# 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.

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.

## OVERVIEW

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.

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, 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.

## QUESTIONS

We are trying to answer questions regarding the pattern of usage of Hubway bikes broken down and stratified by gender, age, time of the day, time of the week, and type of user. We want to see if registered users use the bikes more on weekdays whereas casual users use it more on weekends, indicating a work-related vs. recreational usage factor.

## DATA AND DATA PROCESSING

*Data Sources:*

http://hubwaydatachallenge.org/trip-history-data/

We are collecting our data from the Hubway website. The data is publicly available to all on their website in csv format.

*Data Processing Procedure:*

We have obtained cleaned csv files from the Hubway site. One csv file contains information regarding the specific stations, which we used to map the markers on the Google Map. Another file contains information regarding individual trips including time of departure and arrival and station number as well as demographic data including gender and age for the registered users.

We also obtained a cleaned csv file from a public GitHub repository including semi-aggregated data within 5-minute intervals for the month of August 2012. This data includes the capacity of bikes at each station (remains fixed over time) as well as the number of bikes that have left and come into each station at a given time.

In order to create the density graph for the brush, we needed to create a function that counts the number of trips for each day. We needed to build another data set that is aggregated by day.

## EXPLORATORY DATA ANALYSIS

We initially plotted the stations on a map to see how spread apart they were and discovered that drawing paths for each station to others would grow cluttered very easily. Next we began to explore the demographic breakdown of users to see if there were any discoverable trends.

We also looked at other designs and visualizations for inspiration, such as ones for MBTA and those included in the Hubway Data Challenge so we could get a sense of the data without doing extensive coding, which decreased the necessity for an extensive EDA phase. In addition, we spent time looking at the data and seeing what the best way to represent it would be.

## DESIGN EVOLUTION

We underwent many iterations of our design before we arrived to our current state. We began with the idea that we would be visualizing individual trips on the map visualization, but we decided that that would be a very cluttered visualization. As a result, we shifted our focus to looking at each station as an entity. We wanted our target audience to be more geared towards the interests of the bike users so that also guided our design decision. After our peer feedback design studio, we decided that it would be interesting to encode the stations on the map by a color gradient that measured the fullness of each station at a given point of time, i.e. the number of bikes present in relation to its capacity of bikes. Specifically, we would have a brush feature that would allow the user to select a specified time range during a day. For example, a user may want to see the average fullness around 9am for a station on a selected brush of dates so on the map visualization, the station color gradients would update to show this change. The gradient would be a single color where the saturation of the color would be on a scale from white to fully saturated, corresponding to the fullness of the station’s supply of bikes. We also decided to encode the stations by size according to their bike capacity, which would be constant over time. As a result, the user would be able to get a complete picture of which stations are full and the number of bikes that are generally available at each station, and also whether larger stations are fuller than smaller ones and vice versa. This insight would also prove valuable for Hubway because they would be able to see which stations need more bikes and which could have fewer bikes. For example, it would also allow users to optimize the time that they go to stations if there are repeat users (registered) and find that they often do not find bikes at a certain station.

## IMPLEMENTATION

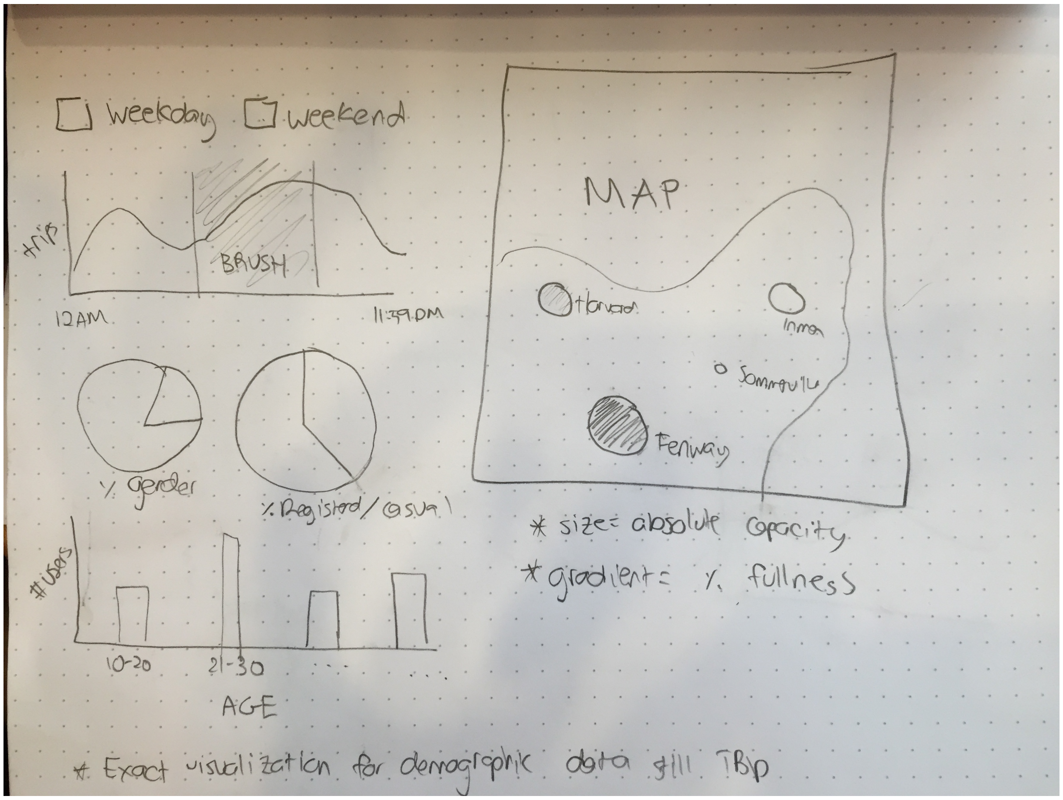

Figure 1: Sketch of final design plan

The main visualization is the map of the Boston area. This map will contain the locations of each Hubway station with the SVG element of each station being encoded by size and color. The size of each station's SVG will be proportional to its absolute capacity for bikes, so this will be a static encoding even as the data filtering changes. The color saturation of each station's SVG element will be proportional to the percent fullness of the bikes in relation to their absolute capacity. In other words, a station who has all their bikes available, (100% full), would have a fully saturated hue of green, and a station with none of their bikes available (0% full), would be not saturated, or white. The next layer of filtering includes the sidebar on the left. First, we have radio buttons controlling weekday or weekend. Below is a map that shows the average number of trips taken over the period of a day. Next are pie charts that show the demographics of the users, which include whether they are a casual or registered user, and if they are a registered user, their gender. Finally, there will be a bar chart displaying the age ranges of the registered users.

The weekday and weekend radio buttons filter the data so that the map updates to show how full a station is on either a weekday or weekend or both, and the same idea for the line graph that display the average numbers of trips over either weekdays or weekends, the pie charts for demographics, and the age bar chart.

The line graph SVG element displays the number of trips over the course of a day. The data that is displayed here is contingent on whether weekend or weekday is checked and then it will display the average number of trips over the span of a day. On this element, you can brush over a certain time range to change the map visualization's fullness at each station, the user demographic pie charts, and the age bar chart.

The demographic pie charts update gender and user breakdown dynamically based on whether the weekday/weekend button is clicked and the brush of the time of day element.

The age bar charts also update dynamically based on whether the weekday/weekend button is clicked and the brush of the time of day element.

We were able to filter the data into buckets of time intervals containing data on each station, and so the next step is to separate this further into buckets by station. We plan to save this array as a json file so it doesn't slow down the visualization upon loading the page by running this each time.

We were also able to calculate the percentages of genders and types of users at each station and push those into array that count them.

Since we have been able to aggregate station trips over time, we can then use this aggregated data to make the line graph of trips over the span of a day.

We have two important functions that we created: sumtrips, which returns an array with the total number of trips in any given time frame and currentnumber, which takes in a station number and returns the number of arrivals and departures during the time frame. As a proof of concept, we have arrays for intervals and interval keys and are only testing a few stations at the moment to make sure the function works so as not to slow down the entire visualization.

**EVALUATION**

So far, we have learned that since we have a lot of data, we needed to narrow our scope of time in order to not slow down our visualization. In addition, since we have a lot of data, it was important to be thoughtful about what to use in our visualizations based on what questions we wanted to answer. Now we have a solid idea of exactly which questions we want to answer and how we can do so effectively with our visualizations. We are still trying to think of more interesting ways to visualize the demographic information containing gender, age, and type of user, since pie charts and bar charts are not as visually appealing.

**CURRENT PROBLEMS**

The following line of code is supposed to allow us to click on each station but it doesn’t work: var layer = d3.select(this.getPanes().overlayMouseTarget, so this is something we are currently working through. In addition, we are working on the optimal placement of each div element on the page.