|  |  |
| --- | --- |
| DS 3001  Homework 3 | Carlos Barcelos  April 12, 2018 |

# Task: Visualization

## Introduction

As a bike share administrator, it is important to understand the use case for your bike share. Riders create massive amounts of data that can be useful in the decision making process for the business side of a bike share. For example, who should the bike share be marketed towards? Are more men using the system than women? If so, the money spent marketing towards women might be wasted dollars. Are certain bike stations being accessed more often than others? Then it might be important to invest in more bikes at that one station.

Through this assignment I will explore Boston Hubway’s bike share data and make an analysis that a bike share manager might want to make in an effort to better understand the user base of the bike share. By visualizing the dataset, I will be able to make quick and meaningful decisions that are in the best interest of the bike share.

## Data Collection and Preprocessing

The dataset that I used is from the city of Boston’s bike share program, The Hubway. Their [system data](https://www.thehubway.com/system-data) is available online organized by month in CSV format. For this homework, I narrowed down my search to explore one year of Boston Hubway trip data. In order to access all of this data, I downloaded it from the Boston Hubway website and [concatenated](https://pandas.pydata.org/pandas-docs/stable/merging.html) it using Pandas. This allowed me to manipulate an entire year’s worth of bike share data.

The columns in the data set and descriptions thereof are as follows:

|  |  |  |
| --- | --- | --- |
| Field | Description | Type |
| Trip Duration | The time (in seconds) the bike was taken out. | Integer |
| Start Time | The date time of the beginning of the rental. | Date Time |
| Stop Time | The date time of the end of the rental. | Date Time |
| Start Station | The stating station ID, name, and location. | Integer, String, Lat, Lon |
| End Station | The ending station ID, name, and location. | Integer, String, Lat, Lon |
| Bike ID | The unique identifier of the rented bike. | Integer |
| User Type | Customer = 24-Hour or 72-Hour Pass user  Subscriber = Annual or Monthly Member | String |
| Birth Year | The year in which the user was born. | Date (Year) |
| Gender | 0 = Not reported  1 = Male  2 = Female | Integer |

I then [dropped](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.drop.html) rows from the data frame that would not be helpful in my visualization. They are the start and end: ‘station name’, ‘station latitude’ , ‘station longitude’. The station name may be a useful metric in the future when considering which station is the most used, but at that time it will be simple to cross reference it from the original dataset.

Because the dataset was cleaned originally by Boston Hubway, the only other task I had to complete in preprocessing was to turn the ‘Start Time’ and ‘Stop Time’ into a more usable