Thao Duyen Tran

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.

There are 16 missing ("NA") values in this dataset, along with 40 wrong values ("00000") and 622 "NULL" values. I also double checked this with the COUNTIF function in Excel and returned the same number of values. All the missing and incorrect values were found in the postal.code column of the dataset.

As the missing values are located in the postal.code column, I do not think it will impact the dataset analysis, unless I were to perform analyses that focused on the riders' ZIP codes. As far as the problem set goes, I do not think I will be doing anything to the incorrect/missing values. As those values are in the ZIP code section, which is not the focal point of the activity, I don't think it would be optimal to remove these values (there are over 600 of them, which is almost 10% of the dataset). The ZIP code data is also impossible to impute.

If there are analyses that include ZIP code data (perhaps for Question 15), they can be removed the incorrect/missing data values and use a smaller dataset (however, the data will still contain well over 6000 entries which would still yield results).

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.

The cutoff values for a z-score of +/- 3 around the trip duration mean are -55.096 minutes and 89.330 minutes. As it is not possible to get negative trip duration, and also the minimum trip duration is a little over 1 minute (which corresponds to a z-score of -0.668), there will only be an upper limit for trip duration.

We believe that this method of identifying outliers can work well for the trip duration observations, despite the median, mean, and third quartile range of hovering around 12 to 20 minutes. The distribution of trip duration is quite diverse While having 87 outliers does sound like a lot, it is acceptable in this case as we have 7528 observations in total. The proportion of our trip duration data that contains outliers is around 1.2%. While it would be ideal if the proportion of outliers was around 0.3%, as that would allow for the removal of outliers 3 standard deviations above the mean. The frequency distributions for trip duration, both from the raw data and from the data with outliers removed (via z-scores) can be found in Appendices A and B, respectively.

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.

The boxplot/IQR method will eliminate outliers that are Q3+1.5(IQR) above the mean, and Q1-1.5(IQR) below the mean, for trip duration, it translates to a lower bound of -10.56 minutes and upper bound of 38.13 minutes. As we cannot use negative numbers in our trip duration variable, we will only be left with an upper bound of 38.18 minutes. The boxplot of trip duration is very oddly shaped, as there are 2 outliers that are very far from the boxplot range, and there is a large number of observations that exists outside the box's whiskers (this can be found in Appendix C).

The IQR method has a lower cut-off limit than what we got with the z-score method, and, as such, I believe that this is not a good method to identify outliers in this variable, generally speaking. Using this method, we will have 483 outliers, which is 6% of the observations, which is much higher than the 1.2% we got for the z-score method.

However, this could be useful if we wanted to make our distribution look more centered (which was not the case with the z-score approach), or if we were looking to analyze data for observations that have short trip durations. The distribution of trip duration with outliers removed by the IQR method can be found in Appendix D; it is still positively skewed, but it is much less so than the distribution of the same variable with outliers removed by the z-score method.

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?

There are a total of 7326 rides in the new data frame with a duration of 60 or less minutes.

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.

The shape of this distribution is positively skewed, with a long right tail and a high peak on the left (around the 5 to 10 minutes point). The distribution can be found in Appendix E. I assume that the distribution is skewed this way because of the Blue Bikes' station locations and the overall layout of Boston. If we were to only look at trips that lasted between 4 to 11 minutes (which is where peaks of the distribution appears to be), we can see that the most two popular end stations are "Roxbury Crossing T Stop" and "Huntington Ave at Mass Art", with 273 and 259 trips ending there, respectively. Both of these bike stations are situated near popular apartment rental areas for students (Appendix F), this is especially true for "Roxbury Crossing T Stop".

Moreover, the Northeastern campus is located in a prime location and is close to both the center of Boston, and the areas West of Boston (which is popular among younger people, especially students), so there would not be much need for students, who are starting at Blue Bike stations near Northeastern, to ride bikes for more than 20 minutes. The average speed of a biker is around 12mph (around 19.5km/h)[1], so a 20-minutes bike ride from Northeastern's campus would cover a radius of 4 miles (around 6.5km), which would cover the majority of Boston. Similarly, a bike ride of 10 minutes (around the peak of the distribution) would cover a radius of 2 miles, which would still get an individual to most places in the Boston downtown area. As such, longer bike rides are most likely to be for recreational purposes.

[1] Eriksson J, Forsman Å, Niska A, Gustafsson S, Sörensen G. An analysis of cyclists' speed at combined pedestrian and cycle paths. Traffic Inj Prev. 2019;20(sup3):56-61. doi: 10.1080/15389588.2019.1658083. Epub 2019 Sep 27. PMID: 31560212.

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.

The contingency table (Appendix G) between start station name and user type shows that:

- The number of subscribers are more than double that of customers.
- The top two most popular start locations, for both subscribers and customers, are "Northeastern University's North Parking Lot" and "Ruggles T Stop at Columbus Ave".
- The least popular start location for customers is "Tremont St at Northampton St", while the least popular start location for subscribers is "Wentworth Institute of Technology".
- Subscribers generally have double or triple the number of trips in each station, compared to customers. However, this is not the case for the "Wentworth Institute of Technology" station, which shows almost identical trip numbers (322 for Customers and 392 for Subscribers).

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.

The contingency table of the average trip duration for each of the 5 start stations (Appendix I) shows that:

- "Tremont St at Northampton St" is the station with the shortest average trip duration (at 12.93 minutes), while "Wentworth Institute of Technology" has the longest (17.48 mins).
- The other three stations had very similar trip durations, and are around a minute of each other.
- The general trend appears to be that Blue Bikes stations that are closer to the center of Boston would have shorter trip durations (as shown in Appendix J), with the exception of Mass Ave T Station, which is one of the closest stations to Boston's city center, but has the second longest trip duration.

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.

The table (Appendix K) shows that Customers, on average, take longer trips than Subscribers. The difference between trip durations is quite big, as the average trip duration for Customers is almost 1.5 times more than that of Subscribers.

◆ ◆ ◆

Probabilities

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 probability of a randomly selected ride to be taken by a Subscriber can be determined by taking the number of Subscribers (5301) and dividing it by total number of observations (7326), which would be around 0.724 (72.4%).

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

The probabilities for the Subscriber user type conditional on Start Station (Appendix L) are as follow:

- P(Subscriber | Mass Ave Station) = 0.723
- P(Subscriber | NEU Station) = 0.757
- P(Subscriber | Ruggles T Stop) = 0.744
- P(Subscriber | Tremont at Northampton St) = 0.735
- P(Subscriber | Wentworth Institute)= 0.549

According to this, the probability of selecting a Subscriber user type is dependent on start station as none of the conditional probabilities equal to 0.724 (which is the unconditional probability for Subscribers). Even though the conditional probability for Mass Ave Station is very similar to the unconditional probability of Subscribers (it would be identical if rounded to 2 significant figures), the fact that we have a conditional probability on any start station that does not equal 0.724 means 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?

Hypothetically speaking, if we were to repeatedly select random samples of 50 rides and calculate the mean trip duration of each sample, we would get an average value of sample means ($\bar{X}$) that will be equal to the population mean (μ). In this case, if we are using our current dataset (with trip durations of 60 minutes or less), we would get a repeated sample mean of about 14.65 minutes, if we were using our original dataset with all observations, we would get a mean of about 17.12 minutes.

If we are assessing the variability between each of the samples of 50 rides, then we would calculate the standard error of, $se(\bar{X})$, which would be the population standard deviation over the square root of observations ($\frac{a}{\sqrt{n}}$). In this case, we would get a standard error of 1.43 minutes for trips under 60 minutes, and a standard error of 3.40 minutes for all observations.

☂

Similarly to the sampling distribution mean for trip duration, the mean for our sample proportion of user type Customer, $E(\hat{p})$, will be equal to the population proportion (p). In this case, the resulting proportion of Customer for observations with trip duration under 60 minutes will be around 0.276, while the proportion for all observations in the start_data dataset will be around 0.292.

The standard error of the sample proportion, $se(\hat{p})$, will be calculated with the population proportion and number of observations, using the equation $\sqrt{\frac{p(1-p)}{n}}$. The $se(\hat{p})$ of Customer will be 0.0632 and 0.0643 ,for trips equal to or under 60 minutes and for all observations, respectively.

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.*

The random sample of 50 rides derived from the trimmed dataframe (with ≤60 minutes trips) had a:

- Sample mean ($\bar{X}$) of 15.237 and sample standard deviation (s) of 9.920 (found with R)
- T-score of 2.010 (can be found in t-table or with R using alpha of 0.025 and 50-1=49 df)
- Standard error of 1.403 (9.920 ÷ SQRT(50))
- Margin of error of 2.819 (2.010 x)
- Upper CI bound of 18.056 and lower CI bound of 12.417 (15.237 ± 2.819)

This confidence interval does include the true population standard means of both my trimmed data (μ=14.65) and original data (μ=17.12). I also double checked my answer using the t.test() function in R and received the same confidence interval bounds.

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

The sample.df observations had a:

- Sample proportion ($\hat{p}$) of 0.26 for Customer user types (13 Customer ÷ 50 observations)
- Standard error of 0.062 (derived by using the equation $\sqrt{\frac{0.26\times(1-0.26)}{50}}$)
- Z-score of 1.96 (can be found using R's qnorm() function or the Z standard normal table)
- Margin of error of 0.122 (1.96 x 0.062)
- Upper CI bound of 0.3816 and lower CI bound of 0.1384 (0.26 ± 0.122)

The confidence interval does include the true population proportion of user type Customer for both my trimmed data (p=0.276) and original data (p=0.292).

☐ Interestingly, when I double checked my answer using the prop.test() function in R, I was not given the same results, but instead was given upper and lower CI bounds of 0.406 and 0.151, respectively. After doing some research, I found that the test done through prop.test() is an Exact Binomial CI method, which is different from what we have been working on, which is the Binomial Confidence Interval Approximation method. It is interesting to see the difference in intervals between the two methods.

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).

The week1 dataframe has 1670 observations, with a mean of 14.557 (mu0), while the week4 dataframe has 1700 observations (with an mean of 15.271)

(The answers for questions 14a and 14b continue on the next page)

(a) State the null and alternative hypotheses to test whether the average trip length during the final week of August (week4) is higherthan 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.

The null and alternative hypotheses to evaluate whether the average trip length of week4 is higher than the average trip length of mu0 are:

$$H_0 : \bar{X}(week4) \leq \bar{X}(mu0) \qquad H_A : \bar{X}(week4) > \bar{X}(mu0)$$

$$H_0 : \bar{X}(week4) \leq 14.557 \qquad H_A : \bar{X}(week4) > 14.557$$

The sample statistics and test statistic for this hypothesis test are as follow:

- Number of observations (n) is 100, and alpha (α) is 0.05 (given)
- Mean ($\bar{X}$) of 16.54 and standard deviation (s) of 10.77 (found with R)
- T stat of 1.841 (found using the equation $\frac{16.54-14.56}{10.77 \div \sqrt{100}}$)
- The p-value is between 0.050 and 0.025 (with 99 degrees of freedom in the t-table)
- Actual p-value is 0.0343 (found using pt() function in R)

Since the p-value is lower than our alpha (0.05), we are able to reject the null hypothesis and conclude that the average trip duration in week 4 is higher than the average trip duration in week 1.

(b)* Without running set.seed() again, re-run the lines of code that select a sample and calculate sample statistics ($\bar{X}$ and s) and the test statistic 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.

We wrote a for loop to re-run the sampling process and get the sampling statistics/test statistic 100 times, and also to make a rejection/non-rejection decision based on the t-value (using the t-critical value as a baseline so that if my t-value is ≥ the critical t-value of 1.660, the code will return a decision to "reject the null hypothesis"). We have included a snippet of my data (25 runs) in Appendix M. According to the data, here is actually a much higher chance of the null hypothesis not being rejected, which is very surprising, considering that for question 14a, we had a decision to reject the null instead.

◆ ◆ ◆

Further Analyses

15. 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.

One simple, yet interesting, analysis we can do is to use the most 6 most popular end stations, and evaluate how a rider from a start station will end there. The contingency table for this analysis can be found in Appendix N. The clustered bar plot for this can also be found in Appendix O. We can see that Northeastern University (start) has the most trips overall, but the greatest number of rides are from riders who begin their journey at Ruggles T Stop and end at Roxbury Crossing T Stop. Reversely, the least popular start and end journey is from Mass Ave T Station, ending at Northeastern University.

Another analysis which involves end stations would be to compare the top 6 most popular end stations weekdays vs weekends, and sorting it by user type. Both of these graphs can be found in Appendices P and Q. On the weekdays, the most popular end station is Huntington Ave at Mass Art and Roxbury Crossing T Stop. On the weekends, the most popular end stations are Roxbury Crossing and Christian Science Plaza.

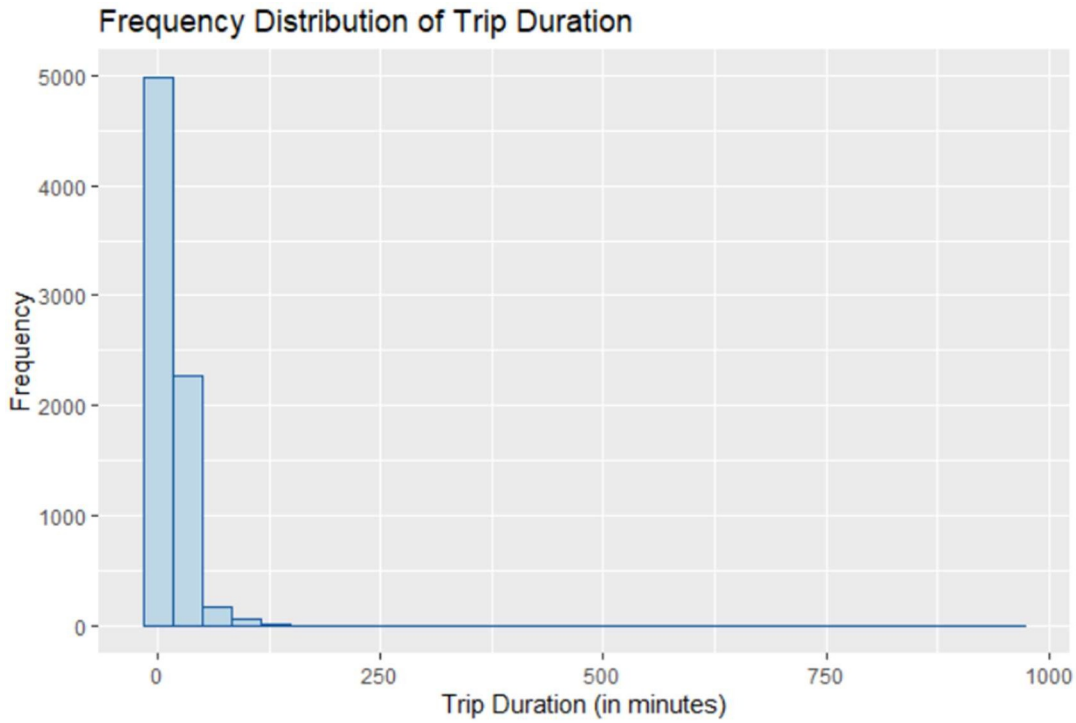
Overall, Blue Bike users will primarily end their rides at four locations, which are: “Huntington Ave at Mass Art”, “Roxbury Crossing T Stop”, “Christian Science Plaza”, and “Brigham Circle – Francis” on both weekdays and weekends. That said, there are still several noticeable differences between the ride trends for weekdays and weekends. On the weekdays, Blue Bike users mostly drop off their bikes at “Northeastern University - North Parking Lot” and “Boylston St at Jersey St”, while, on the weekends, they would drop off at “Landmark Center” and “Brigham Circle” the most. Also, the proportion of users who are non-subscribers (being Customers) is higher on the weekends. The ratio between subscriber and customer is much higher on weekdays compared to the weekends, and most users on weekdays are part of the Subscribers group.

☔

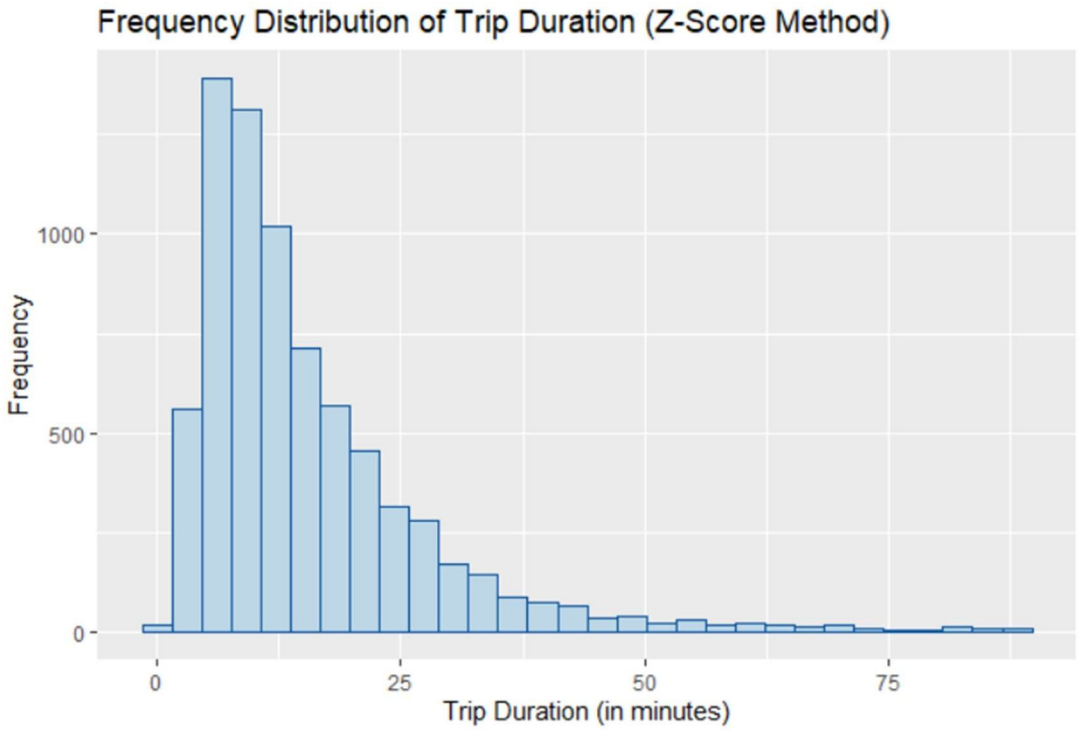
For even more advanced analyses, we can use the longitude and latitude information to geocode the locations of our start/end stations (perhaps with the ggmap() package), and we can map out the start/end stations, while possibly drawing lines to join the start stations on which end stations riders would drop off their bikes most often. We can also use the postal codes of Blue Bike users (after removing all the missing, null, and incorrect values) to assess whether there is a link between a user’s postal code and their likelihood of being a Customer or Subscriber.

Appendix

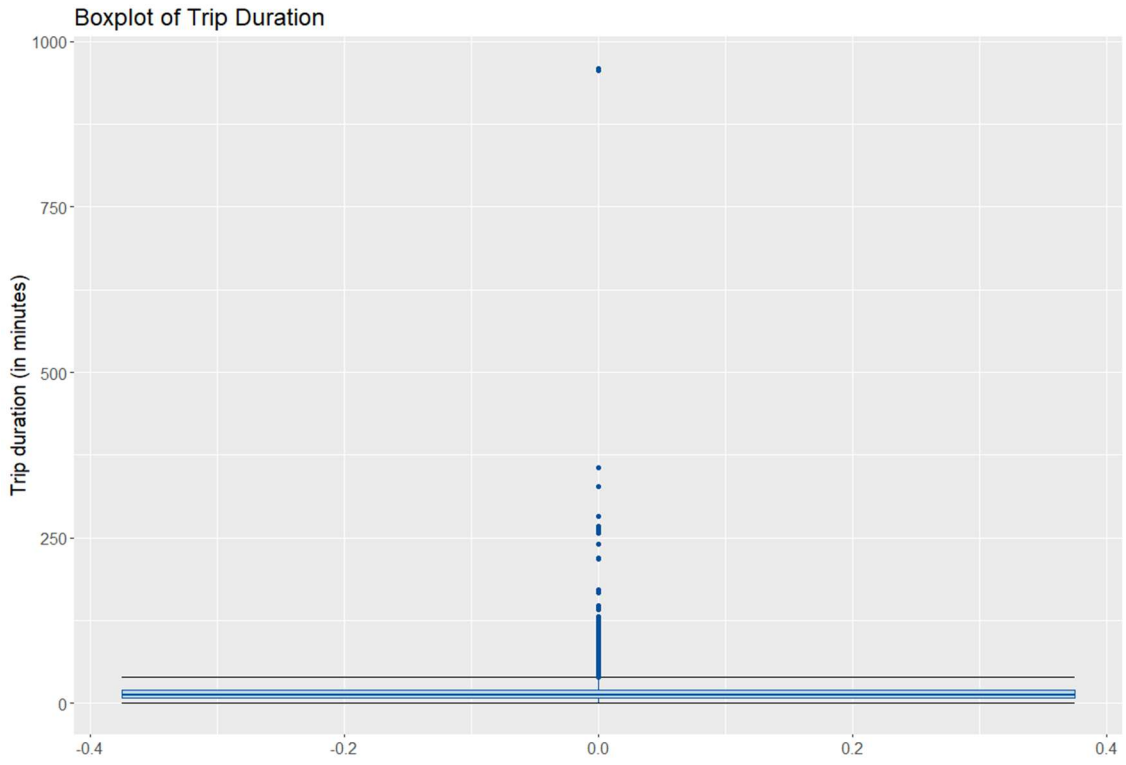
Appendix A



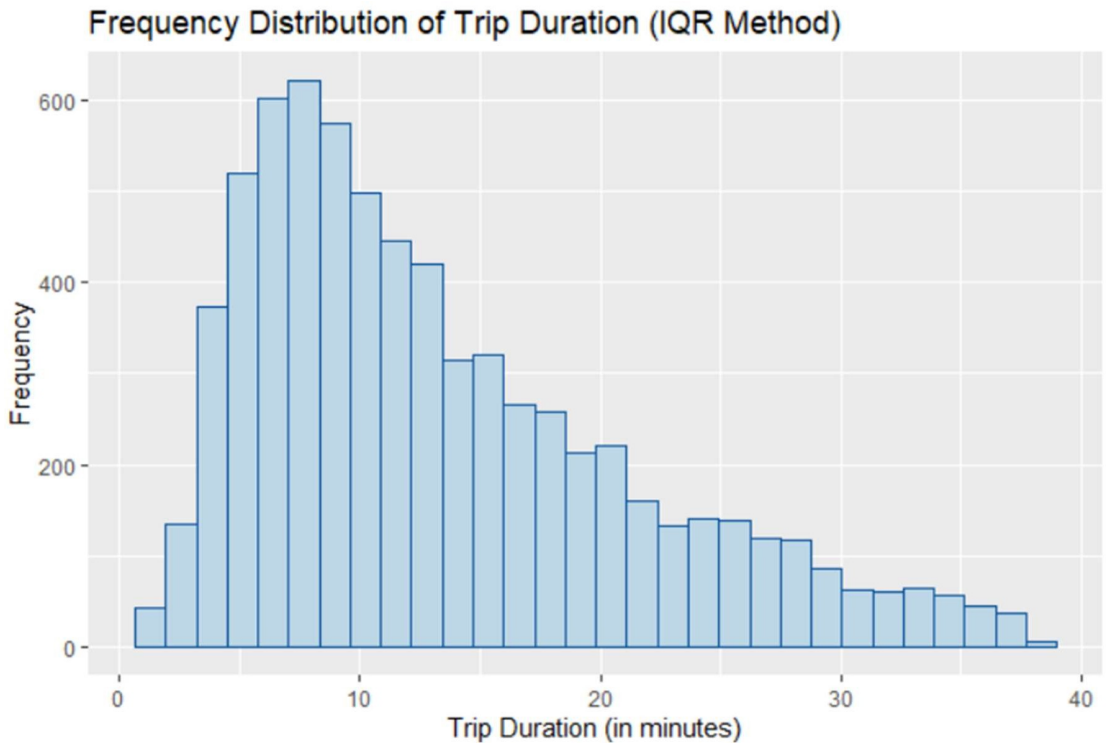
Appendix B



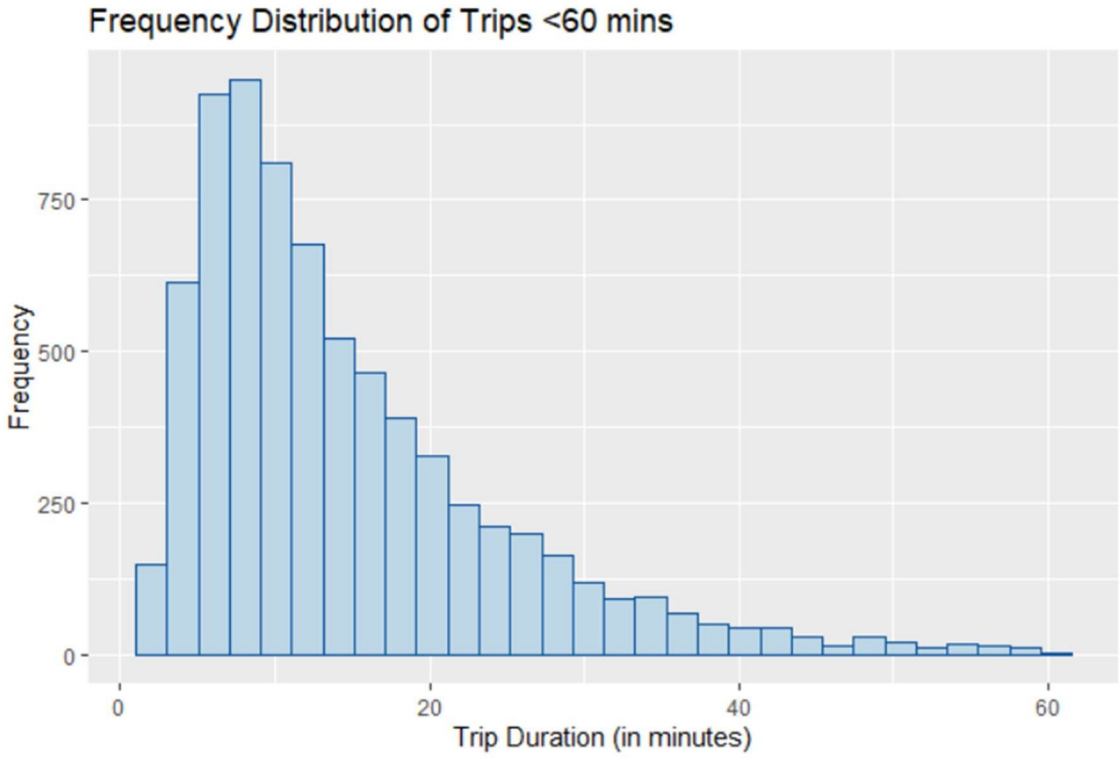
Appendix C



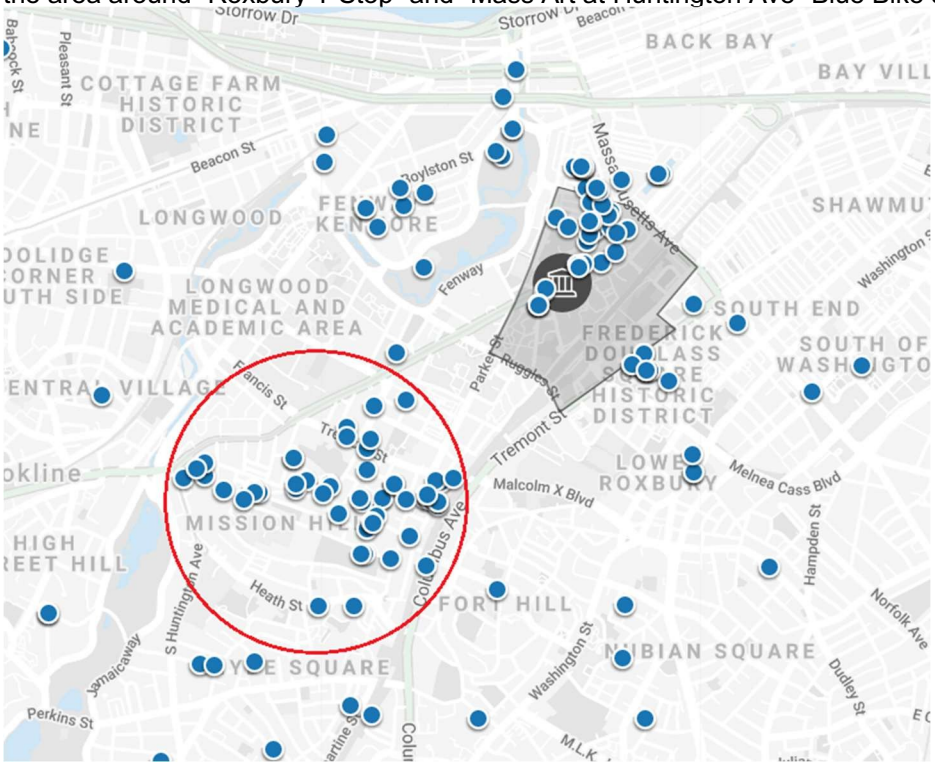
Appendix D



Appendix E



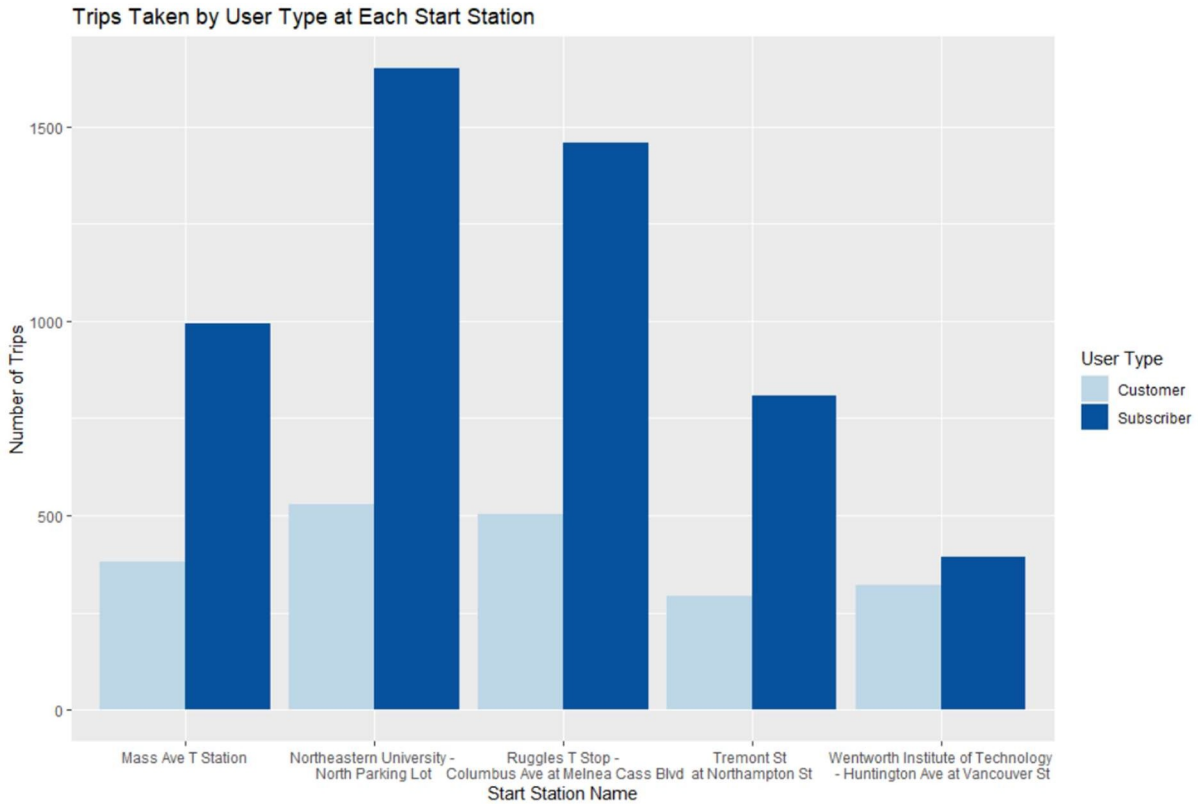
Appendix F – Apartments near Northeastern University (www.aptsearch.northeastern.edu/housing), red circle indicates the area around “Roxbury T Stop” and “Mass Art at Huntington Ave” Blue Bike stations.



Appendix G

	Mass Ave T Station	Northeastern University - North Parking Lot	Ruggles T Stop - Columbus Ave at Melnea Cass Blvd	Tremont St at Northampton St	Wentworth Institute of Technology - Huntington Ave at Vancouver St
Customer	380	529	502	292	322
Subscriber	994	1649	1457	809	392

Appendix H

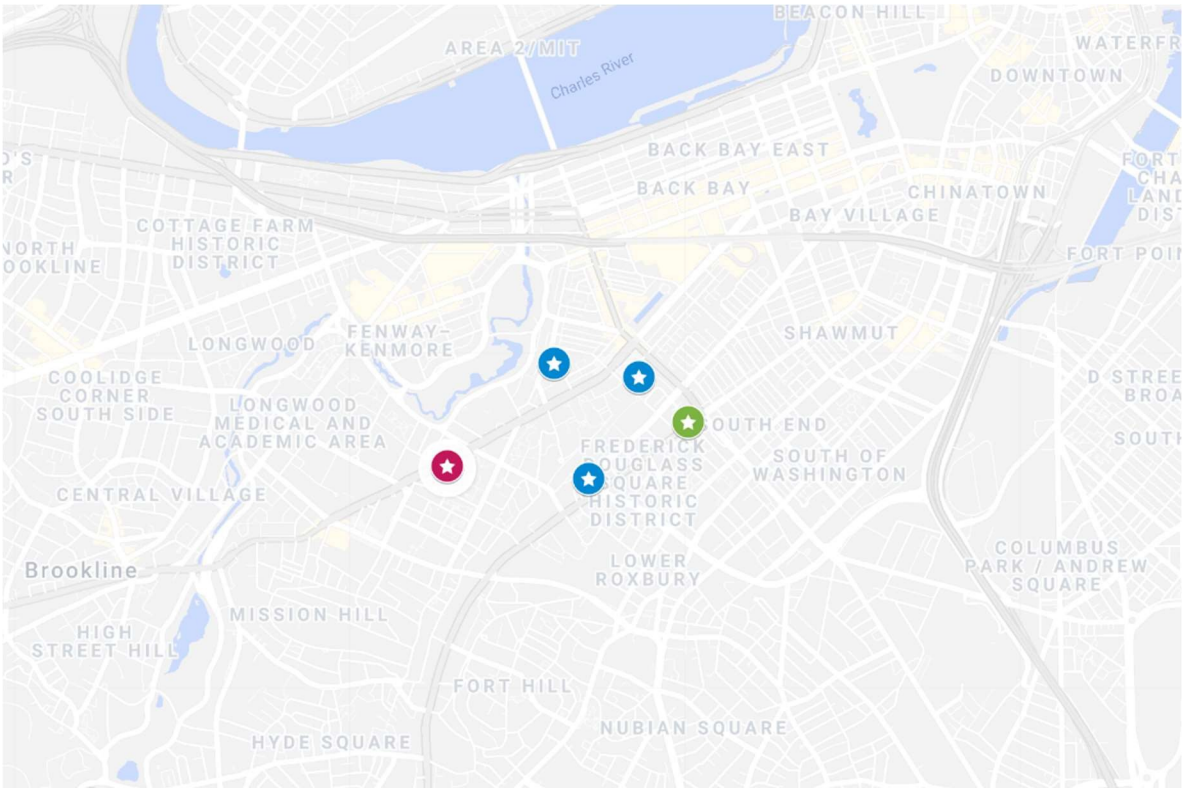


Appendix I

Average Trip Duration for Start Station

Start Station Names	Average trip duration (mins)
Mass Ave T Station	15.068
Northeastern University - North Parking Lot	14.853
Ruggles T Stop - Columbus Ave at Melnea Cass Blvd	14.070
Tremont St at Northampton St	12.934
Wentworth Institute of Technology - Huntington Ave at Vancouver St	17.480

Appendix J – Map showing the location of Blue Bikes start locations. Red star indicates the station with the longest trip duration ("Wentworth Institute of Technology"), and green star indicates the station with the shortest trip duration ("Tremont St at Northampton"). Blue stars are for the remaining three stations.



Appendix K

Average Trip Duration for User Type

	User Type	Average Trip Duration (mins)
1	Customer	19.101
2	Subscriber	12.952

Appendix L

Test of Dependence Between Subscriber and Start Station

	Start Station Name	Subscriber Probability	Dependence
1	Mass Ave T Station	0.723	TRUE, it != 0.724
2	Northeastern University - North Parking Lot	0.757	TRUE, it != 0.724
3	Ruggles T Stop - Columbus Ave at Melnea Cass Blvd	0.744	TRUE, it != 0.724
4	Tremont St at Northampton St	0.735	TRUE, it != 0.724
5	Wentworth Institute of Technology - Huntington Ave at Vancouver St	0.549	TRUE, it != 0.724

Appendix M – Results from 25 runs of a sampling loop to test the hypothesis "is $\bar{X}$(week4) > $\bar{X}$(null)" where the t-critical stat is 1.6604.

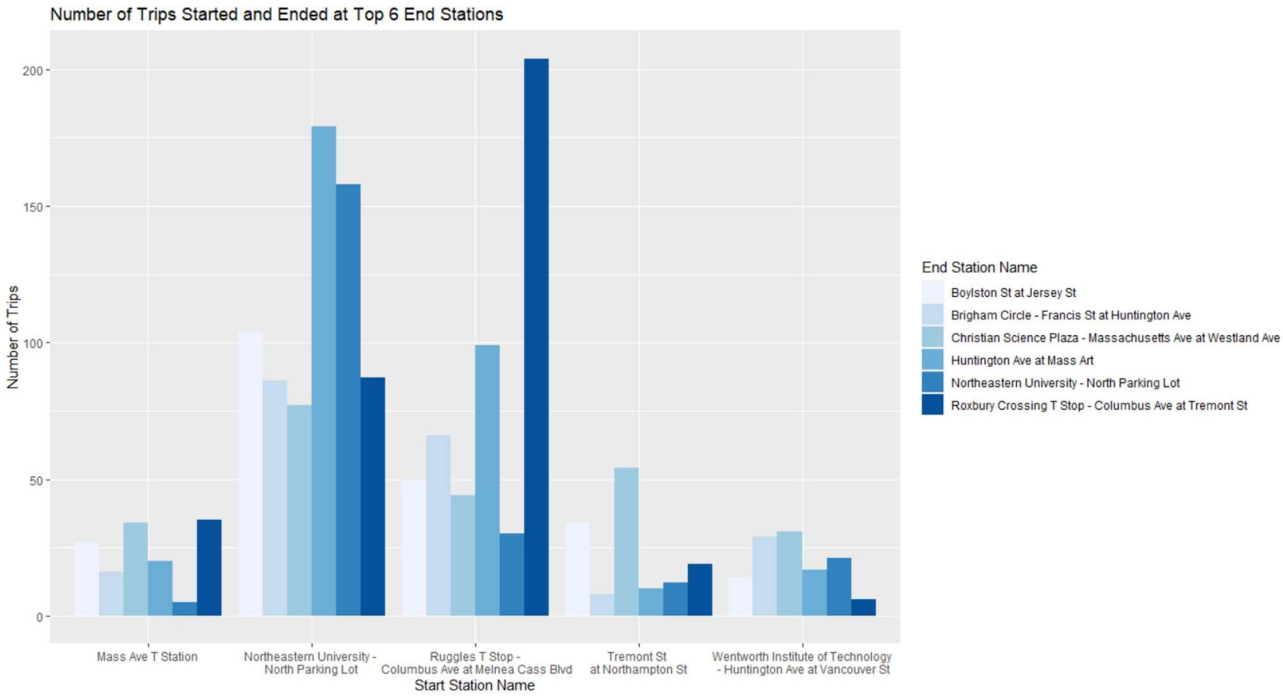
	Sample mean	Sample St.Dev	Sample T-stat	Rejection Decision
1	16.02650	10.833384	1.3568826	Don't Reject H0 (p<0.05)
2	15.38167	9.440202	0.8740594	Don't Reject H0 (p<0.05)
3	15.38367	10.536654	0.7850023	Don't Reject H0 (p<0.05)
4	16.06383	10.930016	1.3790432	Don't Reject H0 (p<0.05)
5	13.25567	9.531682	-1.3647857	Don't Reject H0 (p<0.05)
6	16.17833	11.378572	1.4253075	Don't Reject H0 (p<0.05)
7	14.90267	9.308621	0.3718378	Don't Reject H0 (p<0.05)
8	14.67833	9.375509	0.1299091	Don't Reject H0 (p<0.05)
9	14.85717	10.797122	0.2784351	Don't Reject H0 (p<0.05)
10	14.37883	9.878753	-0.1798846	Don't Reject H0 (p<0.05)
11	17.38583	12.398213	2.2820196	Reject H0 (p>0.05)
12	15.30033	10.832418	0.6866393	Don't Reject H0 (p<0.05)
13	14.70233	10.321534	0.1412546	Don't Reject H0 (p<0.05)
14	16.74800	11.779977	1.8603288	Reject H0 (p>0.05)
15	14.82917	10.094215	0.2700851	Don't Reject H0 (p<0.05)
16	16.33900	11.682758	1.5257211	Don't Reject H0 (p<0.05)
17	13.73317	8.226586	-1.0008651	Don't Reject H0 (p<0.05)
18	16.53100	9.631278	2.0500531	Reject H0 (p>0.05)
19	14.22550	10.377909	-0.3189823	Don't Reject H0 (p<0.05)
20	15.42217	9.928398	0.8718725	Don't Reject H0 (p<0.05)
21	14.88133	9.739804	0.3334733	Don't Reject H0 (p<0.05)
22	16.89767	13.368546	1.7512224	Reject H0 (p>0.05)
23	14.42200	8.604899	-0.1563492	Don't Reject H0 (p<0.05)
24	13.27333	8.621969	-1.4882953	Don't Reject H0 (p<0.05)
25	15.89233	9.887697	1.3509682	Don't Reject H0 (p<0.05)

Appendix N

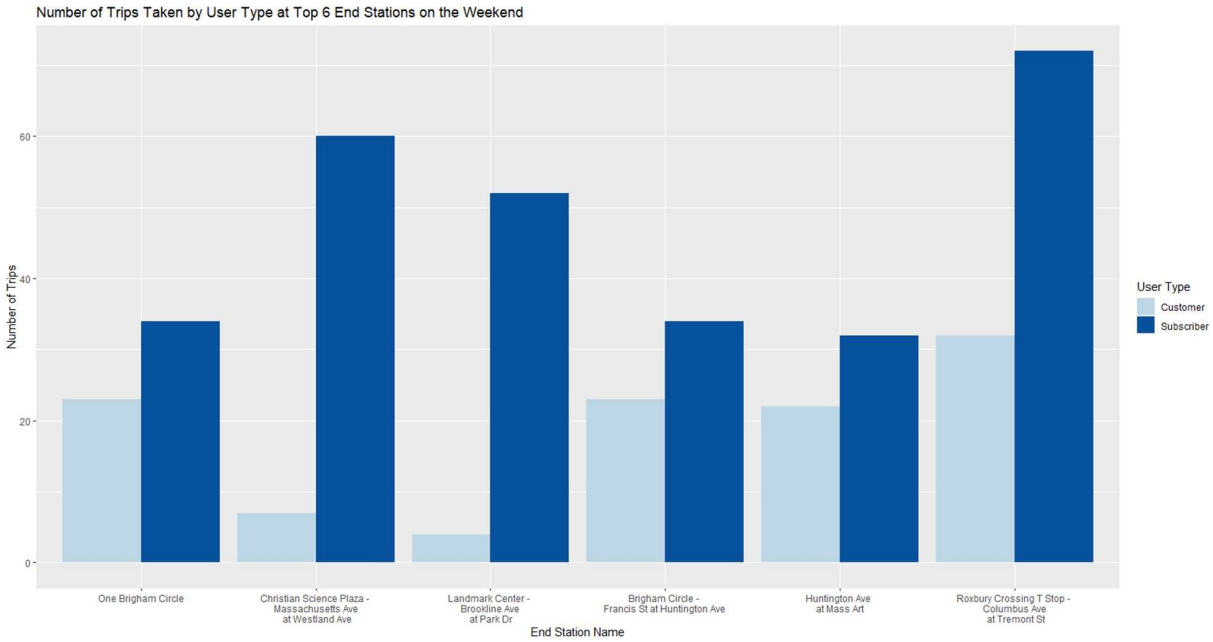
Contingency Table Between Start Stations and the Top 6 End Stations

	Mass Ave T Station	Northeastern University - North Parking Lot	Ruggles T Stop - Columbus Ave at Melnea Cass Blvd	Tremont St at Northampton St	Wentworth Institute of Technology - Huntington Ave at Vancouver St
Boylston St at Jersey St	26	102	50	34	13
Brigham Circle - Francis St at Huntington Ave	16	85	66	8	29
Christian Science Plaza - Massachusetts Ave at Westland Ave	34	74	42	53	31
Huntington Ave at Mass Art	20	179	99	10	10
Northeastern University - North Parking Lot	5	141	29	12	21
Roxbury Crossing T Stop - Columbus Ave at Tremont St	35	87	204	18	6

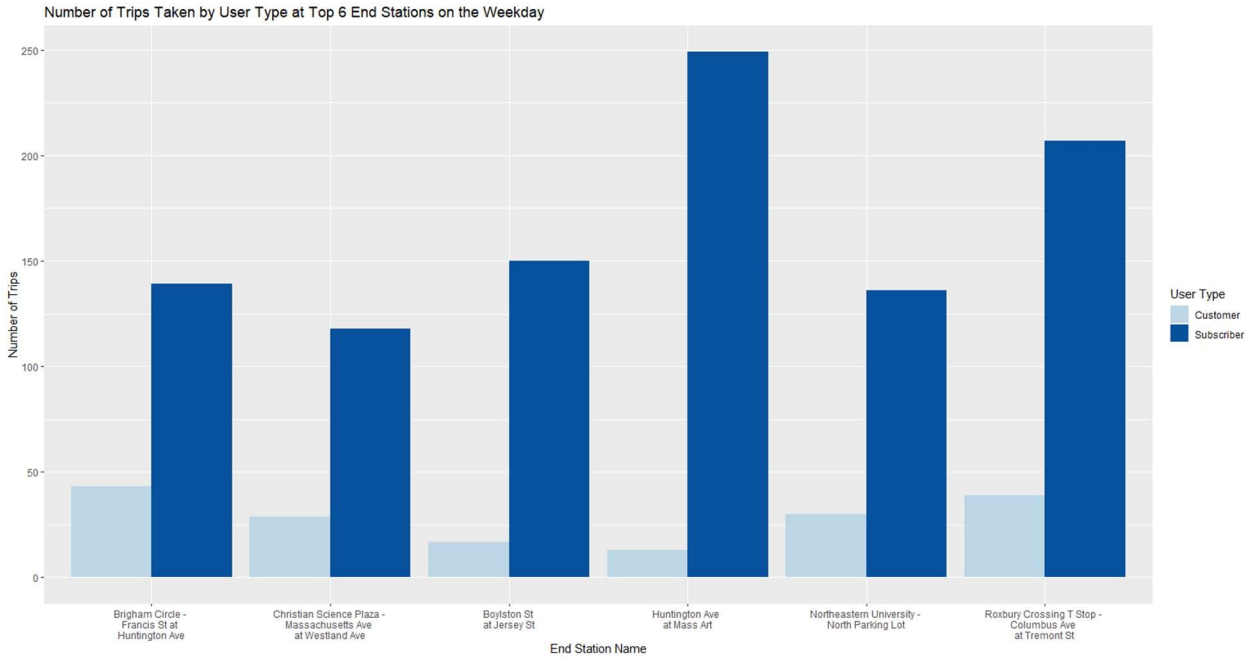
Appendix O



Appendix P



Appendix Q


Number of Trips Taken by User Type at Top 6 End Stations on the Weekday
Number of Trips
250
200
150
100
50
0
Brigham Circle - Francis St at Huntington Ave
Christian Science Plaza - Massachusetts Ave at Westland Ave
Boylston St at Jersey St
Huntington Ave at Mass Art
Northeastern University - North Parking Lot
Roxbury Crossing T Stop - Columbus Ave at Tremont St
End Station Name
User Type
Customer
Subscriber