San Francisco Bay Area Bike Share Visialization

By: Madeleine (netID), Lauren (netID), Chris (cjl248)

We first gathered our raw data from a bike sharing organization based in the San Francisco bay area[[1]](#footnote-1). It was offered in CSV format and in two zip files. The first held data from August 2013 to February 2014 and the second from March 2014 to August 2014. Each of these zip files contained four csv files, which stored extensive data on areas such as bike and dock availability by station, station information like longitude, latitude, name, trip information like length and type of rider, and weather data by date. We decided to focus mainly on the type of costumer, either a subscriber to the bike share network or a casual rider and discern any differences in trip time, as well as variance across the four seasons. Explicitly our variables were decided to be type of rider, either casual or subscriber, date, season and duration of trip.

Our first step after gathering the raw data was to combine the respective csv files of interest to us from the first six and second six months of data. This was completed using excel by deleting extraneous columns of data in each file. In the weather\_data files we left only the “Date” and “Mean\_Temperature\_F” columns. In the trip\_data files we left the “Duration”, “Start Date” and “Subscription Type” columns. The remaining data from each of the weather\_data files were combined into one file by simple copy and paste. The same was done for the remaining columns in the trip\_data files. The rebalancing\_data and station\_data files were discarded.

[TO DO: Madelein: talk about further preprocessing in python]

After deciding to focus on differences participation between the two types of riders, our criteria for choosing further data was that it was explicitly tied to bike users and trips. Therefore data such as longitude and latitude of bike stations, available bikes at stations and similar data were discarded.

In terms of mapping our selected variables to visualizations, we decided to have a stacked bar graph containing a bar for each day of the year as our main figure. The number of users determines the height of each bar for each particular day with casual users on the bottom and subscription users stacked on top. Distinguishing these two groups is a different fill: diagonal strips for subscription users and a plain fill for casual users. Next, we mapped the seasons to the color of the fill for these bars: green for spring, yellow for summer, orange for fall, and blue for winter. Temperature is represented by the saturation of each bar’s color: a warmer temperature results in heavier saturation while colder temperatures are shown by lighter saturation.

[TO DO: How temperature ranges were mapped to specific saturations]

[TO DO: Description of line around circle mapping trip duration if we still decide to do this]

Our visualization focuses on the distinctions in behavior between types of participants in the bike share program and may offer insight to the company on measuring the desirability of having more subscription users. It allows us to see how the ratio of these two numbers changes in warmer and colder seasons in terms of frequency of use as well as length of use.

[TO DO: Once graph is done, add what if anything is surprising about it]

1. <http://www.bayareabikeshare.com/datachallenge-2014> [↑](#footnote-ref-1)