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 bike share 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. We are interested in seeing if there are any interesting trends between the user demographics and the bike usage over time.

OVERVIEW

We are trying to explore how the demand for Hubways changes as a function of time, location, 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, measure the speed and frequency by which bike rentals occur, isolate which stations are full or empty at particular times of the day and see how large (maximum capacity) each of those stations are.

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

Initially, we were motivated to do this project because 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.

In particular, we would like for Hubway riders to use our visualization to be able to tell whether a particular station is full or empty at particular times of the day. This allows them to decide whether they should expect bikes at a station of interest at a given time.

Also we would like Hubway managers to be able to use our visualization to decide how to adjust Hubway station capacities and marketing campaigns. For example, Hubway managers should be able to tell which stations are overused or underused, and determine whether it makes sense to expand their size or try to divert traffic to nearby stations (or even build new stations in a particular area). Also they should be able to determine the demographic breakdown of users at a given station so they know how to adjust marketing campaigns trying to attract users in a particular region.

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 (hubwaydatachallenge.json). 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 (trips_aug12.json)

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 (aggcap_aug12.json). 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. Finally in order to determine percent full values, we needed to create a data set that provided the maximum capacity of each station.

To do this, we created three html files that would reformat and clean the data (eliminate extraneous elements). The first was createaggformated.html, which iterates through aggcap_aug12.json data to create a new new_aggformatted.json file containing an array of objects, each of which contains a time key (hour of day) and then an array of objects with the number of arrivals and departures over the weekend and weekday for each station. We used this file to create the path element for countvis, where the filter/aggregate function would output the total number of arrivals given the specified timeframe, weekday/weekend settings, and station. The second was createdemagg.html, which created a ageagg.json file from trips_aug12.json that also

sorted individual trip information into appropriate time and station objects, which were then further broken down into casual/registered groups and several age groups that were subdivided into male and female categories. We used this file to create the pie and barchart visualizations since pievis and barvis had filter functions that would aggregate data that was filtered by the appropriate time/day/station information. The third was getcapacitydata.html, which used contains javascript code that iterates through the aggcap_aug12.json data to output a stationcapacity.json file with every station and its corresponding maximum capacity. We then used this data in our mapvis.js data processing functions so we could divide the number of bikes present by capacity; in particular our filter/aggregate function would use the new_aggformatted.json data to subtract the total number of arrivals (which would add to the number of bikes from a station) from total number of departures (which would subtract from the total bike number) filtered by time of day and weekend v. weekday). Then it uses the corresponding station capacity information from stationcapacity.json to divide the total bike number by capacity to determine percent fullness. Because the data set was so large, we would have to run each script multiple times altering the intervals_keys values to get the formatted data for several hours at a time before joining files, hence these files will only output data for a subset of hours depending on what is included in intervals_keys.

EXPLORATORY DATA ANALYSIS

We 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. We created some preliminary bar charts and line graphs to get an idea of the trends that were occurring in the data to decide what would be interesting to visualize. For example, we realized that the type of user at a station provided us with interesting insight for understanding why people use the Hubway bikes. As a result, we decided to include the pie chart of the type of user, whether they were a casual or registered user.

DESIGN EVOLUTION

We underwent many iterations of our design before we arrived to our current state. Design decisions were geared towards the fact that our target audience was bike users. 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. 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. When the user selects a specific time on the brush, the time frame selected will appear above the visualization. This makes it easier for the user to see what times he or she is selecting, since it can be difficult to tell from the brush itself. 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 found that a continuous saturation gradient would not allow users to easily distinguish between different levels of fullness between stations, and instead assigned particular saturation levels to grouped stations with similar fullness levels. By using the station dropdown and brush to isolate a particular station and time period of interest, users could use our visualization to see whether bikes would be available at a station they planned to use at

a certain time. In particular, it would help repeat users (registered) to optimize the time by avoiding that do not have bikes at a time the users regularly would visit the station.

In order for our visualization to be useful to Hubway managers as well, we initially decided to encode the stations by size according to their bike capacity, which would be constant over time. We thought that scaling by capacity would allow Hubway to 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 prove valuable for Hubway because they would be able to see which stations need more bikes and which could have fewer bikes. Knowing the relative capacity of these stations would be helpful for Hubway because if a station is relatively empty and also has a small capacity, Hubway should consider increasing its size in order to meet demand. However if the station already has a large maximum capacity and cannot be expanded much further, Hubway should instead invest in trying to divert traffic to other nearby stations. As we moved through the design process, we added the capability to either size by capacity or by number of departure for that station for a brushed time period. We decided to add the departure information because, by knowing how many bikes are taken out of a station at any given time, Hubway could tell which stations are the most popular and therefore generate the most revenue. The percent full data would not be sufficient to give this information since it would be possible for a station to be relatively empty merely because people do not arrive to the station and replenish the bikes, not because it is popular. Meanwhile the arrival data in the path visualization does not allow for a macro-level comparison between stations or give information about departure traffic so departure comparison was necessary.

Additionally we added a legend to clarify to users what fullness level each saturation level corresponded to. We also incorporated this in the pie chart, bar chart, and map to ensure that users could analyze trends for subgroups that may not be apparent in the larger data set. For example, hovering over a particular map legend element will allow users to tell whether full stations are clustered in a particular area or more dispersed. This allows users to tell whether all the stations in a particular region are particularly empty, allowing riders to evaluate where to rent bikes and allowing Hubway managers to decide whether to add new stations in regions where empty stations are clustered.

IMPLEMENTATION

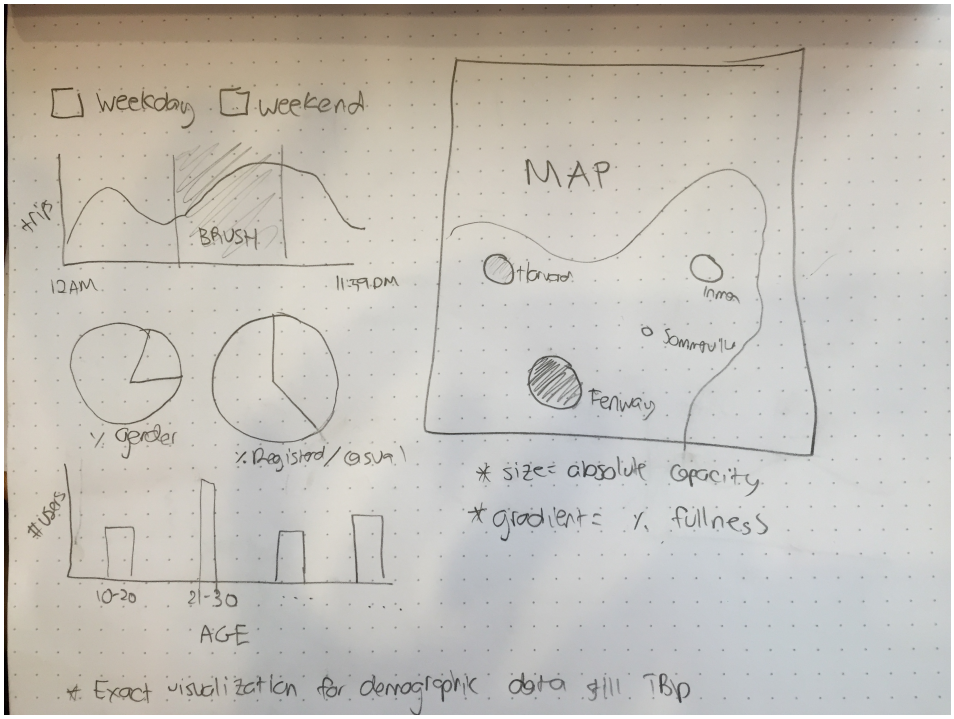

Figure 1: Sketch of final design plan

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 red, 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 (in particular, arrivals to a selected station or to all stations if none are selected) 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 (arrivals in particular) 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.

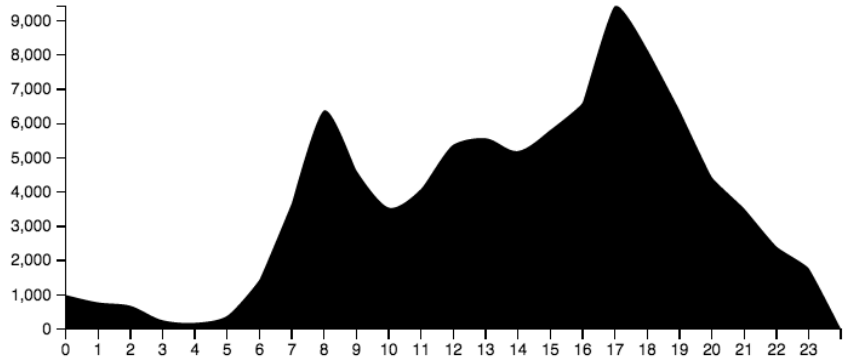

Figure 2: Visualization of the number of arrivals to stations over the period of a day.

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

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

We were able to filter the data into buckets of time intervals containing data on each station, and further filter the data into buckets by station. We saved 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 counts them.

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

As a way to deal with the data being too large to load in and taking too long to render in the visualization, we aggregated and pre-processed the data in such a way that we were only recording the number of arrivals and departures every hour aggregated over the month of August instead of at 5 minute intervals. Although this gives us less of an idea of what happens between the hours, it is a more feasible way for us to obtain and use the data set.

A change in our design occurred while creating the map visualization from only static station sizes based on its absolute capacity to an additional option to scale size based on the number of trips taken at each station (essentially the number of departures from the station; the justification for this design decision is explained in design evolution). In addition, we created a scale for the colors that represent the fullness of each station at the brushed time. A station that is less than 20% full of bikes (or basically empty) is colored white, and gets increasingly redder in intervals of 20% until the 80% to 100% interval, which is colored a dark red.

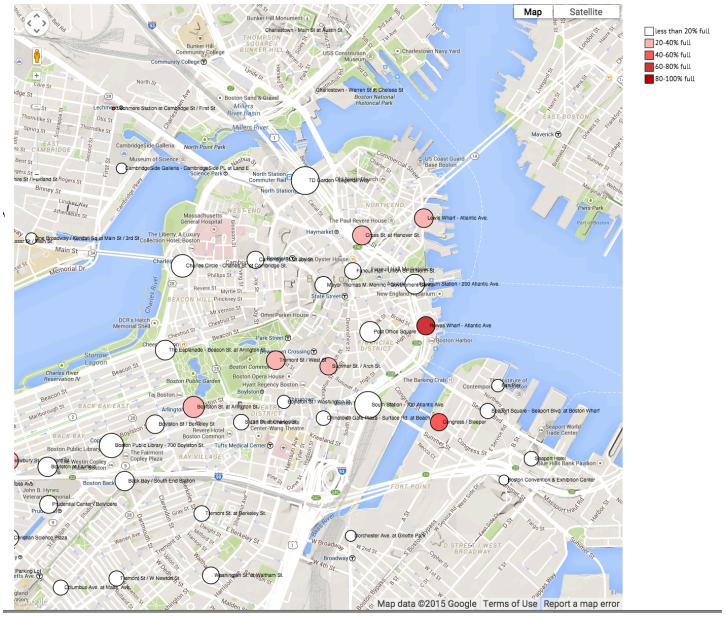

Figure 3: The map visualization showing stations in the Boston area showing their fullness and capacity/departures.

We have a fully functioning brush of trip data over time that interacts with the other views. When one brushes over the aggregate trip data at a particular time of day, the map visualization displays the updated size and saturation of color for each station marker; the user pie graph displays the percentages of casual and registered users over the brushed time and the age visualization displays the number of registered users in each age range based on the brushed time of day.

We opted to create a dropdown menu to filter by each station in lieu of clicking directly on a station on the map. We chose to do this due to difficulty we were having with overlaying the click target on the map. The solution we came up with involves a user selecting a station from a searchable dropdown menu, which triggers all the four visualizations to update accordingly to only show data for that station. In addition, the map zooms to the location of the station. The zooming and relocation of the map makes it easier for the user to see graphically where the station is on the map and the change in the demographic information and count path allows the user to gain better insight for that selected station. In addition, the total capacity of each selected station is displayed above the map visualization to give the user a quantitative sense of how the circle markers are sized relative to each other.

Finally, we added hover features to each of the elements that allow users to isolate a particular data subset of interest by mousing over a legend item. For example, hovering over the “less than 20%” legend element in the map legend will highlight the corresponding markers.

EVALUATION

We 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. As a result, we chose to only incorporate data for the month of August. After completion of our visualization, we noticed a few trends in the data. Specifically, we noticed that for every station there are at least two peek times that can be seen in the capacity visualization occurring at times 7/8am and 7/8pm. These peaks occur for any station selected as well as weekend or weekday data.

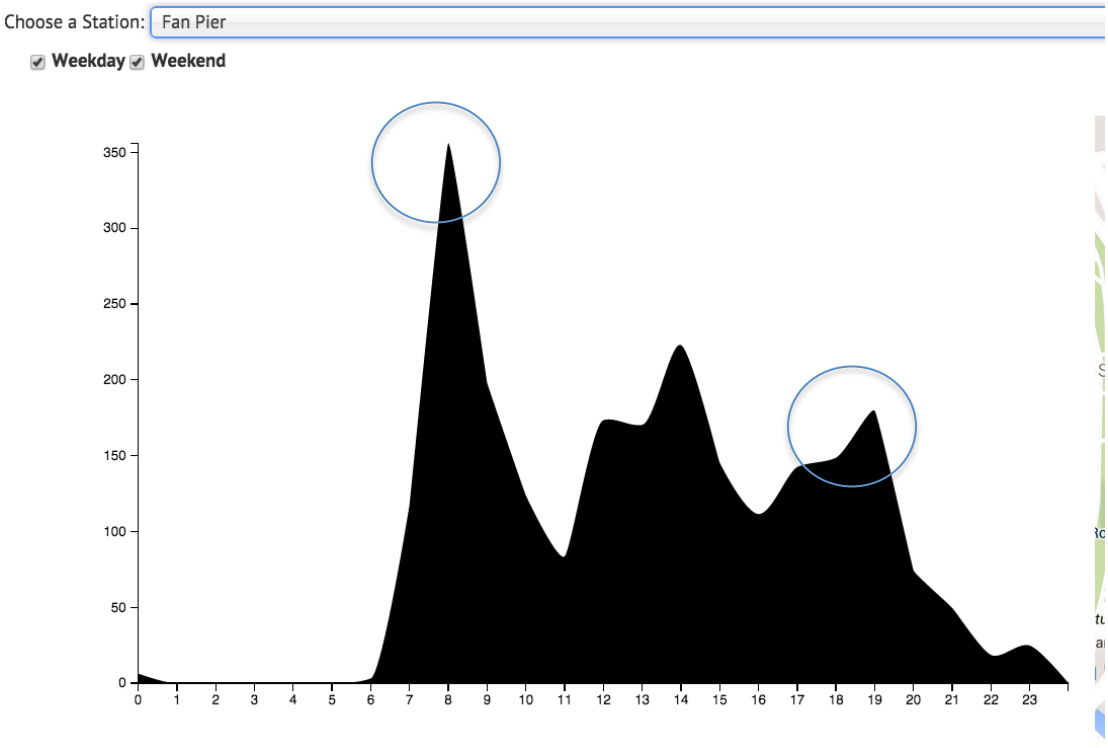

Figure 4: This graph shows the peaks at Fan Pier station around 7-8am and 7-8pm, likely due to commuters going to and from work.

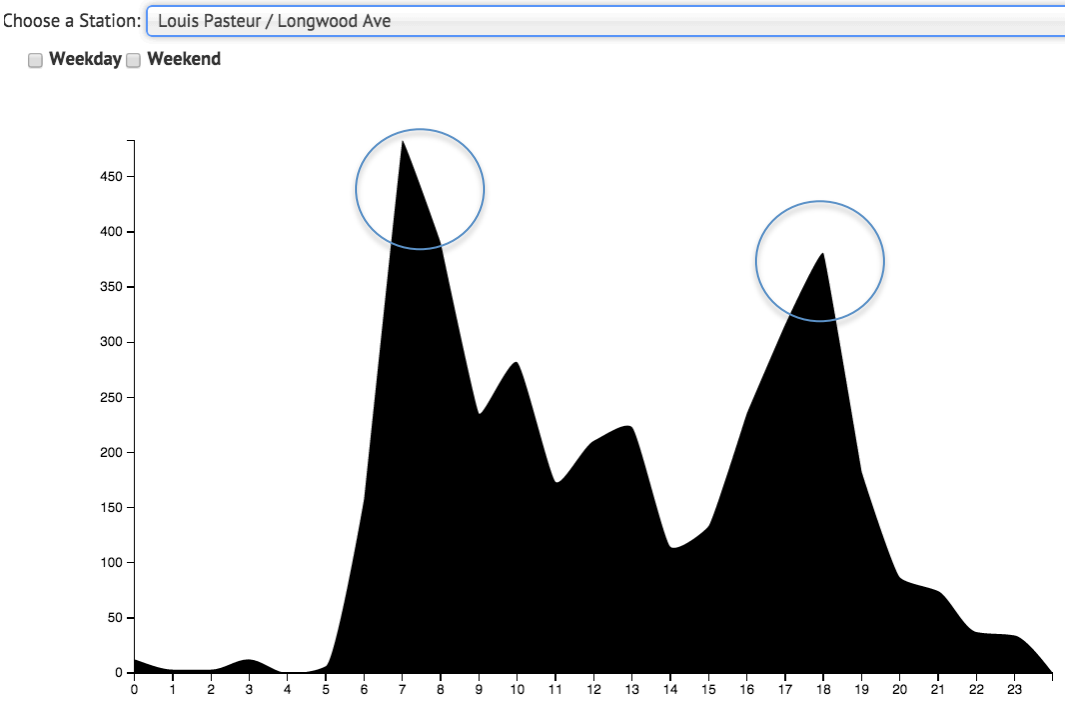

Figure 5: This graph also shows the peaks at Louis Pasteur/Longwood Ave. station around 7-8am and 7-8pm, likely due to commuters going to and from work.

We also noticed that for every station and time brush selected, there were more registered users than casual users.

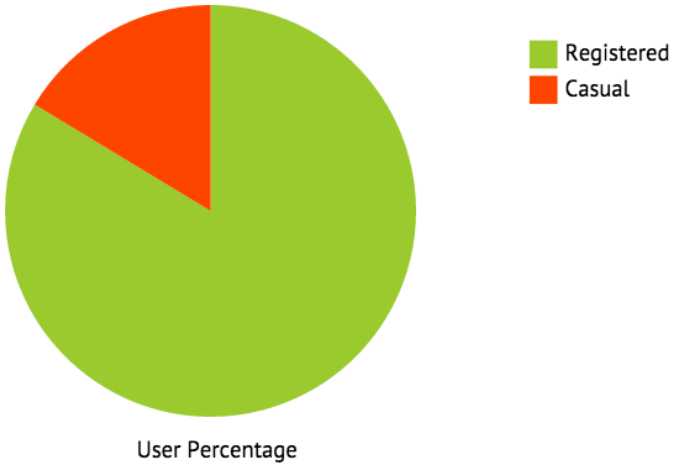

Figure 6: This graph shows the breakdown of registered vs. casual users and updates based on selections and brushes.

There was no trend we noticed with the age groups and gender. We also used our visualization to analyze an individual station, which we explain in the "Story Time" page of our website.

Overall, we thought that our visualization successfully answered our intended questions, which were as follows:

- How does our visualization ensure that Hubway riders are be able to tell whether a particular station is full or empty at particular times of the day?
    - *How can users find out whether a specific station is full or empty at a particular time/day?*
        - Users would between toggle the weekday/weekend checkboxes and brush over the time brush. Then they could use the dropdown menu to search for a station (by typing the station name or by scrolling through and selecting), and the map will reset to zoom in on the relevant station.
    - *How can users use this information to decide whether to try to rent bikes from a particular station at a given time/day?*
        - After setting the weekday/weekend and time filters and selecting a station, users can look at the updated total capacity (above the map) to see whether there are enough bikes left (i.e. a 80-100% full station with a maximum capacity of only 10 is less ideal than a 60-80% full station with a capacity of 50). To more easily compare capacities and percent fullness *between* stations, users can scale station markers by capacity and then observe the percent fullness gradient. This data allows users to choose between two stations that are relatively close to one another based on which station is more likely to have bikes.
        - If a user cannot opt to choose another station even if a station of interest is empty, they can use the path visualization to see when the most arrivals are occurring and plan to visit the station at that time because that is when bikes will be re-entering the station for use.
- How does our visualization ensure that Hubway managers can decide how to adjust Hubway station capacities and marketing campaigns. In particular, how can Hubway managers:
    - *Tell which stations are overused or underused*
        - Users would between toggle the weekday/weekend checkboxes and brush over the time brush. Then they could mouse over the relevant

element in the legend (i.e. less than 20% full if they are trying to isolate the emptiest stations), and the map will highlight the stations with the appropriate fullness levels.

- *Determine whether it makes sense to expand their size or try to divert traffic to nearby stations (or even build new stations in a particular area)*
  - Since the hovering feature allows users to tell whether empty or full stations are clustered in particular regions, Hubway managers can decide to shut down stations where plenty of unused stations exist or create new stations where all stations are overused.
  - By scaling markers by relative station capacity and then using the hover legend feature, Hubway could isolate whether relatively empty stations have small or large capacities. This informs its decision to increase size because if a station is already large then it might not be able to expand much further so Hubway could instead build new stations or divert traffic. Meanwhile if an empty station is small Hubway should consider adding more bikes since there is probably more space available for expansion.
- *Determine the demographic breakdown of users at a given station so they know how to adjust marketing campaigns trying to attract users in a particular region*
  - Scaling the markers by number of departures allows users to determine which stations are the most heavily used and therefore generate the most revenue. Hence Hubway could choose to invest more heavily in marketing campaigns that targeted to the region around that station since it would be confident that there are many Hubway users nearby a heavily used station.
  - The demographic information allows Hubway to determine whether it should target a particular age group or gender more heavily than others. For example if there are many more users using a particular station, campaigns might include college discounts while Hubway should use campaigns including business office discounts in stations with heavy traffic from older groups. Meanwhile if there are more casual than registered users, Hubway could choose to include helmets since casual users are less likely to have their own helmets.

Ways in which we could have improved our visualization:

We could have improved our visualization by having a clickable function of each station on the map. Instead, we opted for a dropdown menu that updates the visualizations. Another way we could have improved our visualization is by having a better website that is more visually appealing with a different background. A final way we could have improved our visualization was by showing a specific trip that a user has taken from station to station in a completely separate visualization.