(Ford GoBike System Data)

by (Diana Henry)

Dataset

This data set includes information about individual rides made in a bike-sharing system covering the greater San Francisco Bay area.

Project overview

This project has two parts that demonstrate the importance and value of data visualization techniques in the data analysis process. In the first part, I use Python visualization libraries to systematically explore a selected dataset, starting from plots of single variables and building up to plots of multiple variables. In the second part, I produce a short presentation that illustrates interesting properties, trends, and relationships that I discovered in my selected dataset. The primary method of conveying my findings was through transforming my exploratory visualizations from the first part into polished, explanatory visualizations.

the structure of my dataset

it contains `183412` rows and `16` columns

Dataset has 16 variables consist of :

- Two integer varaibles (duration_sec and bike_id).

- Seven float variables (start_station_id, start_station_latitude, end_station_longitude, end_station_id, end_station_latitude, end_station_longitude, member_birth_year)

- Seven objects (start_time, end_time, start_station_name, end_station_name, user_type , member_gender, bike_share_for_all_trip)

- There is NAN values in 6 columns: (start_station_id, start_station_name, end_station_id, end_station_name, member_birth_year, member_gender)

The main feature of interest in my dataset

I'm most interested in figuring out how trip duration is dependent on other features such as: age, user type, and gender from the dataset.

Data wrangling

Change `start_time` and `end_time` datatype to datetime Check for duplicates¶ Drop unneeded columns Drop all rows with NAN values for all Columns Change member_birth_year datatype to integer Add new column age to our dataframe

Univariate Exploration

First: Plotting categorical variables

- Plotting gender
- Plotting user_type

Second: plotting numeric variables

- Plotting Age
- Plotting trip duration

Bivariate Exploration

First: Trip Duration and age:(two quatitative variables)¶

Second Trip Duration and Gender: (a quantitative variable and a qualitative variable)

Third Trip Duration and user_type: (a quantitative variable and a qualitative variable)

Multivariate Exploration

First: Plotting age, gender, and trip duration: Second: Plotting user type, gender, and trip duration:¶

Summary of Findings

We can figure out that the proportion of males is around 75% of all observations We can figure out that 90.5% of all users are subscribers Most of users are between 20 and 40 years old We can see that most frequent users aged between 20 and 45. As remark, duration is registered by younger members

- We can notice that:

- Max values of trip duration for females(around 1650 seconds) is longer than the value for males (around 1450 seconds)
- `25%` of females spent around `400` seconds in their trip , but 25% of males spent around `300` seconds for their trip
- `50%` of females spent around `600` seconds in their trip , but 50% of males spent around `500` seconds for their trip
- `75%` of females spent around `850` seconds in their trip , but 75% of males spent around `750` seconds for their trip
- ==> We conclude that females spend more time for their bike trip

- We can notice that:

- Customers has maximum trip duration higher than for the subscriber.
- `25%` of customers their trip duration is less than `500` seconds, on the other hands, `25%` of subscribers spent around `250` seconds for their trip.
- `50%` of customers spent about `750` seconds for their trip, but `50%` of subscribers spent about `500` seconds for their trip
- `75%` of customers spent about `1300` seconds for their trip, but `75%` of subscribers spent about `750` seconds for their trip
- ==> We conclude that customers take more time for their bike trip

Trip Duration is dependendable on the age of the member, when the age between 20 to 45, the trip duration is higher than the elders.

I thought that variables which are user type and gender values having higher value to get higher trip duration but it is the opposite. For gender, value of male members is very high but it got lower trip duration. For user type, value of subscriber members is very high but it got lower trip duration then customer.

We can notice that:

- Number of subscriber is higher than number of customers.
- Trip duration for both user type is decreasing by age.
- For older users (age between 60 and 70) subscribers are higher in trip duration than customers.

For age, duration, and gender: I notice that:

- number of males is higher than females.
- the others leap at an older age (around 60 years) to got 3000 trip duration which is a peak.
- Number of Males is higher than number of Females.
- Trip duration for both gender is decreasing by age.

- For older users (age between 60 and 70) males take higher trip duration than females.

For age, duration, and user_type: I notice that:

- Number of subscriber is higher than number of customers.
- Trip duration for both user type is decreasing by age.
- For older users (age between 60 and 70) subscribers are higher in trip duration than customers.

Key Insights for Presentation

Distribution of Trip Durations: Trip Durations in the dataset take on a very large range of values. Number of Trips values first increases starting from around 8000 values to 12500 values at peak around 600 seconds but then starts to fall below at 2000 values.

Distribution of User Age: In the case of age, the distribution is more concentrated between 20 to 40 years old.