1.5))
```



b. Boxplot

```
par(mfrow=c(1,2))
boxplot(df$price, main = "Boxplot of Prices", ylab = "Price")
boxplot(df$log_price, main = "Boxplot of Log10 Transformed Prices", ylab = "Log10 Transformed Price")
```



#SECTION 3: DATA ANALYSIS

Research question: In Boston, are Uber prices higher than Lyft prices?

#Part 1. Summary of the data by group

a. Uber

```
cat("Summary of Uber data:\n\n")
```

```
## Summary of Uber data:
```

```
summary(sample_uber)
```

```
##        X              name               price           distance    
##  Min.   :  1.00   Length:250         Min.   : 6.000   Min.   :0.040  
##  1st Qu.: 63.25   Class :character   1st Qu.: 8.000   1st Qu.:1.250  
##  Median :125.50   Mode  :character   Median : 9.500   Median :2.140  
##  Mean   :125.50                      Mean   : 9.852   Mean   :2.138  
##  3rd Qu.:187.75                      3rd Qu.:10.875   3rd Qu.:2.815  
##  Max.   :250.00                      Max.   :24.000   Max.   :6.910  
##    log_price       far_distance 
##  Min.   :0.7782   Min.   :0.00  
##  1st Qu.:0.9031   1st Qu.:0.00  
##  Median :0.9777   Median :0.00  
##  Mean   :0.9797   Mean   :0.02  
##  3rd Qu.:1.0363   3rd Qu.:0.00  
##  Max.   :1.3802   Max.   :1.00  
```

b. Lyft

```
cat("Summary of Lyft data:\n\n")
```

```
## Summary of Lyft data:
```

```
summary(sample_lyft)
```

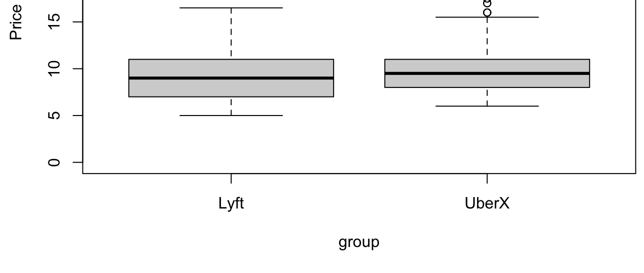
```
##        X              name               price           distance    
##  Min.   :  1.00   Length:250         Min.   : 5.000   Min.   :0.440  
##  1st Qu.: 63.25   Class :character   1st Qu.: 7.000   1st Qu.:1.250  
##  Median :125.50   Mode  :character   Median : 9.000   Median :2.150  
##  Mean   :125.50                      Mean   : 9.638   Mean   :2.162  
##  3rd Qu.:187.75                      3rd Qu.:11.000   3rd Qu.:2.940  
##  Max.   :250.00                      Max.   :22.500   Max.   :5.430  
##    log_price       far_distance  
##  Min.   :0.6990   Min.   :0.000  
##  1st Qu.:0.8451   1st Qu.:0.000  
##  Median :0.9542   Median :0.000  
##  Mean   :0.9694   Mean   :0.008  
##  3rd Qu.:1.0414   3rd Qu.:0.000  
##  Max.   :1.3522   Max.   :1.000  
```

c. Distribution of Prices by Group + Variability

It appears that variability between groups is small relative to the variability in the measurements within groups. This indicates that we are less inclined to conclude that there is a difference between Uber and Lyft prices.

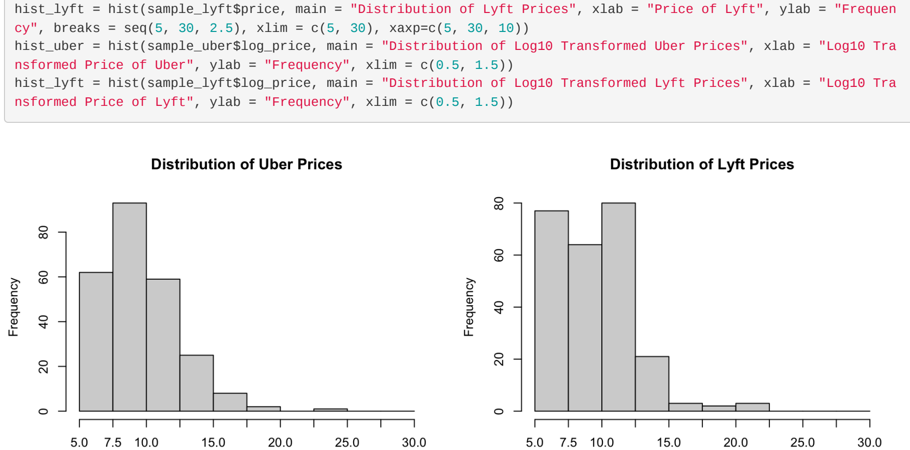
i. Boxplot

```
boxplot(df$price~df$name, main = "Price by Cab Type", xlab = "group", ylab = "Price", ylim = c(0, 30))
```



ii. Histogram

```
par(mfrow=c(2,2))
hist_uber = hist(sample_uber$price, main = "Distribution of Uber Prices", xlab = "Price of Uber", ylab = "Frequen
cy", breaks = seq(5, 30, 2.5), xlim = c(5, 30), xaxp=c(5, 30, 10))
hist_lyft = hist(sample_lyft$price, main = "Distribution of Lyft Prices", xlab = "Price of Lyft", ylab = "Frequen
cy", breaks = seq(5, 30, 2.5), xlim = c(5, 30), xaxp=c(5, 30, 10))
hist_uber = hist(sample_uber$log_price, main = "Distribution of Log10 Transformed Uber Prices", xlab = "Log10 Tra
nsformed Price of Uber", ylab = "Frequency", xlim = c(0.5, 1.5))
hist_lyft = hist(sample_lyft$log_price, main = "Distribution of Log10 Transformed Lyft Prices", xlab = "Log10 Tra
nsformed Price of Lyft", ylab = "Frequency", xlim = c(0.5, 1.5))
```

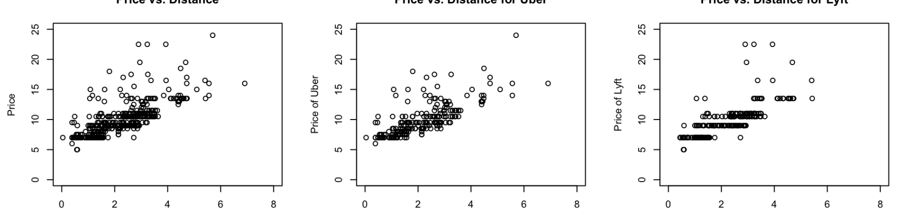


d. Correlation between price and distance

Pearson correlation coefficient between price and distance: 0.7738999 Pearson correlation coefficient between price and distance for Uber: 0.7301435 Pearson correlation coefficient between price and distance for Lyft: 0.8168339

Both Uber and Lyft rides have a strong positive association between price and distance. Since Lyft rides have a higher correlation coefficient than Uber, Lyft prices are more strongly correlated with distance than Uber prices.

```
par(mfrow=c(1,3))
cor = plot(df$distance, df$price, main = "Price vs. Distance", xlab = "Distance", ylab = "Price", xlim = c(0,8),
ylim = c(0, 25))
cor_uber = plot(sample_uber$distance, sample_uber$price, main = "Price vs. Distance for Uber", xlab = "Distance",
ylab = "Price of Uber", xlim = c(0,8), ylim = c(0, 25))
cor_lyft = plot(sample_lyft$distance, sample_lyft$price, main = "Price vs. Distance for Lyft", xlab = "Distance",
ylab = "Price of Lyft", xlim = c(0,8), ylim = c(0, 25))
```




```
r = cor(df$distance, df$price) #Function to get Pearson correlation coefficient
r_uber = cor(sample_uber$distance, sample_uber$price)
r_lyft = cor(sample_lyft$distance, sample_lyft$price)
cat("Pearson correlation coefficient between price and distance:", r)
```

```
## Pearson correlation coefficient between price and distance: 0.7462789
```

```
cat("\nPearson correlation coefficient between price and distance for Uber:", r_uber)
```

```
##
## Pearson correlation coefficient between price and distance for Uber: 0.7381263
```

```
cat("\nPearson correlation coefficient between price and distance for Lyft:", r_lyft)
```

```
##
## Pearson correlation coefficient between price and distance for Lyft: 0.7585541
```

#Part 2. Hypothesis testing for difference in means between Uber and Lyft prices

I will use the two sample t-test to determine whether Uber prices are more expensive than Lyft prices.

Assumptions of Two Sample t-test

a. Independence
   - This assumption is met.
   - Since the data is collected from two different companies, the samples collected from each company is independent.
b. Same measurement
   - This assumption is met.
   - Since we measuring price, they are measured in the same way.
c. Similar distributions.
   - This assumption is met.
   - Looking at the boxplot and histograms above of both Uber and Lyft prices, we can determine that they both have similar distributions.

Performing two sample t-test using the 5 step hypotheses testing procedure:

Step 1: Setting up the hypotheses and setting the alpha level H0 : mu_uber = mu_lyft (the means of both Uber and Lyft prices are the same) H1 : mu_uber > mu_lyft (the mean price of Uber is greater than the mean price of Lyft) α = 0.05

Step 2: Selecting the appropriate test statistic We will use the t-statistic

Step 3: State decision rule Critical value from the standard t-distribution with df = 250-1 = 249 degrees of freedom and associated with α = 0.05. Decision Rule: Reject H0 if t ≥ 1.650996. Otherwise, do not reject H0.

```
cat("Critical value:", qt(.95, df = 249))
```

```
## Critical value: 1.650996
```

Step 4: Compute the test t-statistic and the associated p-value

```
t.test(sample_uber$log_price, sample_lyft$log_price, alternative = "greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  sample_uber$log_price and sample_lyft$log_price
## t = 1.0651, df = 497.62, p-value = 0.1437
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.005647265          Inf
## sample estimates:
## mean of x mean of y
## 0.9796844 0.9693631
```

Step 5: Conclusion Since the t-statistic = 1.0651 < critical value = 1.650996, we fail to reject the null hypothesis. Hence, we do not have significant evidence at the α = 0.05 level to conclude that Uber prices are higher than Lyft prices.

#Part 3. Hypothesis testing for difference in population means between Uber and Lyft prices, adjusting for distance

Since there is a strong correlation between price and distance, we will test for difference in population means between Uber and Lyft while adjusting for distance.

The assumptions for ANCOVA will be the assumptions for both One-Way ANOVA and Linear Regression.

Assumptions of One-Way ANOVA i. Each sample is an independent random sample. - This assumption is met. - Since the data is collected from two different companies, the samples collected from each company is independent. ii. Distribution of the response variable follows a normal distribution. - This assumption is met. - The log10 transformed prices are normally distributed and we will be using it for hypothesis testing. iii. Each group has equal population variance for the response variable. - This assumption is met. - Rule of thumb: The largest sample variance divided by the smallest sample variance is not greater than two. - As seen in the code below, largest sample variance divided by smallest sample variance: 1.05664 < 2.

```
var_uber = var(sample_uber$log_price) #Variance of Uber Prices
var_lyft = var(sample_lyft$log_price) #Variance of Lyft Prices
cat("Variance of Uber prices:", var_uber)
```

```
## Variance of Uber prices: 0.01141416
```

```
cat("\nVariance of Lyft prices:", var_lyft)
```

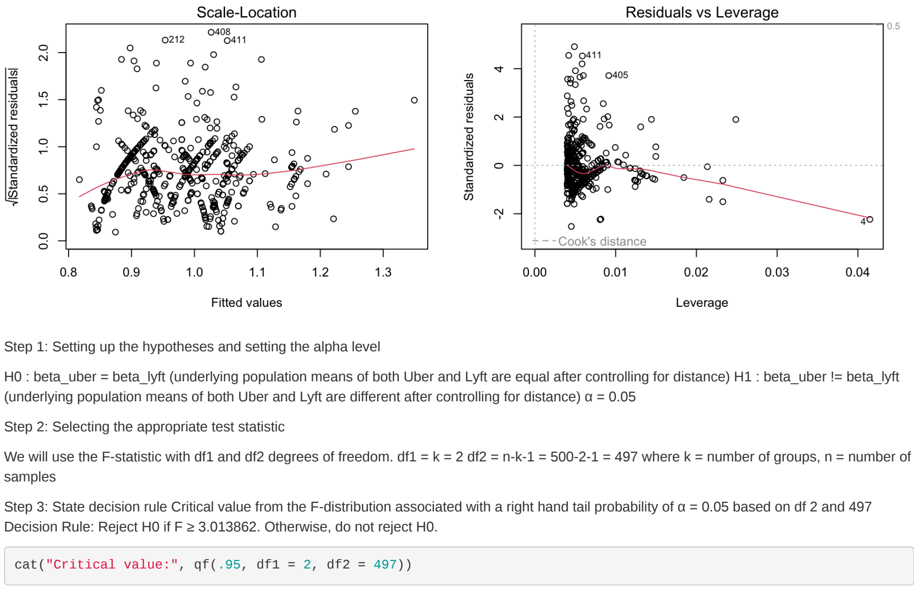
```
##
## Variance of Lyft prices: 0.01206066
```

```
div = var_lyft/var_uber #Largest sample variance divided by smallest sample variance
cat("\nLargest sample variance divided by smallest sample variance:", div, "< 2. Hence, the equal population vari
ance for each group assumption is met.")
```

```
##
## Largest sample variance divided by smallest sample variance: 1.05664 < 2. Hence, the equal population variance
for each group assumption is met.
```

Assumptions of Linear Regression i. The true relationship is linear. - This assumption is met. - Since there is a strong positive linear correlation between price and distance, we can determine that there is a linear relationship. ii. The observations are independent. - This assumption is met. - We can observe from the Residuals vs. Fitted graph that the residuals do not depend on the fitted values. iii. The variation of the response variable around the regression line is constant. - This assumption is not met. - We can see from the Scale-Location graph below that the variance is not constant. iv. The residuals are normally distributed. - This assumption is met. - We can see from the Normal Q-Q graph below that the residuals are fairly normally distributed.

```
par(mfrow=c(2,2))
m2 = lm(data = df, log_price ~ name + distance) #Multiple Linear Regression model, predicting log_price from name
and distance
plot(m2)
```



Step 1: Setting up the hypotheses and setting the alpha level

H0 : beta_uber = beta_lyft (underlying population means of both Uber and Lyft are equal after controlling for distance) H1 : beta_uber != beta_lyft (underlying population means of both Uber and Lyft are different after controlling for distance) α = 0.05

Step 2: Selecting the appropriate test statistic

We will use the F-statistic with df1 and df2 degrees of freedom. df1 = k = 2 df2 = n-k-1 = 500-2-1 = 497 where k = number of groups, n = number of samples

Step 3: State decision rule Critical value from the F-distribution associated with a right hand tail probability of α = 0.05 based on df 2 and 497 Decision Rule: Reject H0 if F ≥ 3.013862. Otherwise, do not reject H0.

```
cat("Critical value:", qf(.95, df1 = 2, df2 = 497))
```

```
## Critical value: 3.013862
```

Step 4: Compute the test statistic and the associated p-value

```
summary(m2)
```

```
##
## Call:
## lm(formula = log_price ~ name + distance, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16753 -0.04083 -0.01233  0.02708  0.32560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.801869   0.007187  111.57   <2e-16 ***
## nameUberX   0.012187   0.005946    2.05   0.0409 *
## distance    0.077486   0.002697   28.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06648 on 497 degrees of freedom
## Multiple R-squared:  0.6251, Adjusted R-squared:  0.6236
## F-statistic: 414.4 on 2 and 497 DF,  p-value: < 2.2e-16
```

Step 5: Conclusion Since the F-statistic = 414.4 > critical value = 3.013862, we reject the null hypothesis. Hence, there is sufficient evidence to conclude that the underlying population means of both Uber and Lyft are different after controlling for distance at the α = 0.05 level.

#Interpretations

a. Least squares regression line log_price = 0.801869 + (0.012187 x UberX) + (0.077486 x distance) Hence, price = 10^log_price = 10^(0.801869 + (0.012187 x UberX) + (0.077486 x distance))

b. Beta Estimate Since the p-value of nameUberX = 0.0409 < α = 0.05, we can conclude that the variable "name" is a predictor in the output of the prices. Since Uber is the reference group, there is a mean difference of 0.012187 increase in log_price, which is an equivalent of a 10^0.012187 = 1.028459 increase in price, if you order an Uber instead of a Lyft, when controlling for distance.

c. R-squared Given that the R-squared of the model is 0.6236, this means that 62.36% of the variation in price can be explained by the cab type and distance.

d. Confidence Interval

```
confint(m2, level = 0.95) #Finding confidence interval
```

```
##                    2.5 %     97.5 %
## (Intercept) 0.7877479277 0.81598909
## nameUberX   0.0005044658 0.02387000
## distance    0.0721883059 0.08278445
```

After controlling for distance, the confidence interval of the beta estimate for Uber variable is (0.0005044658, 0.02387000), which is in log_price. When transforming it back to price, the confidence interval is (1.001162, 1.056501). Hence, we can say with 95% confidence that the true increase in Uber prices compared in Lyft prices is between (1.001162, 1.056501), adjusting for distance.