Data Analysis practice

Exploration and visualization

FordGo Bike dataset analysis

Alejandro Sierra
Email: asierra21@gmail.com

21st December 2020

Contents

1 Introduction

FordGo Bikes, own by the company lift, is a public bike share system that works on collaboration with the Metropolitan Transportation Commision.

The sistem is deployed in California and the West Coast in the United States.

In 2009 the sistem was renamed to Bay Wheels and it has over 2.600 bicycles in over 300 stations.
Their fleet consists of classic bicycles, and hybrid electrically asisted bikes. This last model allows to be used without a dock thanks to a rear-wheel lock.
These bicycles can be used 24/7 for periods that go from a single ride of up to 30 minutes to a day pass covering 30 minutes increment. Additionally, an annual suscription can be purchased, which in this case allows up to 45 minutes trips. The "Bikeshare for All" suscription is a reduced priced option and ca be obtained meeting qualifying terms.

The main goal for this analysis is to answer the following questions:

- What is the user profile in terms of age and gender?
- What is the suscriptor profile in terms of age and gender?

- Which stations have the higher demand?
- Are bikes used more on weekdays o weekends?
- How long do rides last?
- Is the "bike share for all" suscription being used?

2 Data Overview

2.1 Features

This dataset contains inforation of bike travels collected by FordGo bike system. Variables recorded are the following:

- **duration_sec**
- **start_time**
- **end_time**
- **start_station_id**
- **start_station_name**
- **start_station_latitude**
- **start_station_longitude**
- **end_station_id**
- **end_station_name**
- **end_station_latitude**
- **end_station_longitude**
- **bike_id**
- **user_type**
- **member_birth_year**
- **member_gender**
- **bike_share_for_all_trip**

Since their names are clear, no description for each category is considered necessary.

2.2 Data structure

The dataset contains over 180.000 observations of 16 different variables.
Those are both qualitative and quantitative varaiables as shown below.

Qualitative	Quantitative	Ordinal
start_station_id	duration_sec	member_birth_year
start_station_name	start_time	
end_station_id	end_time	
end_station_name	start_station_latitude	
bike_id	start_station_longitude	
member_gender	end_station_latitude	
bike_share_for_all_trip	end_station_longitude	
user_type		

2.3 Null values

The dataset conatins null values for six of the variables.

Variable	Missing values count
duration_sec	0
start_time	0
end_time	0
start_station_id	197
start_station_name	197
start_station_latitude	0
start_station_longitude	0
end_station_id	197
end_station_name	197
end_station_latitude	0
end_station_longitude	0
bike_id	0
user_type	0
member_birth_year	8265
member_gender	8265
bike_share_for_all_trip	0

The columns with the most missing values contain 8.265 missing values. This represents 4.7 % of the whole dataset, therefore, we will choose to remove the rows containing missing values.

3 Data Wrangling

3.1 Remove missing values

As we saw in the previous section, we need to remove missing values in our dataset.
Variables member_birth_year and member_gender both contain 8.265 null observations.

3.2 Unique values

Checking for unique values, we could see that one category presents values that need to be checked.
member_birth_year presents 75 different values, which seems like a large range considering bike riding requires physical abilities.

Minimum value for this category is year 1878, which seems strange considering that this person would be over 100 years.
This will be considered in the next section where will we take a visual approach to this variable performing an univariate exploration and check for outliers.
Additionally, member_gender presents 3 unique values so we need to check it aswell and look for potential misspelling or incongruent information.
In this case the three categories are "Male", "Female" and "Other", which is correct and needs no wrangling for this data.

3.3 Data types

Data types seem correct for almost all of the variables. The only ones that need to be converted in order to be able to work with them correctly are start_time and end_time, which contain the format YYYY/MM/DD HH:MM:SS.SS.
This columns are converted to Datetime64 where we can directly work with year, month, day, hour and seconds precision.

member_birth_year, start_station_id and end_station_id contain decimal values, which do not correspond since they can only take integer values. This dtype are also going to be converted.

3.4 Feature engineering

We are interested in user´s age, so we need to create a new column for this under the name "user_age"

4 Statistics

4.1 Descriptive statistics

Initially, we are interested only in main numeric feature statistics.
So far, we will check duration_sec and member_birth_year.

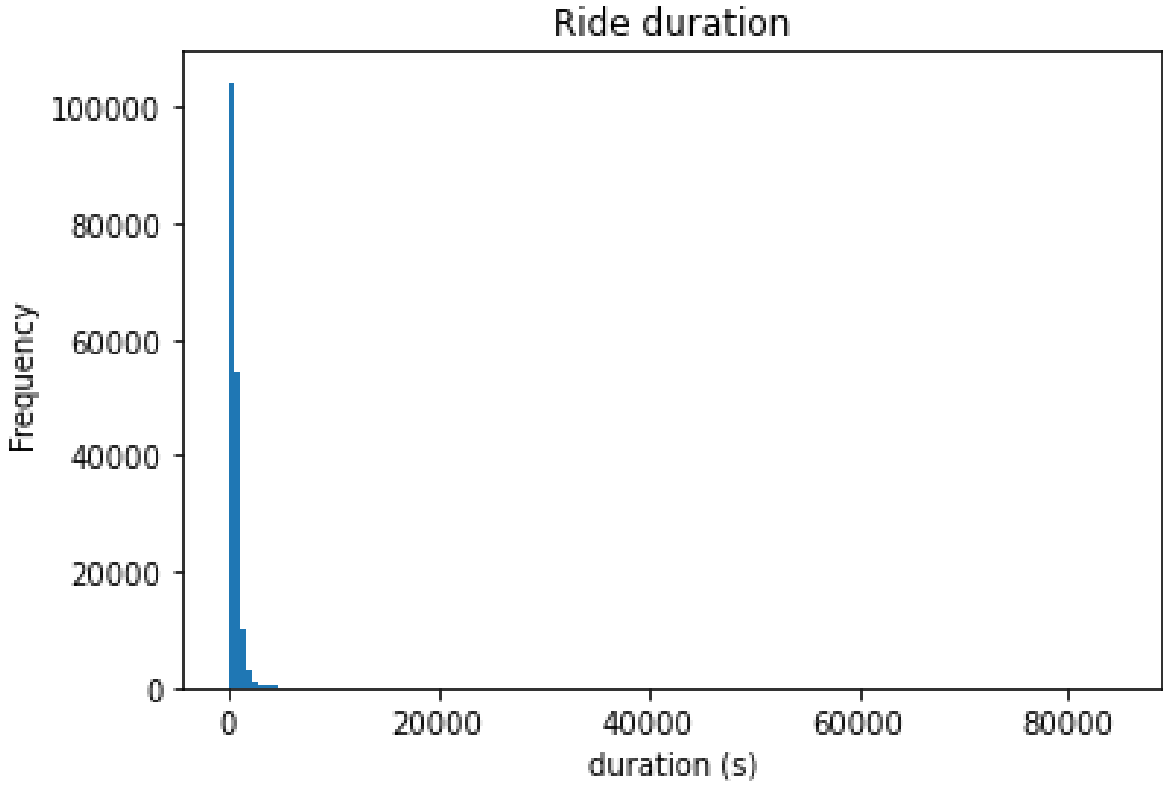
duration_sec:
Mean ride duration is of 704 seconds, which is close to 12 minutes, with a standard deviation of 1642 seconds. This suggests a wide distribution for trips. Range goes from 61 seconds, one minute, to 84.548 seconds, one day.

member_birth_year:
The mean for birth year distribution is 1984 and the median is 1987, which suggests that most users are among the range 30-35 years with a close to normal distribution. Standard deviation is 10 years, and the range, which needs to be checked for outliers goes from 1878 to 2001. Surprisingly, the youngest user is close to 18 years, this may me explained by the bike providers policy of users with at least 18 years old.

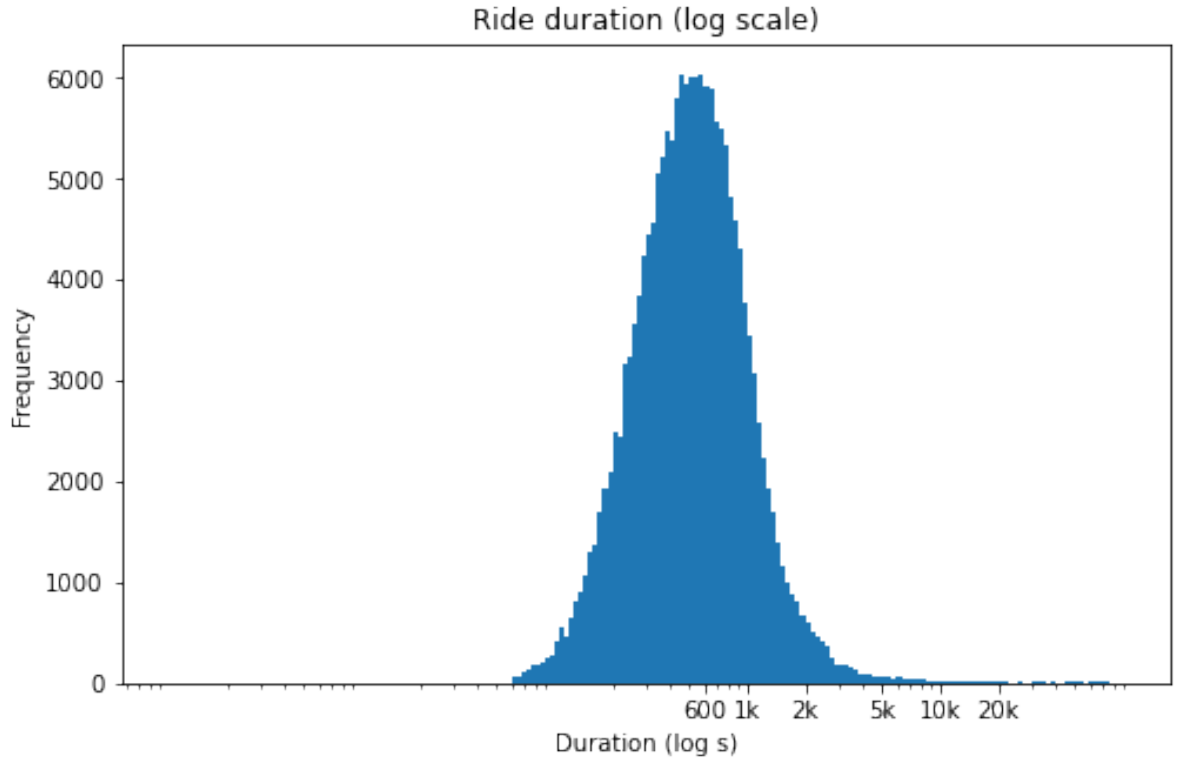
5 Univariate exploration

5.1 duration_sec

We will begin by checking the distribution of the variable.



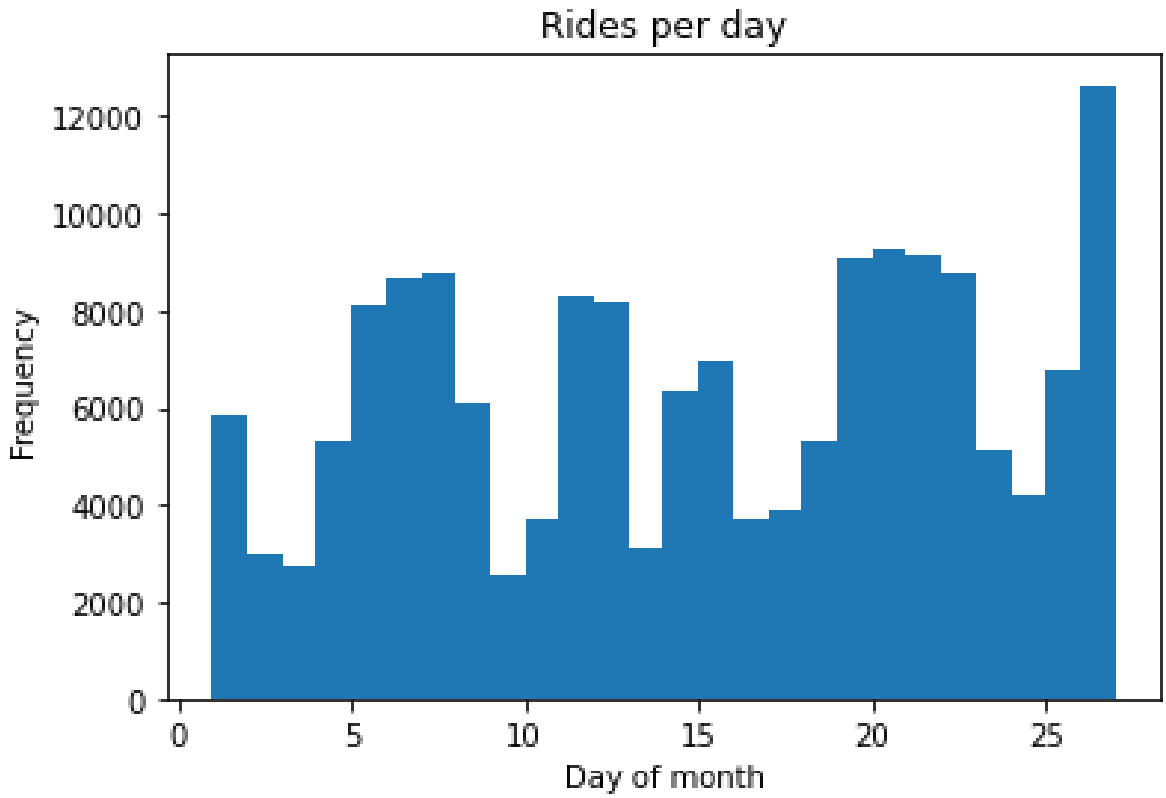
As we see in the plot, and was checked beforehand in the data, there is at least one observation which corresponds to a whole day ride. In this case we need to count how many of this cases we have and then decide what to do with this data.
We are going to count how many observations have a ride duration longer than 12 hours. There are 83 observations that are longer than 12 hours so they can´t be considered outliers.
In this case the best choice is to make a transformation to log scale for this variable.



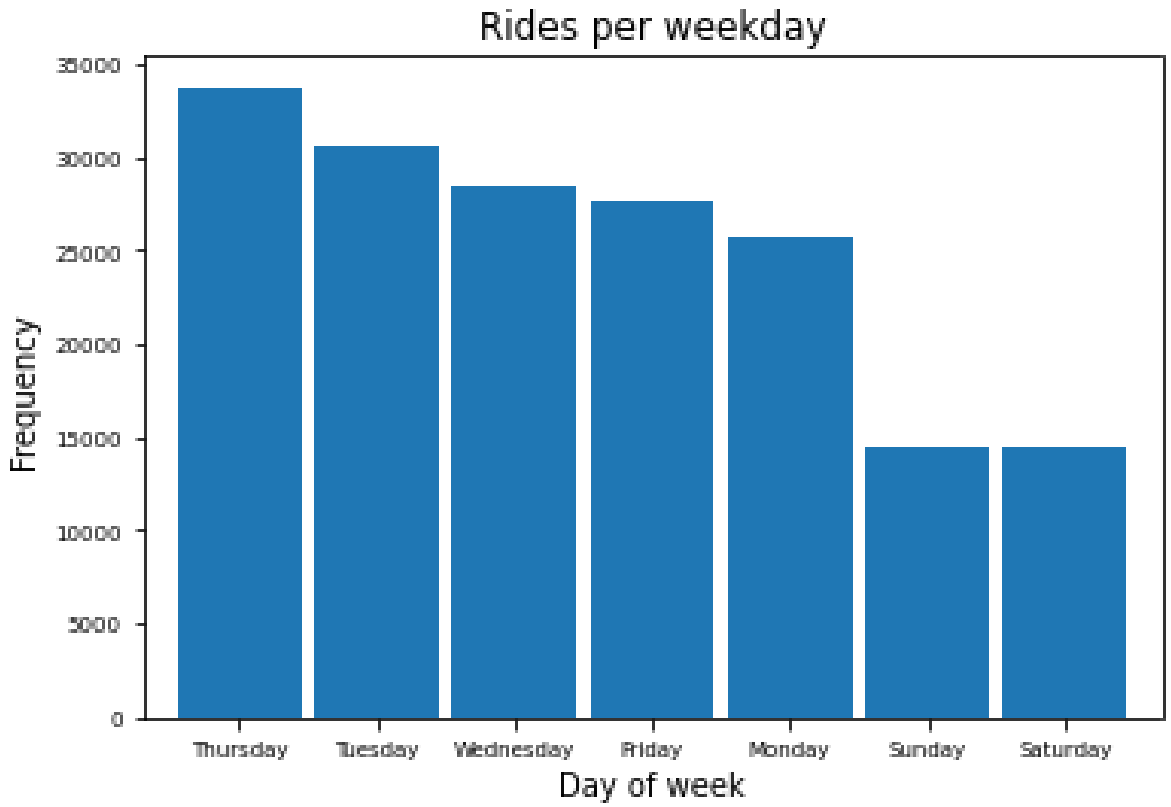
Converting to log scale, we see a normal distribution with a mean ride of 600 sec, 10 minutes.

5.2 start_time

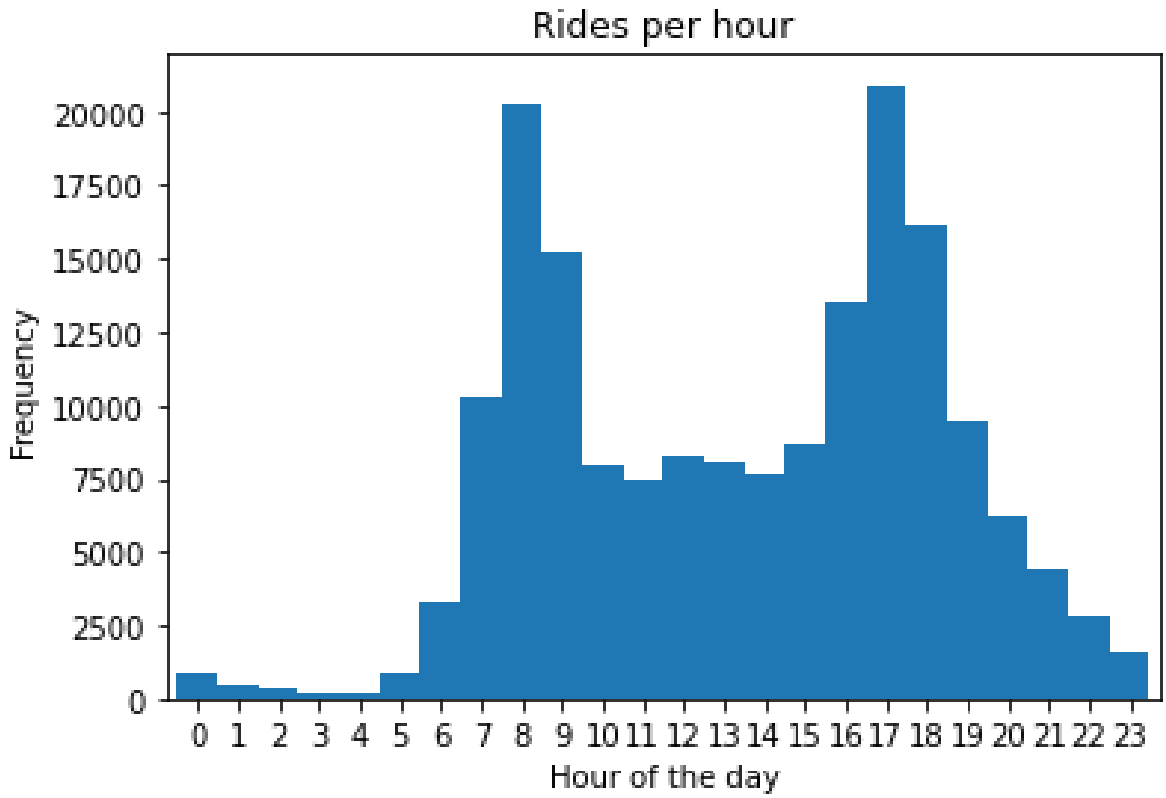
Since the whole dataset is for February 2019, we will not deep into year or month. First of all, we want to see how many rides where made each day of the month.



The plot shows stationary pattern, which suggests that there´s a different behavior on weekdays than on weekends.
We will explore this with the following bar plot.



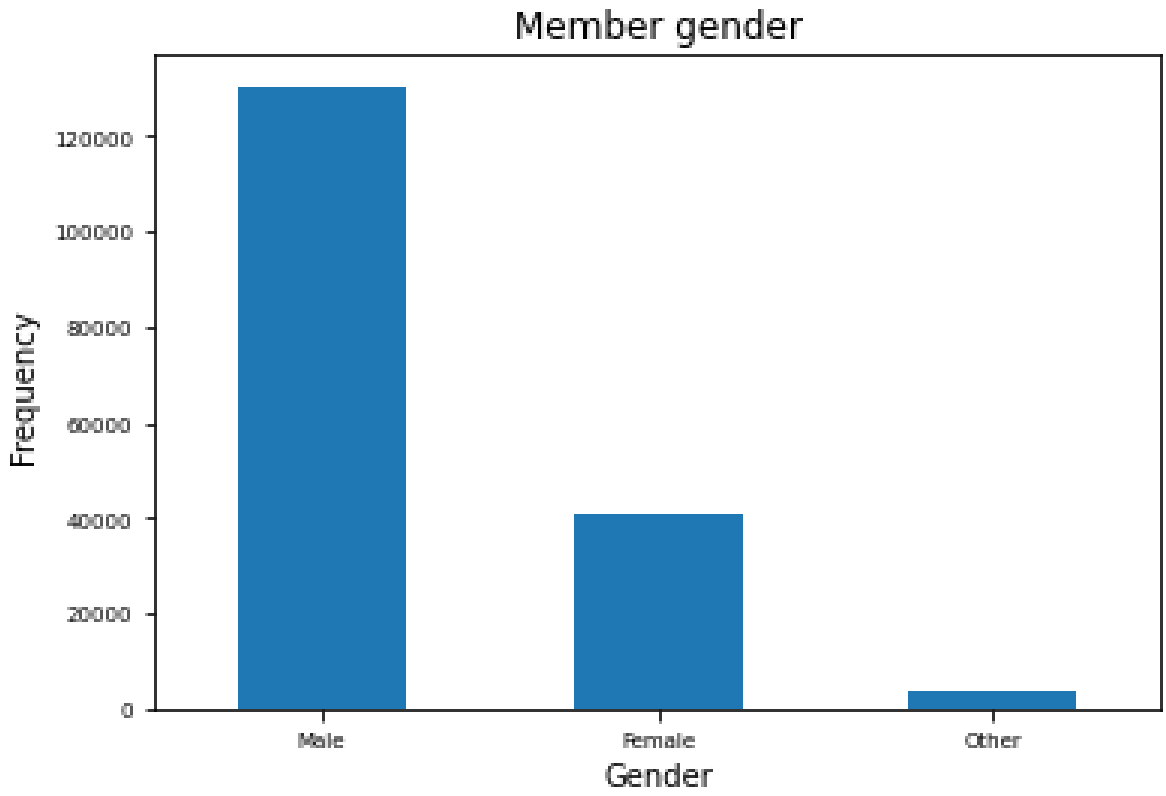
As shown in this plot, it is clear that bike sharing system is mostly used on weekdays and not on weekends.
Now we will se the distribution in daily hours.



The plot is bimodal, with the highest demand for bikes from 7 am to 9 am and from 16 pm to 19 pm hs.

5.3 member_gender

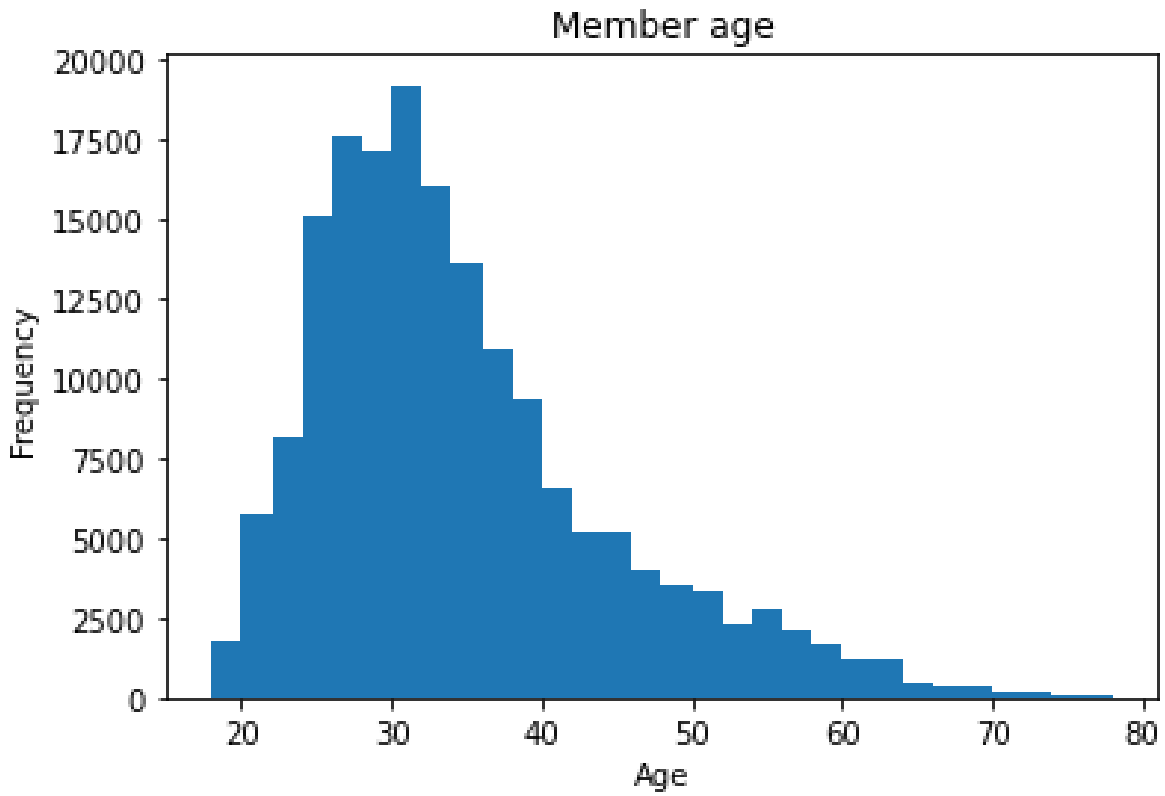
Is there a gender that uses the system more often?.



The plot shows a great difference between gender proportion. Male users represent 74.6 % of the users, while female are 23.3 %. The rest 2.1 % is registered as "Other" gender.

5.4 member_age

Whats the age distribution for bike users?.



Age distribution seems unimodal with main users in the range 28-38 years old.

5.5 user_type

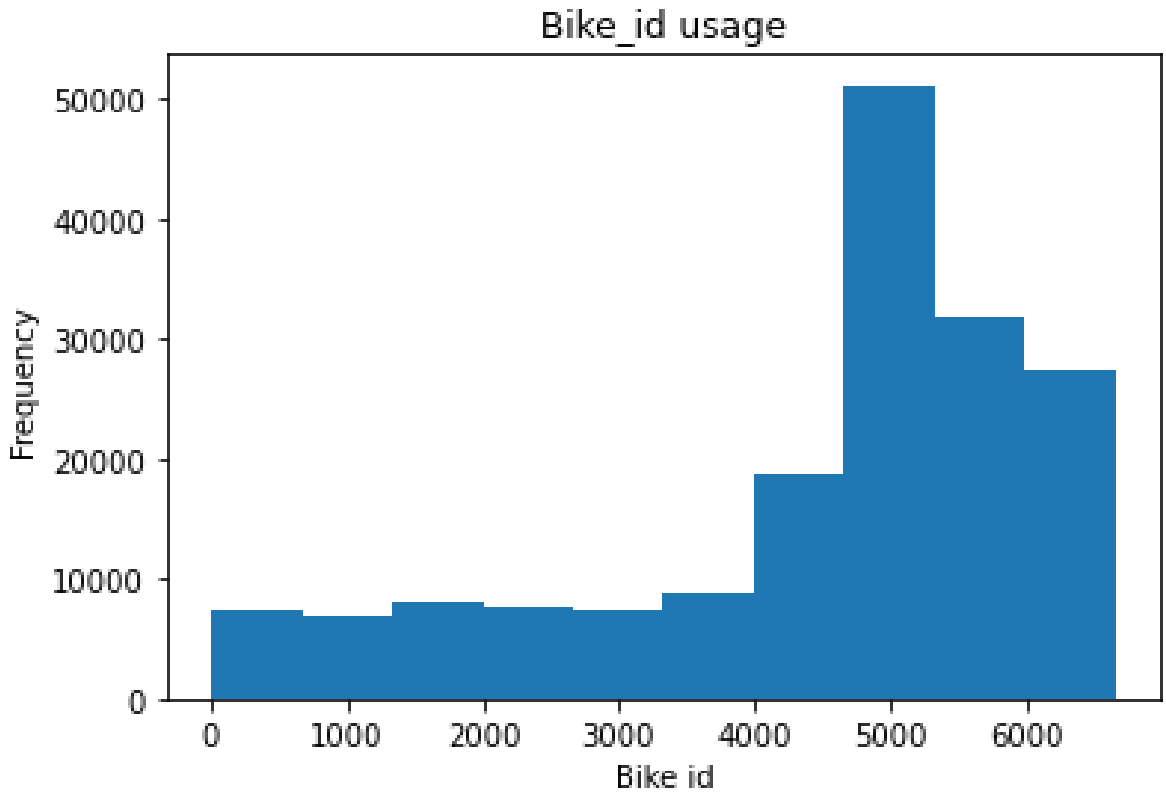
User type can be either customer, or subscriber.
90.5 % of the users prefer the subscription rather than be occasional customer for the bike sharing system.

5.6 bike_share_for_all

Bike share for all represents 9.9 % of the total rides.

5.7 bike_id

In this variable, I expect an uniform usage of every bike. We will analyze how many times each bike was used.



As we can see in the figure, the usage of bikes is not uniform, there are bikes that are considerably more used than others. This could be explained by station location of the bike and demand for the service. It could be beneficial for the company to rotate the bikes among the stations to flatten this curve and give each bike similar usage time.

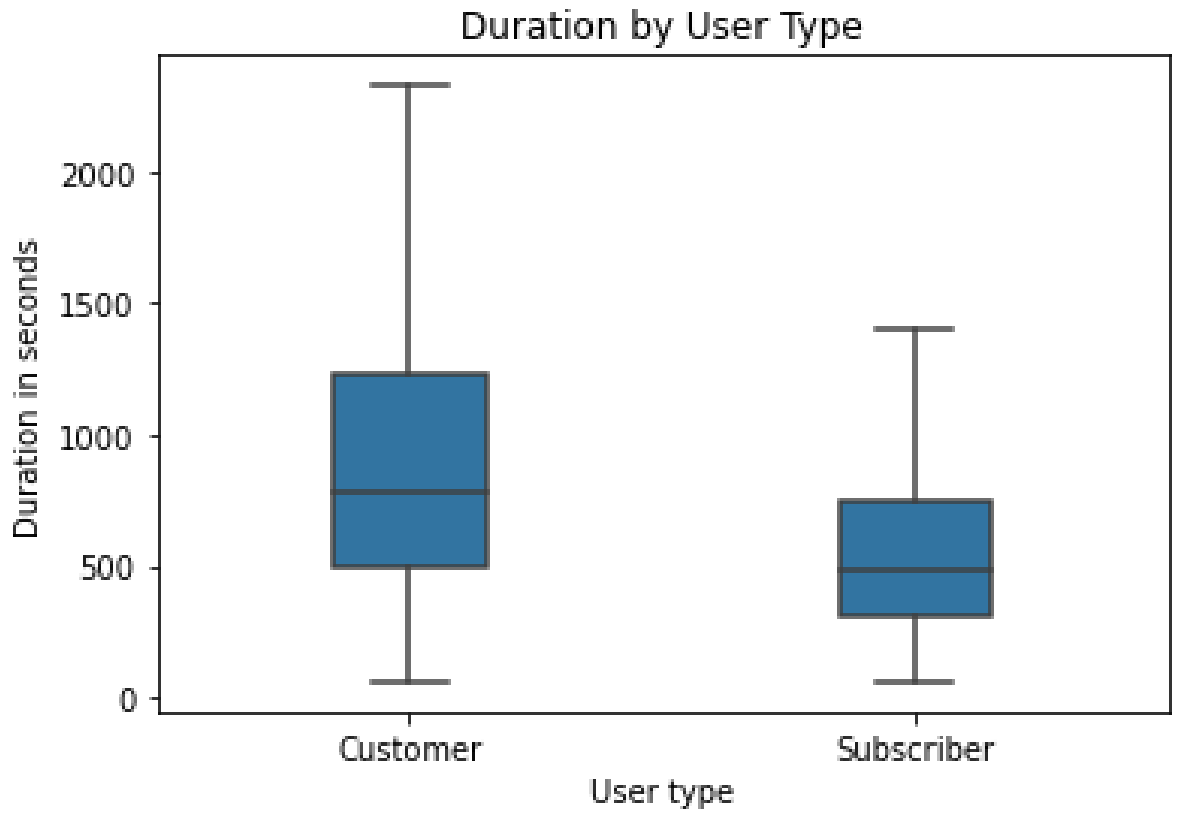
6 Bivariate exploration

6.1 Feature correlation

To start off with, I want to look at the pairwise correlations present between features in the data. Since numeric variables of interest for this are duration and user age, we will not plot the whole matrix in this case.
Surprisingly, this value is of only 0.6 %, so we could say that there is no correlation between age and duration of the ride.

6.2 Duration and count by user type

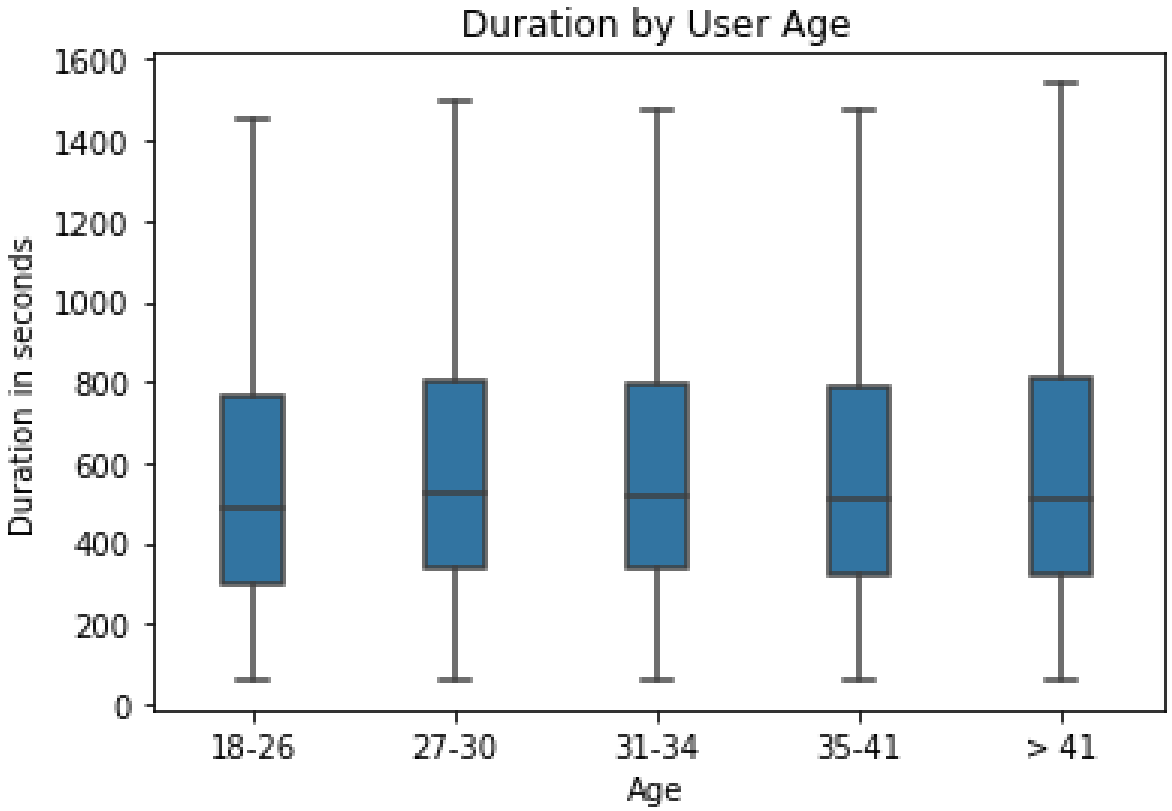
There are two different user types, which are "Customer" and "Subscriber". Subscriber type represent 90,5% of the total rides, while Customer type is only 9.5% of them.
On the other hand, the mean for Subscriber rides duration is 640 seconds and for Customer rides duration is 1310 seconds, more than double.
Let´s plot and see their distribution.



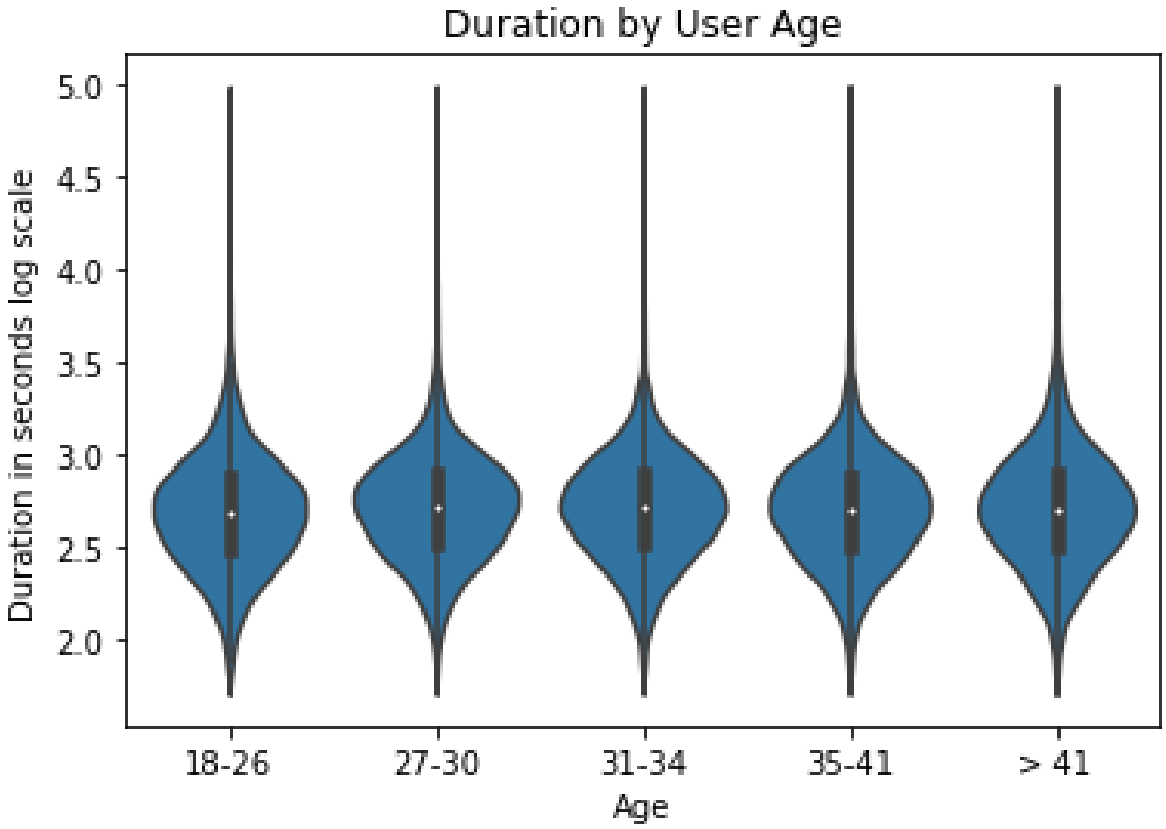
As we can see in this boxplot, and expected by the mean calculation, different user types have different behavior. Customer tends to have larger rides. This has as an evidence, that the first quartile for customer is higher than the mean for subscriber, and the mean for customer is higher than the third quartile for subscriber.
Note that for this boxplot, outliers were removed.

6.3 Duration by user age

We will use in this case a box plot to analyze the different ride duration behaviour according to age.



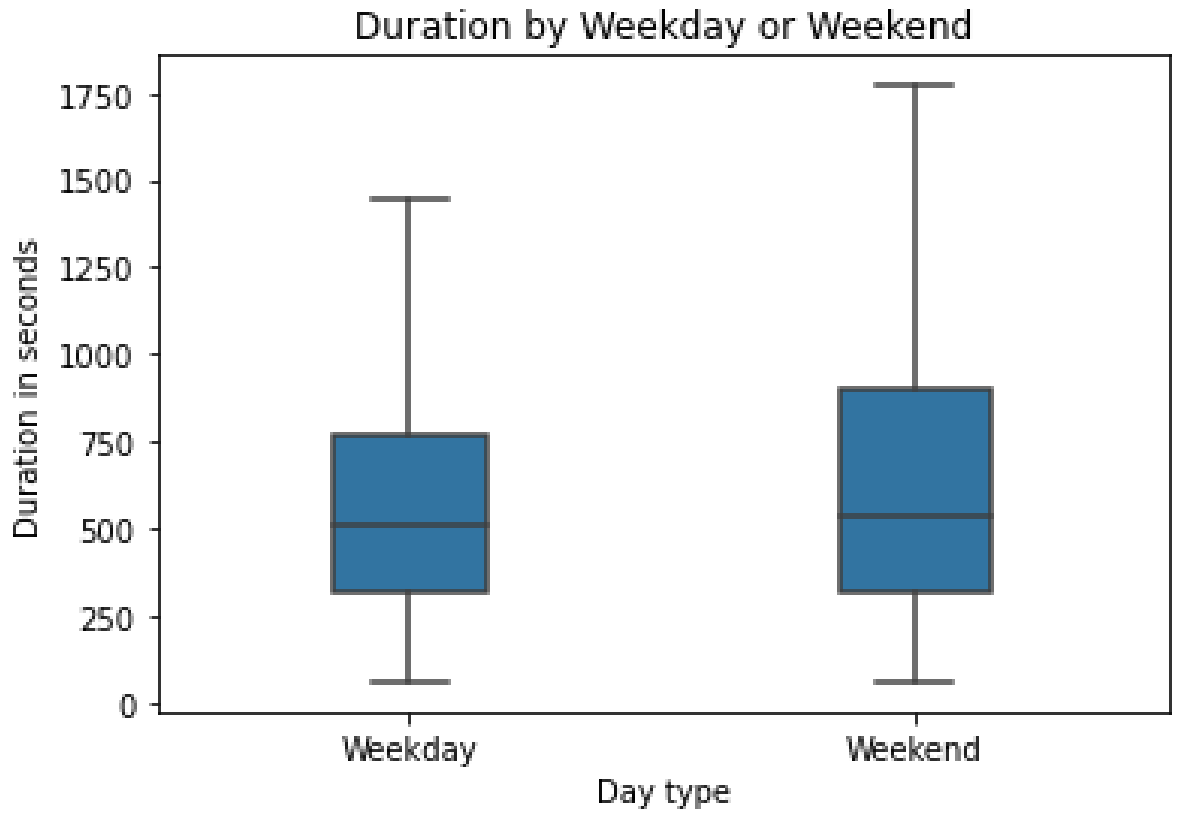
Note that outliers have been removed and bins are determined by quintile cut.
This plot shows that behaviour is not different according to age. We see for every quintile very simmilar box size and quartile statistic values.
Given this, we could explore a little deeper with a violin plot with transformed scale and see if some more information comes up.



This plots shows that every group suggests a normal distribution, but again, we do not see any differences in how much times rides last regarding age of users.

6.4 Duration by weekday or weekend

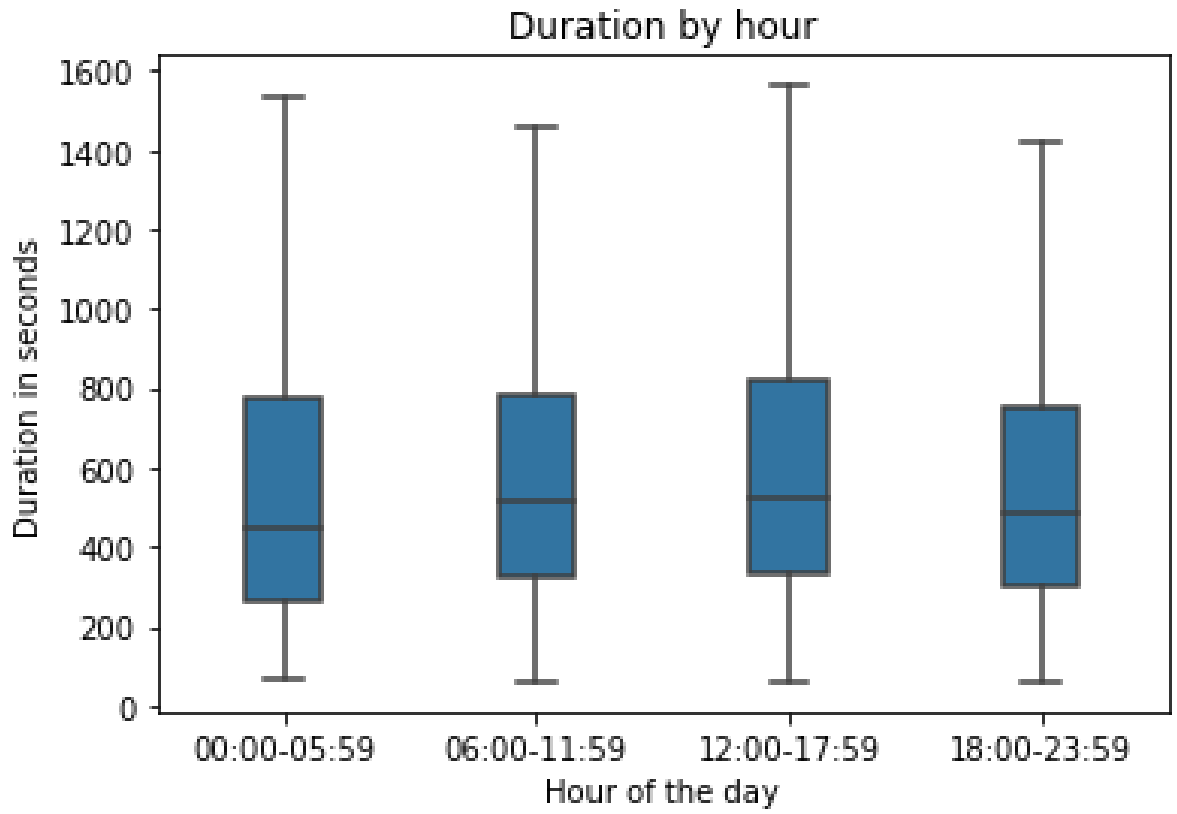
Is there a difference between duration in rides wheather it is made on a weekday or on a weekend?
Again a boxplot will be usefull here.



Even though mean for the first and second quartiles there is no big difference in duration for rides, we can see that weekend rides tend to have a wider range and reach longer duration that on weekdays.

6.5 Duration by hour

I want to see if duration in the morning is simmilar to duration in the afternoon. The limit hours will be 6, 12, 18 and 24 hours.



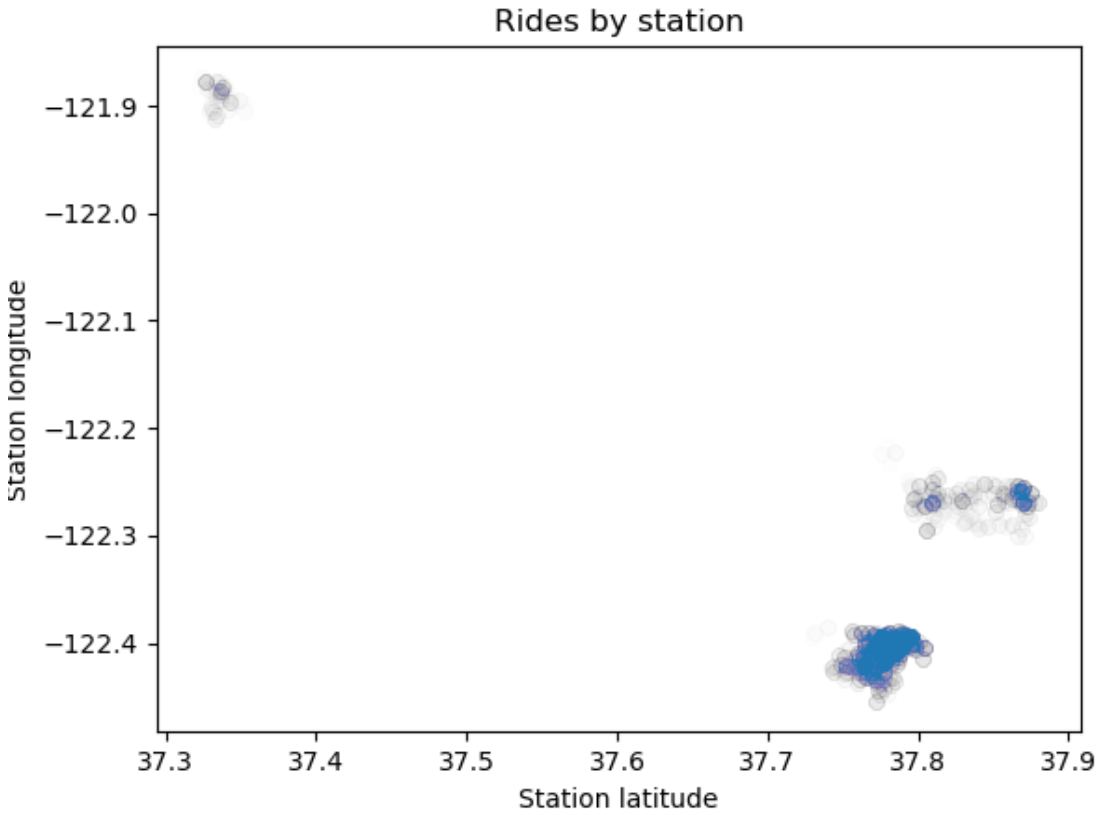
This plot shows that duration mean is lower for trips from 18 to 6 hours. We could say that during the night rides are shorter in mean.

6.6 Stations and rides

There are 329 different stations in the dataset.
First of all we want to see how many rides each station has, both for start and for end. We will use latitude and longitude combined as one unique location variable.

6.6.1 Count

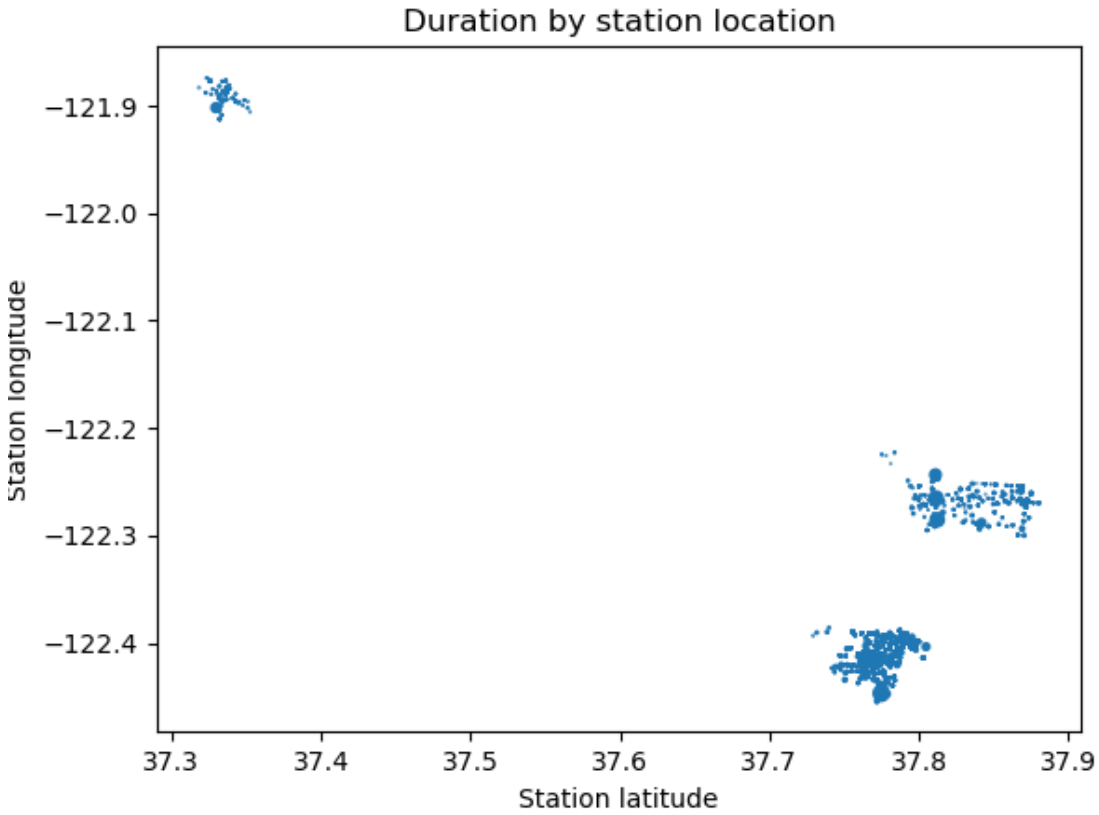
Ride departure divided by station.



As we can see in the scatterplot, there are three main clusters where bike stations are located. Specifically, the south east cluster is the one with more rides and bike usage.

6.6.2 Duration

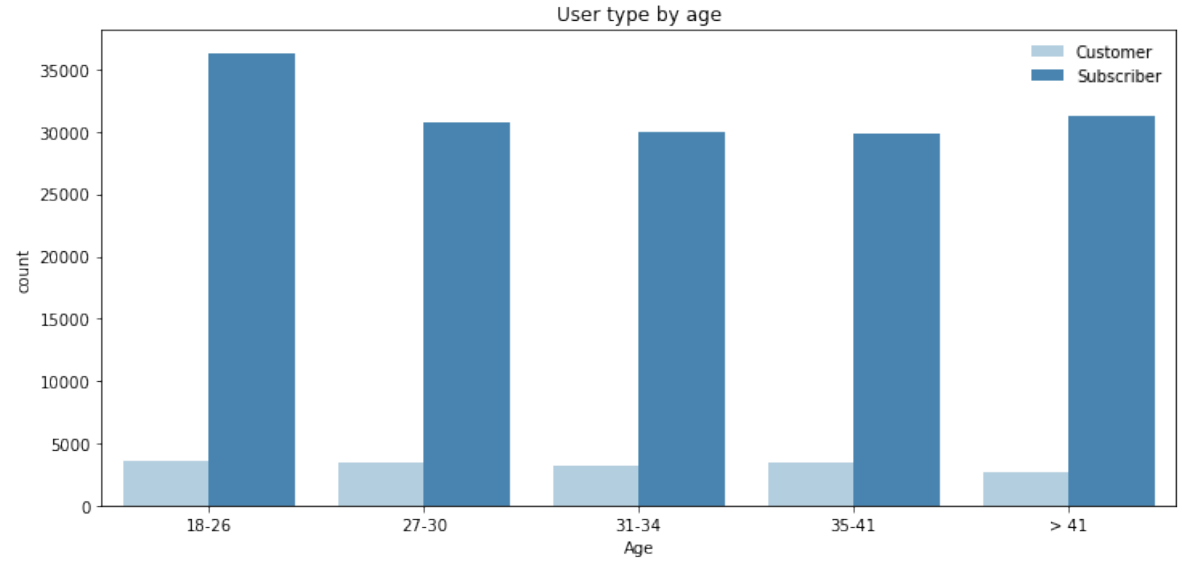
Taking a start station as reference, we are going to see how long those rides last and see if there are stations where duration is longer.

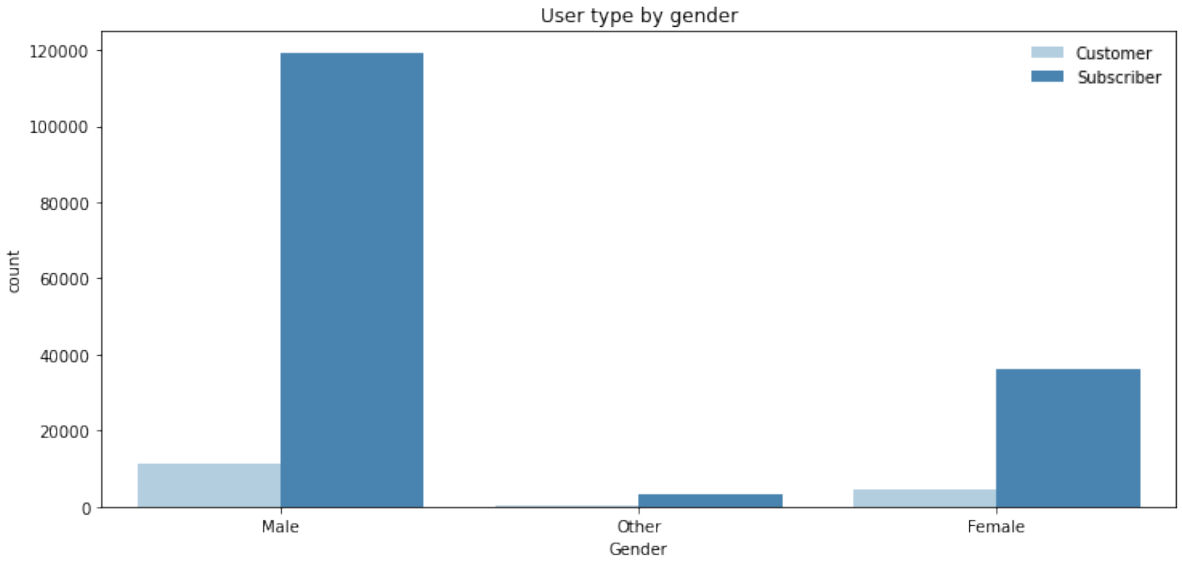


Again, we see that the south east cluster has the longest rides among users.

6.7 Subscriber by age and gender

I want to see the profile of subscriber by age and gender.



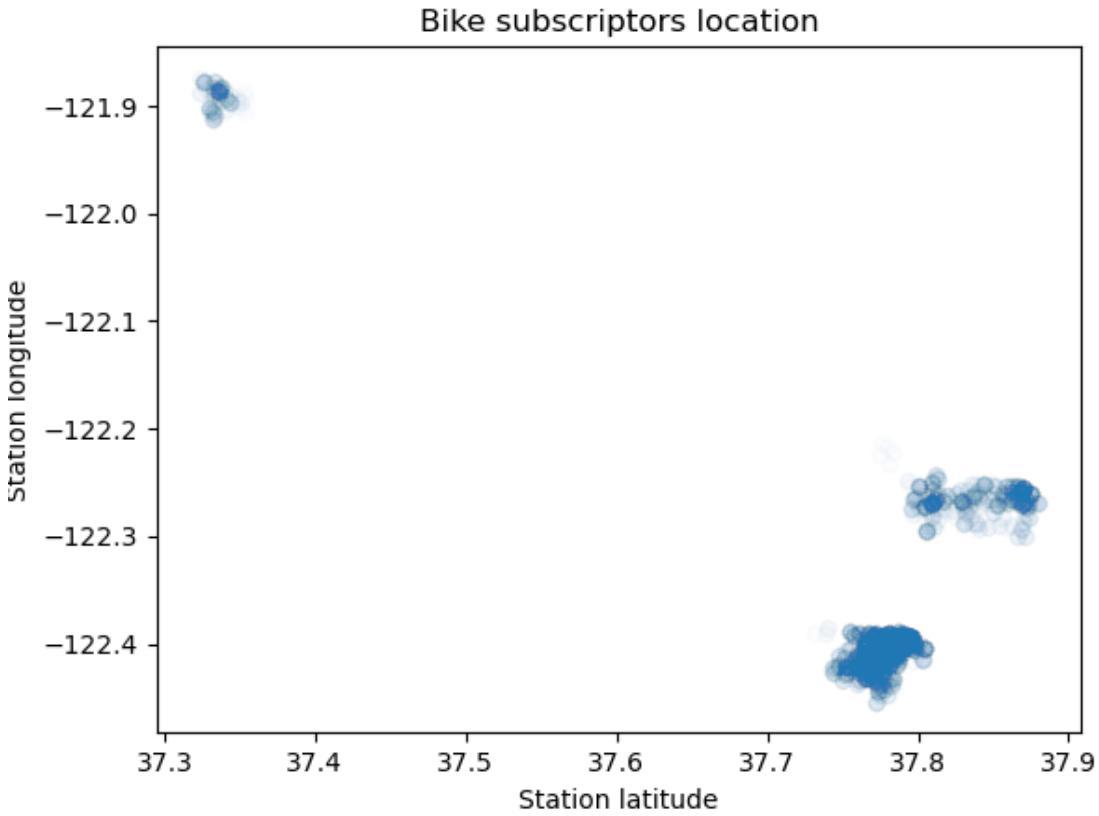


As we see in both barplots, subscribers are higher in number than customers for all age ranges and for all genders.
Also, we see that male subscribers are substantially more than female subscribers, and the highest count in terms of age was for the youngest of users, from 18 to 26 years old.

6.8 Stations and subscribers

6.8.1 Subscribers

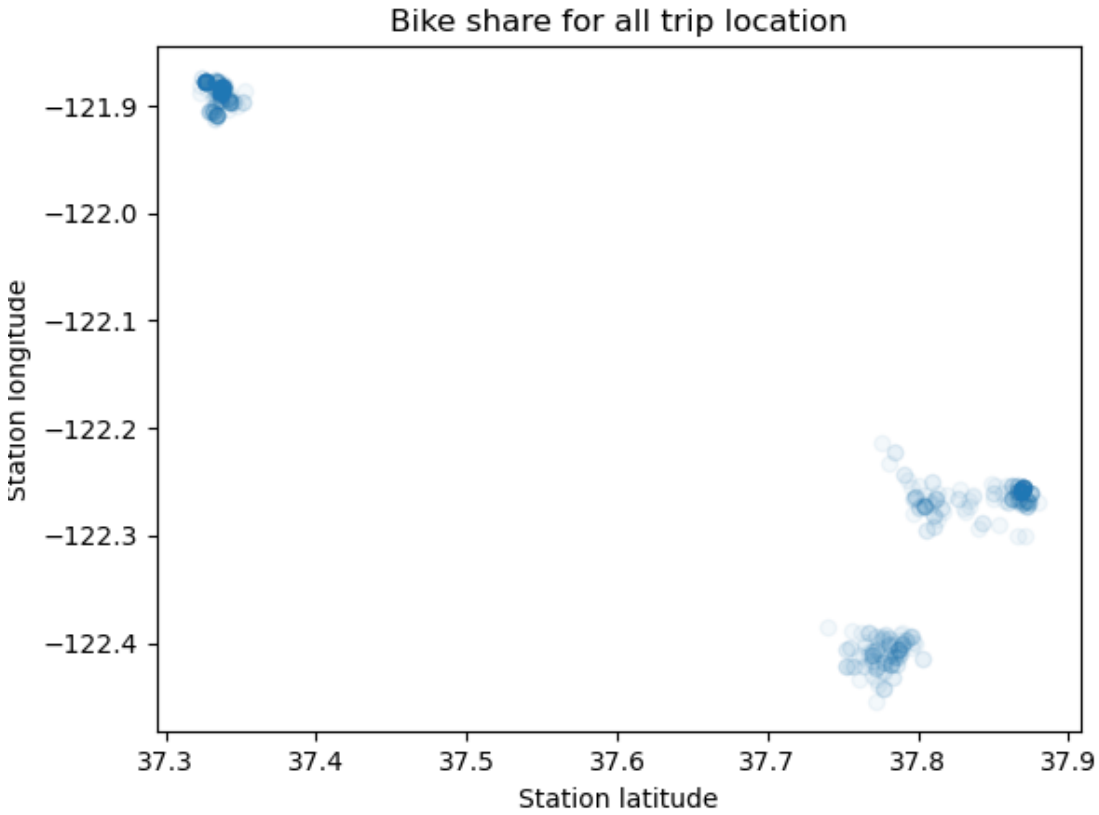
We will plot the stations and show the ammount of subscribers that usually use bikes on each one of them.



Again, the south east region is the one with most subscribers, which was expected since most rides started from there.

6.8.2 Bike share for all

Now we want to see where are the "Bike share for all" subscriptions located.



As shown in the plot, "bike share for all" trips are mostly located in the north west region, although not exclusively, since in the east region many of this subscriptions can also be found.
Finally, the south east region, which we saw presented the most rides and the longer ones, is the region with fewer of this subscription type.

7 Multivariate exploration

7.1 Ride duration across subscription type and weekday

In this section I will consider "Bike share for all" as a subscription type and therefore create a new column with unique values Customer, Subscriber and Bike share for all.

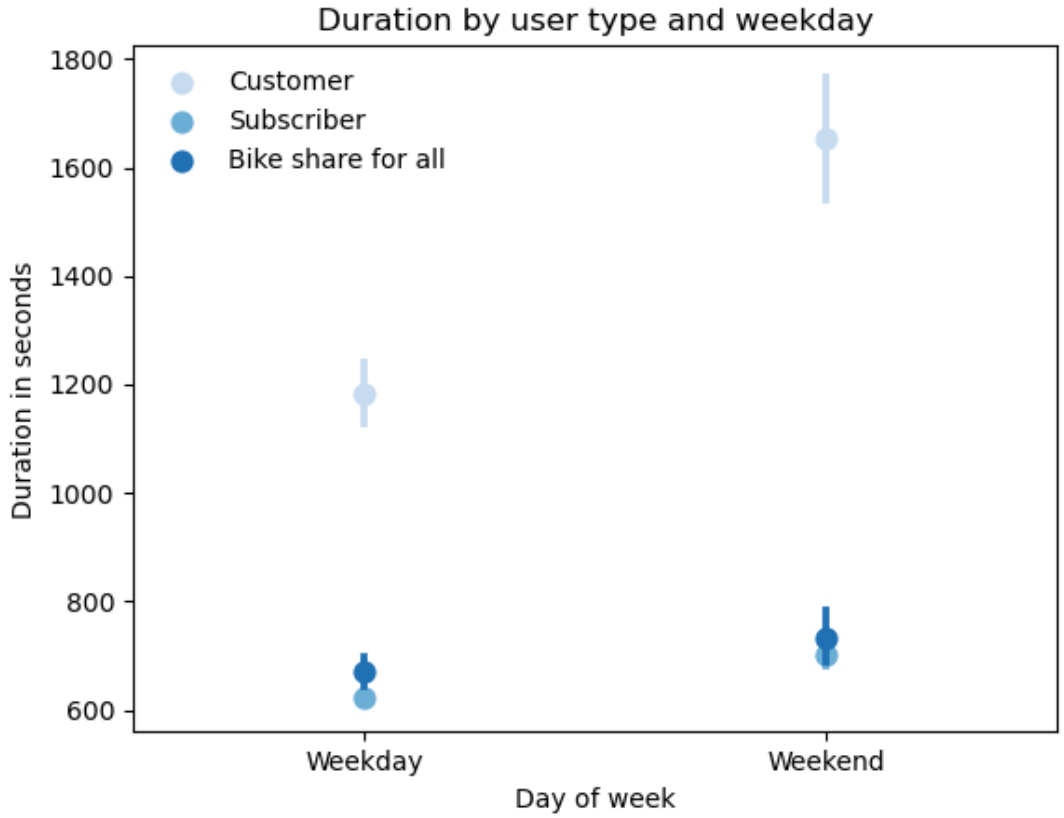


Figure 1: Duration by day and user type

We can see with more detail, how weekend rides usually take longer than weekdays rides. Also, customer users are the ones that make the longer rides, being close to double amount of seconds than subscriber rides.
Among subscribers, bike share for all system is slightly above subscribers in ride duration.

7.2 User age by subscription type and weekday

Let´s explore how age, subscription type and weekday interact among each other.

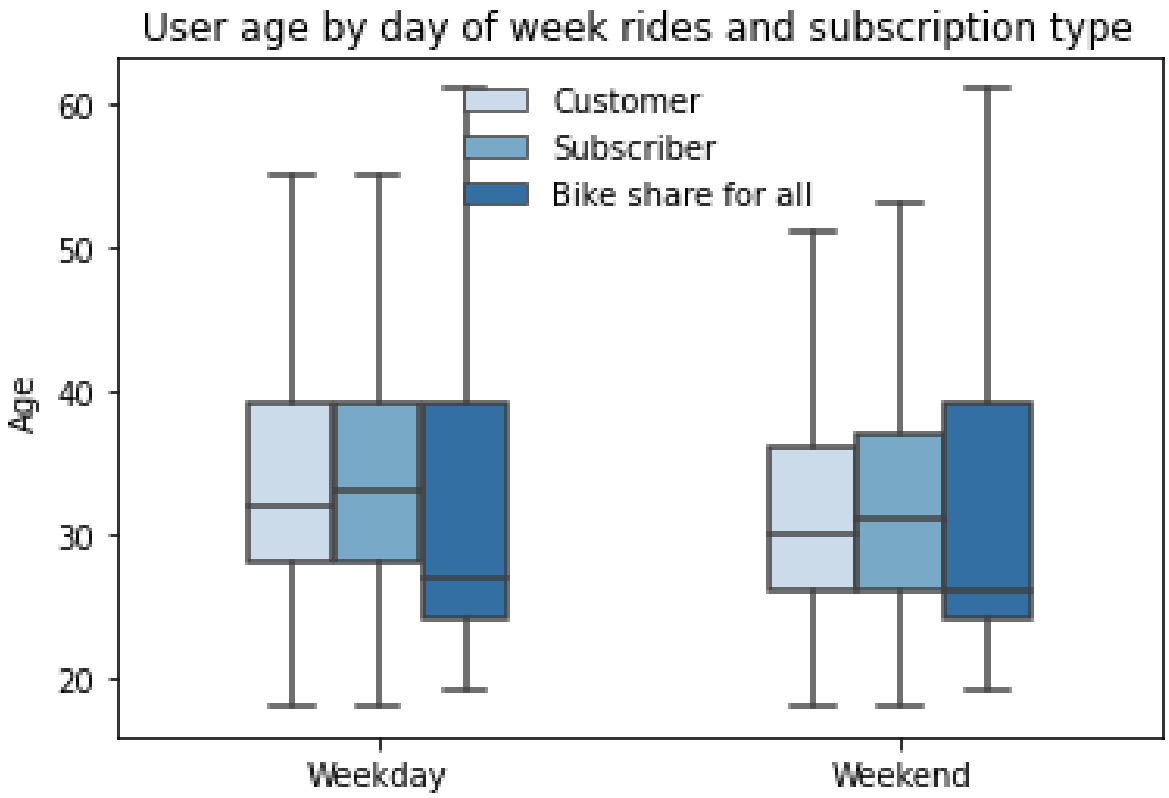


Figure 2: Age by day and user type

We can see here that "Bike share for all" users, tend to be younger in mean but present a wider range of age.
Besides, we can also see that on weekends, bikes tend to be used by younger people in mean.
While bike share for all users don´t show a different behaviour according to weekday type, on weekends customer and subscriber users tend to be younger than on weekdays.

7.3 User age by subscription type and gender

Is gender an important variable regarding user age and subscription type?

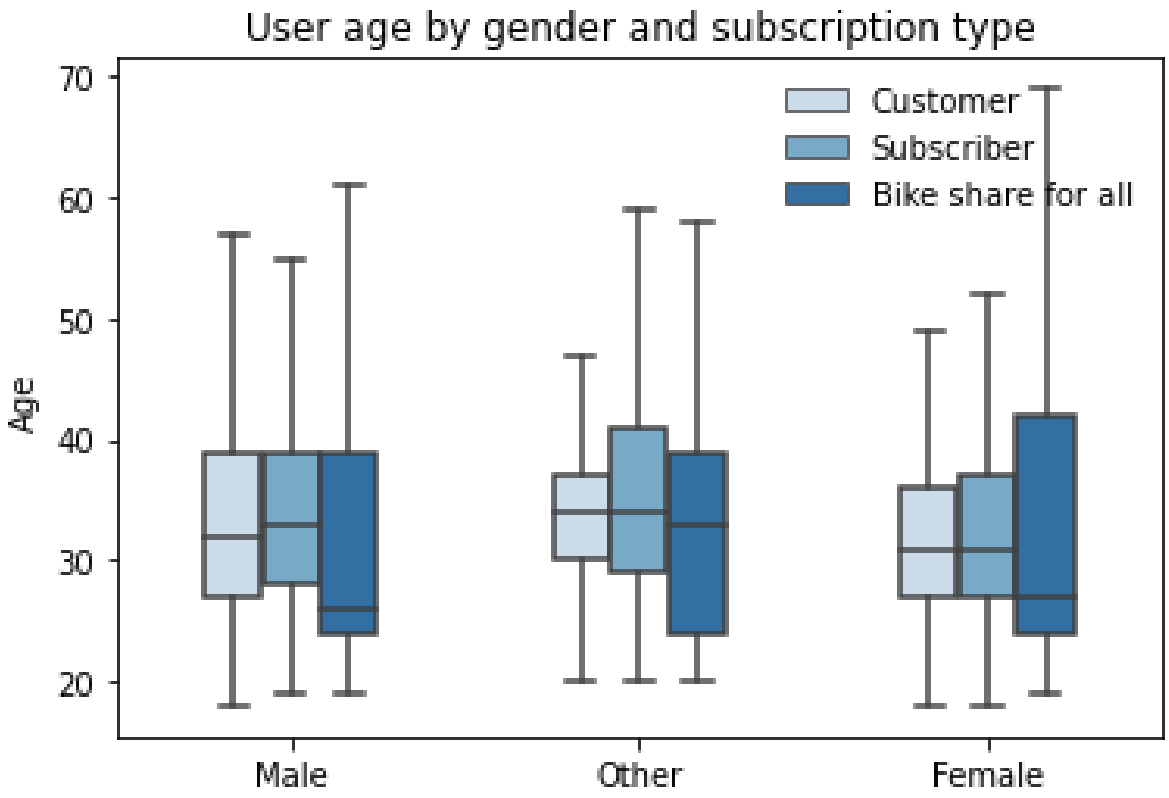


Figure 3: Age by day and user gender

It can be seen in this plot that even though there is no big difference among user age and gender, the range 25-40 years is the most important one in terms of ride counts. Besides, Bike share for all system tends to have younger users but with a wider range. Gender does not seem to be a significant variable in user type behaviour according to age.

8 Conclusions

It would have been very usefull to count with user_id information to be able to deep into behavior of a user after serveral rides, or subscription payment and the patterns that led to it.
We could see that most users choose the subscriber option, and that Bike share for all subscription corresponds to close to 10% of the dataset.
Almost 90% of the rides observed were made by male users.
Age that has the most users is in the range 28-35 years old.
Most rides last for 10 minutes and usually customer rides are longer than subscriber ones.
Same behaviour is observed on weekends, when rides are longer than on weekdays.

We could determine that user profile is a male of 30-35 years old willing to pay for a subscription that rides for ten minutes especially on weekdays and on the south east region.