Ford GoBike System Data

By: Abdelrhman Hamdy

Dataset :

the name of the dataset is “ 201902-citibike-tripdata.csv ”

you can find this dataset in following link :
https://s3.amazonaws.com/tripdata/index.html

the data set "Ford GoBike System Data" consist of 900,000 rows and 15 columns. columns contain features like : gender, age , user type, start trip , end trip, bike ID.

the data was relatively clean, i just made some modifiction :

1 - i checked it through some methods like : .head() , .isnull() , .duplicated() , value_counts() , max() , min() , shape

2 - converted start and stop time columns to “datatime” category and extracted hour, day of week, month, and years which helped to calculated users ages

3 - dropped all values of age that larger than 79 years as it's typo or not the regular users and effect my visualsation.

4 - dropped all values of trip duaration column that larger than 5000s, as it's outliers, and its percentage is only 0.002% of the dataset

Summary of Findings:

the whole project tried to answer two Questions :

1 - who use the service more ?

2 - who use the service longer ?

For the **first question** , here are Findings :

1- ages between 25 ~ 35 are the highest range of ages that use the service , and as the age increase age the usage of service decrease

2- Percentage of men is 74 % against just 21 % for women which is massive difference in gender- note there are about 5% of entries that have unknown gender-

3- in 8 A.M , 5 P.M , and 6 P.M are the rush hours of using the serivce

4 – 95 % of users are Subscribers against just 5% of Cusomers

5- the most days that customers use the service is in Weekends, in contrast Subscribers use the service most in Weekdays and least in Weekends

For the **Second question** , here are some stastics :

1- Although there are specifec ranges of ages that use the service more , there isn't strong relation between ages and the trip duration and all ages is roughly has average duration around 700 s

2- even that men are using the service much more than women, the average trip time for women is higher than men

3- the trip duration average increase between 11 A.M to 5 P.M to be from 700 to 750 seconds

4 – even subscribers are more than customers the average trip duraion for customer is higher at any time and between 1200 and 1300 against 600 and 700 seconds for subscriber

The previous findings will be used in explanatory analysis.

There are other findings that not realted to the above 2 questions , and here are them :

1- duration trip has very large distribtuion of values , so i considerd the values after 5000s are outlliers because it just represent 0.002% of my data and effect my visuialisation, there were some attributes associated with these outliers :

gender : men or unknown gender repesent about 80 % of these outliers

usertype : customers increased from 5% -in all dataset- to 51% in outliers

day : saturday and sunday is the most two days in which outliers can happen

users age : 50 ~ 55 ages is about 40 % of ages range in outliers

2- older people espiacally in 50s like to use the service as customers more than subscribers

3- both men and women has the same age distribution

Explanatory Key insights :

1 - Bike-share service is popular amongst men more than women.

2 - Most bike-share users prefer to subscribe to the service rather than use it when they need it as customers.

3 - Subscribers depend on this service as the main transportation method for going and returning from their workplace.

4 - Most people that using this service are between 25 ~ 35 years old.

5 - Young users tend to subscribe in the service, in contrast, old ones, especially in the 50s, tend to use it when they need it, which makes them customers.

6- The weekends are the least days the service used in.

7- The peak hours of the service usage is when people go and return form their work

8- Generally, if a customer uses the service, he/she will use it longer than a subscriber