Explanatory Analysis

Analysing the factors effecting the frequency and duration of bike rides

Investigation Overview

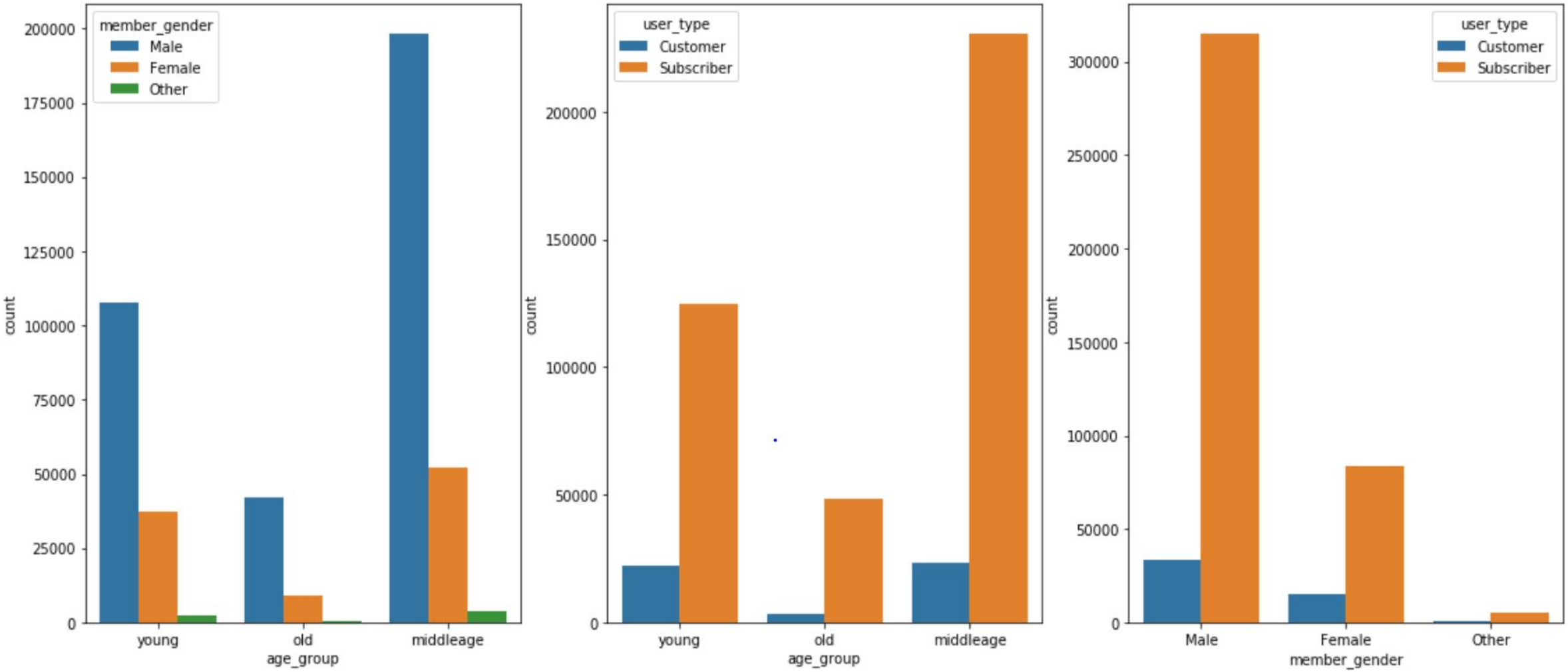
In this investigation, I want to look at how factors such as age group, gender, user type and whether the start station and end station are same will influence the frequency of bike rental and duration taken

Dataset overview

The original dataset consists of duration, start time, end time, information about start station and end station, latitude, longitude, bike id, user type, member birth year and member gender.

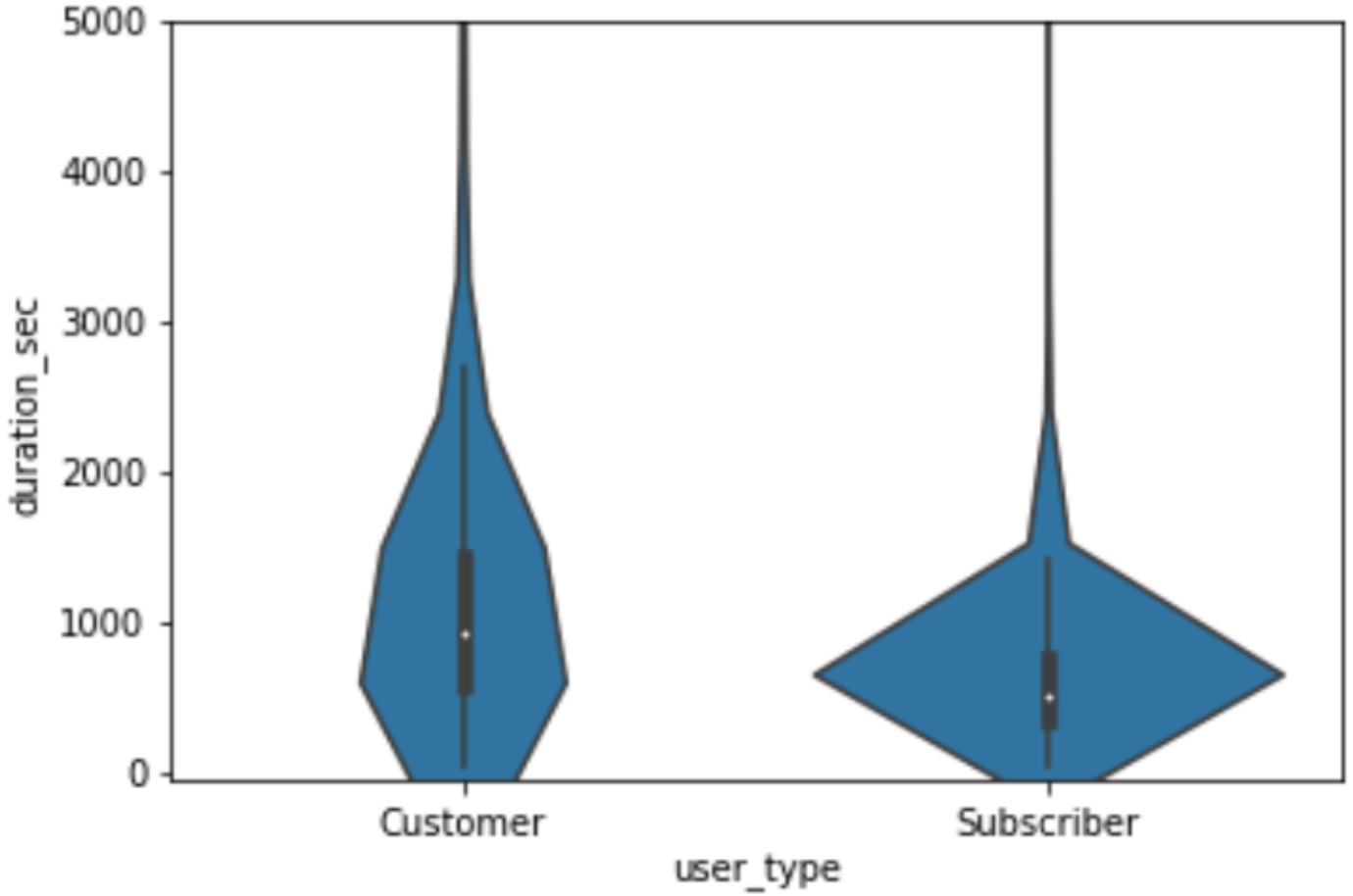
Columns consisting of categorical variables such as age,age group and whether the start station and end station are same are added to the dataset to have a deeper analysis. Age is calculated by subtracting 2017 from member birth year. The age group is categorized as young who are less than or equal to 30,middle aged who are between 30 and 50 and old who are greater than 50.'start=end' column is included to categorize whether the start station being equal to the end station is true or false by comparing the station ids.

The number of people based on gender,age group and user type are counted using count plot.



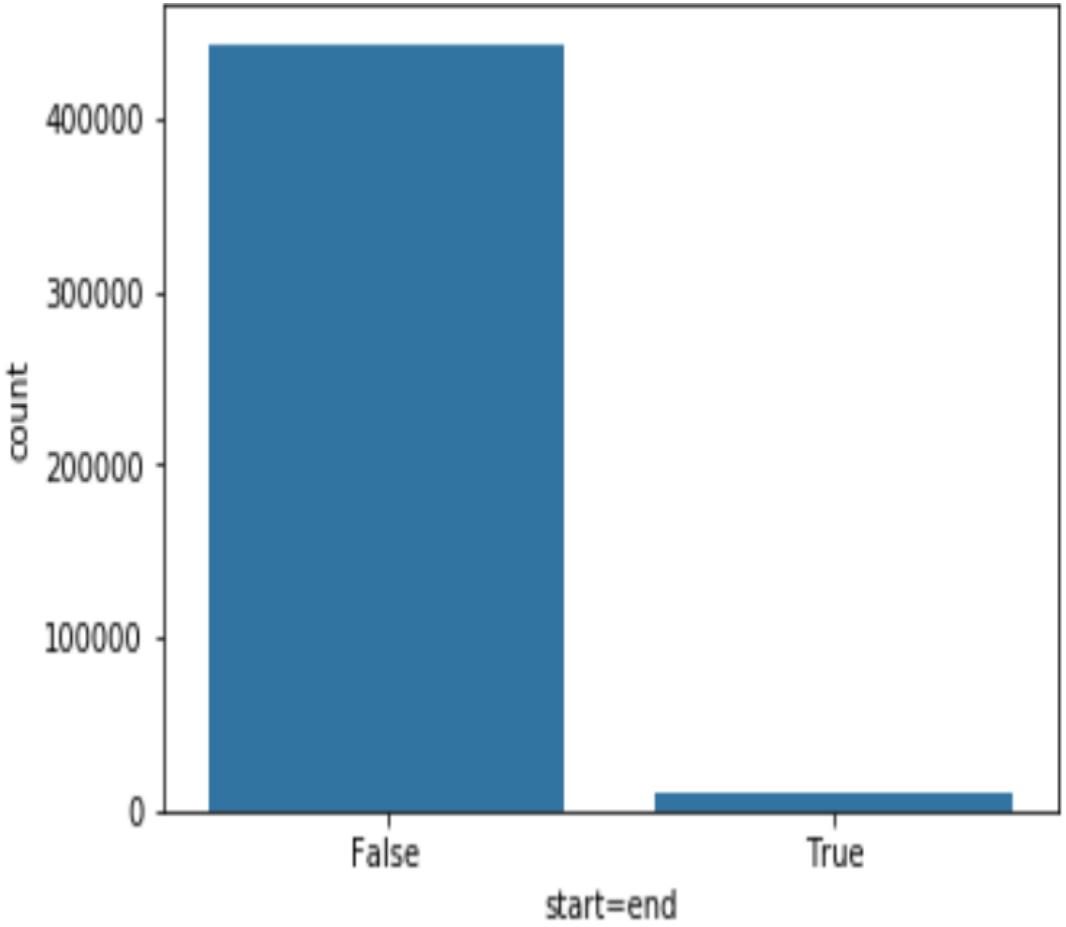
It is found that the users are mostly middle aged men who are subscribers.

The median of both the user types are found using violin plot.



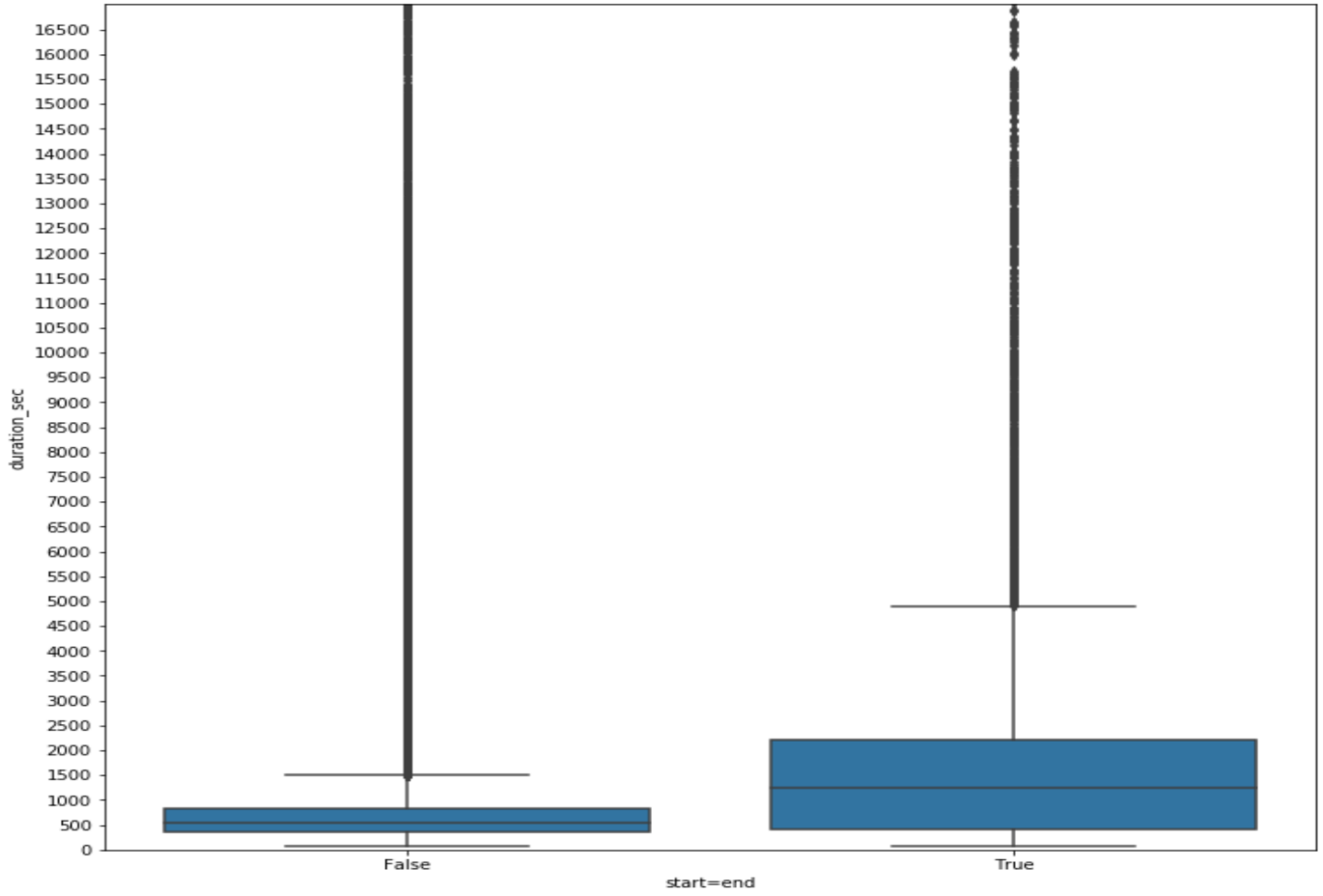
It is interesting to note that the median duration taken by subscribers is much less than the customers. And also the probability density of subscribers and customers are high at around 800 sec.

Count plot is plotted to find the number of people who choose the same end station as the start station.



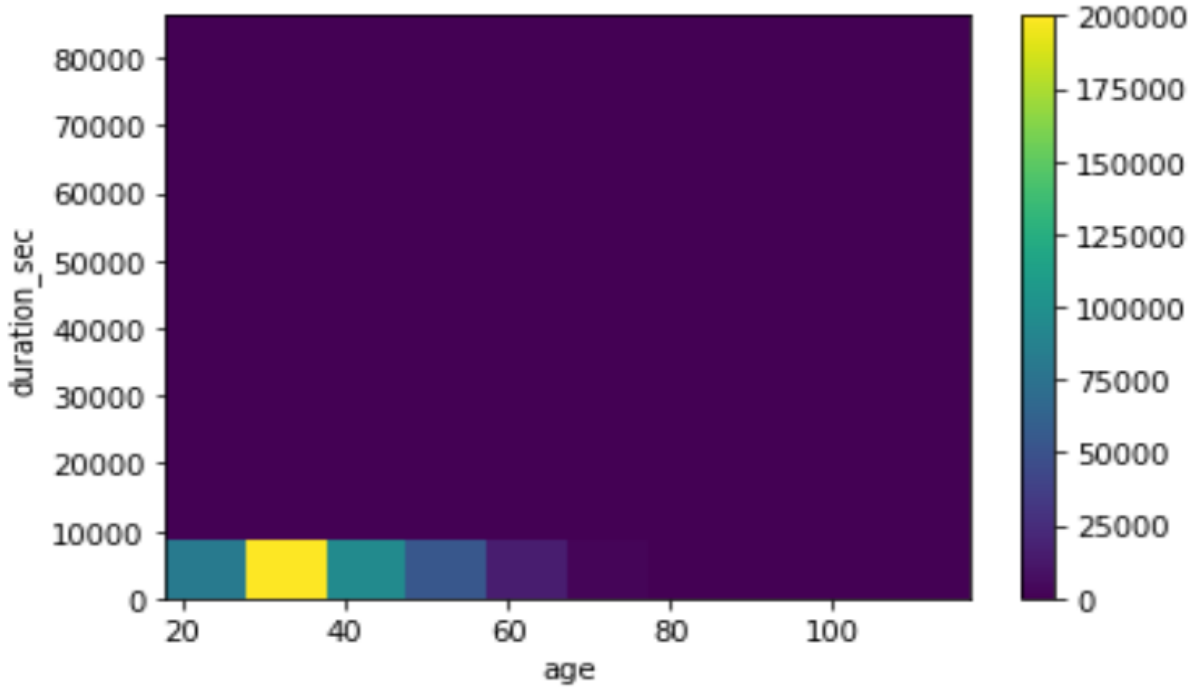
From the count plot it has been found that considerably a large number of people don't choose the same end station and start station.

The median and the outlier for frequency of where the bike is dropped is checked using boxplot.



The median duration of users who drop the bike in same station is higher, inspite of the frequency being lower. This can be attributed to outliers.

A heatmap is plotted to further identify the correlation between age and duration.



It is found that users from age 30-38 correlated to the duration taken have highest density.There is a general negative correlation between age and duration taken.

Conclusion

It has been found that a large section of users are middle-aged men who are subscribers followed by young men who are subscribers. Though the number of subscribers are more the median duration taken by subscribers are less than the customers. Most of the users don't drop the bikes at the same station as to where they start irrespective of their gender, user type and age group. Meanwhile the median and mean duration of users whose end station and start station are same is higher than the duration of users where both the stations are not the same. This can be attributed to the presence of many outliers. Most of the users are particularly between the age of 30-38 who take within 10000 secs.