MET CS 555 Term Project

10 points **(Due December 18, 2022 at 11:59 PM)**

1. Assignment Description

Find a dataset for a research problem of interest, here are some good websites for this

Kaggle Data Science Competitions: http://kaggle.com
UCI Machine Learning repository: http://archive.ics.uci.edu/ml/datasets.php
Google Cloud public data: https://console.cloud.google.com/marketplace/browse?filter=solution-type:dataset&_ga=2.265202890.490000482.1586060190-1118401016.1586060190&pli=1

Describe a research scenario and specify a research question based on data analytic methods that we learned in class, for example methods like, *one and two sample means, t-test, correlation tests, simple and multiple linear regression, ANOVA and ANCOVA, one and two-Sample Tests for Proportions and logistic regression*.

Clean up your data and reduce it to no more than 500 observations if your data set is large. If your task is to build a prediction model, ensure that you split the data into **training and testing datasets at this stage.**

2. Research Scenario Description (no more than 200 words)

Describe your research scenario in no more than 200 words. This is a general description of the use case. Give relevant background information so that the reader can understand your research question. Assume that the reader does not have any specialized knowledge about your topic. Define all acronyms/abbreviations and try to avoid jargon.

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

Uber and Lyft are two competing rideshare companies that provide taxi services to anyone that orders a ride on the apps. For both Uber and Lyft, they provide different types of taxi services. For example, you can order a luxury car or an SUV at a higher price than an economy level ride.

For Uber, their most commonly used type of ride is called “UberX”, which is the basic economy level ride. For Lyft, their most commonly used type of ride is called “Lyft”, which is the Lyft equivalent of “UberX”.

3. Describe the data set (no more than 400 words)

Describe each of the columns of the data set **that are used in your analysis**. Clean up your data before usage (e.g. asses and remove outliers, perform any necessary transformations, ensure assumptions of your analysis are met). Remove unused columns. Describe each step of your data cleaning process and why you did this step. If possible, provide a link to the main data set source.

Link to main data set: https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma

This dataset from Kaggle contains data from Uber and Lyft rides within Boston from November 2018 to December 2018. The original data set has 57 columns and 693,071 tuples. The tuples in the data set are the Uber and Lyft rides people took in Boston. Some examples of the columns in the original data set are timestamp, hour, day, month, price, cab_type, source, destination and temperature.

Data Cleaning

1. Filtering conditions
    a. Condition 1: Since we are only comparing the most commonly used types of Uber and Lyft rides, we filter out all other types of rides. These types of rides are defined in the column cab_type in the data set. For Uber, the most commonly used ride would be 'UberX', and for Lyft it would be just 'Lyft'.
    b. Condition 2: Filtering out rows that have NAs. Although this particular data set did not have any NA values, I still used this for precaution and good practice.

2. Selecting columns
    a. Since my research question is comparing Uber and Lyft prices, I will choose the following columns.
        i. name: This is the type of cab, which is either Uber or Lyft. We need to keep this column to identify whether the datapoint is an Uber or Lyft ride.
        ii. price: This is the price of that particular ride. We need to keep this column to compare prices.
        iii. distance: This is the distance of that particular ride. We keep this column in order to adjust for this variable in analysis.
3. Transformation + Additional columns
    a. 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.
    b. Hence, the following columns are used in the final data set:
        i. name
        ii. price
        ii. distance
        iv. log_price
        v. far_distance
4. Reducing tuples in data set
    a. Since there is a 500-row limit for our project, I took 250 samples from Uber and 250 samples from Lyft. These samples are chosen randomly.

5. Identifying and dealing with outliers
    a. Using the IQR method, I checked for outliers in the numerical columns in the dataset, which are price and distance.
        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.
        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.

4. Research Question (no more than 100 words)

Describe the main research question in one or two sentences. This is similar to the last sentence of our class examples.

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

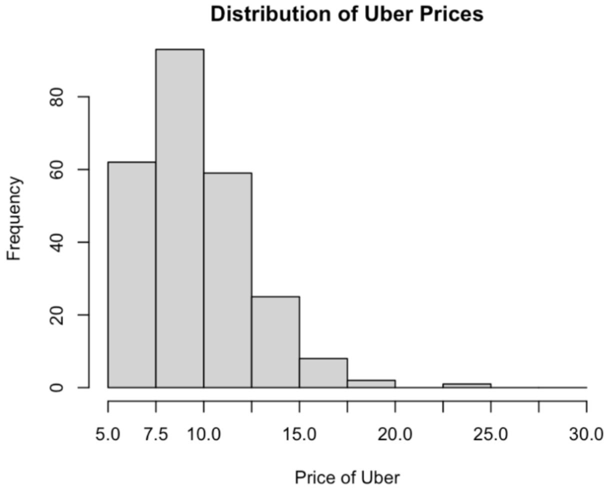
We compare the two commonly used types of rides from both Uber and Lyft, which is the basic type that excludes luxury or SUV rides. This research question will help us answer whether the Uber rides are more expensive than Lyft rides.

5. Statistical Analysis

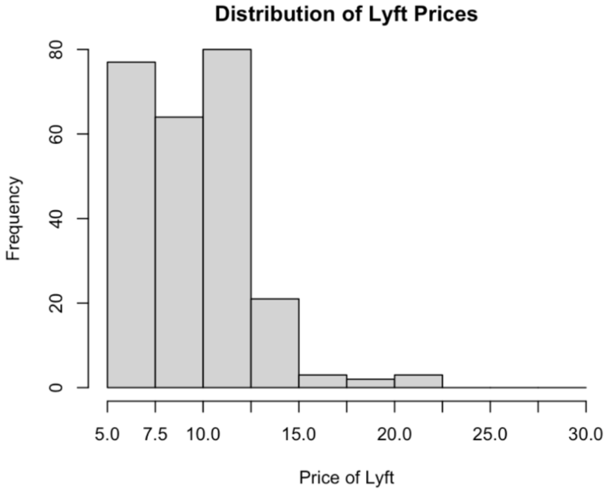
Give at least one main visualization that supports your conclusion. Be sure to use correct axis labels and avoid common mistakes when generating visualizations. **State all assumptions of the statistical technique(s) that you are using and give evidence as to whether or not these assumptions are met.**

Hypothesis testing

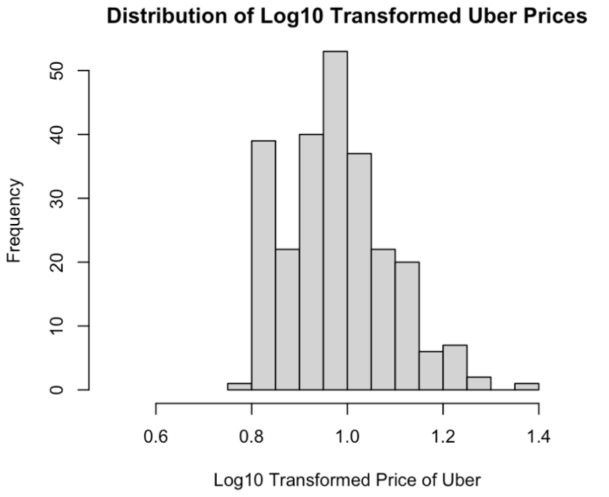
1. Two sample t-test
   a. Assumptions
      i. Independence
         - This assumption is met.
         - Since the data is collected from two different companies, the samples collected from each company is independent.
      ii. Same measurement
         - This assumption is met.
         - Since we measuring price, they are measured in the same way.
      iii. Similar distributions.
         -This assumption is met.
         - Looking at the boxplot and histograms below of both Uber and Lyft prices, we can determine that they both have similar distributions.



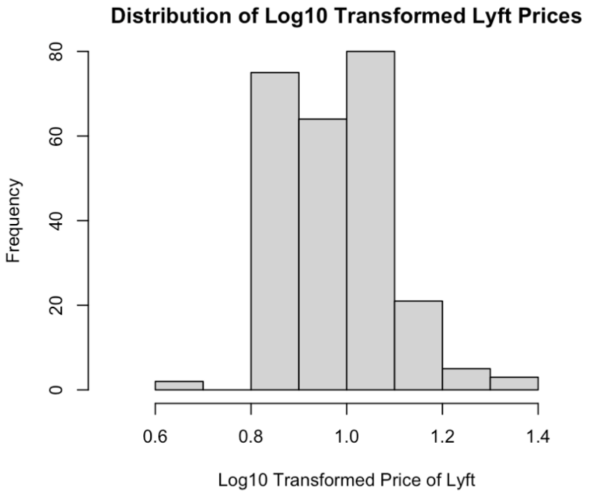
Distribution of Uber Prices



Distribution of Lyft Prices



Distribution of Log10 Transformed Uber Prices



Distribution of Log10 Transformed Lyft Prices

b. 5-step hypothesis testing procedure for two sample t-test

i. 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)
$\alpha = 0.05$

ii. Step 2: Selecting the appropriate test statistic

We will use the t-statistic as the test statistic

iii. Step 3: State decision rule

Critical value from the standard t-distribution with df = 250-1 = 249 degrees of freedom and associated with $\alpha = 0.05$.
Decision Rule: Reject H0 if $t \geq 1.650996$. Otherwise, do not reject H0.

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

Test t-statistic: 1.0651
p-value: 0.1437

v. 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 $\alpha = 0.05$ level to conclude that Uber prices are higher than Lyft prices.

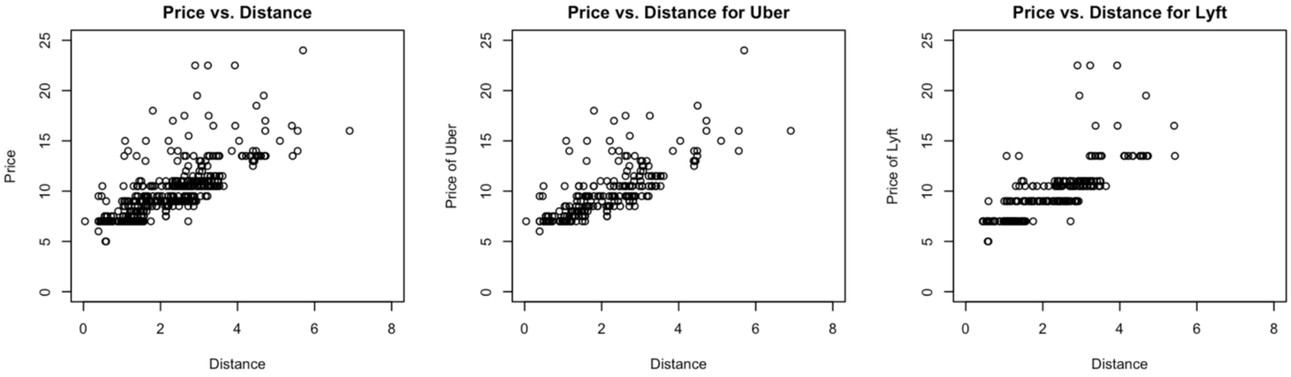
2. ANCOVA adjusting for distance

   a. Covariates

      - The covariate for in our data is distance. As seen in the correlation scatterplot below between price and distance, we can observe that distance has an effect on the dependent variable, price.

   b. Correlation coefficients

      - 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.



   c. Assumptions of ANCOVA

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

      Assumptions for 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.
      - The largest sample variance divided by the smallest sample variance = 1.05664 < 2.

      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 below that the residuals do not depend on

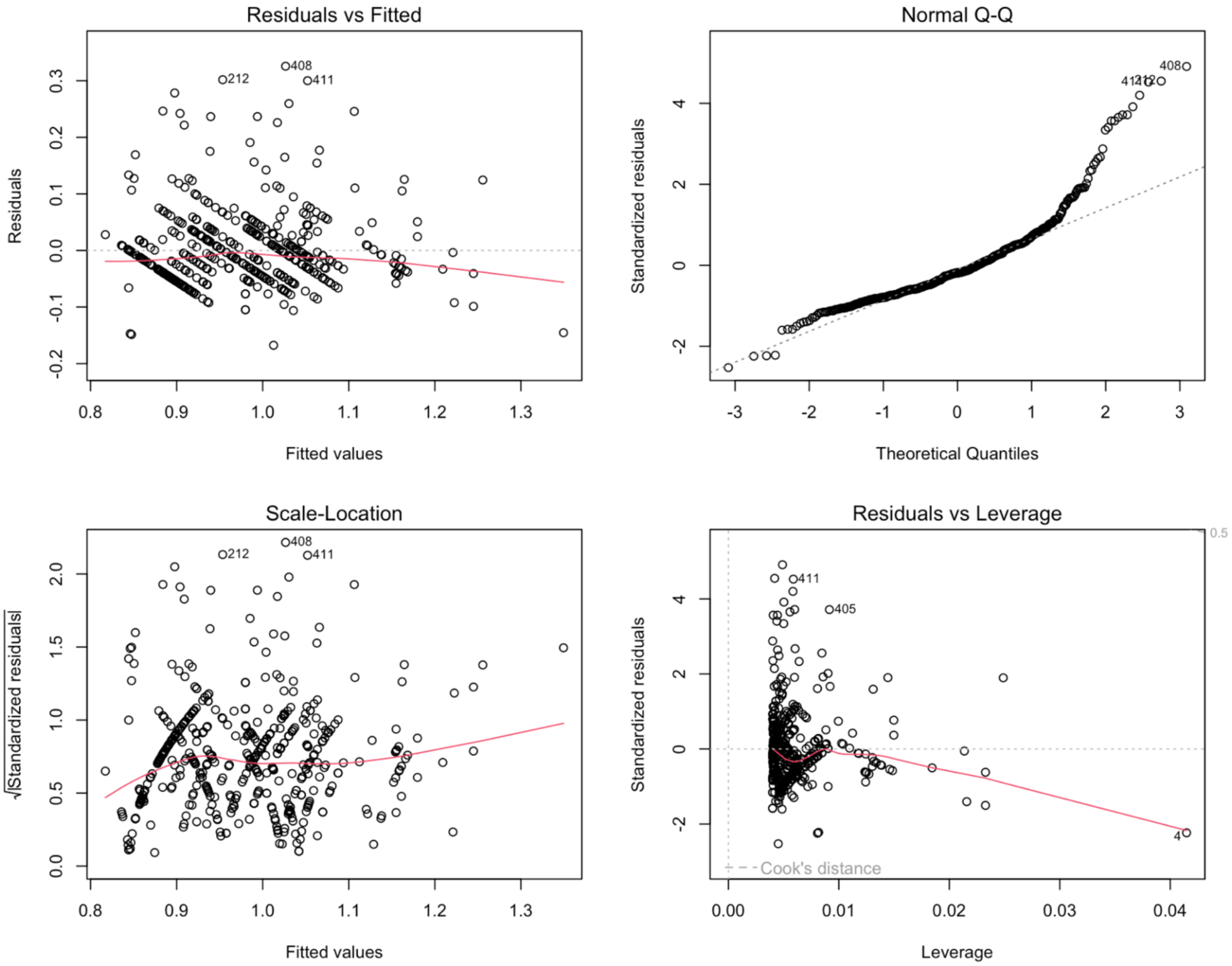
the fitted values.

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.



d. 5-step hypothesis testing procecure for ANCOVA, adjusting for distance

i. 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

ii. 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

iii. 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.

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

Test F-statistic: 414.4

p-value: <2.2e-16

v. 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.

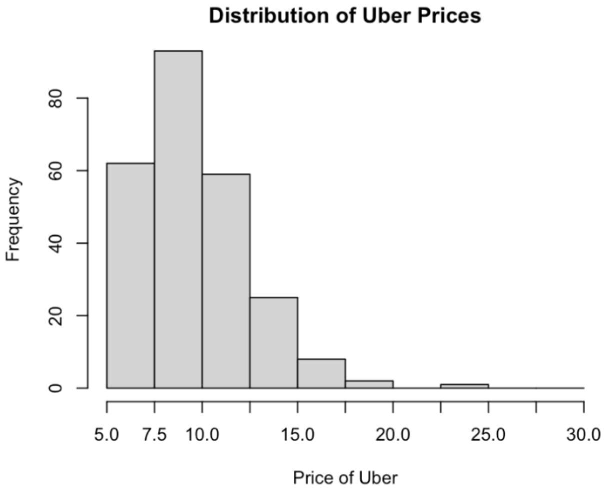
3. Interpretations
    a. Least squares regression line
        - log_price = 0.801869 + (0.012187 x UberX) + (0.077486 x distance)
        - Hence, price = 10^log_price = 10^(0.801869 + (0.012187 * UberX) + (0.077486 * distance))

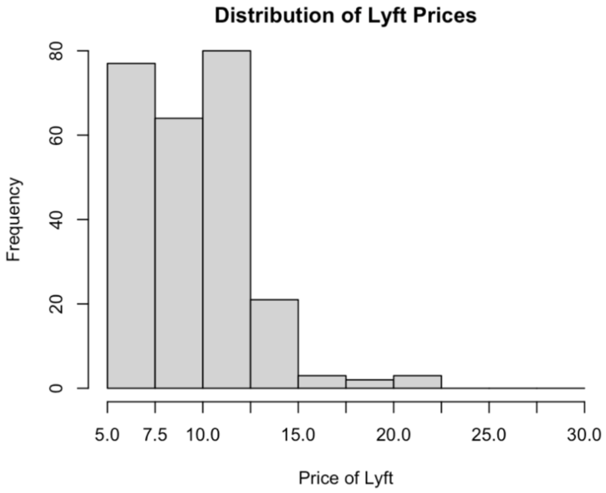
    b. Beta Estimate
        - Since the p-value of nameUberX = 0.0409 < α = 0.05, we can conclude that the variable "name" is a predictor in the output of the prices. Since Uber is the reference group, there is a mean difference of 0.012187 increase in log_price, which is an equivalent of a 10^0.012187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

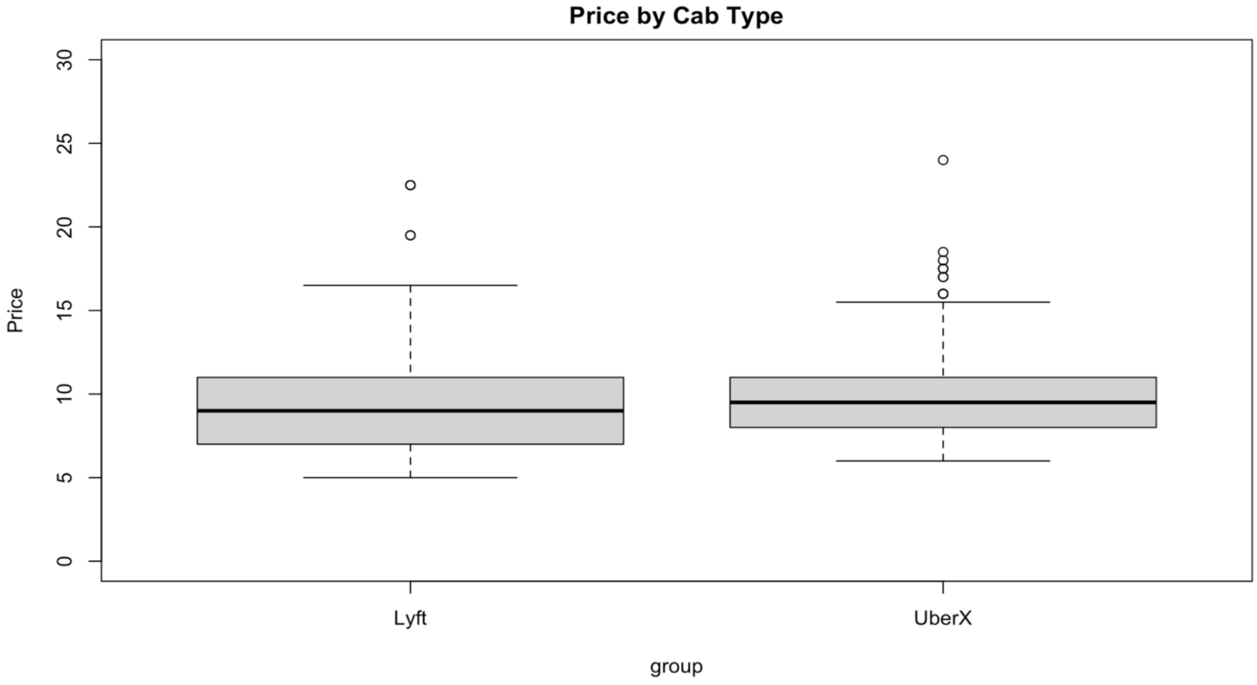
4. Supporting visualization
    a. Distribution of prices by group
        i. Histogram
        - The histogram below shows that the distribution of Lyft prices are slightly more right skewed than the distribution of Uber prices.
        - This indicates that Uber prices may be more expensive on average than Lyft prices.





ii. Boxplot

- The boxplot below shows 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.
- This explains why the two sample t-test failed to reject the null hypothesis and why the p-value for nameUberX was 0.0409, just slightly below our α = 0.05 level. This shows that we were close to failing to reject that the cab type affects the outcome of price, after adjusting for age. This aligns with the visualization below.



6. State Your Conclusion (no more than 100 words)

State the conclusion(s) of your analysis in a way so that a non-statistician can understand.

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

Conclusion:

On average, Uber prices are higher than Lyft prices in Boston, when accounting for distance. When comparing an Uber ride and a Lyft ride of the same distance, Uber rides are $1.028459 more expensive on average. We are 95% confident that for the same distance, Uber charges between $1.001162 to $1.056501 more than Lyft rides.

Solution Submission

1. **Fill up this word file and upload it.**
2. **Upload your data set. This is the data set after cleaning (a small CSV file)**
3. **Submit R code as either a well-commented .R file or as a .Rmd file with associated .html file.**

Grading will be done based on

1. **Originality of selected data set and data analysis approach**
2. **Data Preparation set and cleanup**
3. **General Correctness of data analysis**
4. **Quality of your R code and output results**
5. **Correct final conclusion and useful visualization**