Project: Communicate Data Findings
By: Parvina Pasilova
For: Udacity Data Analyst Nanodegree Program

Read Me

This data set includes information about individual rides made in a bike-sharing system covering the greater San Francisco Bay area. First, we will import all the required packages, then I will overview the dataset, and proceed with data assessment, cleaning and visualizations. Lastly, I will analyze the main findings.

Following steps will be taken in the project:
- Supplement statistics with visualizations to build understanding of data.
- Choose appropriate plots, limits, transformations, and aesthetics to explore a dataset, allowing to understand distributions of variables and relationships between features.
- Use design principles to create effective visualizations for communicating findings to an audience.

Data: Bike-Sharing system in San Francisco Bay area

Each trip is anonymized and includes:
- Trip Duration (seconds)
- Start Time and Date
- End Time and Date
- 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 (Subscriber or Customer – “Subscriber” = Member or “Customer” = Casual)
- Member gender

The dataset requirements:
- include at least 600 observations. (This is the number of rows after tidying your data - see - the bullet points below about tidy data.) include at least eight variables.
- include at least one qualitative / categorical variable. (This can also be engineered / created.)
- include at least one numeric variable.
- be in a tidy format. In a nutshell, tidy data has each row as a single observation and each column reporting a single variable. You can read more about tidy data in Hadley Wickham’s paper [here]. We may need to do some cleaning and reshaping to tidy your dataset, before you actually get started with your exploration.

- be in a common data format. This includes .csv, .tsv, .txt, and .xls. Basically, there should be a reasonable pandas.read_*() function to open up your data in a tidy format as a pandas DataFrame.

Potential research questions:

- 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?

Main findings:

- I found that Subscribers are the main users of bike sharing system. So, the bike sharing system can target customer user type to expand their consumer segment.

- The majority of users are male. For reference, Bike sharing companies could target female users in the future and engage them in using bike sharing services.

- Actually, by plotting the start_station and the end_station I was able to identify the two major stations that bike users start and end their trip. The above results show that San Francisco Caltrain Station 2 and Market St at 10th St are the two major stations; these stations are probably the center of the city thus a lot of the users start or end their trips here and perhaps if the Bay Wheels would like to gain higher number of users can increase number of available bike sharing service in those stations.

- Another observation from above earlier we found out that the most trips started at 8-9 am and 5-6 pm and ended in the same time slots.

- The above plots show that busiest days for bike sharing are Sunday and Saturdays with the duration of the trips between 25~30 mins in average.