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 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 it using Pandas. This allowed me to manipulate an entire year's worth of bike share data. In total, there were 1,313,774 rows in the final, concatenated dataset.

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

Table 1 The fields and descriptions of the dataset used for the following visualizations

I then dropped 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 format. To do this, I used the Python datetime module to separate the 'starttime' into a 'start_date' and 'start_time'; I then repeated this process for the 'endtime'. These datetime objects are much more suitable for data processing.

Goals

Thinking like a bike share administrator, these are the following questions that I want to answer through my visualization:

- What is the distribution of the riders?
    - Men v. Women & Customers v. Subscribers
- What bikes are being used the most?
- What is the most popular time that bikes are being used?
    - Day of week & Month of year

Visualization a single aspect of this dataset would be insignificant. To understand the Hubway bike share system, it is necessary to create multiple visualizations. The visualizations created for each of the three questions above are a start at understanding the system as a whole. Even more visualizations could be created to future subdivide the dataset and understand Hubway bike share system.

Data Visualization

The visualizations that I created were all made using Pandas data processing and matplotlib. My goal was to communicate the health of the bike share program to the system administrator. This being the case, they want to know about their users (Figure 1), their bikes (Figure 2), and when people use the bike share (Figure 3).

Figure 1 describes the user base of the bike share. A pie chart is appropriate because it easily conveys the percentage of the entire usership that is male vs. female or are customers vs. subscribers. The Male vs. Female chart uses the stereotypical colors for boys (blue) and girls (pink) while reserving yellow for unspecified genders.

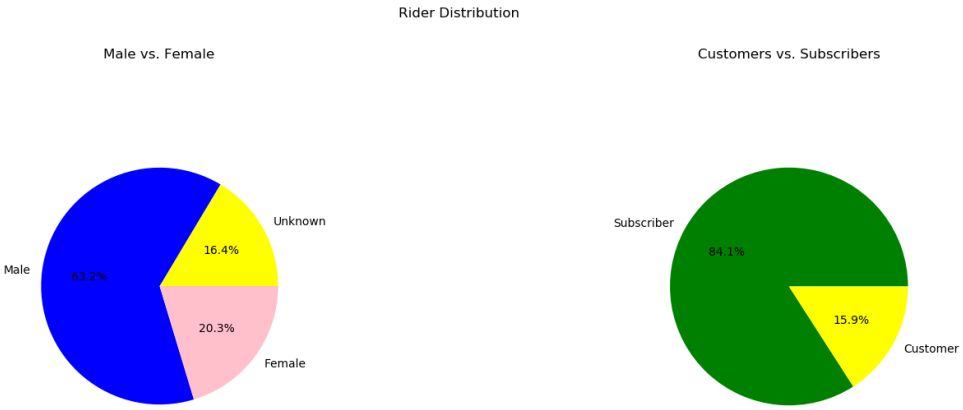

Figure 1 The rider distribution of the bike share

Figure 2 lists the top 10 used bikes by their ID number. While this visualization can quickly show the most used bikes in descending order it does not show the whole picture. To show all the bikes there would be an unreasonable number more bars thus cluttering the chart. Furthermore, this visualization type would benefit from having exact numbers of rides per bike on top of the bar to avoid extrapolation.

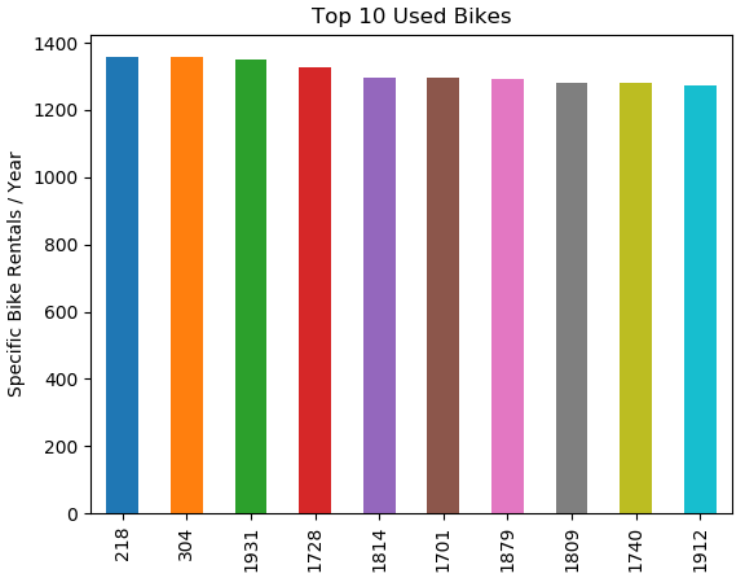

Figure 2 The ten most frequently used bikes (by ID)

Finally, Figure 3 gives specific usage according to the month of the year or time of day. It is a useful histogram to help the bike share administrator understand when users are using the bike share. Utilizing a bar graph helps to show the moving trend of the usage. For example, in the *Total Rentals / Month* chart it shows a clear trend that as time moves towards the middle of the year, and the hotter months, the bike share becomes more used.

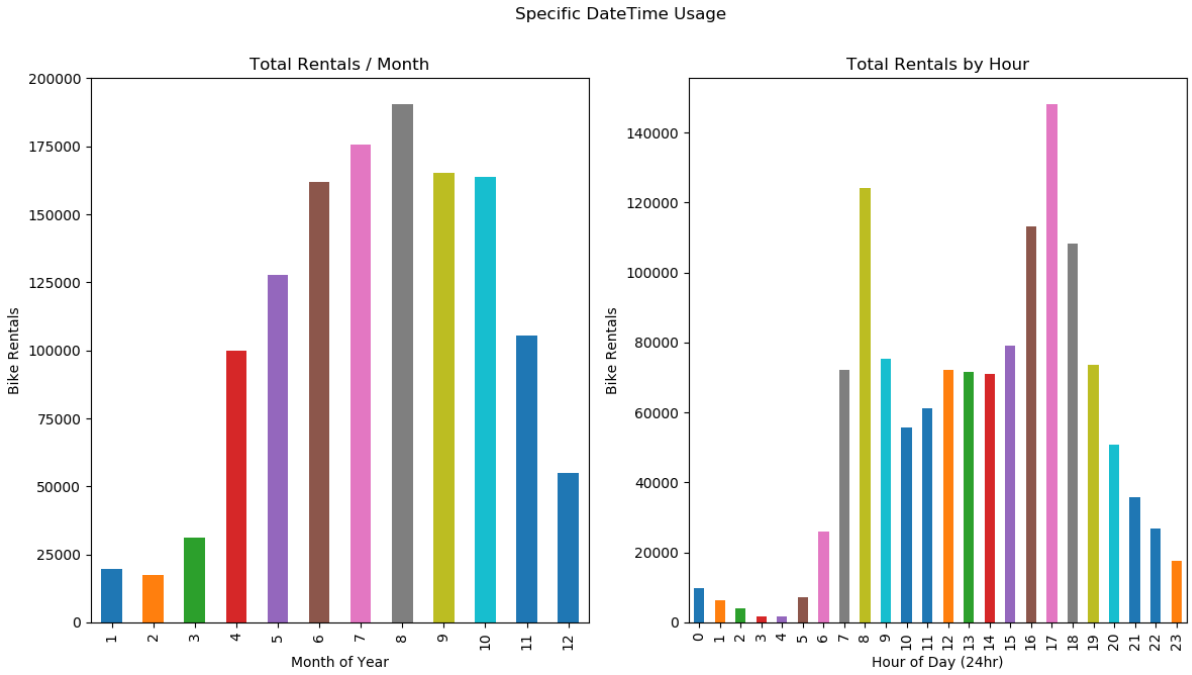

Figure 3 Specific date and time usage of the bike share