Data Visualization Project
Process book

TOPIC: HOW DOES YOUR CITY COMMUTE?

PROJECT MEMBERS:

LAMA ALBARQAWI

MANISH ROY

RUSHIT SANGHRAJKA

Table of Contents

Overview and Motivation

For most of us biking around the neighborhood was synonymous with growing up and becoming independent. As we grow older much of this relationship with our bikes fades away owing to various challenges like longer commute time, logistical aspects concerning maintenance etc. While many continue biking and also use it to commute for work, this segment largely fluctuates caused by various factors.
The prominence of Bike sharing as a service, leverages the interest for biking while diminishing the logistical challenges that usually turn riders away from pedaling away.
Although there exists data and studies, we were quite drawn to notion of augmenting commuting patterns by churning these data sets.
The pivotal data in our approach being the data collated by the Bike-Sharing Service providers. We aim to derive sense of these chunks of data by leveraging efficient visualization techniques, and find implicit correlations of other variables that dictate biking habits.

Related Work

We were inspired by many previous projects (of much bigger scale) that attempt to study bike usage and present them through intuitive visualizations.
Some of these are centered around studying Bike-data in cities like New York, Boston etc.
The techniques learnt in class, along with the functional understanding of the elementary aesthetics act as a precursor to our project trajectory.

Driving Questions

We started out with a naive understanding of what motivates people to bike, what implicit/explicit factors dictate biking habits, etc.
Having worked with the data and seeing how it pans out in our visualizations, significantly augmented our perception of our goals.
An interesting direction of work, that we aim to explore is extrapolation of our understanding to changing data set ranges over time. As the entire process is an exploratory one, each iteration of analysis morphs the underlying trajectory of our project.

Data

The source of our stations datasets is: https://github.com/BetaNYC/Bike-Share-Data-Best-Practices/wiki/Bike-Share-Data-Systems

We started the cleaning up process for the datasets of our scope at that point, which were for 5 Cities (Boston, Chattanooga, NYC, Washington DC, Columbus). The cleaning up was done in several steps, ordered as below:

- We have used Talend Open studio for Data Integration, to read the csv files for each city and fetch the total number of trips grouped by each station ID, in both of the source and destinations sides. The resulted 2 csv files per station from the previous step were sorted with respect to the corresponding number of trips in a descending order, the goal of this sorting was to understand the behavior of the bike trips and how the trips were distributed across the stations, with a special focus on the top 7 stations, which we are going to study in further detail during our project implementation. After that, a new set of Talend processes were created to fetch all the trips records that have used any of the top stations. The goal of this step was getting a clean dataset for the top 7 stations for each city to be used later, upon selecting one of these stations on the map and reflect this info on the stations graph.

- For the purpose of Drawing the initial map, which should include all the stations without focusing only on the top 7 stations, we needed a new way to read the needed fields (stationId, stationName , stationLongitude, stationLatitude). This was done using javascript and d3 code, which reads csv and construct a JSON object. This JSON object is then saved into a file to be read later while drawing the map.

- For testing the Line charts and the stacked Bar charts during development, we needed to create a new 2 csv datasets, that contain:
  A- days of the year along with the corresponding count of bike trips for that day.
  B- hours of each day of the year, with the count of bike trips for that hour.

Exploratory Data Analysis

In order to gain deeper insights and add more reliability to our resulted datasets and also to make sure we are on the right track, we have searched for the top 7 stations using Google maps, and checked their actual positions. We had a hypothesis that these 7 stations that have the highest numbers of trips between them must be close to each other up to some limit, the hypothesis was proved correct after the google maps study. This has influenced the project by changing the way of visualizing the stations graph, we have noticed that the top 7 stations were not constant and they were changing depending of the time (hours, day and month) chosen. So, we have decided to get the top 7 stations for the circular graph dynamically based on the selection of the time chart.

Design Evolution

Initially Proposed Design

We considered various visualizations in order to depict the data. In our proposal, we finalized on the following design:
When the web page is launched, a map for the US and Canada will be displayed. Cities chosen for our project are the only ones that are going to be displayed on the map (as dots) and user has the option to interact with them. When the user clicks on a specific city (for example New York), it gets highlighted, and below the map area a detailed bike station detailed visualization will be rendered for the selected city. For each selected city 7 small bubbles will represent the bike stations in that city. The dots will be placed in a circular shape. A ribbon connection between 2 stations will be created to show numbers of bikes going from station 1 to station 2. The width of the connection will be dependent on the number they are representing. Upon user hovering over the connection the actual number of bikes will be displayed as a tooltip notification.
Three horizontal bars under the map and above the city stations visualization will be shown. They would have the following details:

- The first one displays the months from June to December,
- The second bar shows the days from 1 to 30 and
- The last one shows the hours from 12am to 11pm.

Default values for each of these 3 attributes will be used to draw connections between stations once a city is clicked, in order to give the user a head start on how he/she can interact with the visualization. Then the user can select the time and date that he/she is interested in seeing stations status during it. Based on the user selection the ribbon connections will reflect the respective data.
Four small labels at the top of the stations visualization will be visible on the web page, contains categories on which the user can select data to be visualized with respect to the selected category. We are planning to work on the following categories; Gender, Customer Type, Age

group, Trip duration. User will be allowed to filter data based on one of those categories one at a time. User selection and filtration for the data will split each connection (ribbon) into two connections. Each connection (after category selection) will be presented in a specific color. A color map guide will be displayed for the user under the chart. For example, if the user chose Gender as a category, all the connections between the bike stations will either split into two connections or change the color of the connections. If connection between A and B has only male cyclists that selections will change the connection from black into blue. If connection A and B have both (female and male) cyclists, the connection will split into two connections which their width would be proportional to the count of the cyclists' gender. Each connection will have the gender color (blue for male, pink for female) as well.

Updated Design

As we continued with the project, we started coming across various issues. Based on feedback from the instructor and the TAs, we decided to update our idea to reflect the feedback.

- Efficient use of screen space: With our current design, we leave a lot of space on screen. We decided to add in another charts to visualize information; 2 line charts. When the user selects an edge "a relationship between 2 stations" of his interest from the circular graph, a line chart will show up to show the distribution of the trips between these 2 selected stations over time; where x axis as you might expect will present the selected time interval, and y axis presents the counts of trips.
- We have also decided to add a calendar view, to enable the user to select specific days that are not sequential and seeing the trips that are travelling during them (in a 24 hour interval per day) between the selected 2 edges from the circular graph.
- If the user wants to check how an edge trips counts is divided with respect to a specific category, he/she can see this info by clicking the category button after selecting an edge, this will trigger the bar chart. The stacked bars will be divided proportionally using colors to indicate different category values.
- Focus on a single city: The reason behind this is that our current design doesn't contribute to comparison of bike usage across different cities. Moreover, the map view showing the cities is not efficient use of space. It would be more beneficial to have the map view focused on one city. Hence, we decided to look at one city for the research study.
- A heat map is added to be the first thing that appears after the time charts, it is triggered on clicking the time update button. The heat map shows the trips density for all stations during the selected interval, with a legend showing the color map. The heat map shows vertically the days of the week, and horizontally the months. This chart performs an overview of the whole state of the bike trips across all stations during the interval selected.

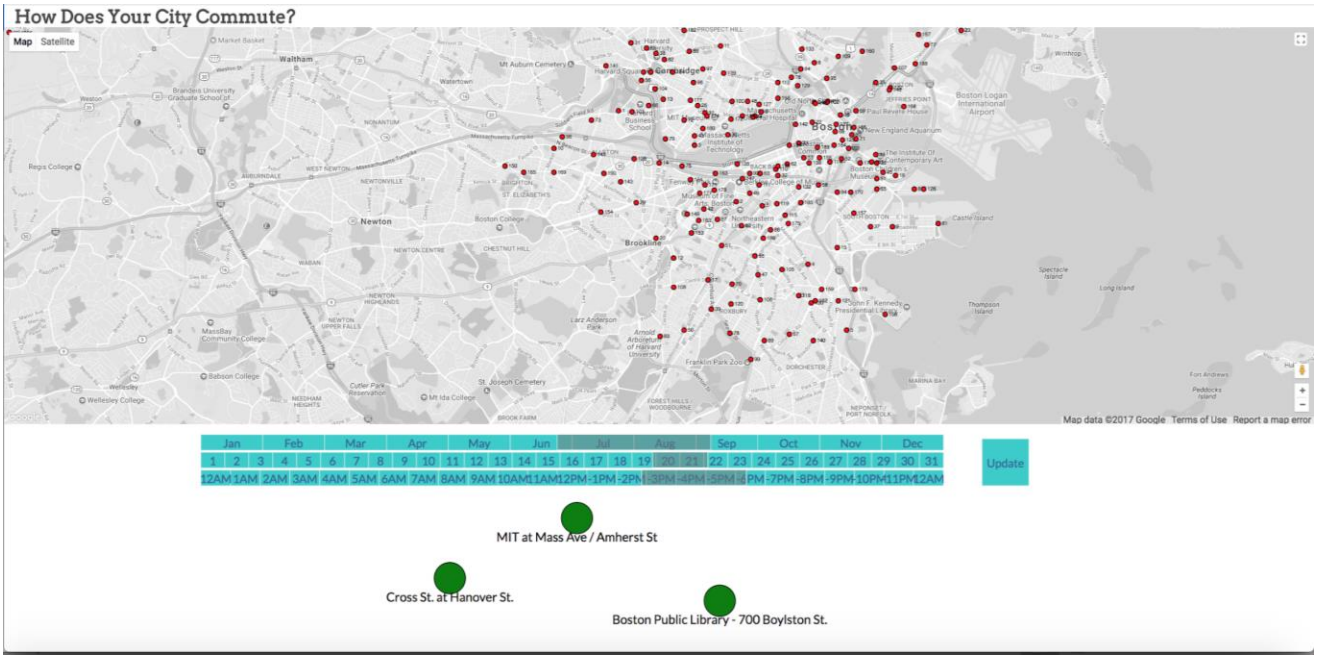
Implementation

For the first milestone:
We have currently implemented the map view, which shows the map of Boston and the stations on the map. We have also implemented the time bar, which allows filtering by months, days and hours.
We make use of an “update” button next to the time chart so that the user can make the filtering selections and then update the dataset. This was done to prevent filtering of data after every selection made by the user (the filtering is time intensive, so we attempted to reduce the number of filter calls through this workaround).
We have also implemented the barebones of our circular chart. Here, the circular chart currently shows seven circles and the stations that they represent.

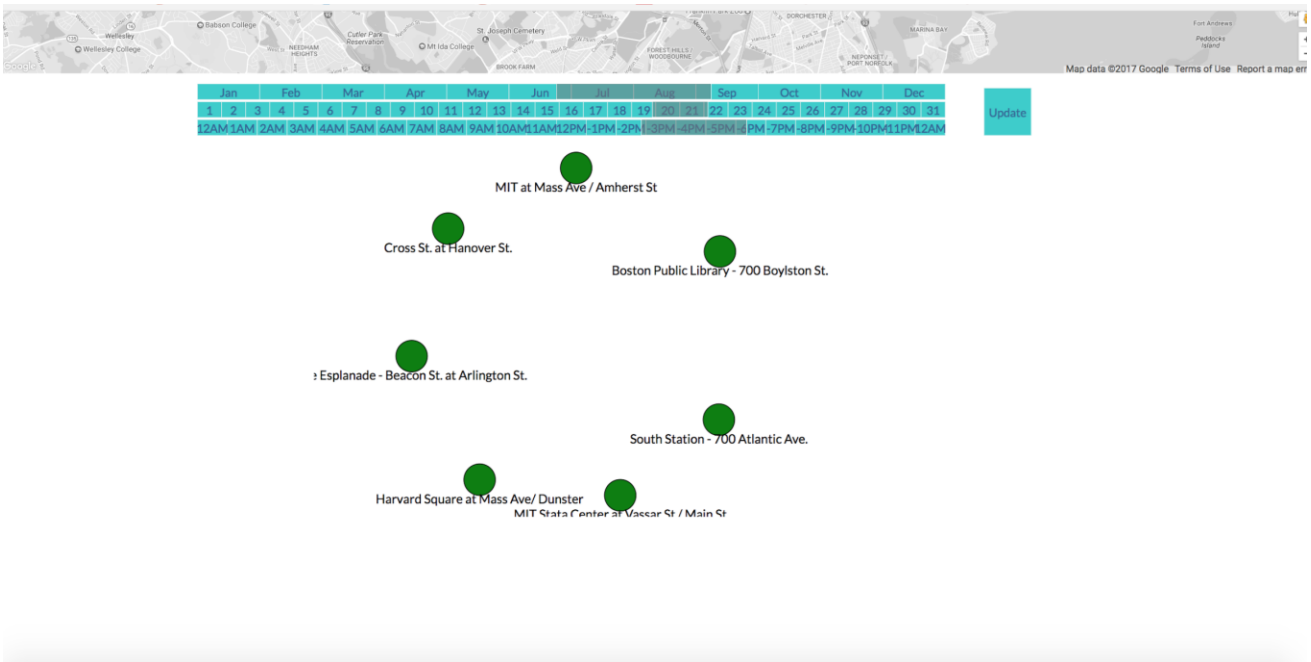
For the second milestone:

- We have updated the map view, it is now implemented using Google Maps APIs with the appropriate selections for styling the map and the stations.
- The time charts are updated, we agreed to remove the days horizontal bar and use the brushing over the month to enable the user to select the days as a range of the month. This option will enhance usability and would make the selection simpler and faster.
- The edges for the circular graph are now retrieved appropriately once the update button is clicked, and the categories sections are implemented.
- The 2 line charts are completed and behave as planned and discussed in the previous sections. We have added tooltips to ensure data accessibility.
- The calendar view is implemented to give the user the ability to choose more than one day discretely.
- The stacked bar chart is completed, and running as planned, to show how different classes of the same category are distributed within the one edge trips.
- The heat map is also completed, its valuable in showing an overview for the status of all bike trips once the user selects his/her time of interest. Before diving in each station in the following visualizations.

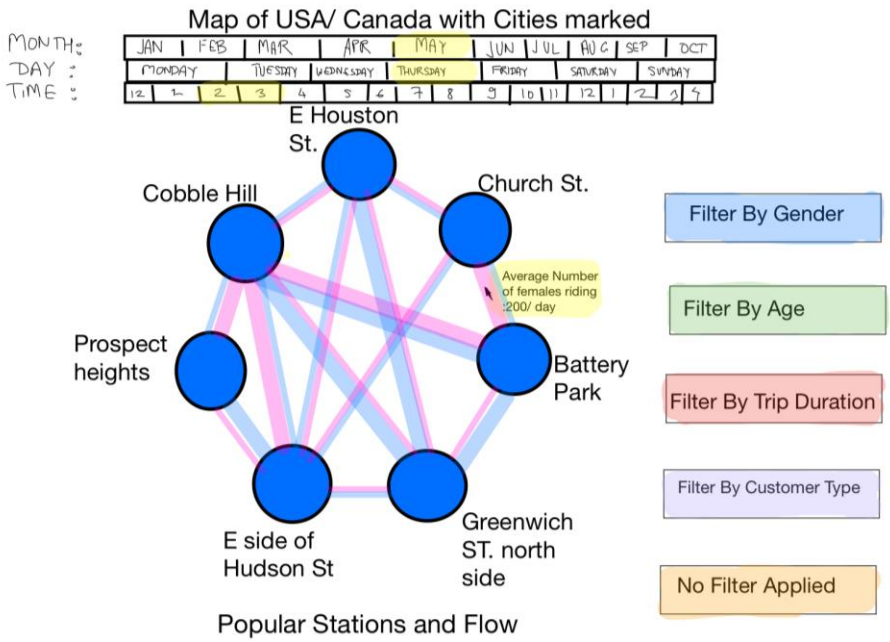
Screenshots of our implementation are shown below.



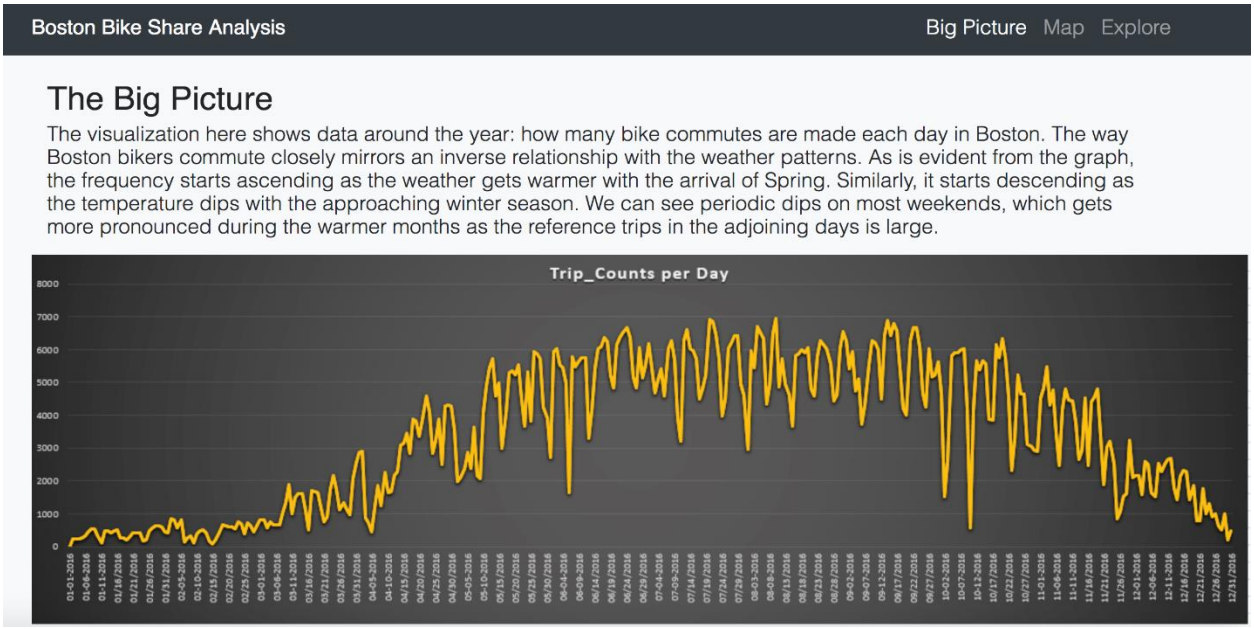
Screenshot 1



Screenshot 2


Map of USA/ Canada with Cities marked
MONTH :
DAY :
TIME :
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
MONDAY
TUESDAY
WEDNESDAY
THURSDAY
FRIDAY
SATURDAY
SUNDAY
12
1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
4
E Houston St.
Cobble Hill
Church St.
Average Number of females riding :200/ day
Prospect heights
Battery Park
E side of Hudson St
Greenwich ST. north side
Popular Stations and Flow
Filter By Gender
Filter By Age
Filter By Trip Duration
Filter By Customer Type
No Filter Applied

Final Presentation



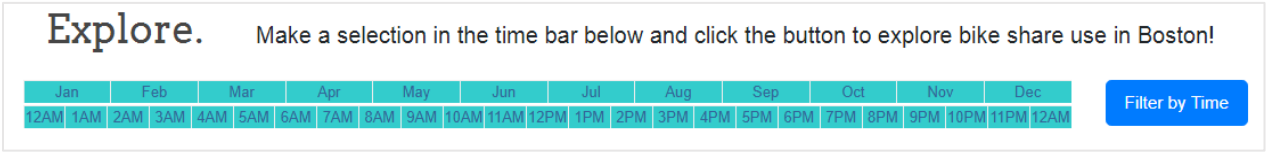
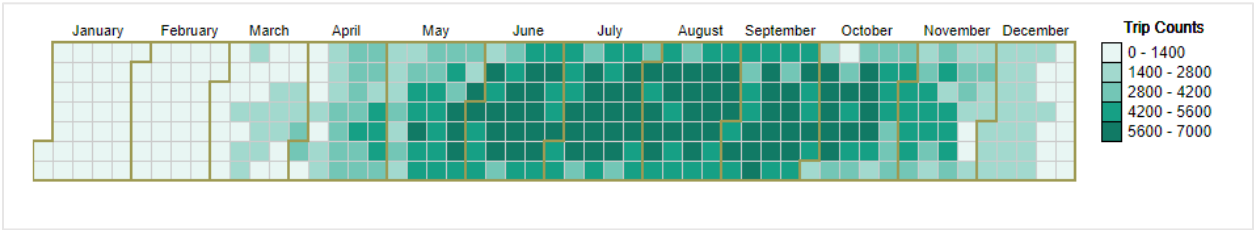


Figure 1: Time Chart Selection



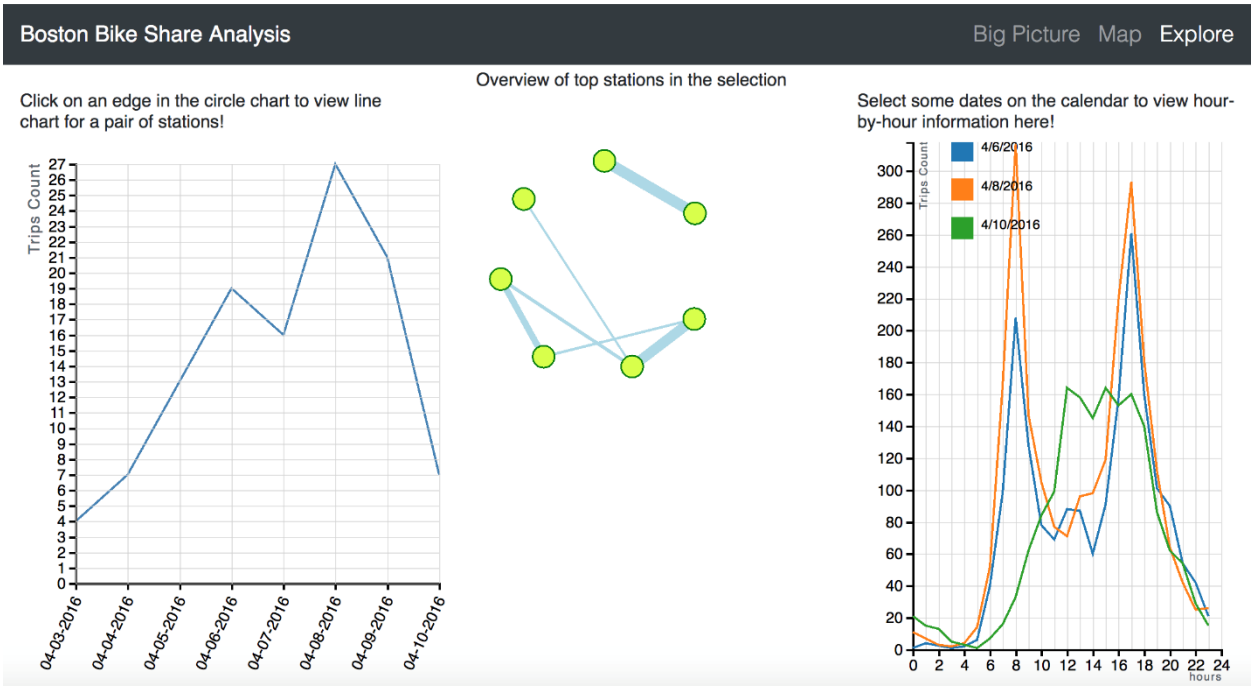


Figure 2: Heat Map projected on the Calendar

Evaluation

We learnt a great deal about the data. We learnt about the problems related to data munging, and problems arising out of the size of the data sets. JavaScript didn't do a great job of filtering the data, so we struggled figuring out how we can fix that.

The challenge of scaling the application to bigger datasets would require an application server to host the data retrieval process instead of using JS. Instead of overarching our project scope, we intend on focusing on visualizations for a smaller subset of data.

Another aspect that became obvious through the iterative process of design and evaluation are the nuances of the data representations through different visualizations. We would delve deeper and reconcile the same.