Bike Rentals : Final Presentation

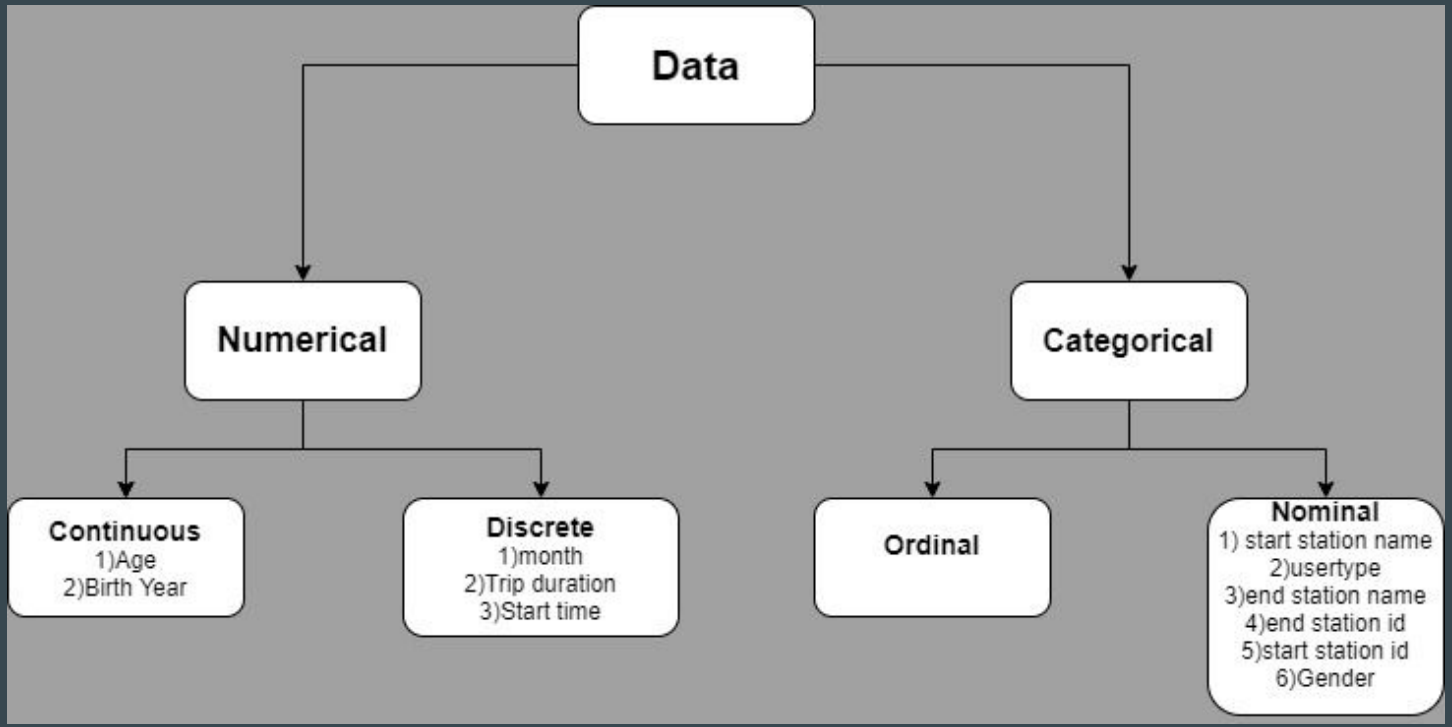
Presented On : 26/05/2021

Presented By:
Sarvesh Meenowa
Amr Mohamed
Neha Devi Shakya
Khushi Chitra Uday

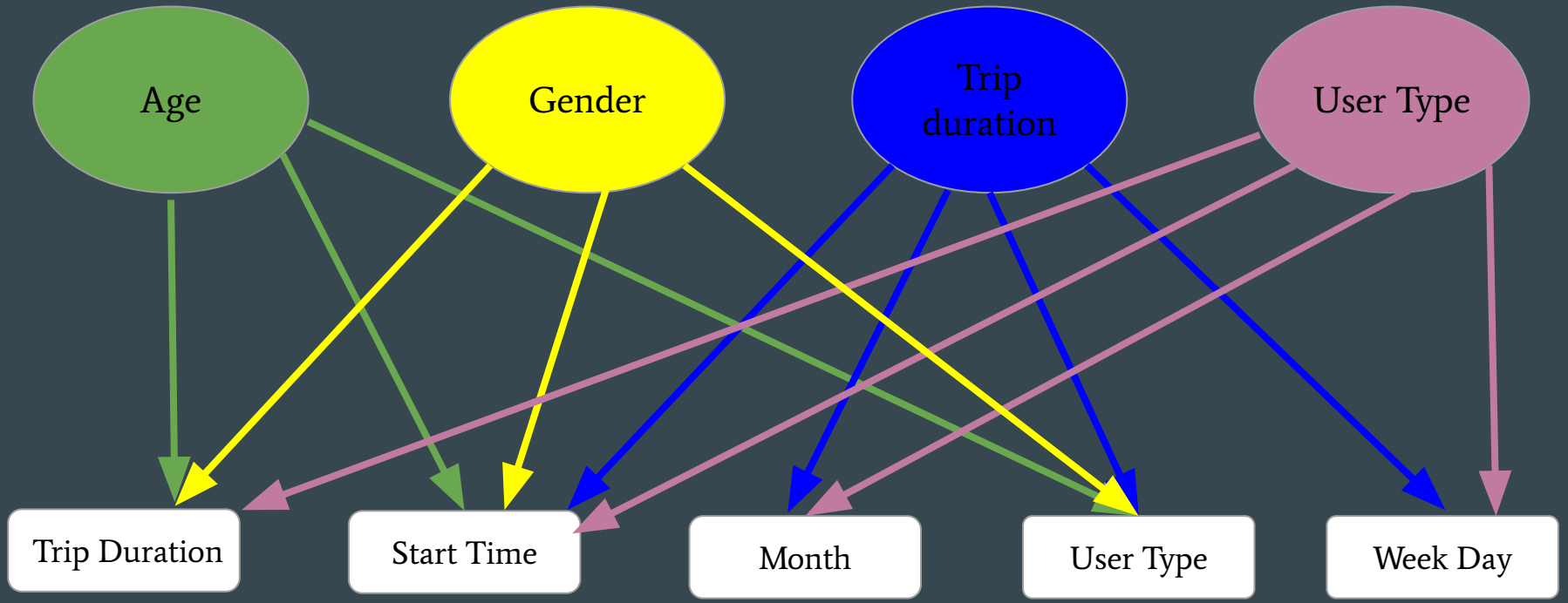
Overview of the Dataset

- Bike rental from San Francisco in 2019
- 2522771 observations with 12 variables

Classification of Data

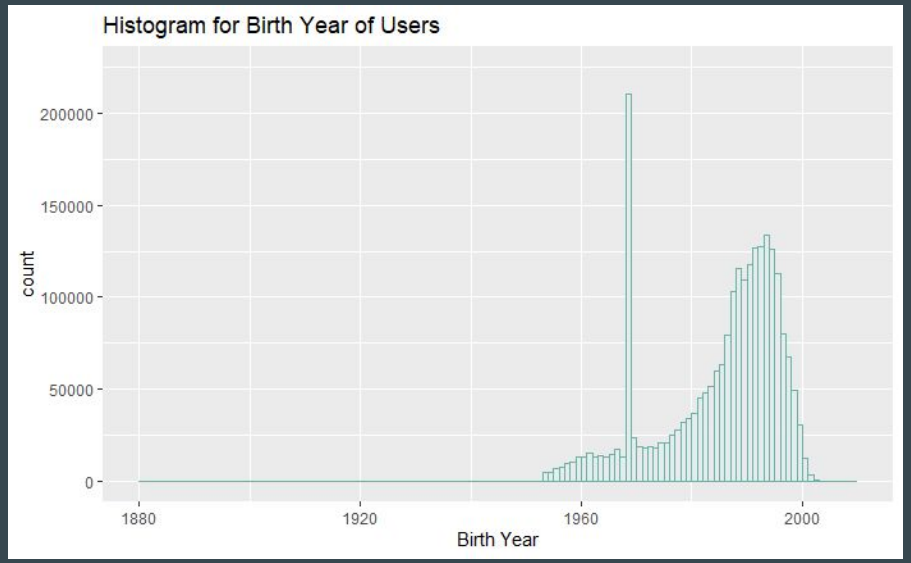


The preliminary different relations between the columns

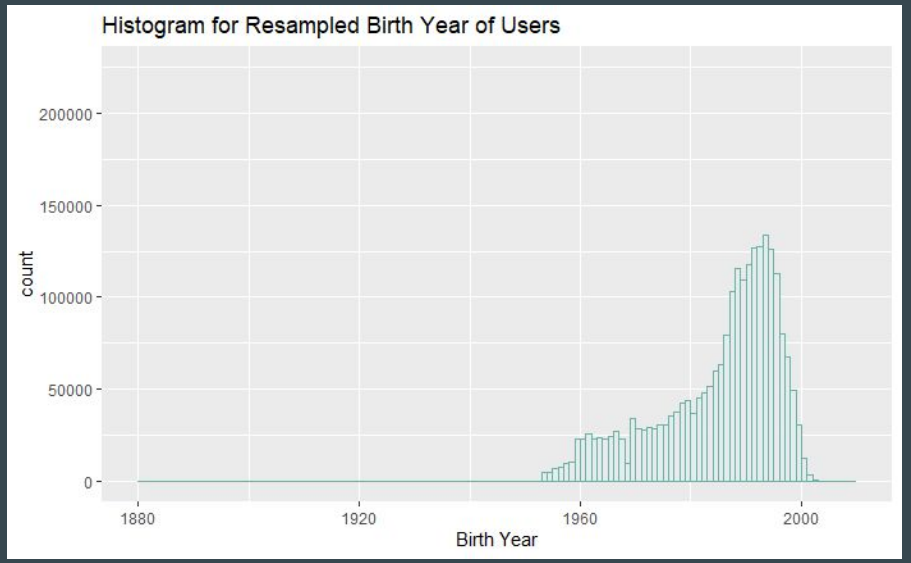


Data Cleaning

Before and After Resampling the Birth Year

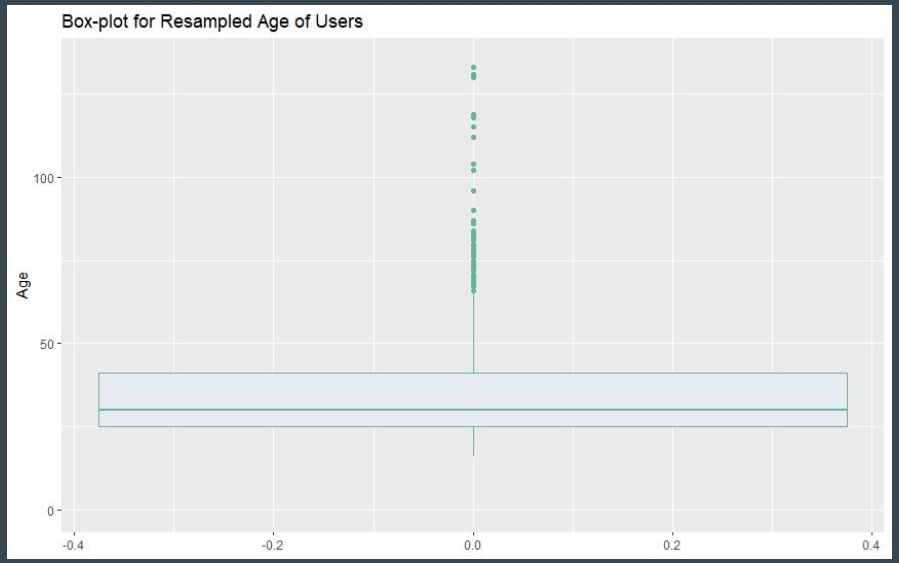


Histogram for the variable Birth Year of the User before Resampling the data



Histogram for the variable Birth Year of the User after Resampling the data

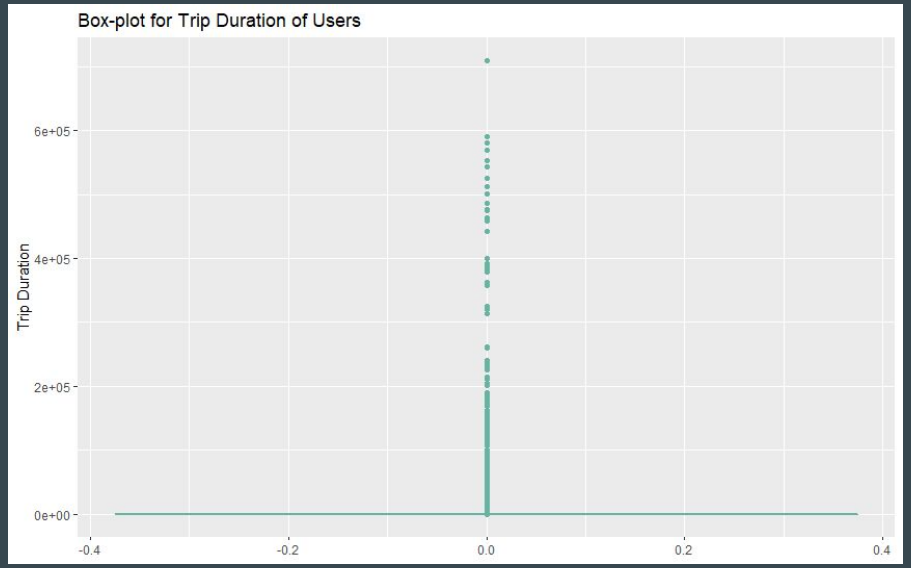
Box-Plots Before and After Removing Outliers in Age



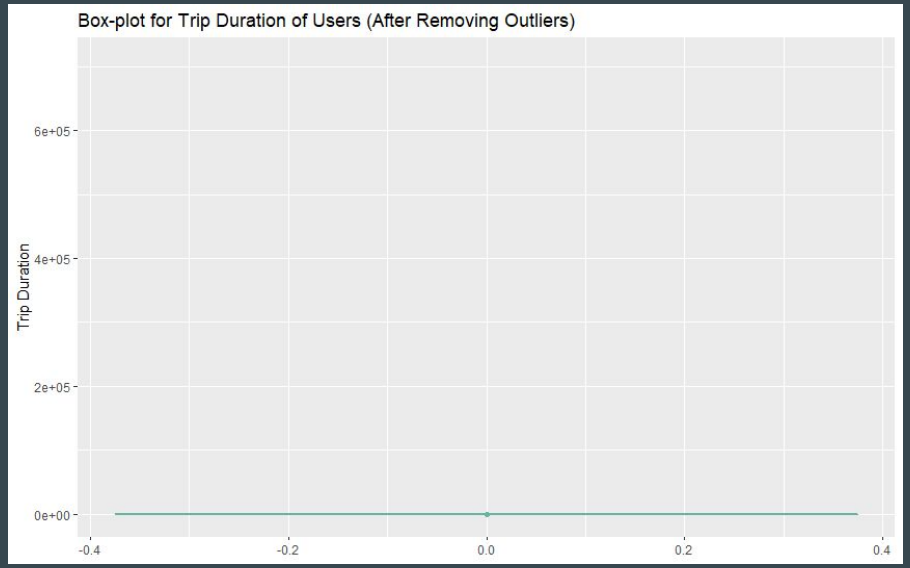
Box Plot for the Resampled Age column Before Removing the Outliers

Box Plot for the Resampled Age column After Removing the Outliers

Box-Plots Before and After Removing Outliers in Trip Durations



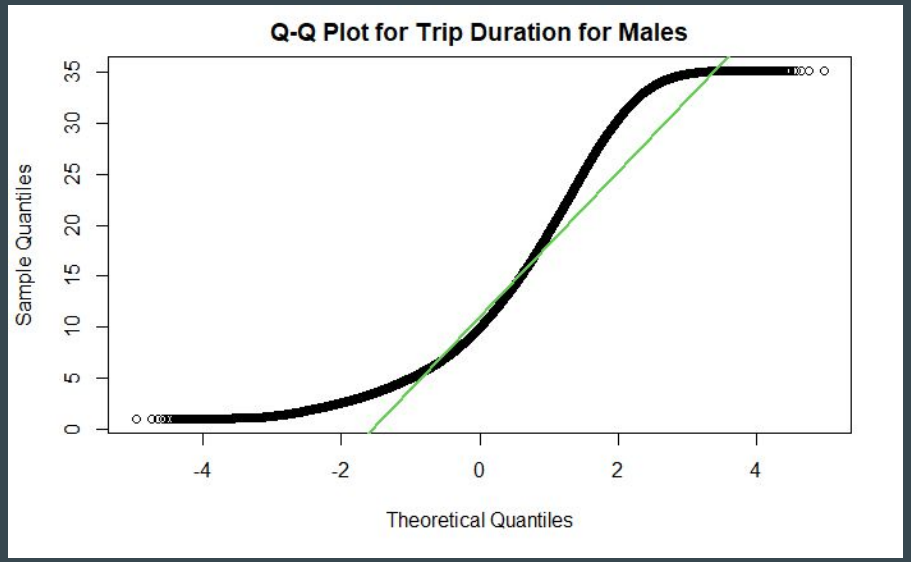
Box Plot for the Trip Duration column Before Removing the Outliers

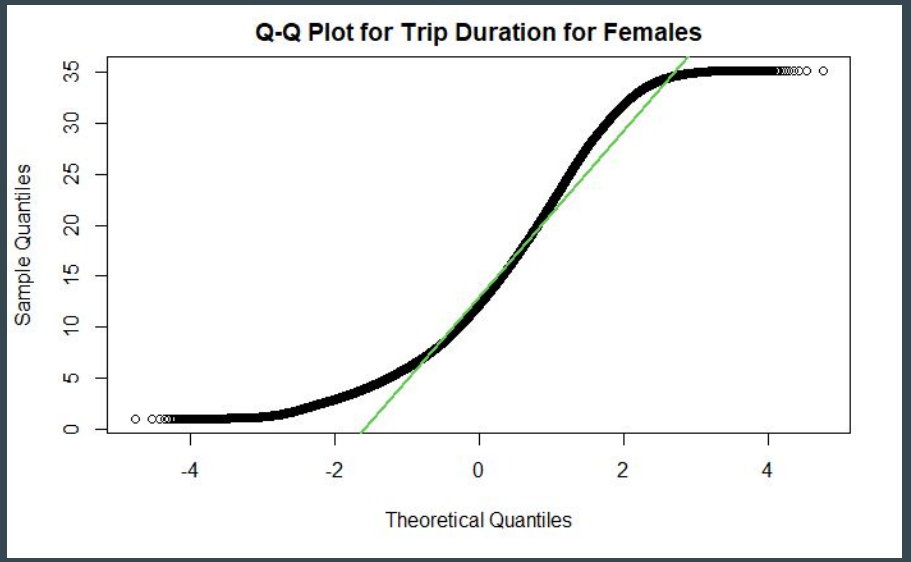


Box Plot for the Trip Duration column After Removing the Outliers

Statistical Tests

Q-Q Plot for the trip duration of each gender





Trip Duration for males and females - Wilcoxon 2 Tail Test

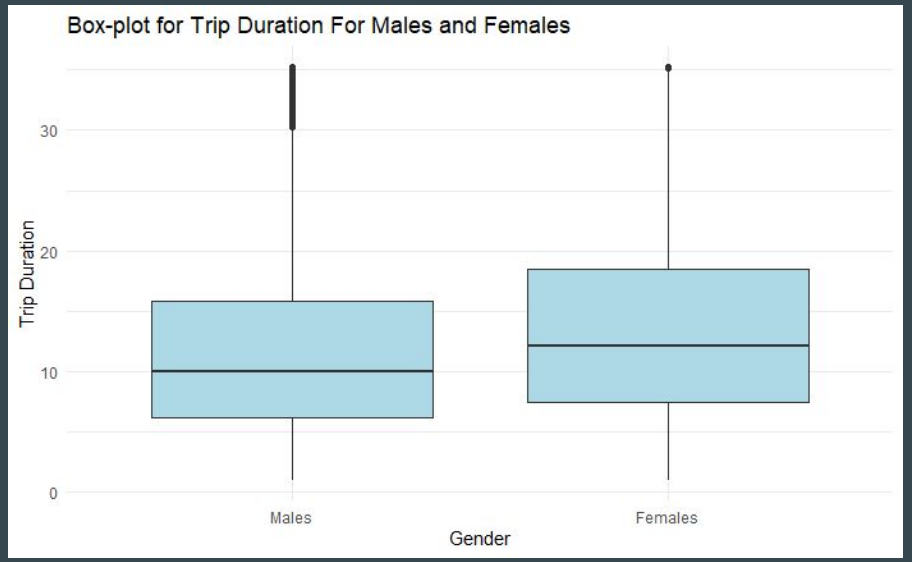
The Q-Q plots show that both the distributions do not follow a normal distribution.

So, we perform the Wilcoxon Test with the following null and alternative hypothesis:

- ➔ H0 : the 2 groups are similar
- ➔ H1 : the 2 groups are different

p-value =2.2x10^(-16) < 1% significance level

Reject H0 and conclude the trip durations are significantly different for males and females.



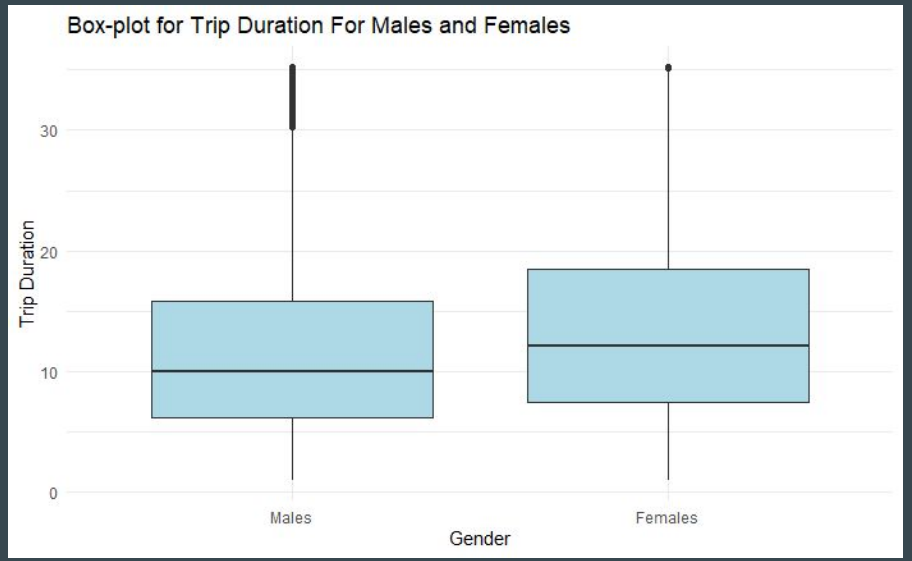
Trip Duration for males and females - Wilcoxon 1 Tail Test

The boxplot shows the trip duration for females is higher than that for males. To test this we can use the following:

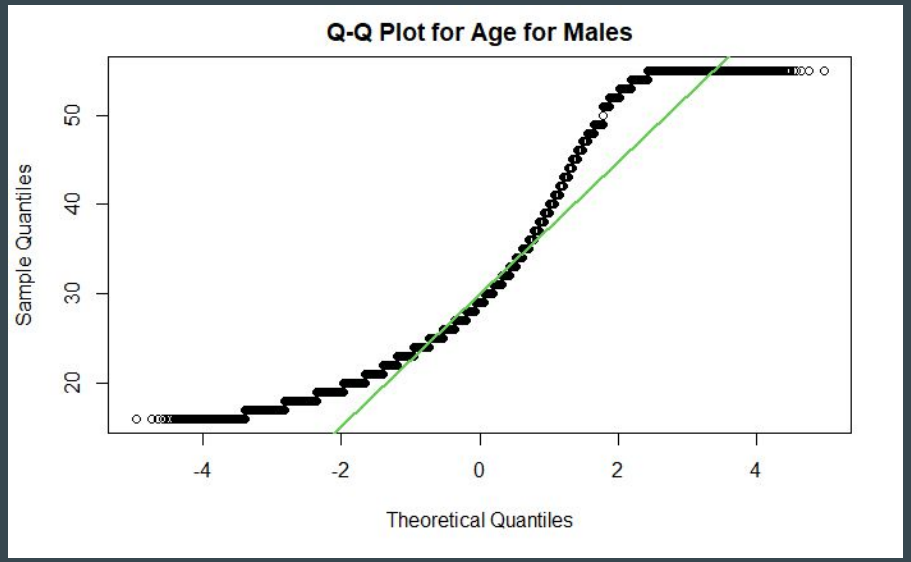
- ➔ H0 : male trip duration > female trip duration
- ➔ H1 : male trip duration < female trip duration

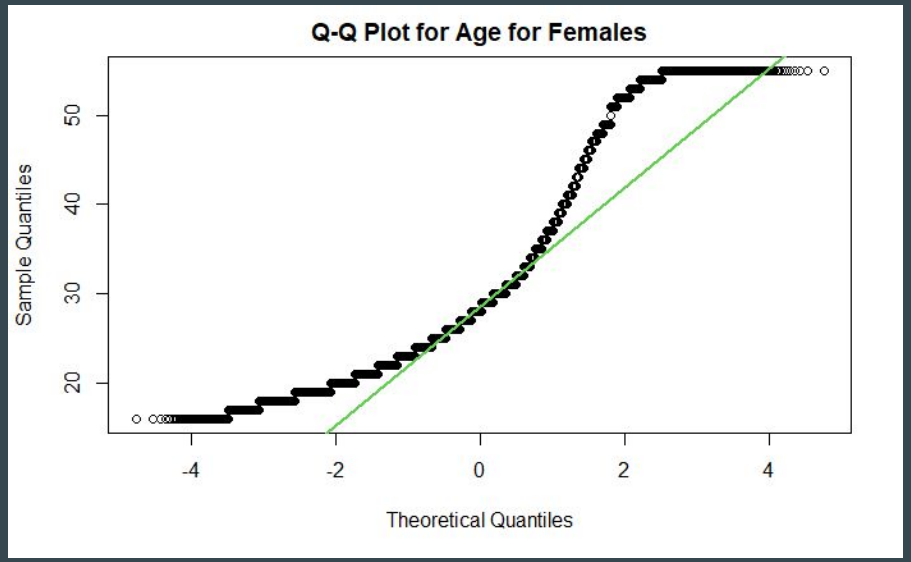
p-value =2.2x10^(-16) < 1% significance level

Reject H0 and conclude males have a significantly lower trip duration than females.



Q-Q Plot for the Age of Each Gender





Age of males and females - Wilcoxon 2 Tail Test

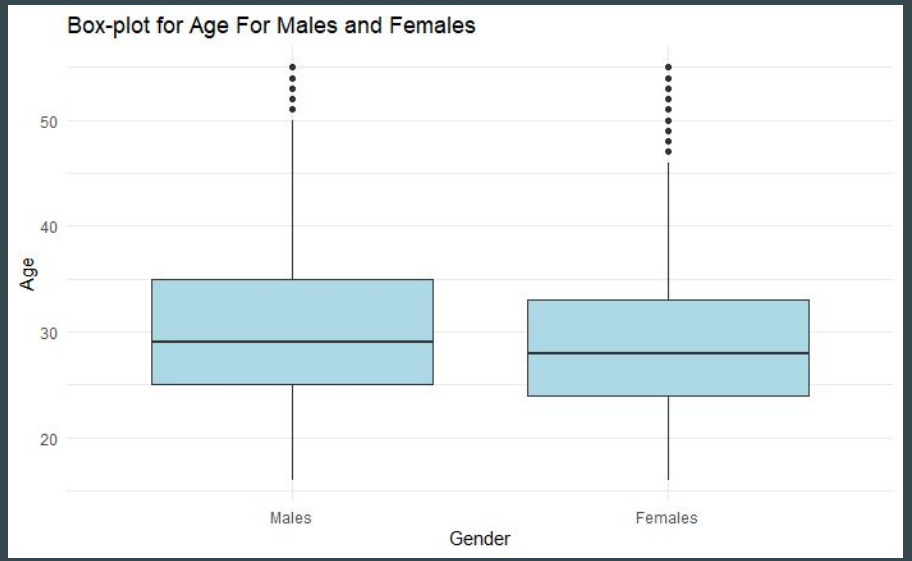
The Q-Q plots show that both the distributions do not follow a normal distribution.

So, we perform the Wilcoxon Test with the following null and alternative hypothesis:

- ➔ H0 : the 2 groups are similar
- ➔ H1 : the 2 groups are different

p-value =2.2x10^(-16) < 1% significance level

Reject Ho and conclude the ages are significantly different.



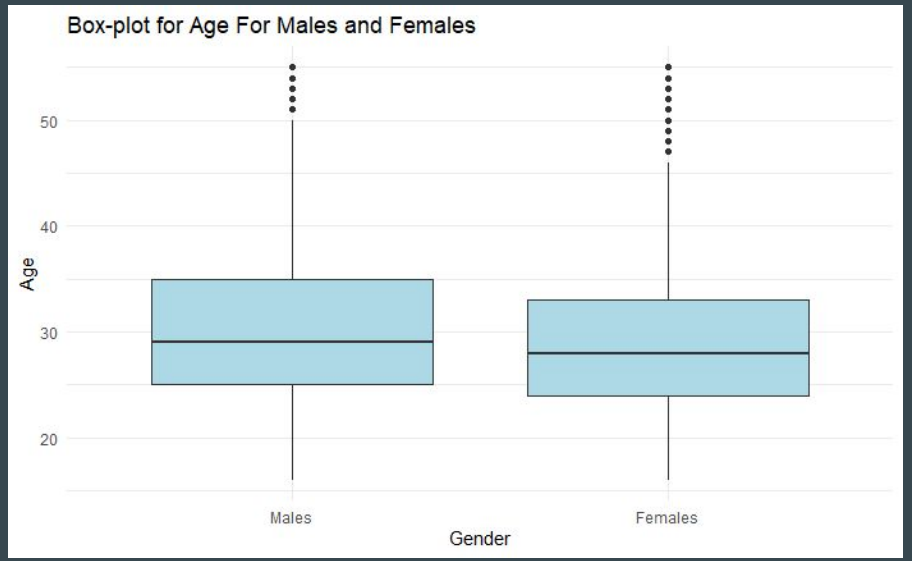
Age of males and females - Wilcoxon 1 Tail Test

In the boxplot we see that the age for females is lower than that for males. To check this we can use the following null and alternative hypothesis:

- ➔ H0 : male age < female age
- ➔ H1 : male age > female age

p-value =2.2x10^(-16) < 1% significance level

Reject Ho and conclude males have a significantly higher age than females.



Linear Regression

Linear regression for a random sample of 8k observations

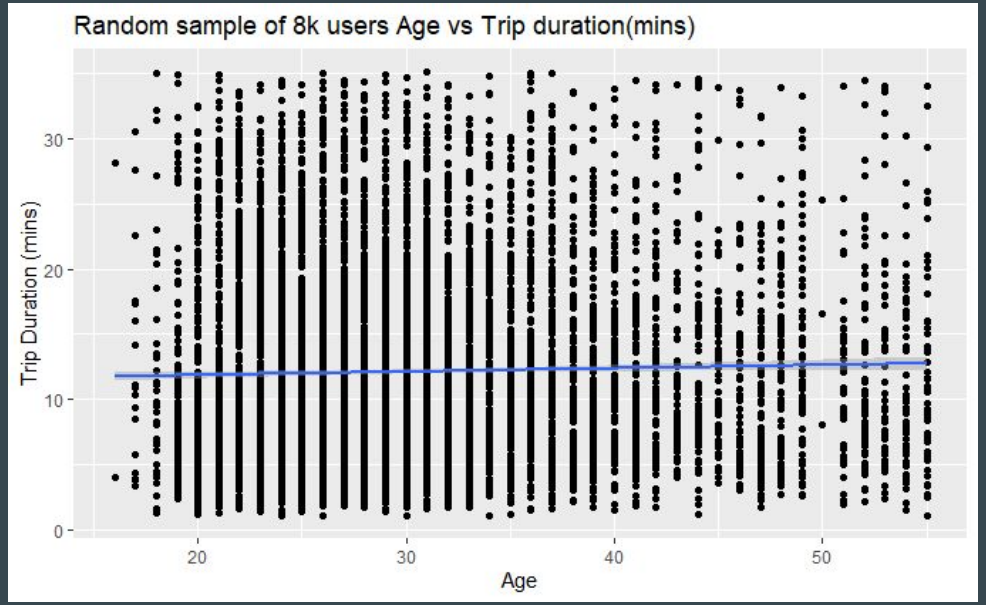
We took a sample of 8k observations at random from the dataset.

We can see from the best fit line slope that as the age of the user increases by one year, his expected trip length increases slightly by 0.029 minutes

Y-Intercept	slope
11.38954	0.02896

t(2002501) = **45.41** at p < 2e^-16 ⇒

Reject the null hypothesis , the linear relationship is statistically significant between Age and Trip Duration

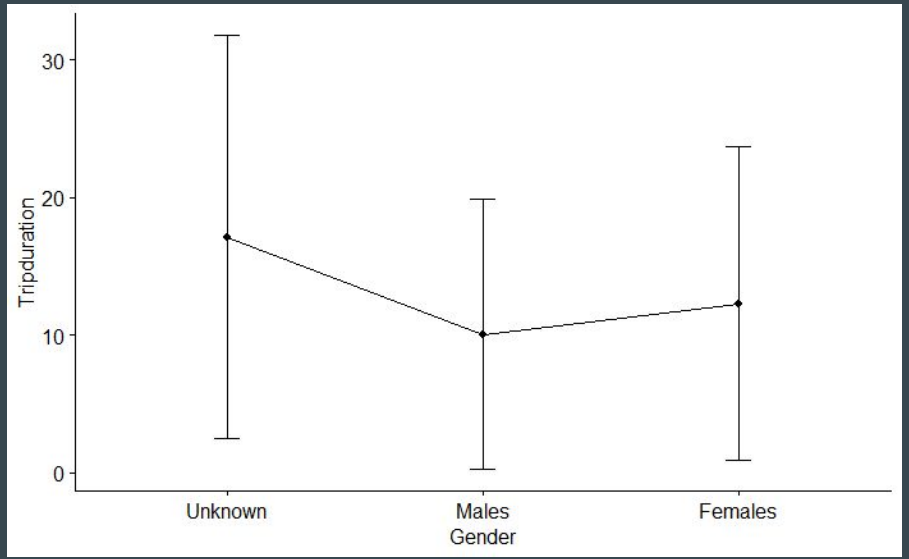


ANOVA/Kruskal-Wallis Test

Kruskal-Wallis Test for Trip Duration and Gender

We know from the previous parts that the data is not normally distributed. So we cannot use ANOVA thus we use Kruskal-Wallis test.

We get a p-value less than the significance level 0.05, so we conclude that there is a significant difference between the different genders.



Pairwise Wilcoxon Test

- Having known that there is a difference between genders, to know between which ones ⇒ apply pairwise Wilcoxon Test

-The pairwise comparison shows that all groups are significantly different from each other as the p-values are less than 0.05.

Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data: Trip Duration and Gender

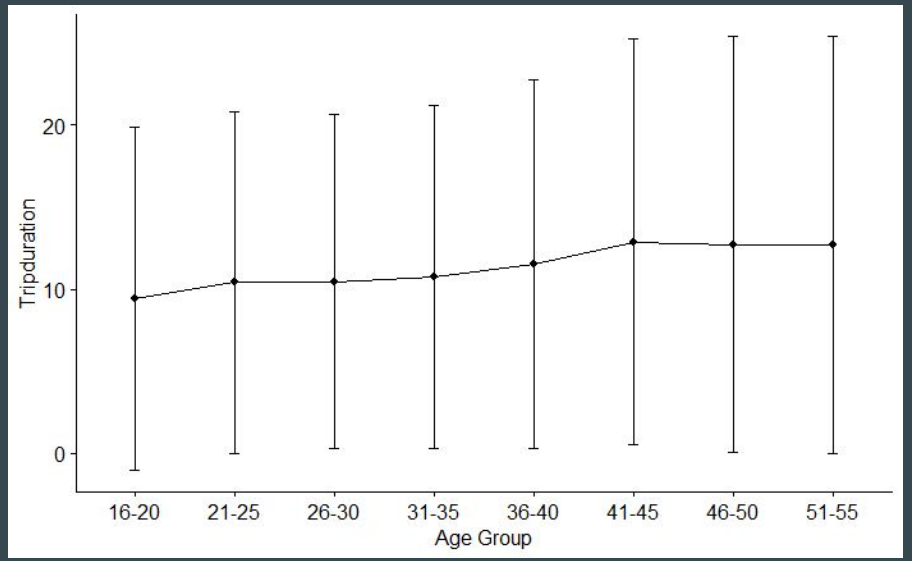
P value adjustment method: BH

	Unknown	Male
Male	<2 x 10^(-16)	-
Female	<2 x 10^(-16)	<2 x 10^(-16)

Kruskal-Wallis Test for Trip Duration and Age

We know from the previous parts that the data is not normally distributed. So we cannot use ANOVA thus we use Kruskal-Wallis test.

We get a p-value less than the significance level 0.05, so we conclude that there is a significant difference between the different age groups.



Pairwise Wilcoxon Test

- Having known that there is a difference between age groups, to know between which ones ⇒ apply pairwise Wilcoxon Test

-The pairwise comparison shows that all groups are significantly different from each other as the p-values are less than 0.05.

Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data: Trip Duration and Age
P value adjustment method: BH

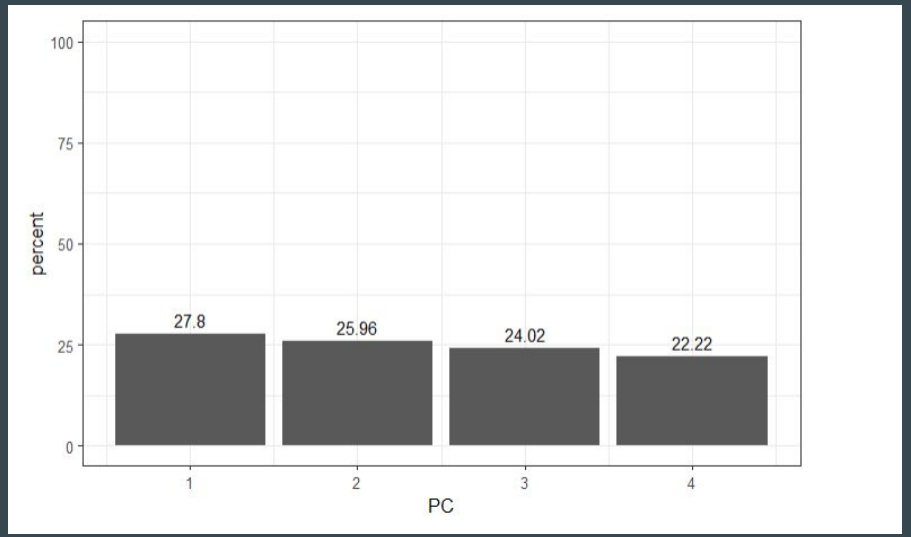
	16-20	21-25	26-30	31-35	36-40	41-45	46-50
21-25	<2e-16						
26-30	<2e-16	<2e-16					
31-35	<2e-16	<2e-16	<2e-16				
36-40	<2e-16	<2e-16	<2e-16	<2e-16			
41-45	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16		
46-50	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	0.013	
51-55	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	9.3e-08	0.024

PCA

Variance

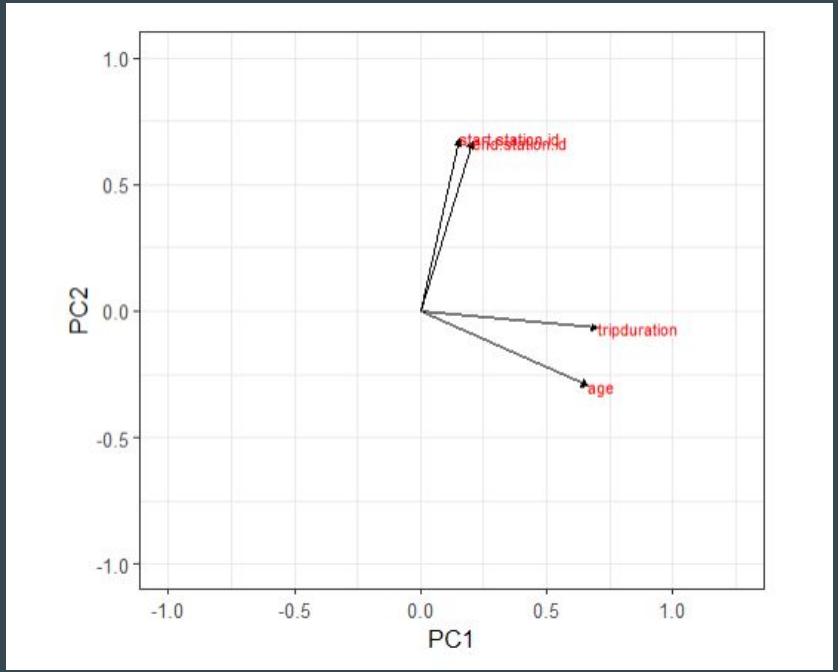
The percentage of variance explained by each Principal Component is the following:

PC 1	27.8%
PC 2	25.96%
PC 3	24.02%
PC 4	22.22%



Biplot

From the graph we can see that the age of the user has a higher impact on the trip duration than the start and end points of the trip.



Findings

- Female users tend to have longer trip durations
- Male users have higher age on average than female users.
- As users gets older, they tend to have shorter trip durations.
- Trip durations also correlates with the starting and ending stations,however not very statistically significant.
- On the other hand, the starting and ending stations are positively correlated to each other.

Thank You

...