(Ford GoBike System Data)

(Ali Mohamed Attia)¶

Table of Contents

Intro

This data set includes information about individual rides made

in a bike-sharing system covering the greater San Francisco Bay area.

Questions to be answered¶

When are most trips taken in terms of time of day, day of the week, or month of the year?

How long does the average trip take?

Does the above depend on if a user is a subscriber or customer?

How could Age and gender affect duration and frequency of trips?

At first let's explore the shape and contents of our data

	duration_sec	start_time	end_time	start_station_id	start_station_name	start_station_latitude	start_station_longitude	end_station_id	end_statio
0	52185	2019-02-28 17:32:10.1450	2019-03-01 08:01:55.9750	21.0	Montgomery St BART Station (Market St at 2nd St)	37.789625	-122.400811	13.0	Commercia Montgomer
1	42521	2019-02-28 18:53:21.7890	2019-03-01 06:42:03.0560	23.0	The Embarcadero at Steuart St	37.791464	-122.391034	81.0	Berry St at
2	61854	2019-02-28 12:13:13.2180	2019-03-01 05:24:08.1460	86.0	Market St at Dolores St	37.769305	-122.426826	3.0	Powell St B Station (Ma at 4th St)
3	36490	2019-02-28 17:54:26.0100	2019-03-01 04:02:36.8420	375.0	Grove St at Masonic Ave	37.774836	-122.446546	70.0	Central Ave St
4	1585	2019-02-28 23:54:18.5490	2019-03-01 00:20:44.0740	7.0	Frank H Ogawa Plaza	37.804562	-122.271738	222.0	10th Ave at St

```
## remove unwanted columns
df.drop(['start_station_latitude','start_station_longitude', 'end_station_latitude', 'end_station_longitude','bike_share_for_all_trip'],
axis=1, inplace=True)
df.head()
```

Wrangling Steps

Add column for duration in minutes to be easily presented on charts

Make column for age

Remove null values in birth year to calculate the age

Convert duration min from float to int¶

Separate start time and end time into date column and time column

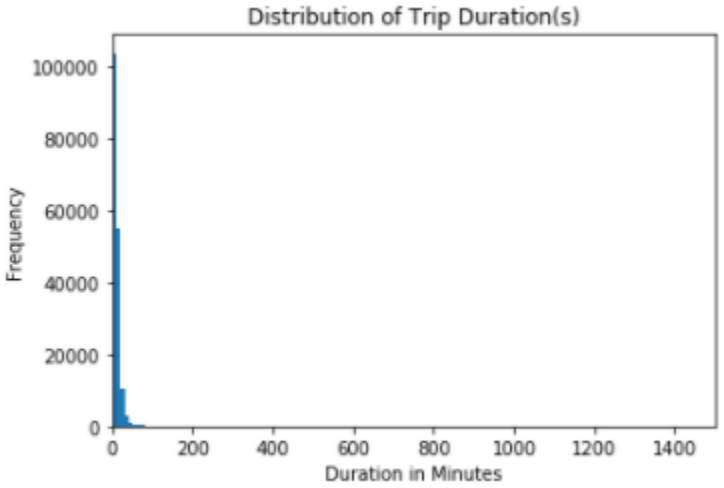
Convert member birth year into int

Separate Days from start date to identify the weekdays with high frequency of using bikes

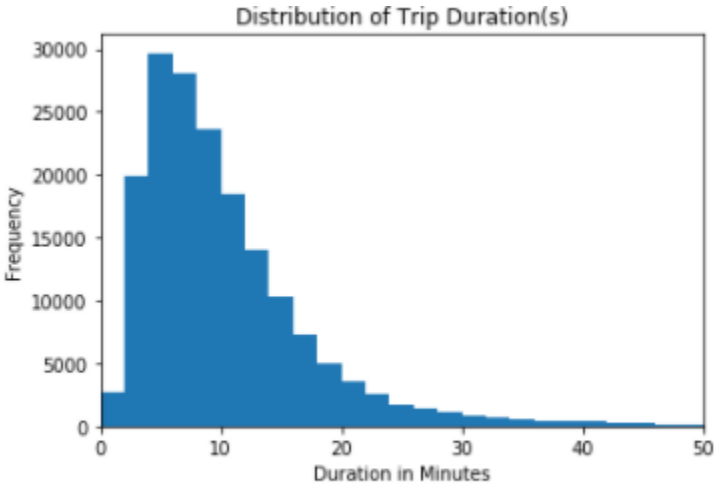
Univariate Exploration¶

We are going to making histogram to get the relation between trip duration and frequency to remove outliers

---

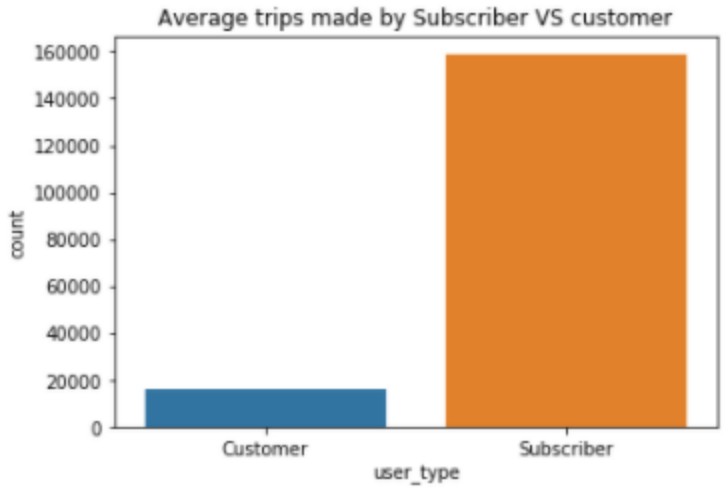


From this graph we find that most trips are from 1 minute to 50 So we are going to make a new plot without outliers



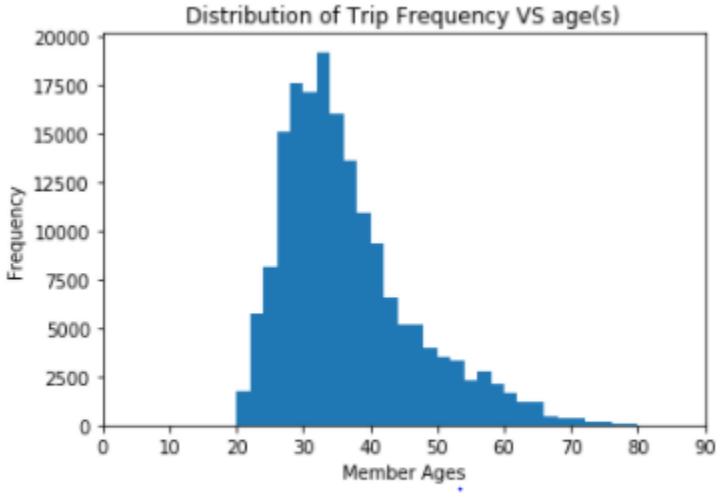
most frequent trips are from 2 minutes to 20 minutes

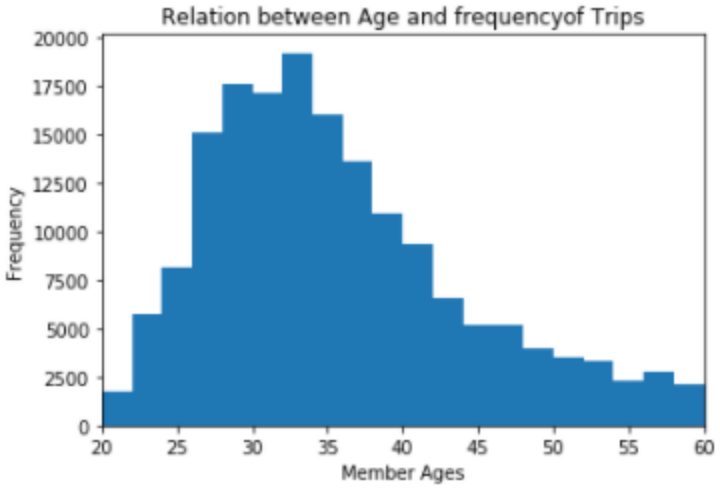
**Identifying the major category that use bikeshare



Subscriber clients are have the largest share of using bikeshare

Identifing Relation between Clients age and Trips frequency





The most frequent segment is between 25 and 45

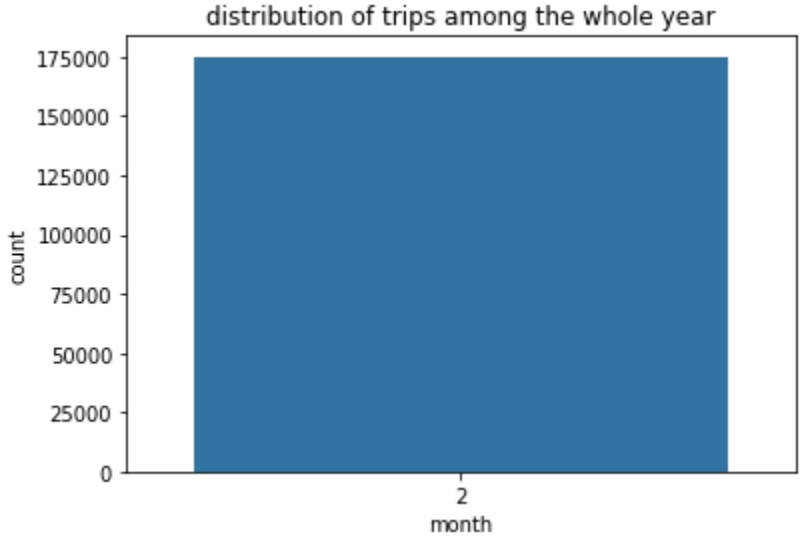
---

let's think about the top 10 locations from which the client start his trip

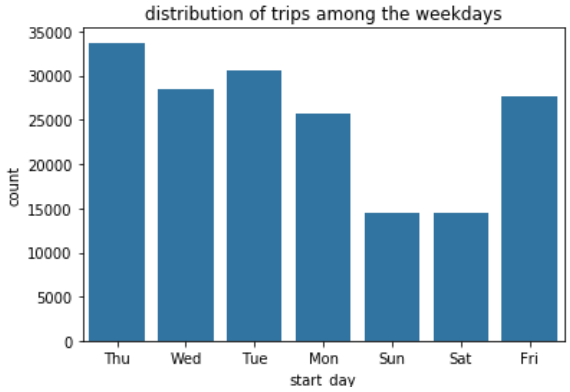
```
start_station_name
Market St at 10th St                                         3649
San Francisco Caltrain Station 2  (Townsend St at 4th St)    3408
Berry St at 4th St                                           2952
Montgomery St BART Station (Market St at 2nd St)             2711
Powell St BART Station (Market St at 4th St)                 2620
San Francisco Caltrain (Townsend St at 4th St)               2577
San Francisco Ferry Building (Harry Bridges Plaza)           2541
Howard St at Beale St                                        2216
Steuart St at Market St                                      2191
Powell St BART Station (Market St at 5th St)                 2144
```

We should increase the bikes located in these locations to avoid loss of clients

```
distribution of trips among the whole year
```




```
distribution of trips among the weekdays
```

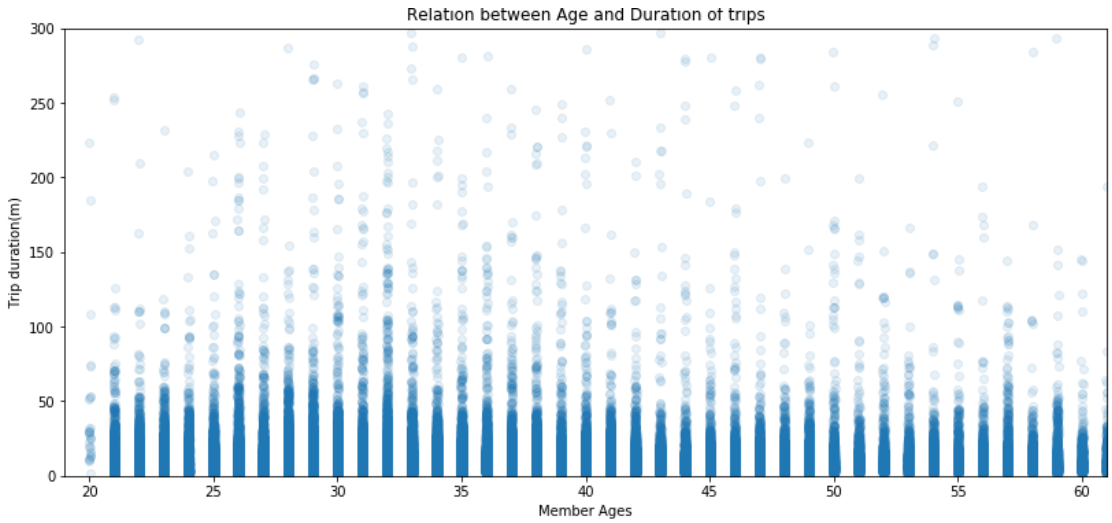


Bivariate Exploration¶

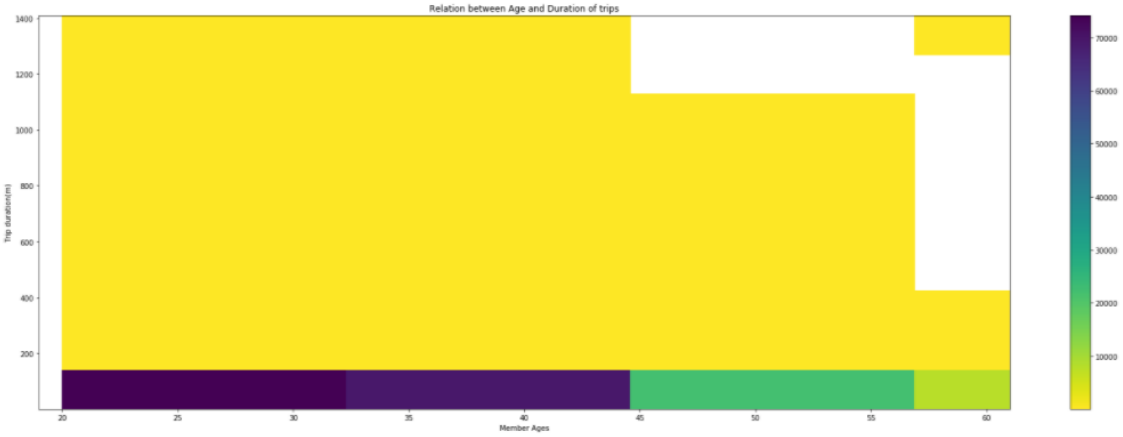
**Relation between gender and Trip duration considering other factors

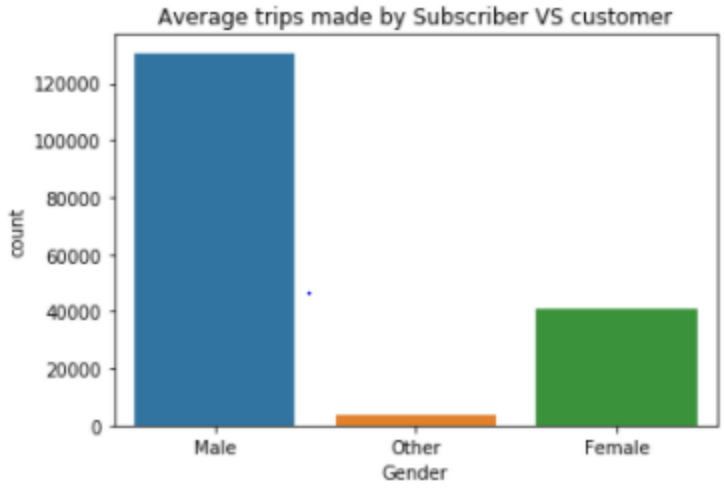
Does the age of User affect the distance of the trip?

We are going to start with scatter plot

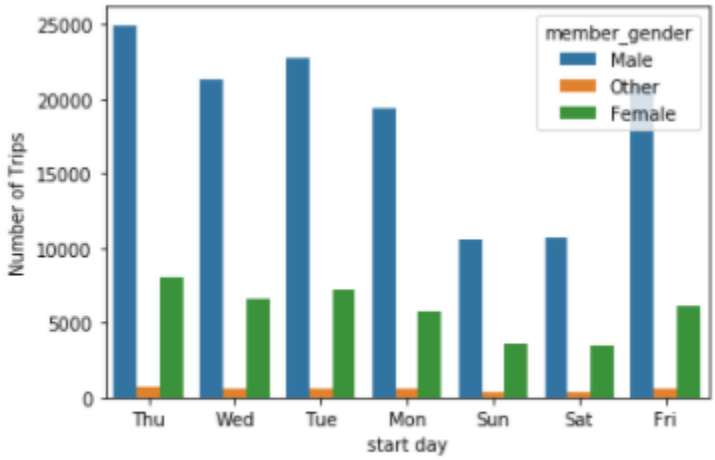


The heat map will be more representable than scatterplot in this situation



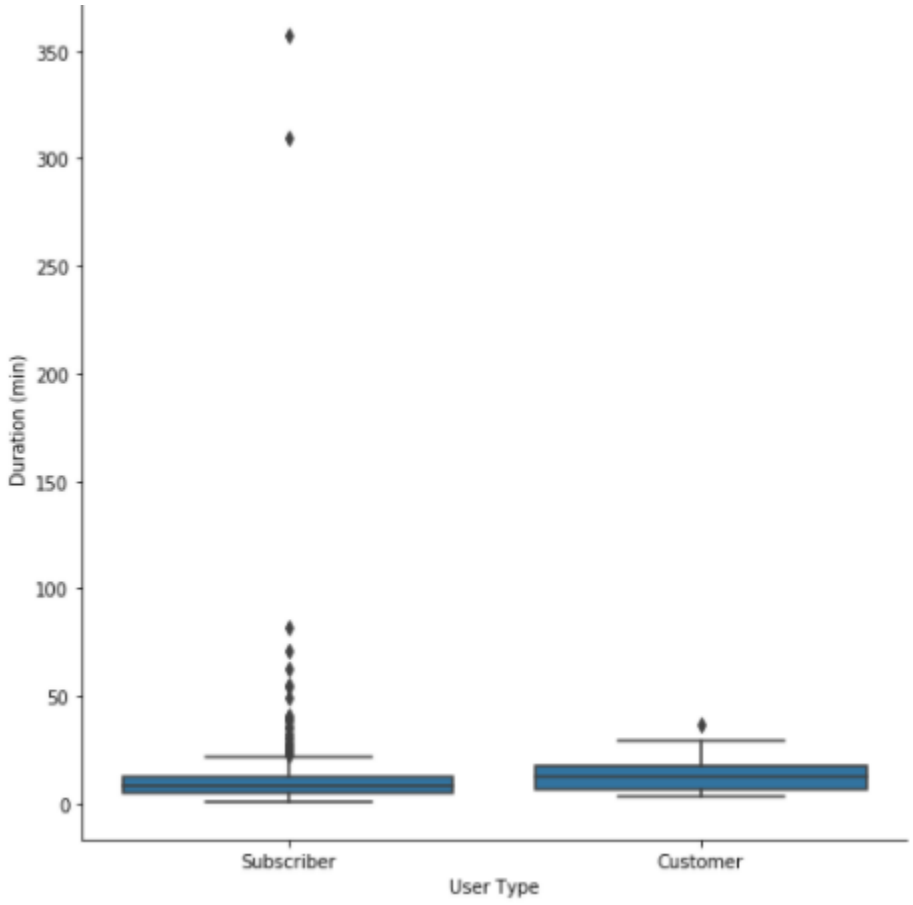


Male have more frequent trips with similar proportion to females and others

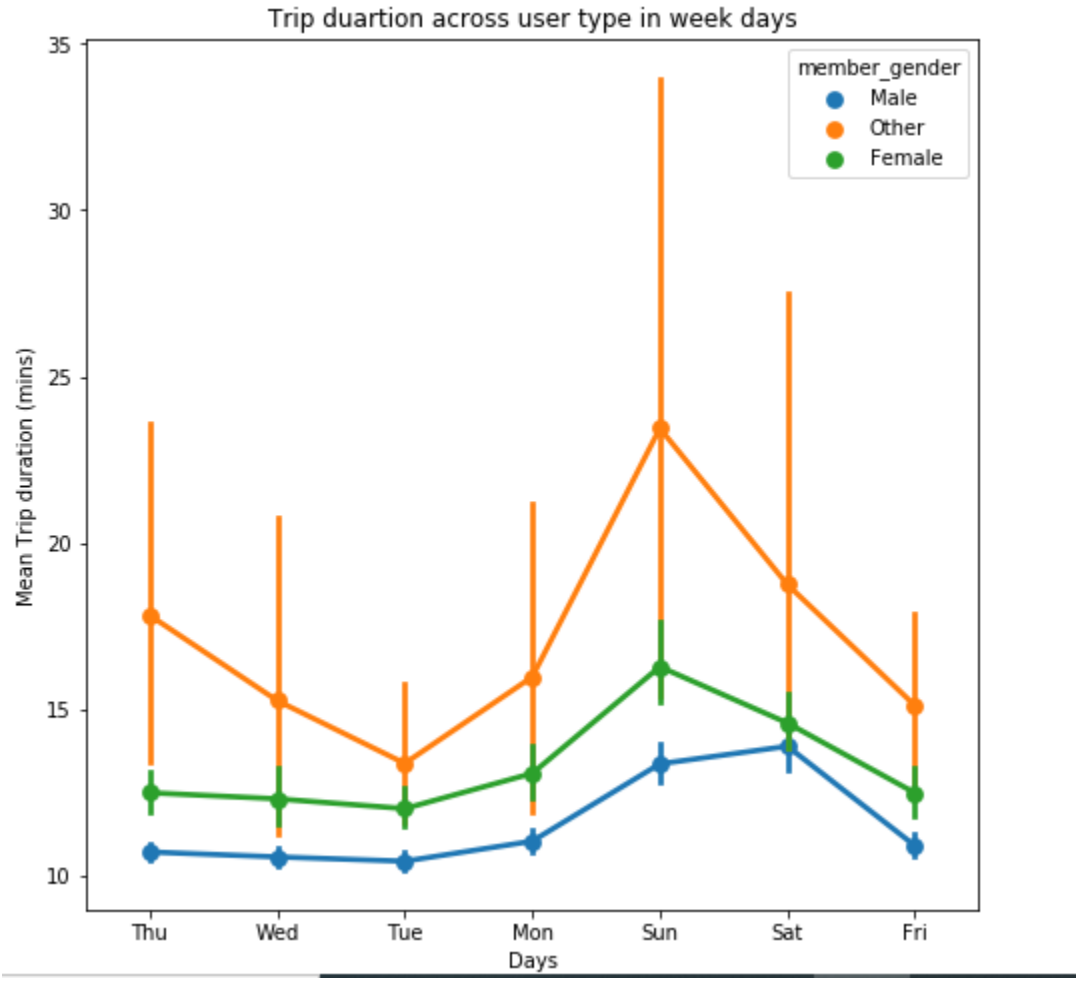


Males Use bikes more frequent than females

**The relation between the User type(Subscriber/Customer) and the time spended on the bike

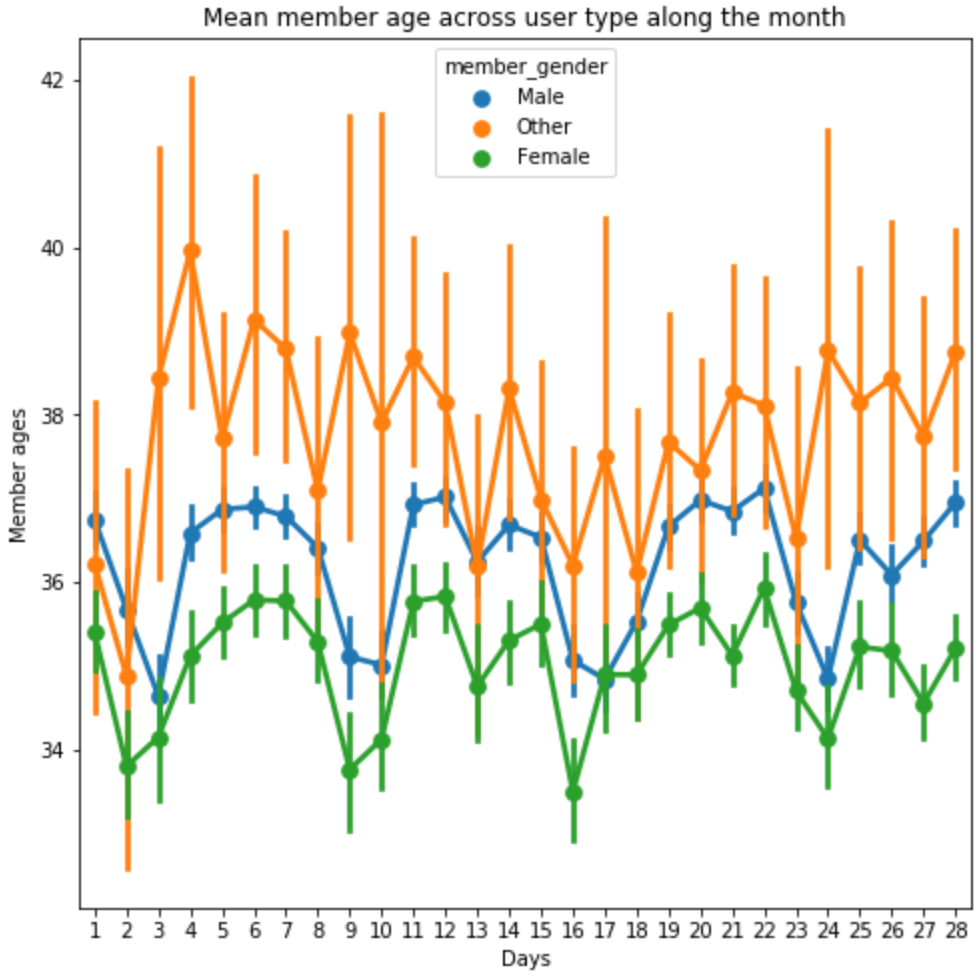


Multivariate Exploration¶



Others and females have have greater duration than males which might be due to driving slower than Males

Alsofrom this graph we conclude that the trip duration increase in holidays which might be beacause customers go in picnics or driving slowly as there is no need for rush for jobs



Females have the least mean ages Than others and Males in Using bikes

https://stackoverflow.com/questions/35364601/group-by-and-find-top-n-value-counts-pandas