Foundations of Data Analytics
Problem Set 1

Group members: *Abhinaya Muthukrishnan*
Pooja Karumanchi
Praveen Sai KumarVJ

1. **Missing values:** 16 Missing values found in "Postal Code"
Null Values: 622 Null Values found in variable "Postal Code"
Incorrect Values: Incorrect values are found in the variable "Postal code"

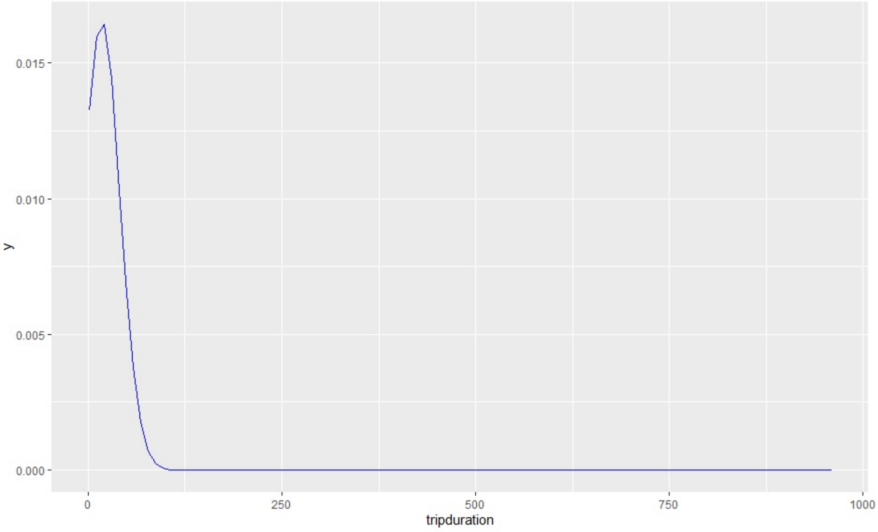
The variable postal code is the one that contains null, missing, and incorrect values. The values of the variable postal code are not used to perform any operations later in our analysis and hence has no effect on the output. We can approach this by replacing the incorrect values with most common values or looking through the data we can assign corresponding values based on previous data.

2. **Outlier using z – score:**
For z score method, data point is identified as outlier if it lies beyond 3 standard deviations from the mean. The cut off values using z-score method are found to be:
   - Left cutoff = $\mu - 3\sigma$ = -55.0964
   - Right cutoff = $\mu + 3\sigma$ =89.3302

If any datapoint is lesser than -55.0964 or greater than 89.3302 it's considered an outlier under this method.

To find if this method is well-suited, we need to find the underlying distribution. For this we plot the probability density curve for the variable tripduration.

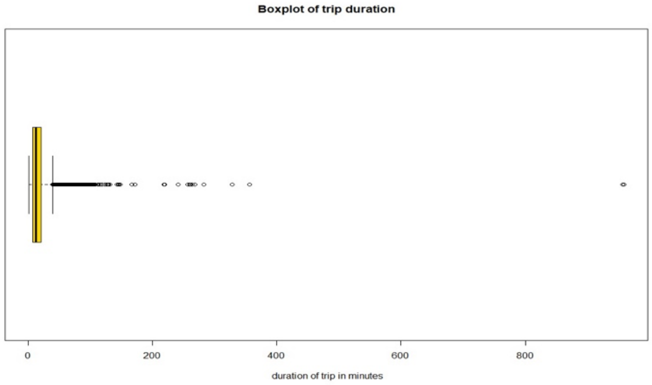
The curve is positively skewed hence the variable is not symmetric.
Thus, the z score method is not appropriate to find the outliers for the tripduration data.



3. **Outlier using IQR Method**:

   We find the Q1 and Q3 Values and use the formula given as
   - IQR=Q3-Q1=12.63
   - Leftcutoff is Q1-1.5(IQR)= -11.483
   - Rightcutoff is Q3+1.5(IQR)= 39.05

If any datapoint is lesser than -11.483 or greater than 39.05 it's considered an outlier under this method. From question 2(Refer to the graph posted above) we know that the underlying distribution is positively skewed. This is also confirmed by the construction of a boxplot. Hence the IQR method is better suited than the Z score method to find outliers in tripduration data. From the box plot we see that there are large number of outliers on the right hand side.



Subsetting data to take only rides with duration one hour or less:

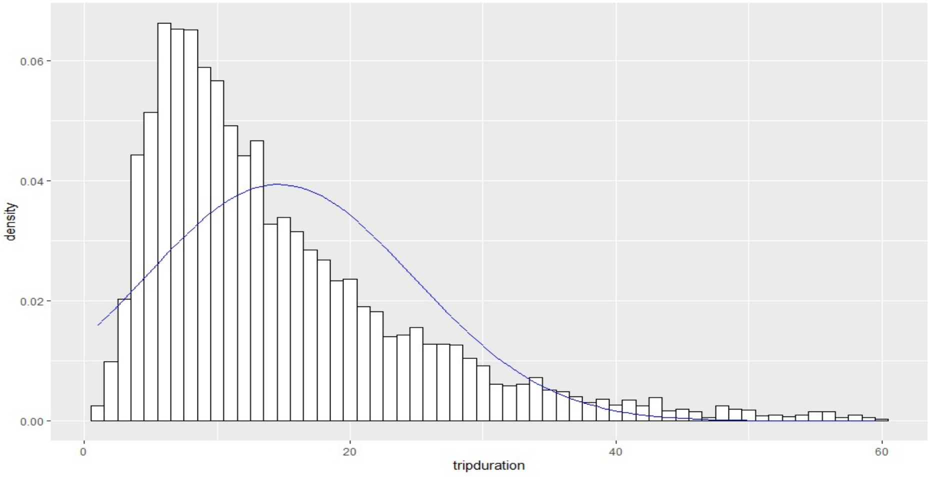
We use the function subset in R to subset a data with specifications that tripduration variable takes values lesser than or equal to 60minutes.The resulting new data frame with rides of up to 60 minutes is named as **"duration_60min"**

4. No of rides in the subsetted data frame = 7326

5. **Histogram and shape of distribution for tripduration variable:**

   From the histogram we see that the right end tail is long indicating positive skewness' of the tripduration variable. Also, for such a distribution the mean will lie to the right of the median value.

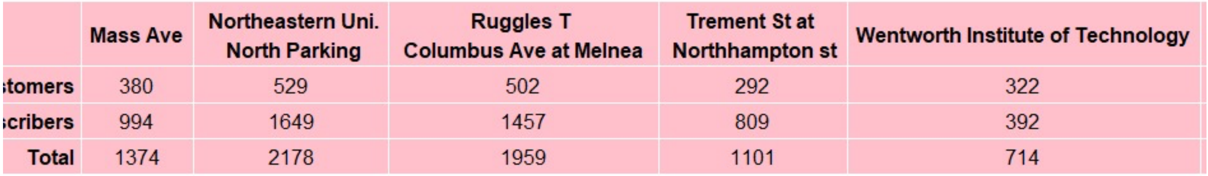
   Reason: This shape of the distribution is because there are smaller number of large trip duration values in the data set. Conversely, we can say that the tripduration variable has greater number of small values. Thus, the center of the distribution is more towards the left with a long right end tail.



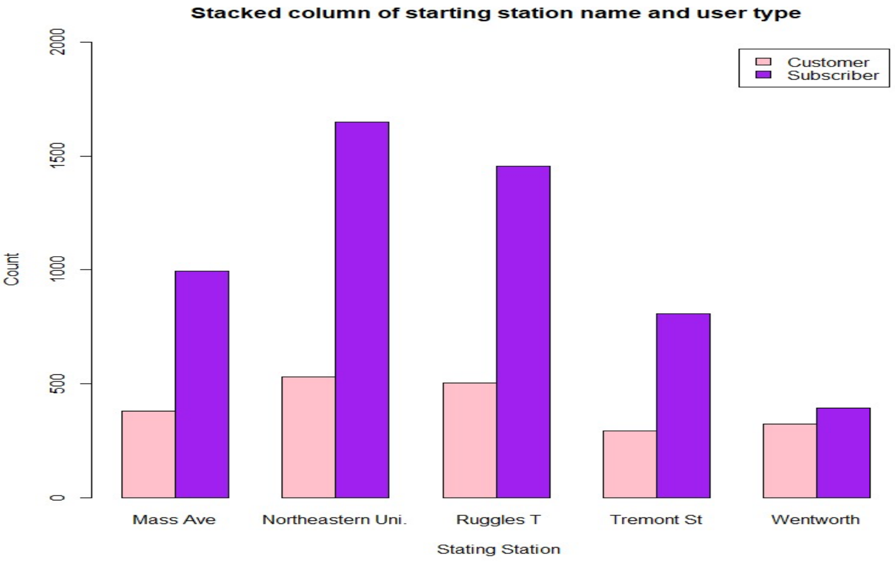
6. Contingency table and Clustered Column chart:

From the table and the chart, the following patterns are found

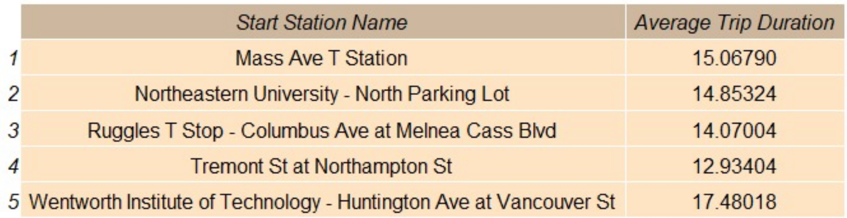
- Across all start stations, the number of customers is consistently lower than the number of subscribers.
- The number of riders who are customers and subscribers are both highest for the start station northeastern university-North parking lot compared to other start stations



	Mass Ave	Northeastern Uni. North Parking	Ruggles T Columbus Ave at Melnea	Trement St at Northhampton st	Wentworth Institute of Technology
stomers	380	529	502	292	322
scribers	994	1649	1457	809	392
Total	1374	2178	1959	1101	714

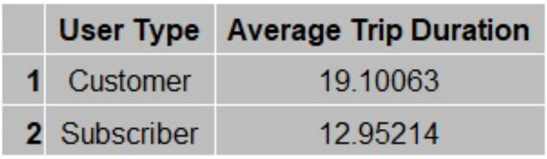


7. From the table we observe that
    - The average trip duration is highest for rides starting from Wentworth institute of technology – Huntington Ave at Vancouver St station.
    - The average Trip duration is lowest for rides taken from Tremont St at North hunter St station
    - For other three stations the average tripduration has very little variation.



	Start Station Name	*Average Trip Duration*
1	Mass Ave T Station	15.06790
2	Northeastern University - North Parking Lot	14.85324
3	Ruggles T Stop - Columbus Ave at Melnea Cass Blvd	14.07004
4	Tremont St at Northampton St	12.93404
5	Wentworth Institute of Technology - Huntington Ave at Vancouver St	17.48018

8. This table shows that the average trip duration for user type customers is higher than for subscribers.



	User Type	**Average Trip Duration**
1	Customer	19.10063
2	Subscriber	12.95214

9. The Probability that a randomly selected ride was taken by a subscriber is given by unconditional probability of subscriber
    - If event S = Ride taken by a subscriber
    - P(S) = 0.7235872 ~ 0.724

10. No, the probability of selecting a subscriber is not independent of start station

This can be proved by comparing the conditional and unconditional probabilities of selecting a subscriber

- Event S= Ride taken by subscriber
- Event M = Ride taken from Mass Ave T station
- Event N = Ride taken from NEU North parking lot station
- P(S) = 0.7235872 ~ 0.724
- P(S|M) = Count of riders who are subscribers and start from Mass Ave/count of rides starting from Mass Ave = 0.7234 ~ 0.723
- P(S|N) = Count of riders who are subscribers and start from Northeastern /count of rides starting from Northeastern = 0.7571 ~ 0.757
- Since P(S|M) is not equal to P(S)

- P(S|N) is not equal to P(S)

It is proved that probability of selecting a subscriber is not independent of start station.

11. Instead of taking repeated samples we can use population parameters to find the sampling distribution of sample mean and sample proportion.

a. μ=14.652
n=50
σ=10.116
E(X bar)= μ=14.652
SE(X bar)= σ/sqrt(n)=10.116/sqrt(50)=1.430

b. p=0.276
E(p bar)=p=0.276
SE(p bar)=sqrt(p*1-p/n)
SE(p bar)=sqrt(0.276*(1-0276)/50)=0.0632

12. A random sample of 50 rides from the data is created using the sample function in R and stored as sample.df.

For the data sample.df:
x_bar=14.972
s =10.421
alpha = 0.05
n=50
$t_{alpha/2,df} = t_{0.025,49} = 2.00975$

95% confidence interval is given as:
= x_bar +/- t $_{alpha/2,df}$*s/sqrt(n)
= [x_bar - t $_{alpha/2,df}$*s/sqrt(n) , x_bar + t $_{alpha/2,df}$*s/sqrt(n)]
= [14.972 - 2.00975* (10.421/ 7.071) , 14.972 + 2.00975* (10.421/ 7.071)]
= [12.011 , 17.934]

- The true population mean is μ=14.652.
- YES, the true population mean does lie in the 95% confidence interval constructed using the random sample of 50 rides.

13. For the data sample.df:
p_bar=0.28
alpha = 0.05
n=50
$Z_{alpha/2} = Z_{0.025} = 1.96$

95% confidence interval is given as:
= p_bar +/- Z $_{alpha/2}$*sqrt(p_bar*(1- p_bar)/n)
= [p_bar - Z $_{alpha/2}$* sqrt(p_bar*(1- p_bar)/n) , x_bar + Z $_{alpha/2}$* sqrt(p_bar*(1- p_bar)/n)]
= [0.28 – 0.1246 , 0.28 + 0.1246]
= [0.156 , 0.404]

- The population proportion is p=0.276
- YES, the 95% confidence interval constructed using the random sample of 50 rides does include the population proportion.

14.

(a) $\mu_0 = 14.55654$

Null Hypothesis: The average trip length during the final week of August is lesser than or equal to the average trip length during the 1st week of August
H_0: $\mu <= \mu_0$

Alternative Hypothesis: The average trip length during the final week of August is greater than the average trip length during the 1st week of August
H_1: $\mu > \mu_0$

With set.seed (999) the sample of 100 rides is taken from week 4 data and stored as sample_withseed. We have done this using a for loop to loop 100 times.

For this sample;

x_bar = 16.53983
s = 10.77362
n= 100

t statistic = (x_bar - μ_0) / (s/sqrt(n))
= (16.539 – 14.557) / (10.774/10)
= 1.840882 ~ 1.841
t- critical = $t_{0.05, 99}$ = 1.660391 ~ 1.660

Since, **t statistic > t critical** value we "**Reject the null hypothesis**" and we conclude that there is enough evidence to say that the average trip length during the 4th week of August is greater than the average trip length during the 1st week of August.

(b) When taking 100 different samples and repeating the hypothesis testing procedure for each one, we get the following result in R.

```
[1] 1.356883
[1] "Do not Reject Null Hypothesis"
[1] 0.8740594
[1] "Do not Reject Null Hypothesis"
[1] 0.7850023
[1] "Do not Reject Null Hypothesis"
[1] 1.379043
[1] "Do not Reject Null Hypothesis"
[1] -1.364786
[1] "Do not Reject Null Hypothesis"
[1] 1.425308
[1] "Do not Reject Null Hypothesis"
[1] 0.3718378
[1] "Do not Reject Null Hypothesis"
[1] 0.1299091
```

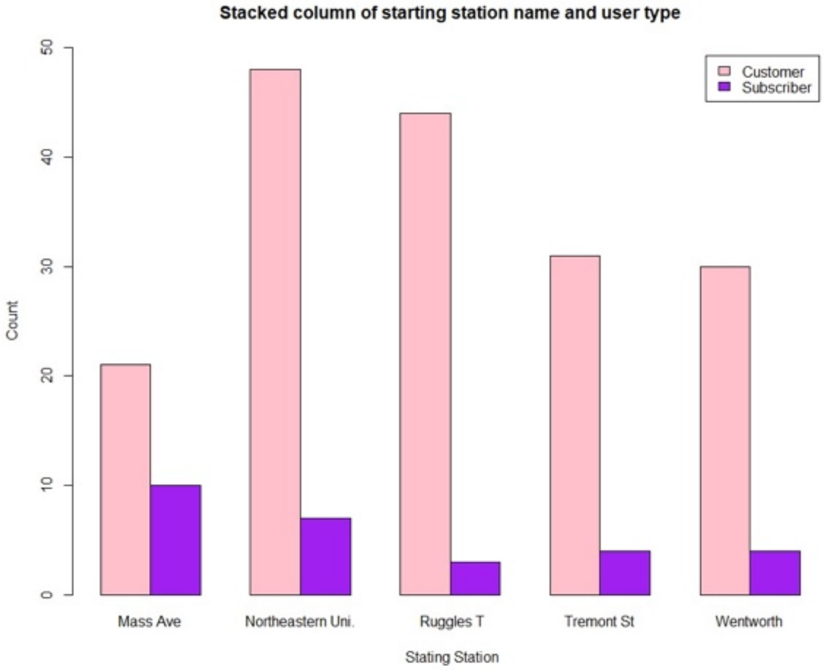
```
[1] "Do not Reject Null Hypothesis"
[1] 0.2784351
[1] "Do not Reject Null Hypothesis"
[1] -0.1798846
[1] "Do not Reject Null Hypothesis"
[1] 2.28202
[1] "Reject Null Hypothesis"
[1] 0.6866393
[1] "Do not Reject Null Hypothesis"
[1] 0.1412546
[1] "Do not Reject Null Hypothesis"
[1] 1.860329
[1] "Reject Null Hypothesis"
[1] 0.2700851
[1] "Do not Reject Null Hypothesis"
[1] 1.525721
[1] "Do not Reject Null Hypothesis"
[1] -1.000865
[1] "Do not Reject Null Hypothesis"
[1] 2.050053
[1] "Reject Null Hypothesis"
[1] -0.3189823
[1] "Do not Reject Null Hypothesis"
[1] 0.8718725
[1] "Do not Reject Null Hypothesis"
[1] 0.3334733
[1] "Do not Reject Null Hypothesis"
[1] 1.751222
[1] "Reject Null Hypothesis"
[1] -0.1563492
[1] "Do not Reject Null Hypothesis"
[1] -1.488295
[1] "Do not Reject Null Hypothesis"
………
```

Form the results we find that the conclusion is mixed when we take 100 different samples. For **majority** of samples, we conclude **“Do not reject the null hypothesis”**. This contrasts with our conclusion in the previous section when we take only a single sample.

15. Analysis 1

We can comfortably assume that longer trips are usually leisure trips or adventure trips and will last longer than 60 minutes. One interesting analysis will be to compare user type for different start stations when trip duration is greater than one hour.

- For this we first subset the data to include points with tripduration greater than 60 minutes
- The bar chart shows that when only trip duration higher than 1 hour is taken, the number of customers is considerably higher than subscribers across all stations.
- This is in line with our intuition that longer adventure trips won’t be taken often and hence it would not require a subscription. Hence most riders who take trips for greater than 1 hour will purchase a one-time pass and hence their user type will be customers.



Analysis 2:

We can construct a contingency table between “start.sation.name” and “end.station.name” using the original data set . Using the intersection values from the table we can find the route that is most popular among riders. We can also find the conditional probabilities that given a rider starts his ride from Northeastern University-North parking lot, the probability of the rider reaching different end stations. This will give the most popular destinations for riders taking blue bikes from Northeastern University.