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 and name, trip information like length and type of rider, and weather data by date. We decided to focus mainly on the variation in number of riders each day through out varying seasons and temperatures. Our variables included number of riders and temperature per day as well as season throughout the year.

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 “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]

When deciding exactly what data to use from the bike share program we looked for variables that had sufficient range as to provide interesting conclusions but that could be easily and interestingly manipulated.

After deciding to focus on differences in ridership through out the year our criteria for choosing further data was mainly that it was related to the number of riders. Therefore data such as longitude and latitude of bike stations, available bikes at stations and similar data were discarded.

In terms of presenting our data, we decided to use a sunburst-like visualization, which serves as the foundation for our yearly ridership numbers. The ridership variable is then mapped onto bars for each day of the year. The number of users determines the height of each bar for each particular day. 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. Our third variable, 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: Madeline, what temperature ranges map to what saturation. This can be as simple as giving the step size]

Our visualization focuses on the distinctions in behavior of riders in relation to daily temperatures as well as seasonal trends. Patterns in ridership may prove useful to the bike share program by offering insight on indicators of high participation, which can help them to stock certain stations in advance. While numbers of riders rise in the warmer months and fall in the cooler months, as expected, perhaps more interesting is the fact that ridership is much higher during the week and tapers off on the weekends. This trend can be seen year-round regardless of season and seems to suggest a more functional use of the system such as for commuting to and from work, as opposed to mainly recreational reasons.

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