Problem set1

Group member:
Shiming Zhao(MISM6202-05)
Haozhe Liu(MISM6202-02)

Foundations of Data Analysis for Business

PROBLEM SET 1

Use the file ‘start-data.csv’ for all questions. Begin your work on this problem set by importing the file ‘start-data.csv’ into RStudio.

Bluebikes is a public bike share system with stations located throughout the Greater Boston area. Users can purchase a monthly or annual membership that offers unlimited rides of up to 30 minutes, or buy one-time-use passes for either a single 30-minute ride or unlimited 2-hour rides within a 24-hour period (the “adventure pass”). Rides that extend beyond their time limit incur additional charges. After purchasing a pass or membership, riders use a kiosk or mobile app to unlock a bike at any station in the system, bike along their desired route, and dock the bike at any station in the system that has open spaces available. Bluebikes staff occasionally redistribute bikes across the various docking stations in the city in order to optimize bike and dock availability.

The Northeastern Student Affairs team is interested in understanding current usage statistics and trends for the Bluebikes stations closest to campus. The team has identified five stations (*Mass Ave T Station; Northeastern University North Parking Lot; Ruggles T Station; Tremont St at Northampton St;* and *Wentworth University*) and collected data for all rides originating at or terminating at these five stations during the month of August. Data captured include ride date, trip duration in minutes, start and end station information (name, internal station ID, and latitude and longitude coordinates), the internal ID number of the bike used in the trip, user type (*Subscriber* if the user has a membership; *Customer* if the user purchased a one-time pass) and the rider’s ZIP code. Student Affairs has asked you to help with some preliminary data analysis for the rides originating at the stations near campus, which are contained in the file ‘start-data.csv.’

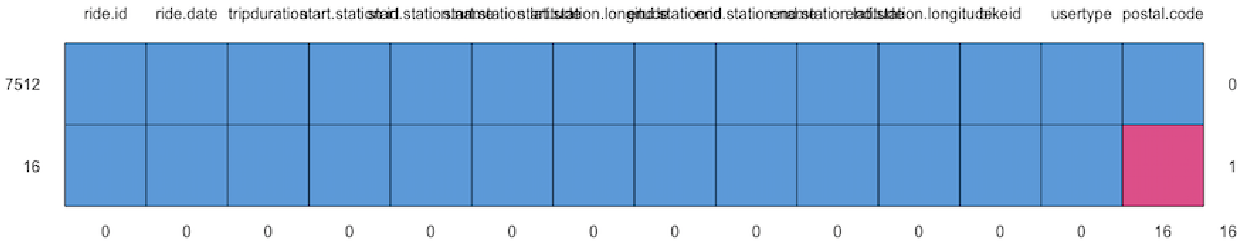
Use RStudio to answer the following questions. Provide your written answers, along with any relevant tables and charts, in a **single PDF file**. Any charts or tables included in your report should be properly labeled and formatted for an audience of company executives. Screenshots from an R output are not appropriate for this assignment. Additionally, your answers to each item listed below must be clearly numbered in order to receive full credit. You should also submit a **single .R script file** with your code for the analysis. *No other file formats will be accepted for this assignment, and you should submit just one file of each type.*

Data Wrangling.

1. Which variables, if any, appear to contain missing, null, or incorrect values? After reviewing the rest of the assignment instructions, briefly describe the impact you expect missing values to have on your analysis for this problem set, and your approach to handling these data errors.

1) Which variables, if any, appear to contain missing, null, or incorrect values?

After writing code in R, we find the column "postal.code" has 16 NA values and 622 NULL values, 40 incorrect values,(00000). (Note: these NULL values are not missing values and the current postal code range of the United States (US) is: 00001 – 99950. The lowest postal code area is 00001.). The process of pinpointing missing, null, or incorrect values is shown in Rscript. The picture below shows us a straightforward visualization of where the missing data are. The pink square area represents 16 missing values of our data.



2) The impact of missing values to have on our analysis for this problem set.

Based on the problem set, we find our analysis mainly uses several variables such as trip duration, usertype, and start station, so those missing values in postal.code column will not make a big influence on the relationships among above predictors. However, if we remove all rows containing missing values or NULL values, then it will make some bias for the rest of variables' relationships.

Besides, if we choose to use postal.code as a predictor to analyze the arrangement of Bluebikes station when answering question15, it is essential for us to change missing data to a value. Otherwise, NAs will prevent our calculation such as the probability of each zip code.

3) The approach to handling these data errors.

There are some common ways for us to deal with missing data and NULL data.

First, we can delete all rows with missing data or NULL data. However, this is a bad strategy because it also deletes all other essential information of rows we may use in the following analysis as well.

Second, we can delete rows with missing data or NULL data in specific variables. This method is similar to the first one, because there is only one incomplete column in our dataframe. In the end, both methods have the same result after removing the missing value. We don't like the way to omit all incomplete rows.

Third, we can change missing data and NULL data to the value we need. This is a suitable method for us to take because it doesn't affect other variables' relationship, and we will not lose any important information about the missing data rows. In Rscript, we use replace function to fill NAs and NULL with "unknown" so that we keep the integrity of data.

Fourth, we can use a method called **imputation,** which is to calculate the mean/median of the non-missing values in a column and then replace the missing values within each column separately and independently from the others. It can only be used with numeric data, so we don't use this method.

2. Using the z-score method (with z-score threshold of +/- 3), what are the cutoff values for identifying outliers in the 'tripduration' variable? Do you believe this method of identifying outliers is well-suited to the 'tripduration' data? Explain.

1) Using the z-score method (with z-score threshold of +/- 3), what are the cutoff values for identifying outliers in the 'tripduration' variable?

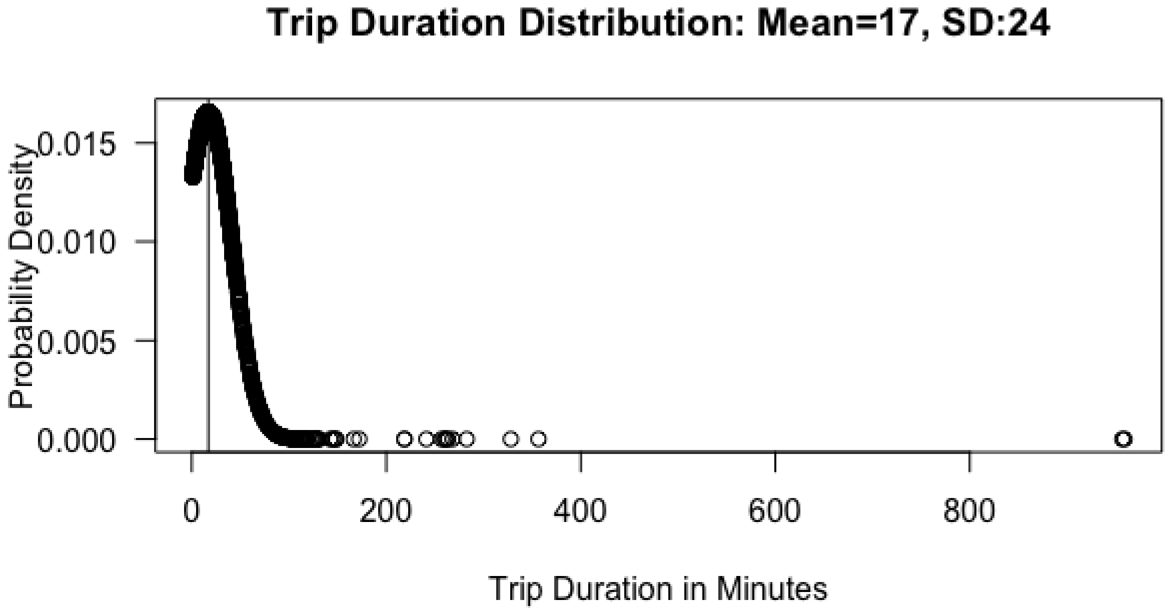
In order to use the z-score method, the first thing we need to do is to calculate the parameters of the formula below.

$Z\ score\ = \frac{x-\mu}{\sigma}$ After calculating in R, we get the mean of "trip duration" is 17.12, and standard deviation is 24.07. And then we create a new dataframe of "tripduration" and calculate individual z-scores based on z-score formula.

After running in R, we find the **upper limit of cutoff value is 89.33023**, and **lower limit of cutoff value is -55.09637** given z-score threshold of +/- 3. What's more, there are 87 outliers beyond the cutoff values.

2) Do you believe this method of identifying outliers is well-suited to the ‘tripduration’ data? Explain.

We don’ t believe that the z-score method is a credible way to identify outliers. When we draw the density distribution of “tripduration”, we find that the curve we get is not normal distribution, instead, it is more likely right sked.



This graph is drawn by R. Because z-scores are reliable indicators of outliers when the distribution is relatively bell-shaped and symmetric, so for “tripduration”, z-score is not well-suited. Nevertheless, we are better served identifying outliers in this case with a boxplot.

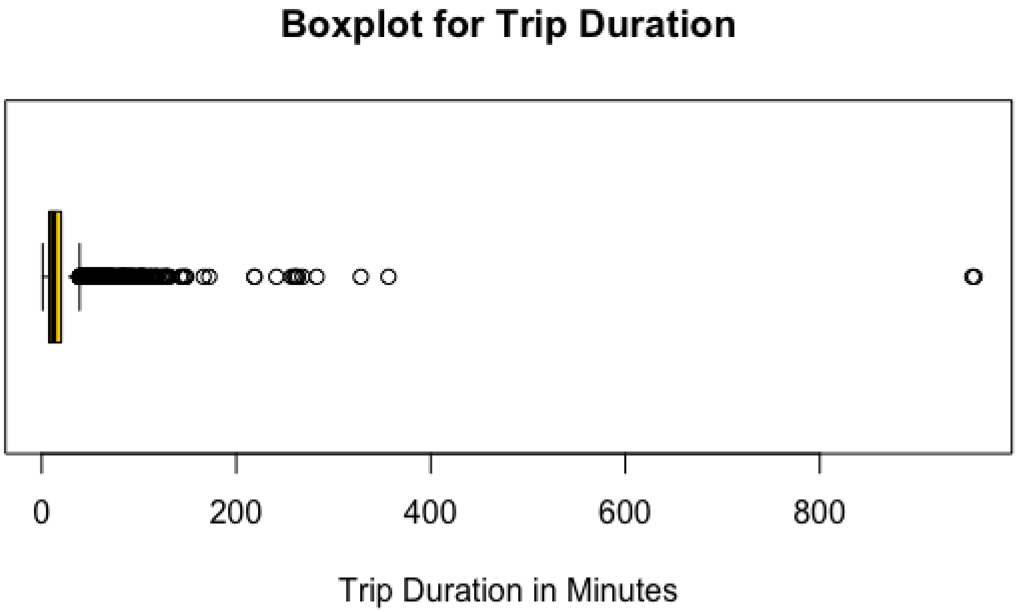
3. Using the boxplot/IQR method, what are the cutoff values for identifying outliers in the ‘tripduration’ variable? Do you believe this method of identifying outliers is well-suited to the ‘tripduration’ data? Explain.

1) Using the boxplot/IQR method, what are the cutoff values for identifying outliers in the ‘tripduration’ variable?

By calculating in R, we can identify Q1=7.47, Q3=20.1, IQR=12.633. Thus, the cutoff value for upper bound is Q3+1.5*IQR = **39.05**, and lower bound is Q1-1.5*IQR= **-11.48333**. The concrete calculation is in Rscript.

2) Do you believe this method of identifying outliers is well-suited to the ‘tripduration’ data? Explain.

Compared with the Z-score method, we think **IQR method is well-suited to the ‘tripduration’ data**. First, the boxplot of ‘tripduration’ is not symmetric, and can be used to informally gauge the shape of the distribution. Second, it can be more effective when the distribution of variables is not bell-shaped. By counting the number of outliers, we find by using IQR, there are 455 outliers, which is more than 87 outliers when using Z-score, so our data will be more reflective after removing the outliers.



4. Student Affairs has decided that only rides of one hour or less should be included in your analysis, as they are most interested in commuting trends and longer rides are more likely to be for leisure purposes. Subset the data, creating a dataframe with only rides that are up to 60 minutes in duration. How many rides are in the resulting dataframe?

1) Subset the data, creating a dataframe with only rides that are up to 60 minutes in duration.How many rides are in the resulting dataframe?

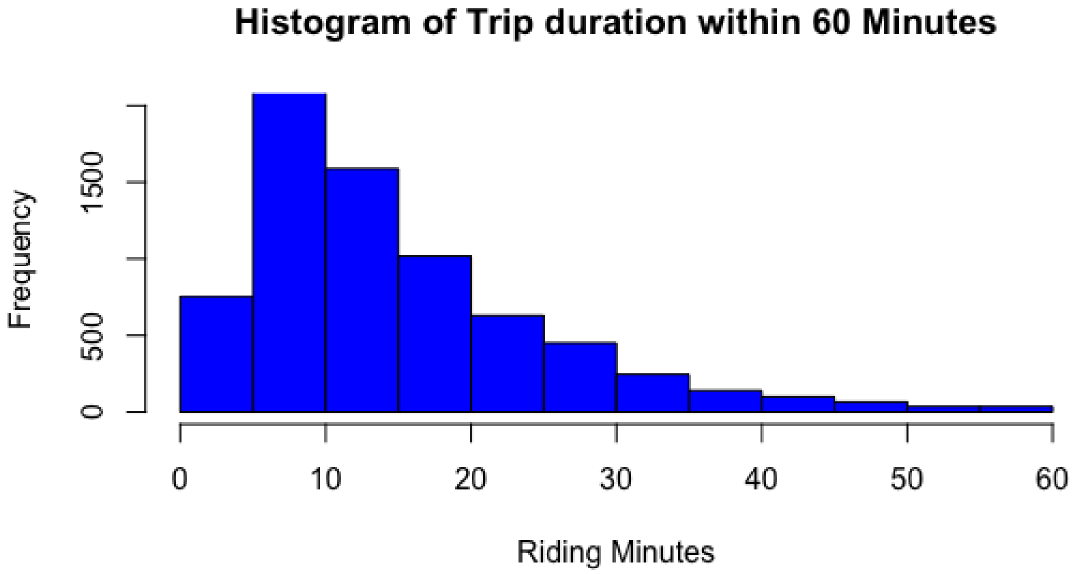
After creating a new dataframe with only rides that are up to 60 minutes, there are **7326 rides** in the resulting dataframe.

For all remaining analyses, use only data for rides that are 60 minutes or less in duration.

Visualization & Descriptive Statistics.

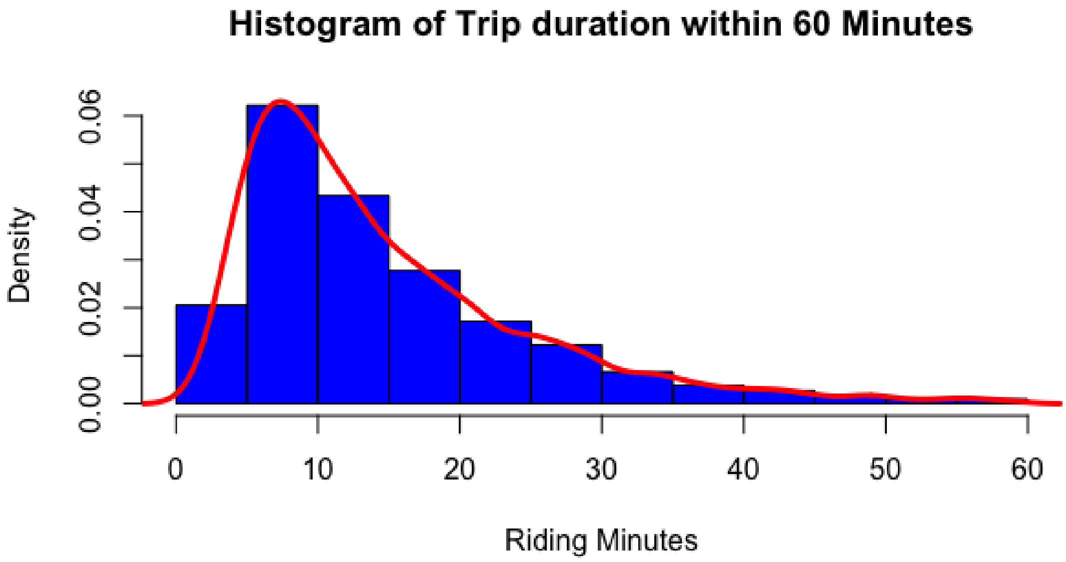
5. Create a histogram for the ‘tripduration’ variable. Describe the shape of the distribution and provide a brief (1-2 sentence) explanation of why the observed distribution shape might be expected for the ‘tripduration’ variable.

1) Create a histogram for the ‘tripduration’ variable. (Building progress shown in R)



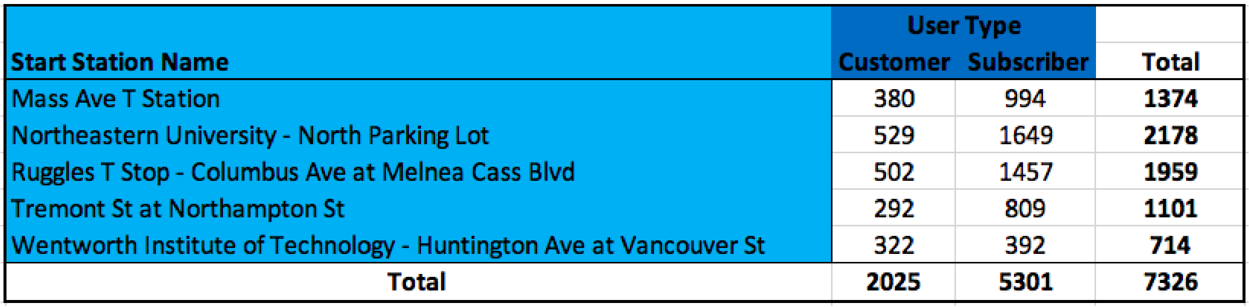
2) Describe the shape of the distribution and provide a brief (1-2 sentence) explanation of why the observed distribution shape might be expected for the ‘tripduration’ variable.

   The shape of the distribution is right skewed. In other words, it has a positive skewness coefficient. Before subsetting our data, the column “tripduration” has many observations at the right side of mean, including many extreme values or outliers. Even if we select the range of 60 minutes, there also is a lot of big data located at the right side of the median. Thus, this distribution is corresponding with our expectation.



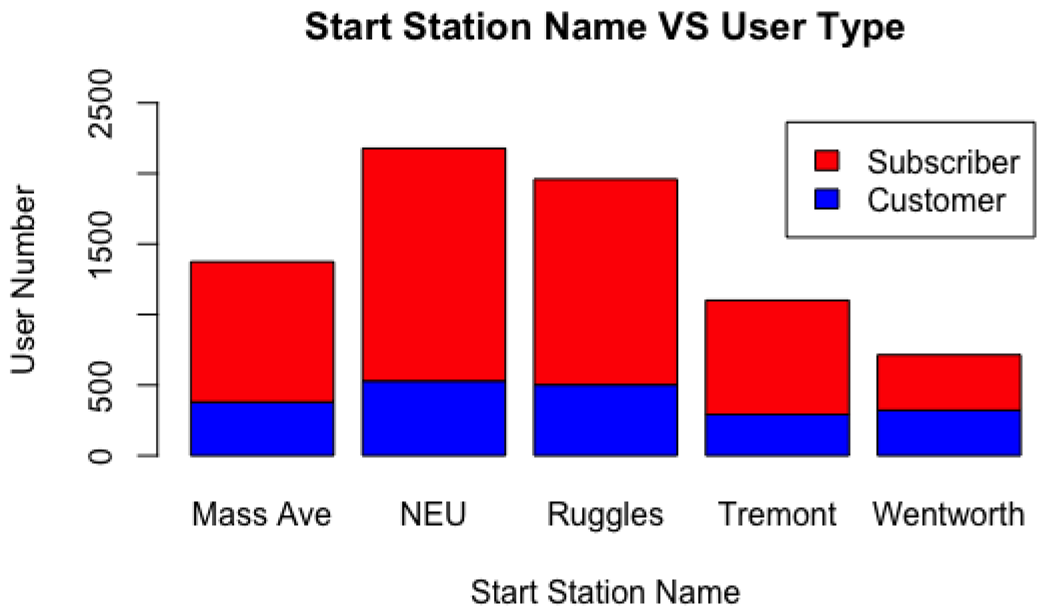
6. Create a contingency table and an accompanying stacked or clustered column chart to summarize and visualize the variables 'start.station.name' and 'usertype.' In 1-2 sentences, describe any noteworthy patterns or insights you observe from this table and chart.

1) Contingency table (it is created by R, but I use excel to beautify this table)

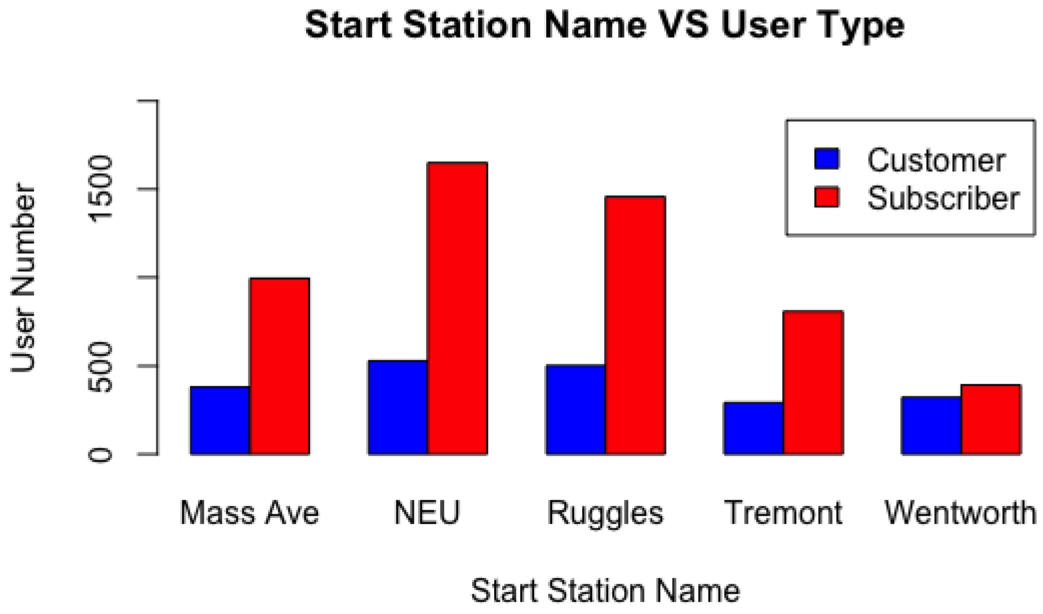


	User Type		
Start Station Name	**Customer**	**Subscriber**	**Total**
Mass Ave T Station	380	994	**1374**
Northeastern University - North Parking Lot	529	1649	**2178**
Ruggles T Stop - Columbus Ave at Melnea Cass Blvd	502	1457	**1959**
Tremont St at Northampton St	292	809	**1101**
Wentworth Institute of Technology - Huntington Ave at Vancouver St	322	392	**714**
Total	**2025**	**5301**	**7326**

2) Stacked chart(created and exported by R)



3) Clustered Column Chart (Created and exported by R)

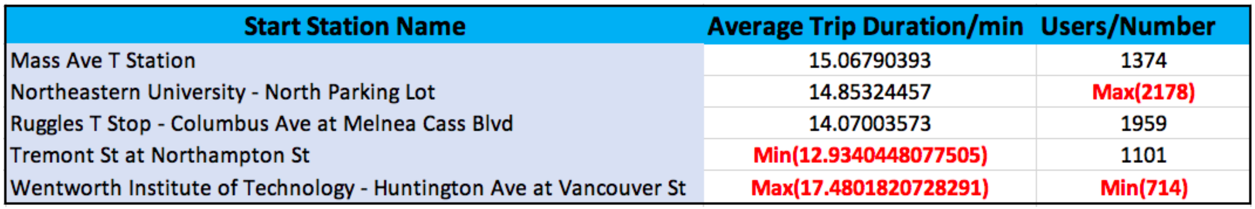


4) describe any noteworthy patterns or insights you observe from this table and chart.

Compared to columns “Start Station Name” and “User Type” with just raw data, contingency table and stacked chart presents the results of the start station and user type in a much more informative format. We can readily see that of the 7326 riders, 5301 of them are subscribers, taking up to 72.36% of the whole group of people. This means that the membership of Bulebikes is more attractive for customers and this marketing strategy seems successful. However, there do appear to be some differences depending on the start station. People were more likely to ride the bike from Northeastern University than from Wentworth institute of Technology, because according to the stacked chart, the total number of Northeastern University users seems twice than Wentworth’s users. Thus, it would be wise for the Bluebikes to reduce the number of bikes around Wentworth and to increase bikes around NEU.

7. Create a table showing the average trip duration for rides originating at each of the start stations in the dataframe. In 1-2 sentences, describe any noteworthy patterns or insights you observe from this table.

1) Create a table showing the average trip duration for rides originating at each of the start stations in the dataframe.



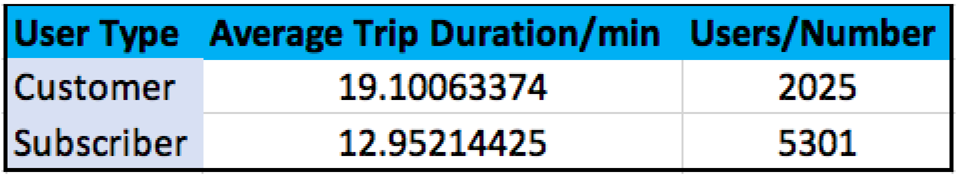
Start Station Name	Average Trip Duration/min	Users/Number
Mass Ave T Station	15.06790393	1374
Northeastern University - North Parking Lot	14.85324457	Max(2178)
Ruggles T Stop - Columbus Ave at Melnea Cass Blvd	14.07003573	1959
Tremont St at Northampton St	Min(12.9340448077505)	1101
Wentworth Institute of Technology - Huntington Ave at Vancouver St	Max(17.4801820728291)	Min(714)

2) Describe any noteworthy patterns or insights you observe from this table.

From the table, we can see that people starting from “Wentworth Institute of Technology - Huntington Ave at Vancouver St " tend to have the longest trip durations, 17.48 minutes compared to other four start stations. Instead, people starting from “Tremont St at Northampton St " spend least time, 12.93 minutes, with bluebikes. This also illustrates question6’s stacked chart, where Wentworth start station has the least number of people riding the Bluebikes, because people might think that riding for more than 15 minute is not suitable for people to continue taking bikes.

8. Create a table showing the average trip duration for rides taken by each ‘usertype’ in the dataframe (*Customer* and *Subscriber*). In 1-2 sentences, describe any noteworthy patterns or insights you observe from this table.

1) Create a table showing the average trip duration for rides taken by each ‘usertype’ in the dataframe (*Customer* and *Subscriber*)



User Type	Average Trip Duration/min	Users/Number
Customer	19.10063374	2025
Subscriber	12.95214425	5301

2) Describe any noteworthy patterns or insights you observe from this table.

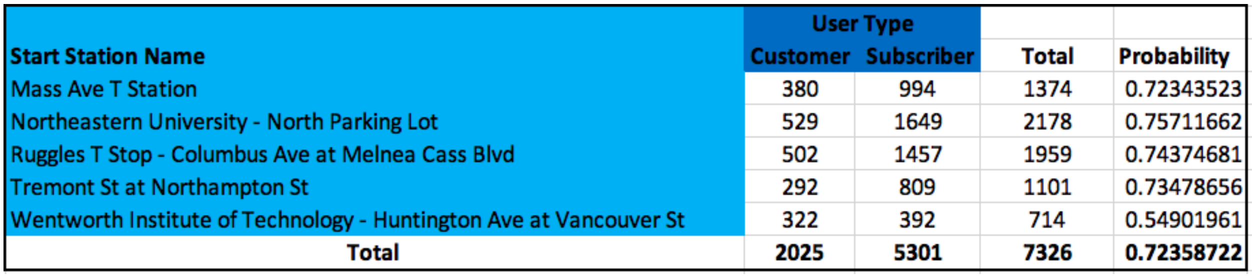
By comparing the trip duration between customers and subscribers, we find that customers have longer riding time than subscribers. The difference in spending on the road is high up to 6.05 minutes (19-12.95). The reason why there is such a big difference between them is that users purchased a one-time pass, so they cherished their travel with bluebikes, which, in part, caused a higher average trip duration.

Probability.

9. Suppose you randomly select a ride from the data set. What is the probability that the selected ride was taken by a Subscriber, as defined by the 'usertype' variable?

The total number of users is 7362, among which there are 5301 subscribers. Thus, the probability that the selected ride was taken by a Subscriber is 5301/7362 = 0.7235872. (Calculated by R)

10. Is the probability of selecting a Subscriber independent of the start station? Briefly describe, citing at least two conditional probabilities in your explanation.



	User Type			
Start Station Name	**Customer**	**Subscriber**	**Total**	**Probability**
Mass Ave T Station	380	994	1374	0.72343523
Northeastern University - North Parking Lot	529	1649	2178	0.75711662
Ruggles T Stop - Columbus Ave at Melnea Cass Blvd	502	1457	1959	0.74374681
Tremont St at Northampton St	292	809	1101	0.73478656
Wentworth Institute of Technology - Huntington Ave at Vancouver St	322	392	714	0.54901961
Total	**2025**	**5301**	**7326**	**0.72358722**

No, the subscriber is not independent of the start station. The unconditional probability of the selected ride taken by a subscriber is 0.7235872, but conditioning on different start stations, the conditional probabilities range from P(Subscriber|Mass Ave)=0.723435226 to P(Subscriber|Wentworth)=0.549019608. Because all the probabilities conditional on any start station differ from0.7235872, we can conclude that the variables are not independent.

Sampling Distributions.

11. Treating the data you have collected as the population of all rides taken from the selected stations in August, suppose you repeatedly selected random samples of 50 rides and calculated mean trip duration and proportion of users of type ‘Customer’ in each sample. (a) What are the mean and standard deviation of the resulting sampling distribution of sample mean? (b) What are the mean and standard deviation of the resulting sampling distribution of sample proportion?

1) What is the mean and standard deviation of the resulting sampling distribution of sample mean?

$$E\left(\overline{X}\right) = E(X) = \mu.$$

Based on the formula above, we calculate that the population mean of trip duration is 14.65167, so the mean of the resulting sampling distribution of sample mean is 14.65167 as well. (Calculated by R)

$$se\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}.$$

Based on the formula above, we calculate that the population standard deviation of trip duration is 10.11615, so the standard deviation of the resulting sampling distribution of sample mean is 10.11615/sqrt(50) = **1.430639**.

2)What are the mean and standard deviation of the resulting sampling distribution of sample proportion?

$$E\left(\overline{P}\right) = p$$

Based on the formula above, we calculate that the proportion of customer type is 0.2764128, so the mean of the resulting sampling distribution of sample proportion is 0.2764128 as well. (Calculated by R)

$$se\left(\overline{P}\right) = \sqrt{\frac{p(1-p)}{n}}.$$

Based on the formula above, we calculate that the standard deviation of the resulting sampling distribution of customer proportion is sqrt((0.2764128*(1-0.2764128)/50)= 0.06324694.

Statistical Inference.

12. Choose a random sample of 50 rides from the data and store these observations in a new dataframe called ‘sample.df.’ Estimate a 95% confidence interval for population mean trip duration based on your sample. Does the confidence interval include the true population mean trip duration? *Show your work.*

1) Use R to calculate the 95% confidence interval for population mean trip duration based on the ‘sample.df’:

   The 95 percent confidence interval: **10.70638 and 16.04496**

2) The mean of the true population is 14.65167, which is included in the 95% confidence interval of the sample.

13. Using the observations in ‘sample.df,’ calculate the sample proportion of rides taken by user type ‘Customer’ and estimate a 95% confidence interval for the population proportion of ‘Customer’ user types. Does the confidence interval include the true population proportion? *Show your work.*

1) Use R to find the number of ‘Customer’ in the ‘sample.df’, there are 11 ‘Customer’, then calculate the 95% confidence intervals of the sample are **0.1199448 and 0.3633110.**
2) The true population proportion 0.2764128 is located in the 95% confidence interval.

14. Calculate the average trip duration for all rides taken during the first week of August (8/1-8/7) and store this value as ‘mu0.’ Next, create a dataframe called ‘week4’ that includes only rides taken during the final week of August (8/25-8/31). Run the command `set.seed(999)` and then take a sample of size 100 from the ‘tripduration’ variable in the ‘week4’ dataframe (i.e., sample from week4$tripduration).

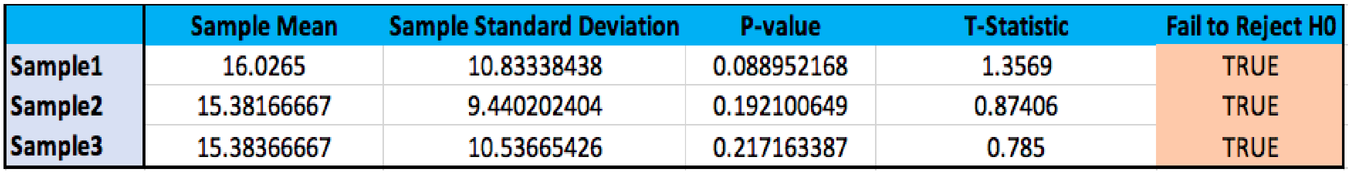
(a) State the null and alternative hypotheses to test whether the average trip length during the final week of August is higher than the average trip length in the first week (mu0). Then, perform the hypothesis test using significance level α = 0.05. Show your work and clearly state your decision in the test.

Before we get started, let's calculate **mu0=14.55654**.

1) H0: x<=mu0, the average trip length during the final week of August is less than and equal to the average trip length in the first week
2) Ha: x>mu0, the average trip length during the final week of August is higher than the average trip length in the first week
3) At α = 0.05 df=99, we can calculate t = 1.8409 and p-value = 0.03432 by R.

   Because of P-value = 0.03432 < α = 0.05, so we reject H0, and conclude that the average trip length during the final week of August is higher than the average trip length in the first week given the confidence level of 95%.

(b) Without running set.seed() again, re-run the lines of code that select a sample and calculate sample statistics (xbar and s) and the test statistic (tstat) several times. What do you observe in repeating the sampling and hypothesis test procedure? *Optional: Write code to take 100 different samples, calculating the test statistic for each one, and briefly summarize the resulting hypothesis test decisions.*



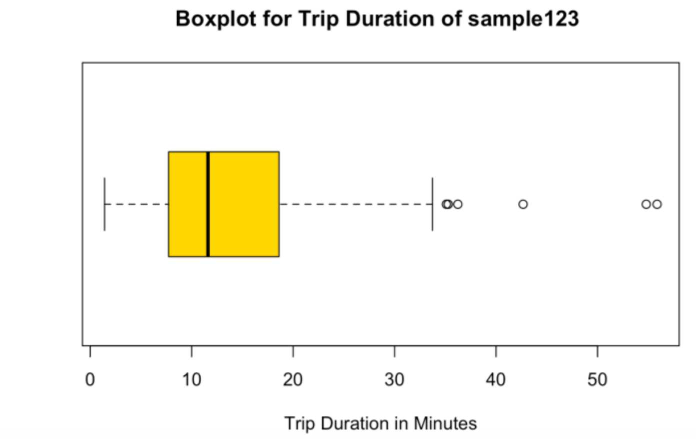
	Sample Mean	Sample Standard Deviation	P-value	T-Statistic	Fail to Reject H0
Sample1	16.0265	10.83338438	0.088952168	1.3569	TRUE
Sample2	15.38166667	9.440202404	0.192100649	0.87406	TRUE
Sample3	15.38366667	10.53665426	0.217163387	0.785	TRUE

1) We created three samples without the setseed(999) and computed the sample statistics and the test statistic of them, and created the table by R and butified it by Excel.
2) We found that all samples' p-value are greater than 0.05, so all samples fail to reject H0 and do not have sufficient evidence that average trip length during the final week of August is higher than the average trip length in the first week (mu0).

15. *[Open ended]*. What other analyses would you be interested in performing with this data? Think about the data fields you have and what would be interesting to know. In 1-2 short paragraphs, describe at least two specific analyses that could be done using the variables in this dataset. *Optional: Perform these analyses and create visualizations to show the results!*

1) First analysis: We want to analyze whether the difference between the sample and the overall data is very large, to determine whether the sample is a good way to analyze the overall data.

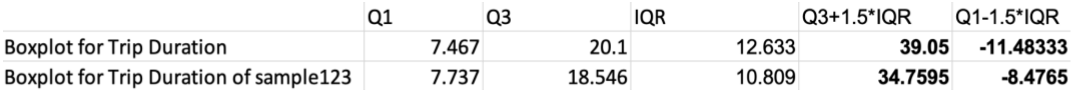
Run the setseed(123) and take a sample(called ‘sample123) of size 100 from the ‘tripduration’ variable in the ‘Bluebikes.max60m’ dataframe, then create the boxplot for the trip duration of the sample. compare it with the boxplot in question3, what do you find on these outputs?



The cutoff values for identifying outliers in the ‘tripduration’ variable:

By calculating in R, we can identify Q1=7.737, Q3=18.546, IQR=10.809. Thus, the cutoff value for upper bound is Q3+1.5*IQR = **34.7595**, and lower bound is Q1-1.5*IQR= **-8.4765**. The concrete calculation is in Rscript.

Use Excel to compare these two boxplot



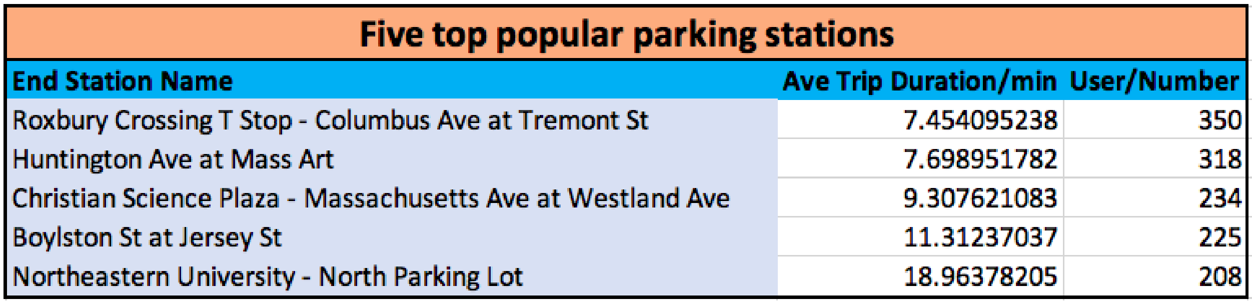
	Q1	Q3	IQR	Q3+1.5*IQR	Q1-1.5*IQR
Boxplot for Trip Duration	7.467	20.1	12.633	**39.05**	**-11.48333**
Boxplot for Trip Duration of sample123	7.737	18.546	10.809	**34.7595**	**-8.4765**

By comparing these two boxplot, boxplot of sample are easier to understand than the boxplot from Question 3 because there many data are over than 60 minutes in the boxplot from Question 3, this means that we can ignore some extreme values to sample first; and we conclude that the difference between the two is not very large, which also shows that sampling methods to analyze or predict the overall data is feasible, sampling can also reduce the occurrence of some extreme data, such as Q3's boxplot has too much extreme data to make it difficult for us to analyze or predict.

In general, sampling is the great way to predict or analyze data, but we have to take multiple samples to analyze the overall data, because a sample will be quite different from the overall data.

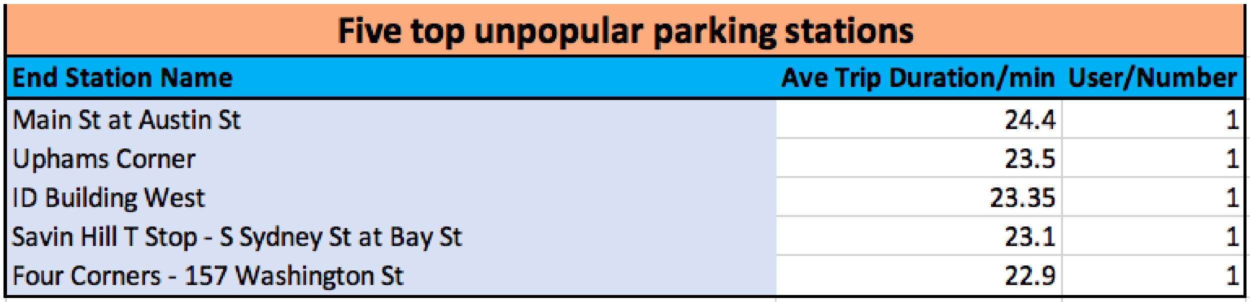
2) Second analysis: We want to analyze the effects between variables and find the relationship between them and then give Bluebikes some suggestions.. For example, we want to analyze the relationship between the average trip duration and the number of users.

We also use the dataframe where the trip duration is less than 60 minutes. Let’s create a table showing the average trip duration for rides ending at each of the end stations in the dataframe.We focused on the top five popular end stations and five unpopular end stations. The tables are created by R and butified by Excel.



Five top popular parking stations

End Station Name	Ave Trip Duration/min	User/Number
Roxbury Crossing T Stop - Columbus Ave at Tremont St	7.454095238	350
Huntington Ave at Mass Art	7.698951782	318
Christian Science Plaza - Massachusetts Ave at Westland Ave	9.307621083	234
Boylston St at Jersey St	11.31237037	225
Northeastern University - North Parking Lot	18.96378205	208



Five top unpopular parking stations

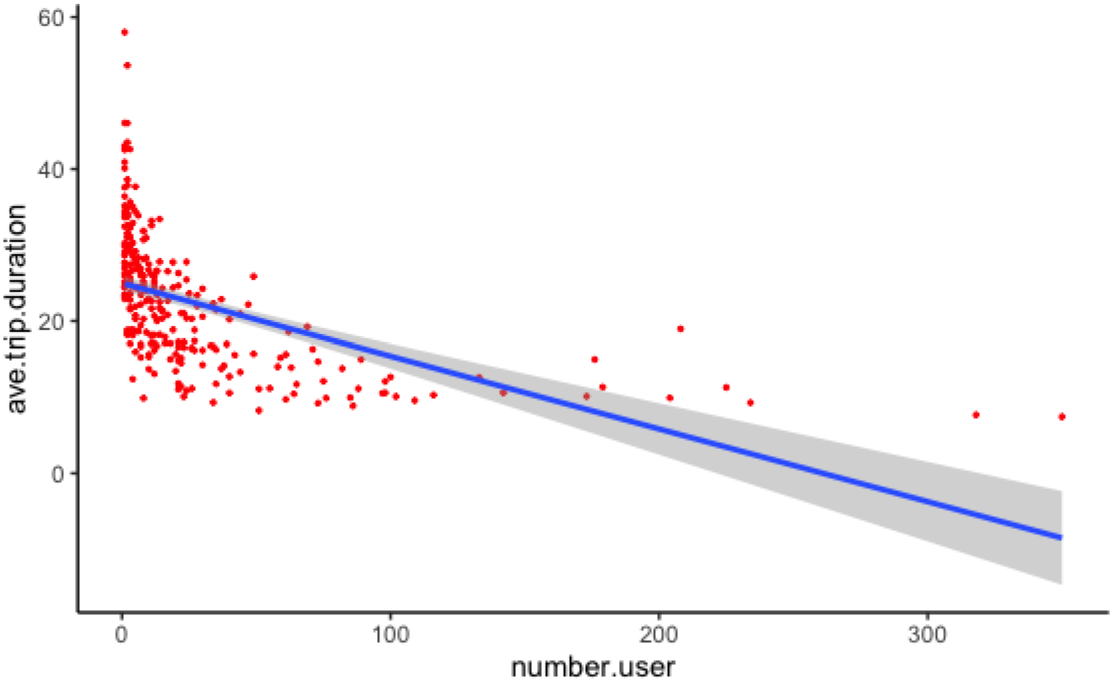
End Station Name	Ave Trip Duration/min	User/Number
Main St at Austin St	24.4	1
Uphams Corner	23.5	1
ID Building West	23.35	1
Savin Hill T Stop - S Sydney St at Bay St	23.1	1
Four Corners - 157 Washington St	22.9	1

Here, we can see that top popular parking stations like Roxbury, and Huntington has more than 300 people choosing to park. However, some stations like Four Corners are least likely for people to park. Besides, we believe the reason why people are inclined to go to these popular stations is because they spend less time riding on the road compared with those unpopular stations, where people have to spend more than 20 minutes.

Then, we are interested in exploring some potential linear regression between trip durations and number of users.

First, we verify the linear relationship between trip durations and number of users. The t value of number.user is -9.948, which is low enough to reject the H0, so there is a strong negative relationship. (t statistic is calculated by R)

Second, we draw the relationship. This graph below is created by R. There is a negative relationship between the number of users and average trip duration. The more users in the end station, the less time people spend during the trip. Based on this case, we might suggest that Bluebikes build more parking stations around those popular stations.


ave.trip.duration
number.user