CS171 FINAL PROJECT PROPOSAL: VISUALIZING BIKE SHARE DATA

By Michelle Choi, Alina Ranjbaran, and Sam Udolf

INDEX

BACKGROUND AND MOTIVATION

With the growing movement towards decreasing fuel emissions from vehicles, there has been a greater push towards carpooling, public transportation, and especially, the good old bicycle. As students frustrated by living in a city where car and public transportation can be difficult, a bicycle can be a great solution and a much faster alternative to walking. Boston is fortunate to be a city that has publicly available bicycles for rent, a system called Hubway. Each of us have had a positive experience with Hubway, whether it was purely functional to get to a location more quickly, or simply for pleasure in order to explore the Boston area. As a result, we are very interested in breaking down the factors that contribute to the demand and usage of Hubway bikes and visualizing the factors in new ways to better understand them.

Additionally, we aim to compare the Hubway data with that of DC's comparable Capital Bikeshare program. We want to be able to compare the ways that the bikes are being used in both cities, depending on their stations, types of users, and weather conditions. We think that there will be some interesting comparisons to be drawn between these two cities, as they are relatively similarly sized cities with surrounding suburban metropolitan areas.

We believe that there are many intricacies at play with bikeshare usage and interesting trends to be discovered in the data. Since there are so many ways that a person can use these bikes, it makes for a very compelling study for creating visualizations.

PROJECT OBJECTIVES

We are trying to explore how the demand for Hubways changes as a function of day, time, location, weather, and type of user. Specifically, we would like to learn whether or not more Hubways are used on weekdays versus weekends, by casual vs. registered users, and measure the speed and frequency by which bike rentals occur. We will compare Boston Hubway data against DC Capital Bikeshare in order to tell a more compelling story and understand the variables of the data relative to each city.

This project would benefit the average user of Hubway looking to gain a bit of insight into the busy times that people are trying to rent bikes in order to optimize usage and also figure out which bike stations are less frequently taken from or more frequently brought to. This project would also be interesting to Hubway and Capital Bikeshare, as it would give them information about how their bikes are being used in order to provide better services that are tailored to how each city, and potentially each station, uses their bikes. For example, if one station is mostly used by casual users, Hubway may want to add a helmet hub there because those users are less likely to actually bring their own helmet with them as opposed to registered users. We can also learn a bit about the social implications of riding on the shared bikes to see if users are more likely to ride in groups and the age and gender of such users.

DATA AND DATA PROCESSING

Data Sources:

http://hubwaydatachallenge.org/trip-history-data/
https://www.capitalbikeshare.com/trip-history-data

We are collecting our data from the hubway and capital bike share websites. We will aggregate the capital bike share data for the two years we want to look at, July 2011-July 2013. The data is publicly available to all on their website in csv format. Weather data would need to be obtained using scraped data using Open Weather Map API or Weather Underground API for the desired time periods we are looking at.

Data Processing Procedure:

We have obtained cleaned csv files from the Hubway and Capital Bikeshare sites. A small amount of wrangling will need to be done to combine the files for DC, as they are currently in separate csv's by quarter. We may elect to scrape weather data to analyze the effects of a day's weather on bike usage using an API either for Open Weather Map or Weather Underground. We may also elect to interact with the Hubway API to query the data directly to create custom parts of the data sets instead of using the entire csv file for the data. We plan to derive quantities from the data that measure the way that bikes are rented and used over time for specific stations and demographics of users.

VISUALIZATION

We plan to use Leaflet to create a map visualization with a hover feature over each station to obtain specific information regarding the demographic information of riders and amount of traffic of bikes to and from that station (Figure 1). There will also be a feature to look at the stations by day and hour, in order to include a time element to the visualization. In addition, there will be a way to visualize demographic information about the bikers as an aggregate separate from the map in order to get a sense of the types of riders who use the bikes. This feature (Figure 2) will be on a separate tabbed page from Figure 1 and will provide an overview of the bikers as opposed to the stations.

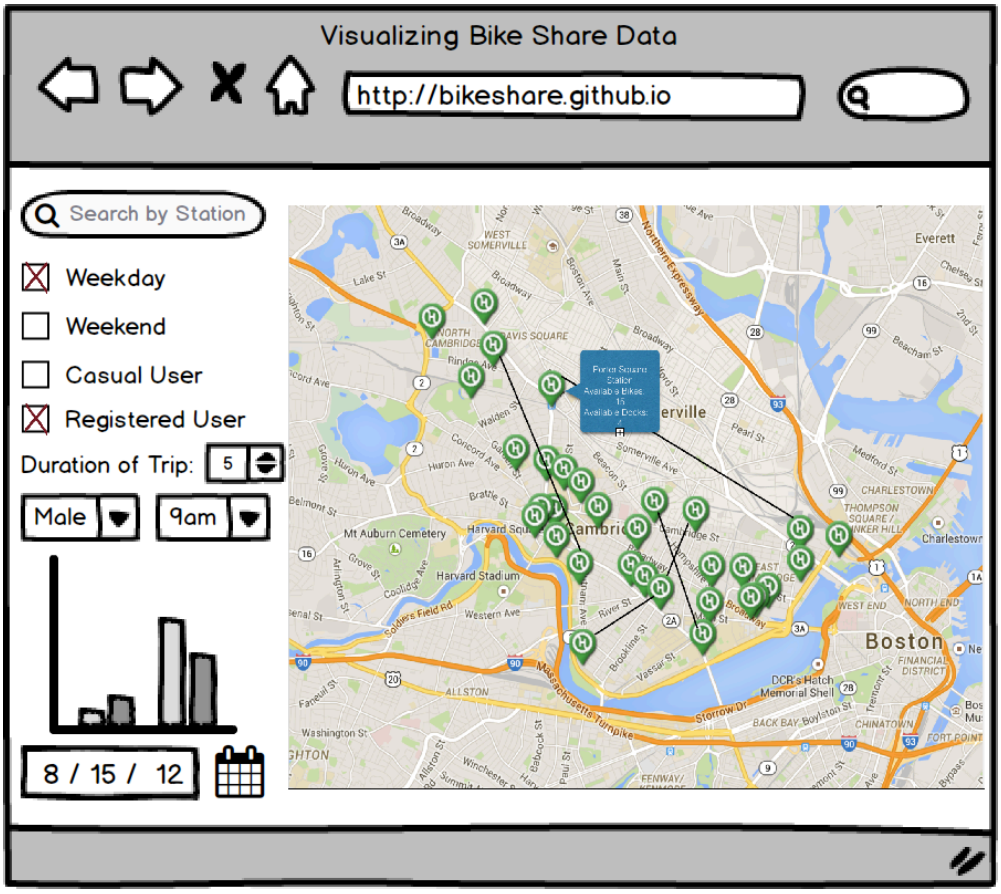

Figure 1: A sample website of the visualization and a few of the features available

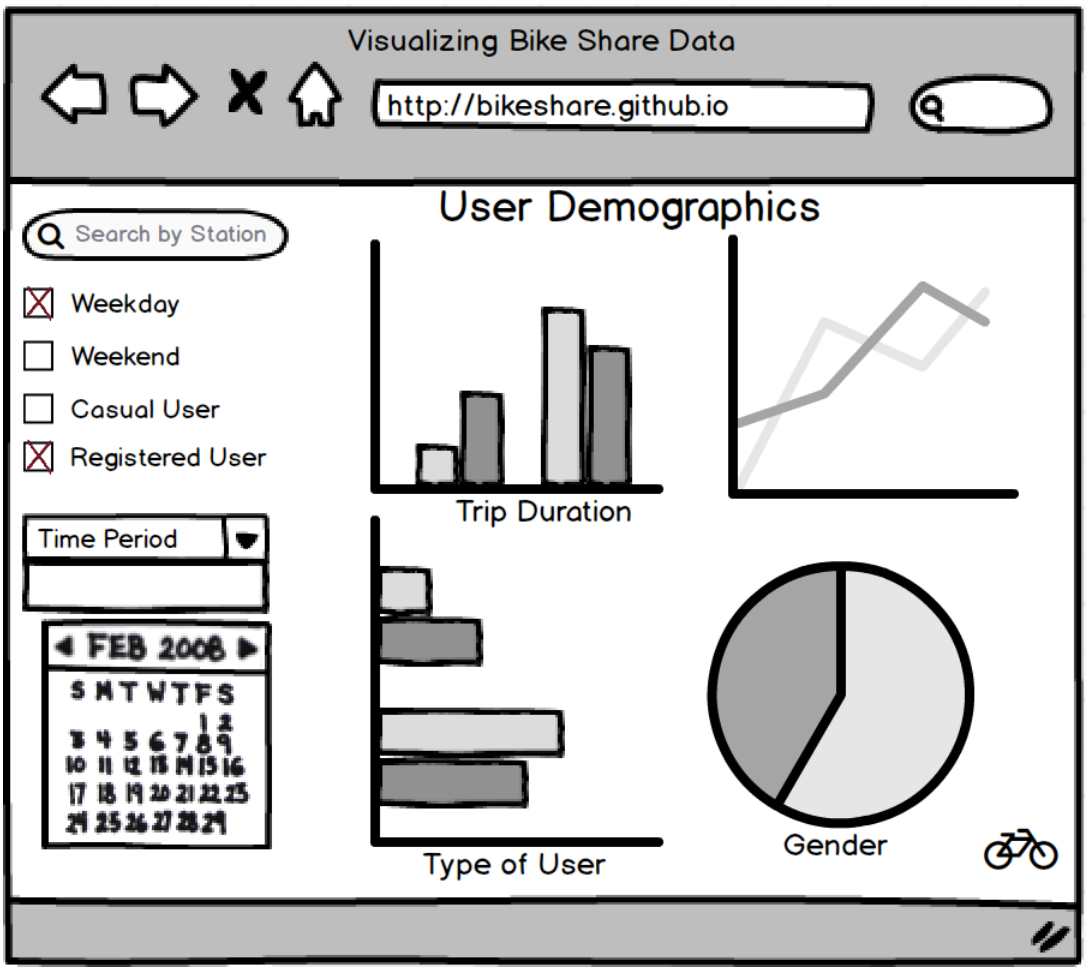

Figure 2: A sample website of the in-depth user demographics view.

MUST-HAVE AND OPTIONAL FEATURES

Must-Have Features:
We will create a map that indicates people's traveling routes: their starting station, ending station, with the ability to filter by trip duration, time of day, type of user, gender, and day. A user will be able hover over each line to get all the demographic information about that user and their trip. In addition, the user may hover over each station to get information on the number of bikes at the station at the time and how many bikes have gone in and out that day.

In addition, selecting a station would yield a separate graph to update usage by brushed time period in a chart, being interactive with the other demographic information to see cross-sectional data for a particular station to compare how different stations are doing.

Optional Features:
We would ideally also like to create a movie-type simulation that shows a typical day with a sped-up visualization of how each bike on a given day traveled from station to station. A comparison of the two cities, Boston and DC, would be secondary, depending on how much time it takes to implement the visualization for Boston first, because that is our main interest.

PROJECT SCHEDULE

- April 8
    - Alina- Collect, aggregate, and scrape all data
    - Sam- Finalize the structure of the code (i.e. MVC setup, coordinate how all the different filters work together, decide whether different types of visualizations are necessary for different sets of filtered data)
    - Michelle- Do the first map visualization of routes (i.e. decide how the data will be formatted with Alina and write the code to represent user paths on a map)
- April 12
    - Alina- Complete process book to current state (flesh out overview, related work, questions, data, and exploratory data analysis based on research up to this state)
    - Sam and Michelle- Have map visualizations working with the formatted data set we put together during the week prior (set up map background and place different filter buttons on the index page; finalize aesthetic decisions for layout of different features)
- **April 17- Ensure process book and data are prepared for milestone 1 submission
- April 19
    - Michelle- Flesh out design evolution, implementation, and evaluation for process up to that point in the process book
    - Sam- Get our top priority filter visualizations (i.e. trip duration, starting point, etc) to work
    - Alina- Finish visualization of gender and weather
- April 26
    - Everyone- Work on uploading all the different parts onto the website and ensuring that broadcasting works; finish any missing elements of the process book or putting in new information based on progress on the project)
- May 3
    - Everyone- Final touches (edit process book, debug last errors, clean up extraneous code)
- **May 5- Submit final project

***Deadlines for milestone submissions*