# Process Book: Bike Sharing

## Team Members


1. Siwadon Saosoong, ssaosoong@dons.usfca.edu
2. Surada Lerkpatomsak, slerkpatomsak@dons.usfca.edu


## Project URLs

Source code: https://github.com/idewz/bike-vis
Website: https://bike-vis.firebaseapp.com

## Background and Motivation

Public bikesharing has grown tremendously over the last ten years as governmental and non-profit organizations have recognized it as a means of increasing transportation accessibility and mobility, reducing vehicle miles travelled (VMT), and having positive impacts on public health (Shaheen, 2014).

The Bay Area Bike Share, as known as Ford GoBike, has become popular and kept expanding, ultimately covering approximately half the city - from 700 to 7,000 service bikes, across the East Bay and San Jose areas in addition to San Francisco itself. The system is new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through this system, users are able to easily rent a bike from a particular position and return back at another position. Users can unlock bikes from a variety of stations throughout each city, and return them to any station within the same city. Users pay for the service either through a yearly subscription or by purchasing 3-day or 24-hour passes. Users can make an unlimited number of trips, with trips under thirty minutes in length having no additional charge; longer trips will incur overtime fees. Today, there exists great interest in this system due to its important role in traffic, environmental and health issues.

Apart from interesting real world applications of the bike sharing system, the characteristics of data being generated by this system make it attractive for the research.

## Objectives

In this project, we will focus on the explorative analysis of GoBike open dataset. Our goal is to illustrate how interactive visualization can help to open the data that is produced in smart cities to a wider group of people. We will generate a high-level understanding of the data and identify potential relationships amongst variables by focusing on visualizing these relationships and patterns to make it easier to understand. For this exploration, several hypothesis that could be uncovered such as:

- Q1: Are there more riders on the weekdays or weekends?
- Q2: Are there more customers or subscribers using the service?
- Q3: How does the distribution of trips look like (across time, days of week, dates and certain stations)?
- Q4: Which are the stations that tend to be the most popular?

We believe that interactive visualization can help us to engage people, and to interactively explore and understand collected data, in order to make life more comfortable, safer and sustainable.

## Related Work


1. Tyler Field. 2014. “The submission winner of Bay Area Bike Share Open data Challenge for Best Analysis”. http://thfield.github.io/babs/index.html
2. SA Shaheen. 2015. “Bay Area Bike Share Casual Users Survey Report” http://innovativemobility.org/wp-content/uploads/2015/05/Bay-Area-Bike-Share-Final-Casual-User-Report.pdf
3. Ben Hamner. 2016. “San Francisco Bike Trip Map from August 2013 to August 2015” https://www.kaggle.com/benhamner/sf-bay-area-bike-share
4. Bike Visualization, “Bike sharing visualization and analysis website” https://www.visualization.bike/

## Data

Ford GoBike's public trip data from https://www.fordgobike.com/system-data. They provide anonymized trip data since 2017 in CSV format with the following information:

- Trip Duration (seconds)
- Start Time and Date
- End Time and Date
- Start Station ID
- Start Station Name
- Start Station Latitude
- Start Station Longitude
- End Station ID
- End Station Name
- End Station Latitude
- End Station Longitude
- Bike ID
- User Type
- Member Year of Birth
- Member Gender

As we can see, it includes quantitative, categorical and temporal fields. This should be sufficient for us to create variations of good visualizations. Since they provide a lot of data, we probably start with the latest data in March 2018. This data is provided according to the Ford GoBike Data License Agreement.

## Data Processing

The dataset from the original source is quite clean, but we have noticed missing values in gender and birth year fields for some records. These may need to be filtered out for some visualizations.

One thing that might cause a problem is the number of records. The 2017 dataset contains more than 500k records. Each 2018 monthly dataset contains roughly 100k records. And we could use a subset of the original dataset.

For the start and end time fields, it might be possible to split them into smaller fields, day and hour, because it is unlikely that we are going to use all the time information.

Additionally, we only know the birth year of riders, not the exact birth dates. If we are going to show riders' ages, it won't be accurate.

We should also consider pre-computing all the data we need for our visualizations to reduce the workload on the client side. This could be done beforehand using a python or Node.js script.

We started with the data in March 2018, however, it is quite large. To increase the development speed, we decided to create a mini dataset containing only 10,000 records by using shuf.

Then we extracted stations information from the original file with a shell script into stations.csv. And also replaced *User Type* and *Gender* fields with ids to reduce the size of the CSV file.

- Male = 0
- Female = 1
- Other = 2
- Undefined = -1

- Subscriber = 0
- Customer = 1

We also removed all double quotes from the csv file. These steps result in mini.csv with the size of 744 KB, only 3.2% of the original file size.

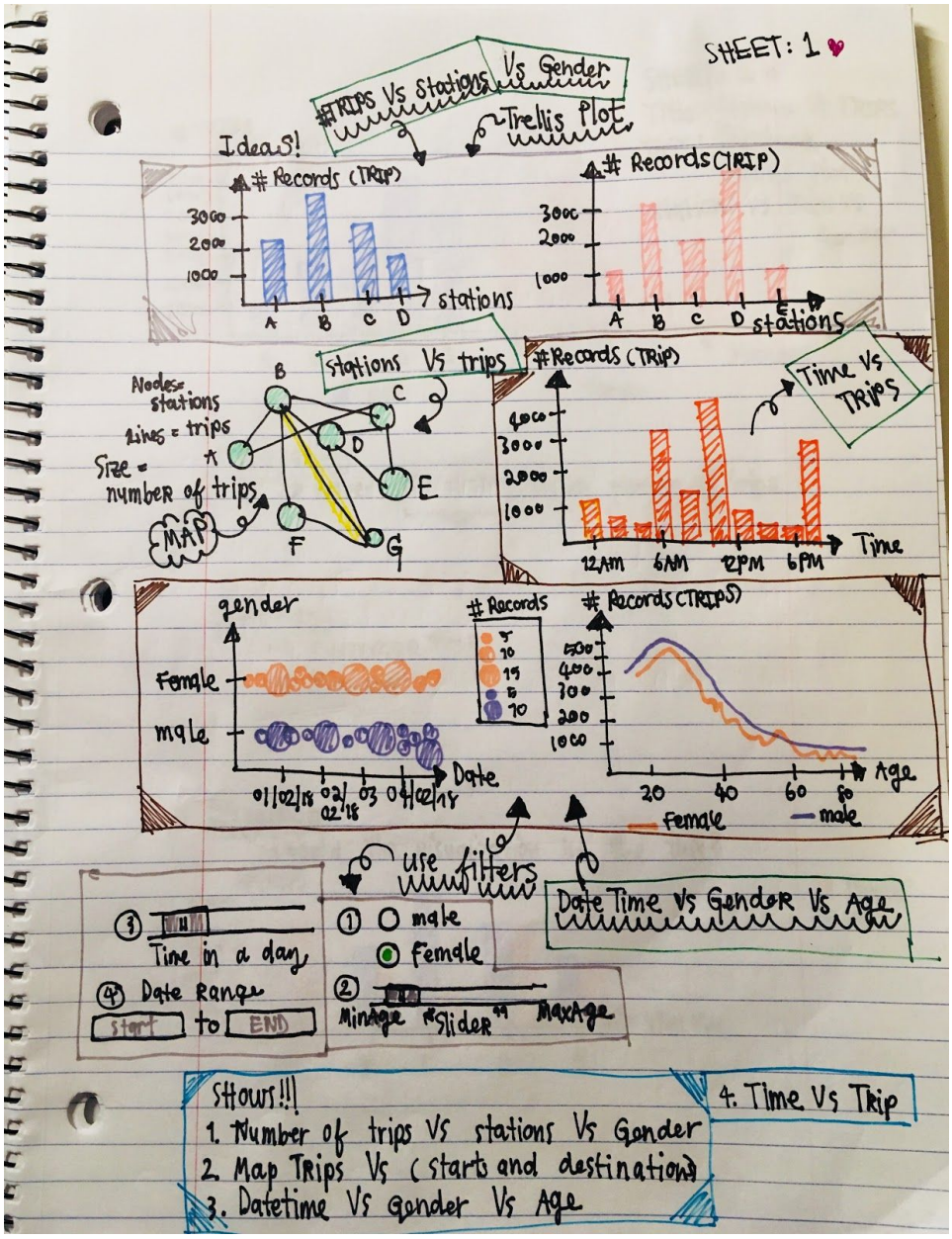
Finally, we did the same thing with 2017 trip data and got 2017.csv down from 117 MB to 38 MB.

We use the whole dataset because we want to create better interactive visualization experience. Viewers can filter the data without waiting for the new set of data from the server. This leads to the longer load time at the beginning.

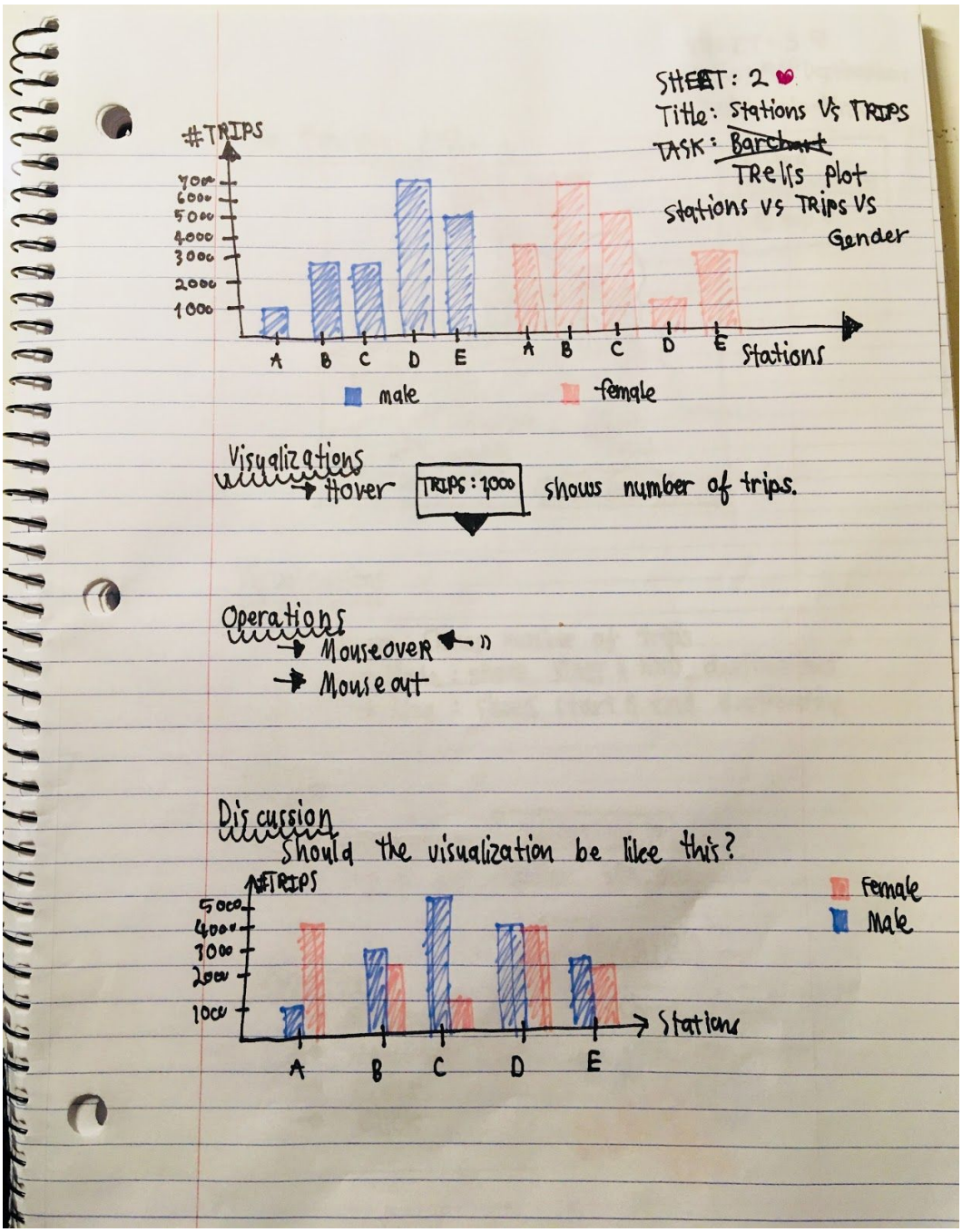
# Design Evolution

## Sketches

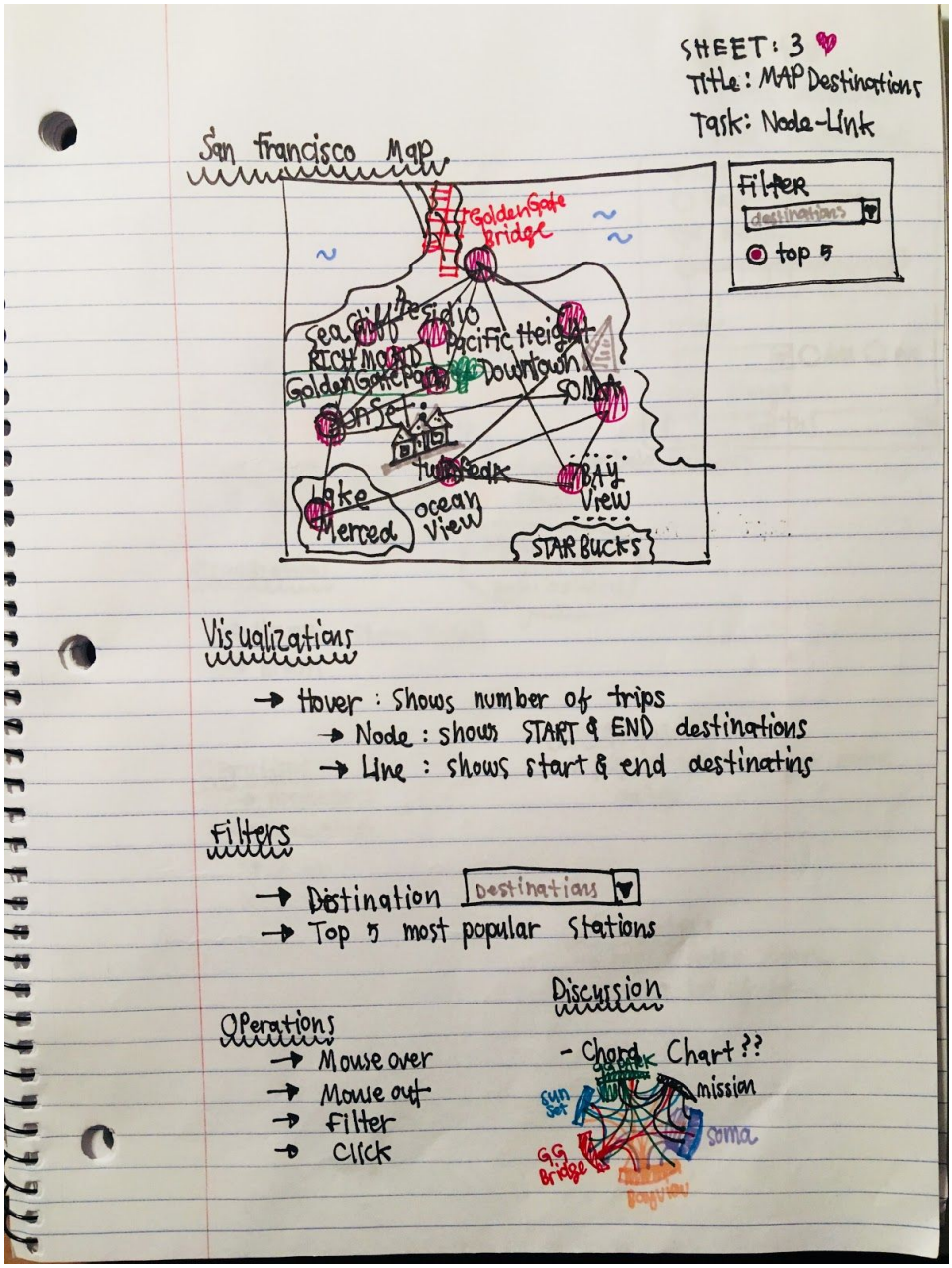
We used Five Design Sheet (FdS) methodology to create information visualization interfaces. In the first sheet, we brainstormed and sketched ideas what visualizations that we would like to show. We got 3 visualizations which were bar chart, node-link and scatter plot. In each visualization would be placed in each sheet. The final sheet is a visualization that we would like to implement. As you can see in the following images:



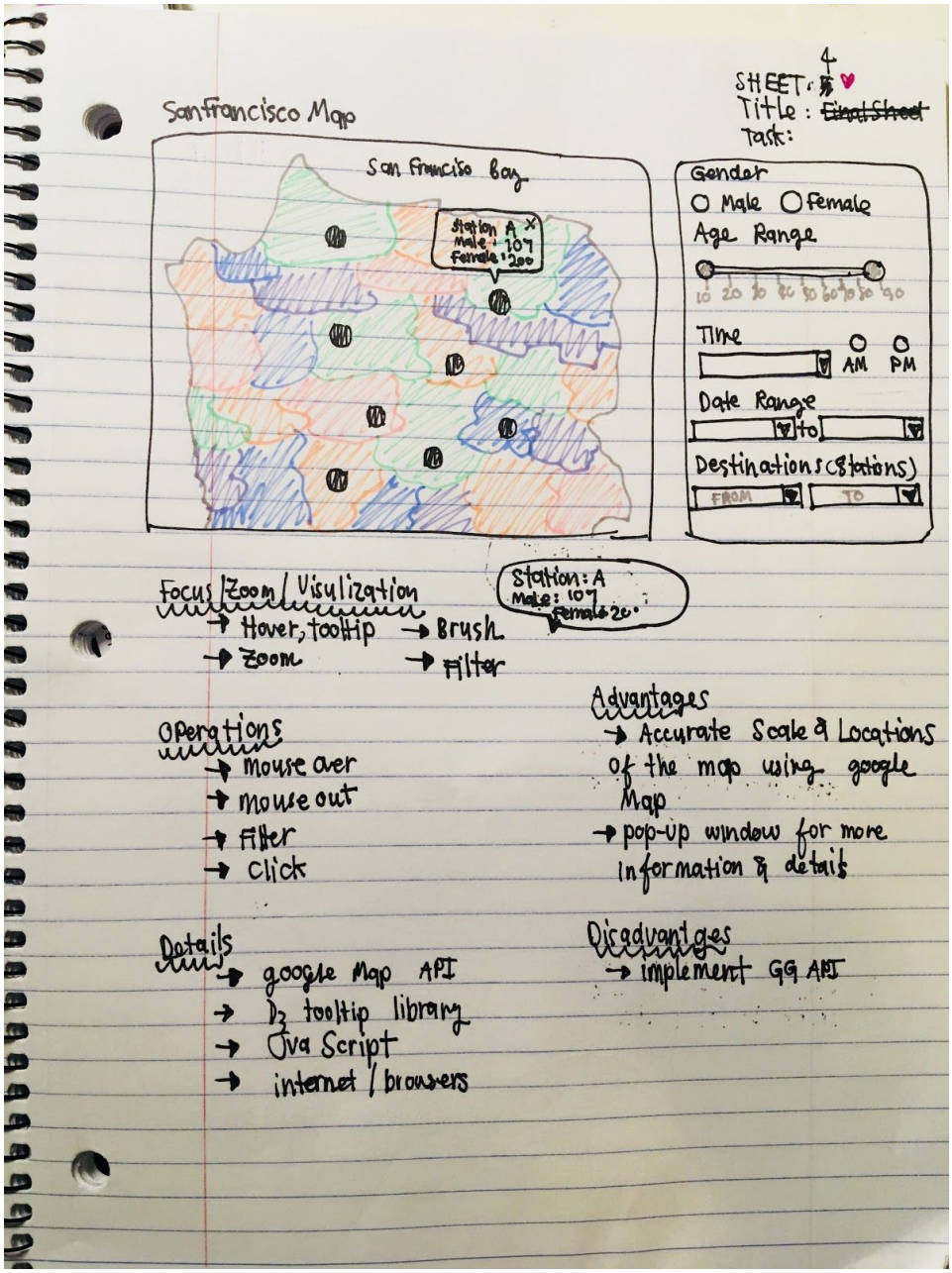

Sheet 1 : Data visualization ideas



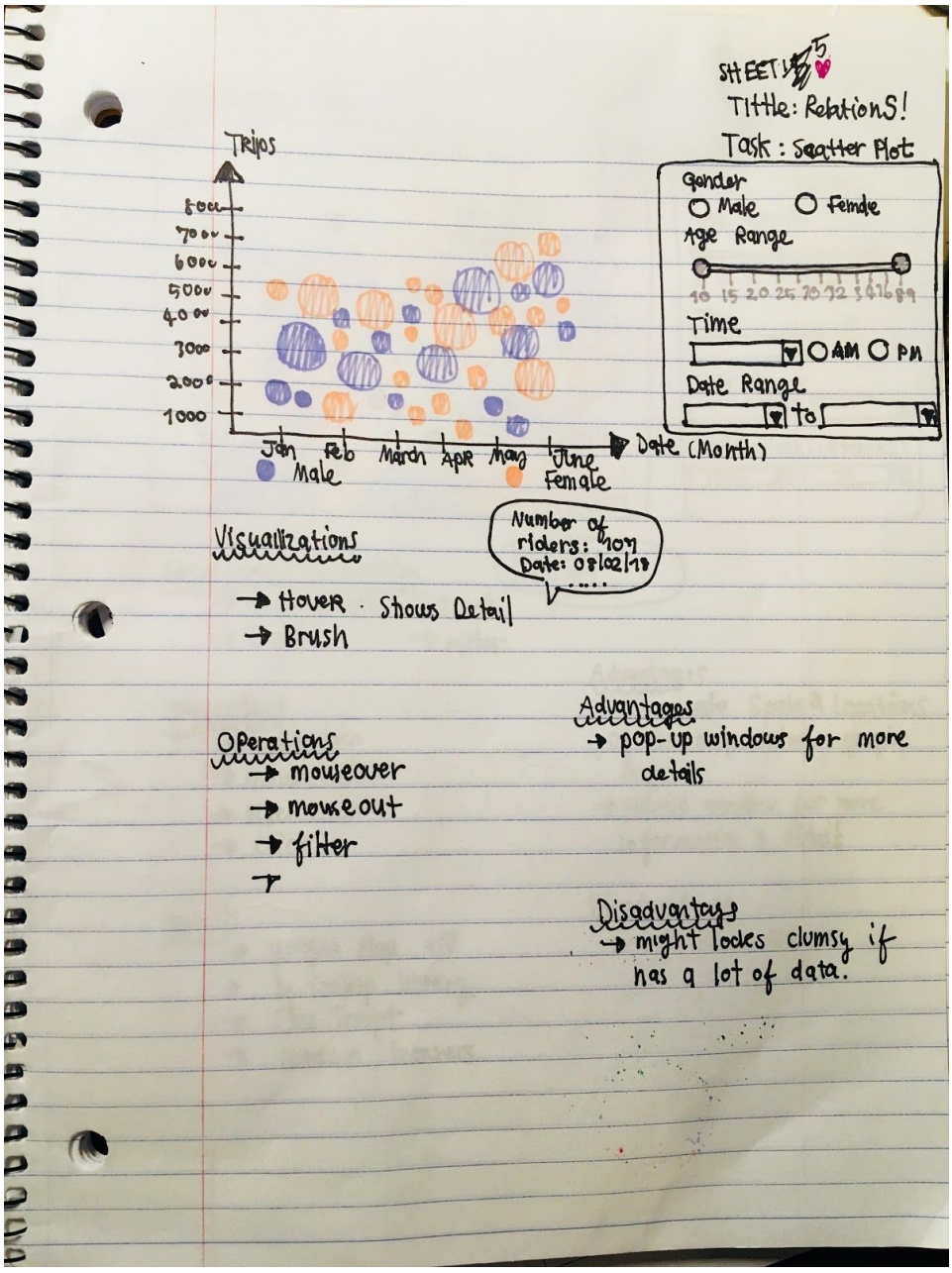

Sheet 2 : Bar chart shows relationship between gender and number of trips.




Sheet 3: Node-link on Google Maps shows occured trips between stations.




Sheet 4: Scatter plot on Google Map shows stations on Google Map which you can filter by gender, age, time, date and destinations.




Sheet 5: Scatter plot shows how many trips occurred at certain time. You can filter by gender, age, time and date.

## Prototypes


Figure 1-1: Example of bubble chart from The Simpson's Paradox

We wanted to create an interactive version of bubble chart like the Simpson's Paradox above to provide answers to questions Q1 (Are there more riders on the weekdays or weekends?) and Q2 (Are there more customers or subscribers using the service?) in which colors encode membership types and positions encode different categorical fields such as hours, days and genders.

We created the first prototype of this chart using D3 force-directed graph layout which is quite simple. However, there are several problems we have discovered along the way. Firstly, we had to group our data into smaller groups since we could not represent each individual record as a single bubble. This made things more complicated, for example, the transition between different combination of filters. Secondly, the force layout has unpredictable behaviour especially with many dots. We had to stop the animation manually. Learning how to control the force is important to make a good visualization, even though we are not Jedi. Due to the rising complexity in our design, we decided not to implement this.

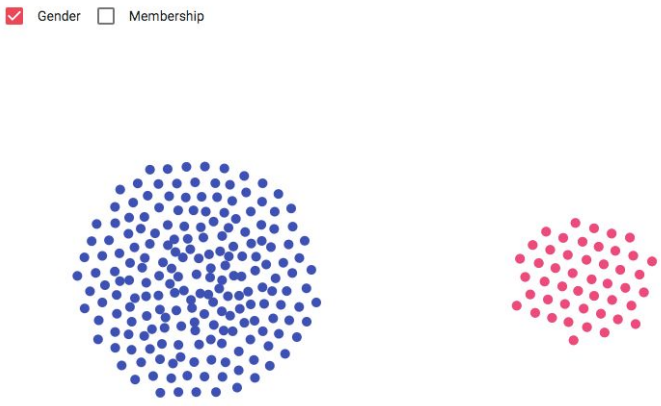



Figure 1-2: A prototype of bubble chart

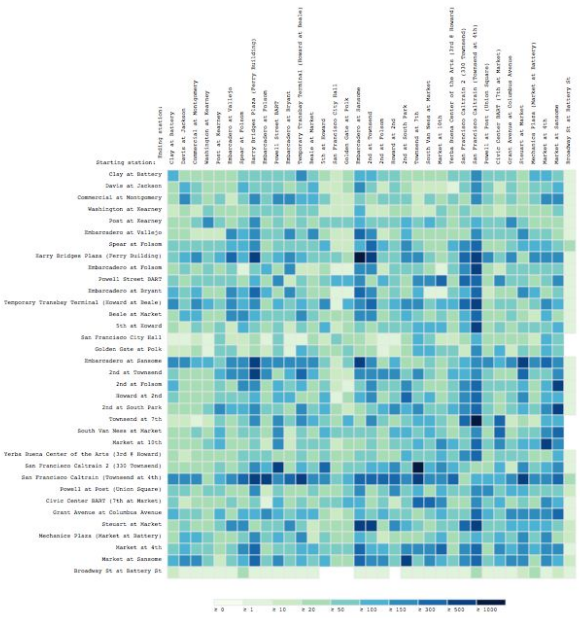



Figure 2-1: Example of matrix heatmap

We wanted to use the matrix representation to visualize the distribution of trips to answer questions Q3 (How does the distribution of trips look like?) and Q4 (Which are the stations that tend to be the most popular?)

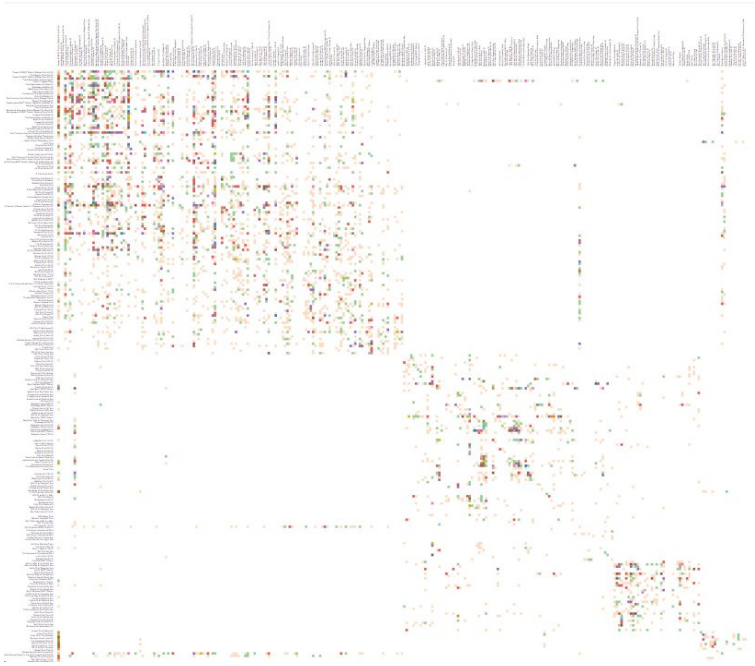


Figure 2-2: Stations heatmap

On the heatmap diagram of systemwide rides we see that activity tends to be grouped into squares. These are trips that took place within city boundaries, and we notice that not many riders go beyond their starting city. An exception to this are 70 rides went from one city to the other. The problem was we had more than 295 stations in the dataset which made the data presented looked cluttered and difficult to understand. After we looked in the dataset, we decided to change to map-based representation because we had latitude and longitude of each station.

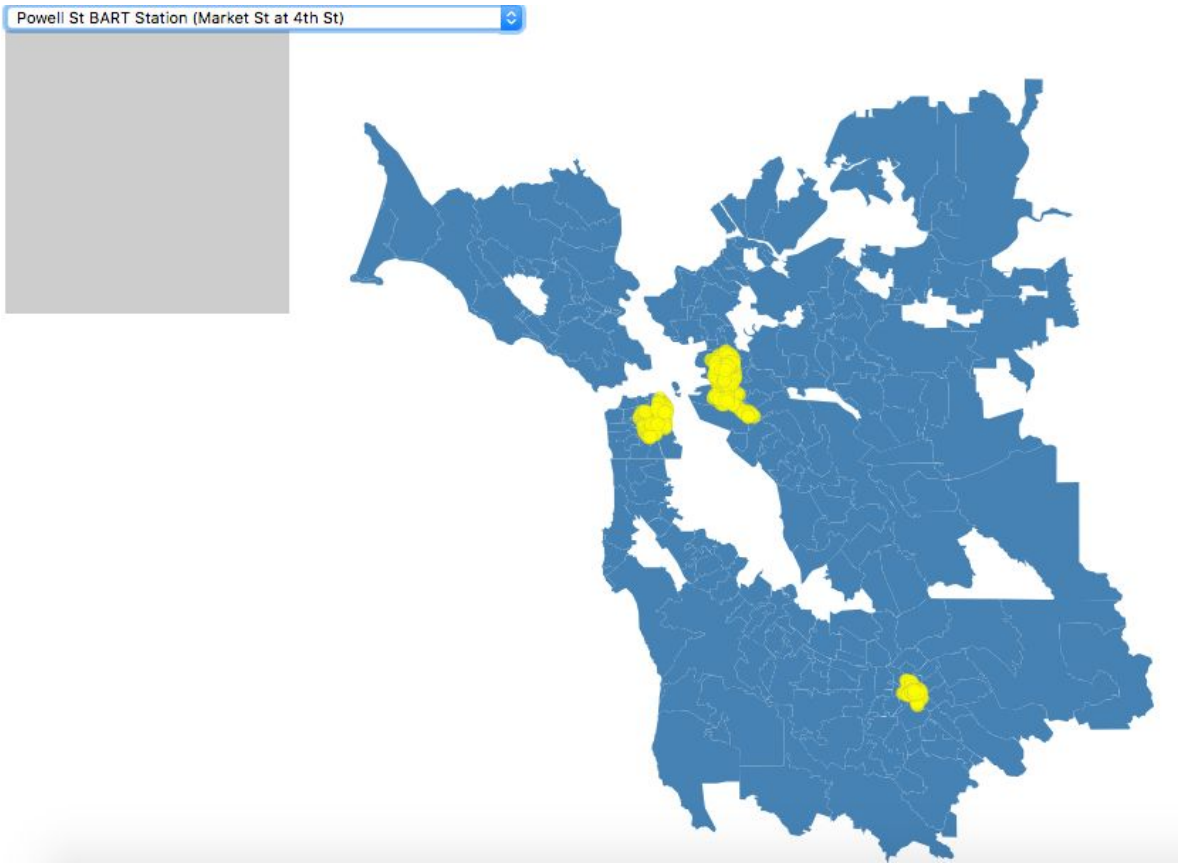



Figure 3-1: Choropleth Map Prototype

First, we implemented Choropleth map which we got inspired by in-class lab. We used Bay-Area geojson dataset to plot the map and then we plotted all stations on it. The map seemed to be a very good choice this time unfortunately it took us a lot of time to implement interactions and we were not able to produce meaningful analyses. We did some research and found that Google Maps JavaScript API would be a better idea because it already provided a basic map, allowed us to customize the map with our content and modified the map using their services and libraries.

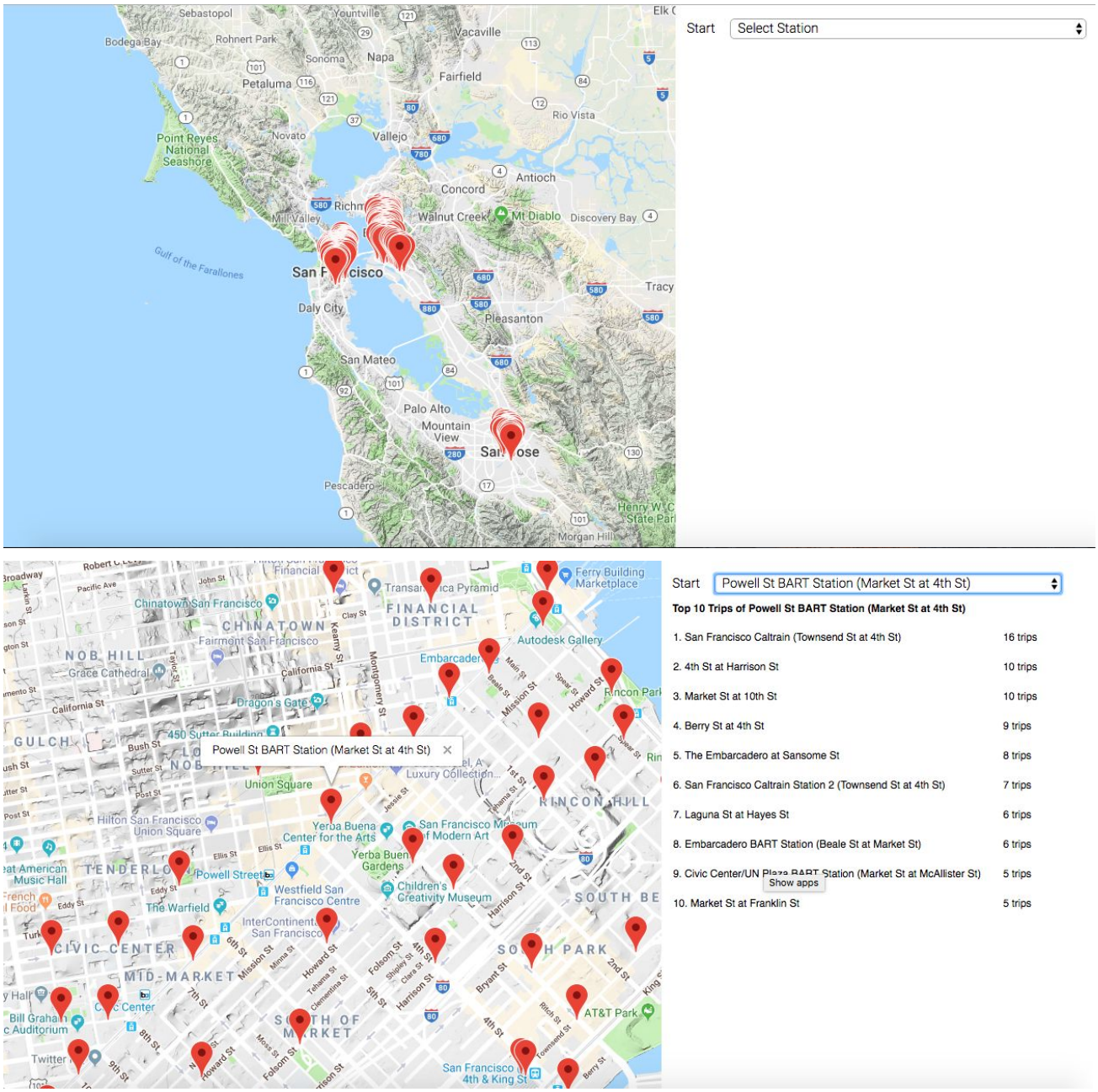



Figure 3-2: Google Maps API implementation

We used latitude and longitude points to create markers which each marker represented each station. We managed to add drop-down list to visualization that would enable the user to filter trips based on selected station. The drop-down list integrated well with the map, in such a way that users were able to select interested station and able to see the distribution of trips which answered the Q3 question. In addition, since we were able to filter the data in term of stations, we could add the most top 10 trips analysis which answered the Q4 question. After the the final presentation, our professor suggested us to draw arcs which showed trip distributions of each station when we moused over the station marker so we decided to implement this feature.

# Implementation / Final visualizations

Regarding the complexity of the bubble chart, we broke it down into multiple charts instead of encoding everything into a single force-directed graph. We decided to replace it with multiple less-fancy charts including a matrix, nested maps and bar charts. Later, we put everything in a single dashboard with a range slider that can filter data based on selected time range and cards displaying summarized statistics of number of trips, number of stations and average trip duration.

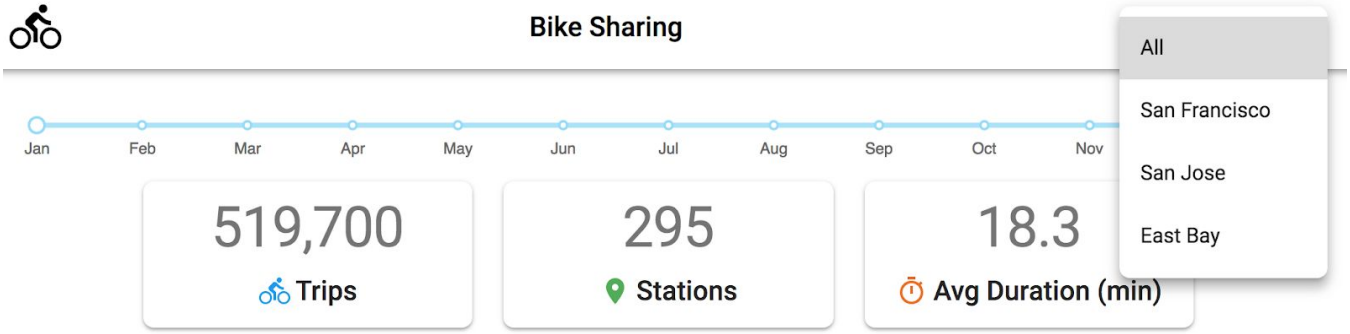



Figure 4: Website header with range slider, big numbers and region selection drop-down list

After conducting the stations heatmap and a Google search on the stations, we realized that the data consists of rides and stations across different cities which are San Francisco, San Jose and East Bay. Hence, we constructed a drop-down list on top-right of the website to examine each of the cities.


Figure 5: Nested maps for gender and membership type

We think a single-level nested map would be a good alternative of pie chart or vertical bar chart for showing ratio of data since we only have 2 to 3 categories for each map.

We already have simple bar charts showing trip distribution by time. Later, we realized that we could combine those together to create a single time matrix. However, we still keep both bar charts as references.

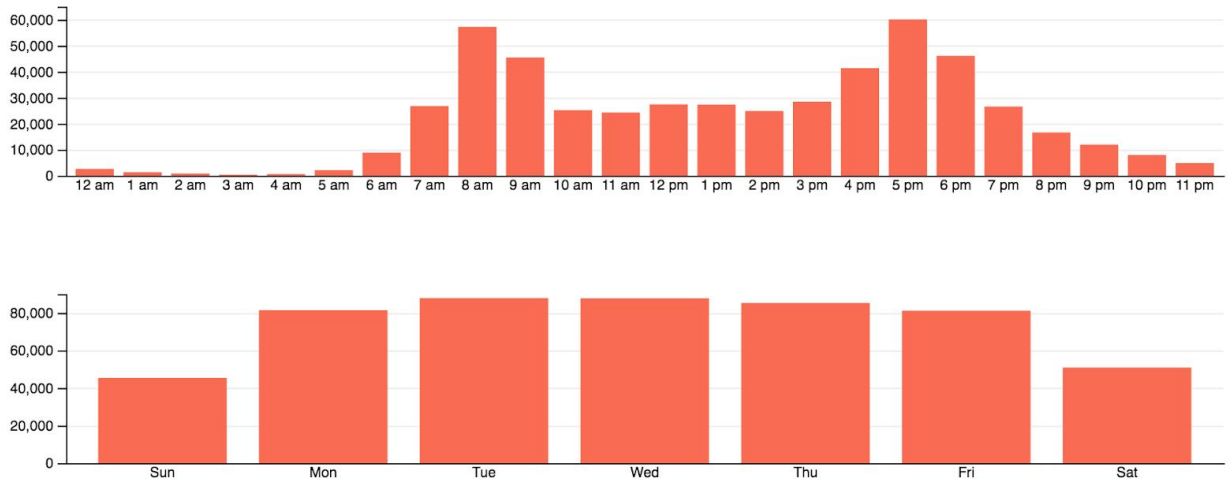



Figure 6: Bar charts showing trips by day of week and hour

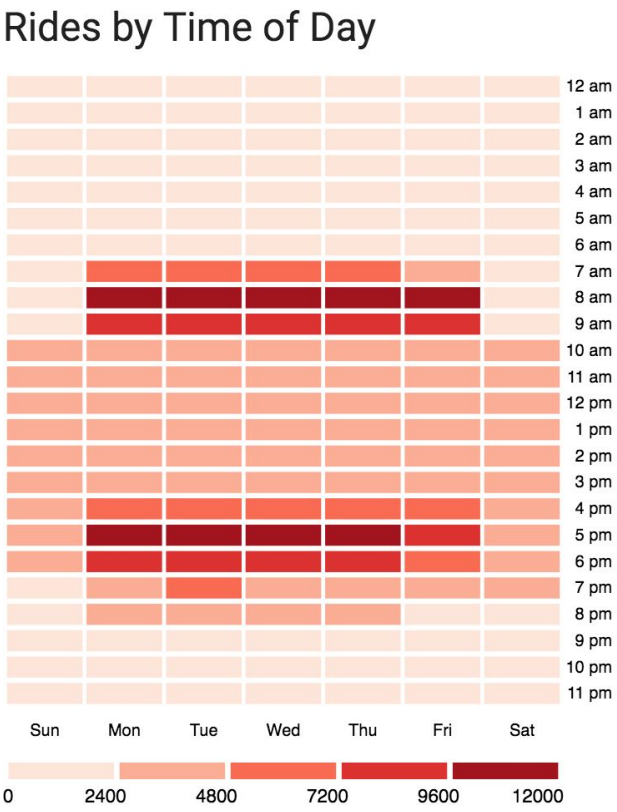



Figure 7: The distribution of trips in matrix

We made sure that every bar and cell have a tooltip showing corresponding information and also highlight it with dotted border or shifting color.

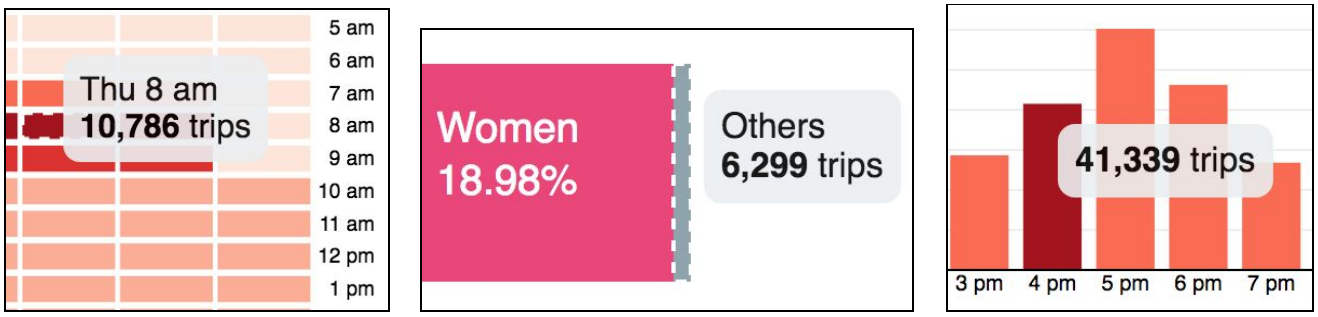



Figure 8: Tooltips and highlightings in different visualizations

Google Maps API provides easy integration of map interactions. We also implemented marker clustering which is good for displaying a large number of markers on our map.

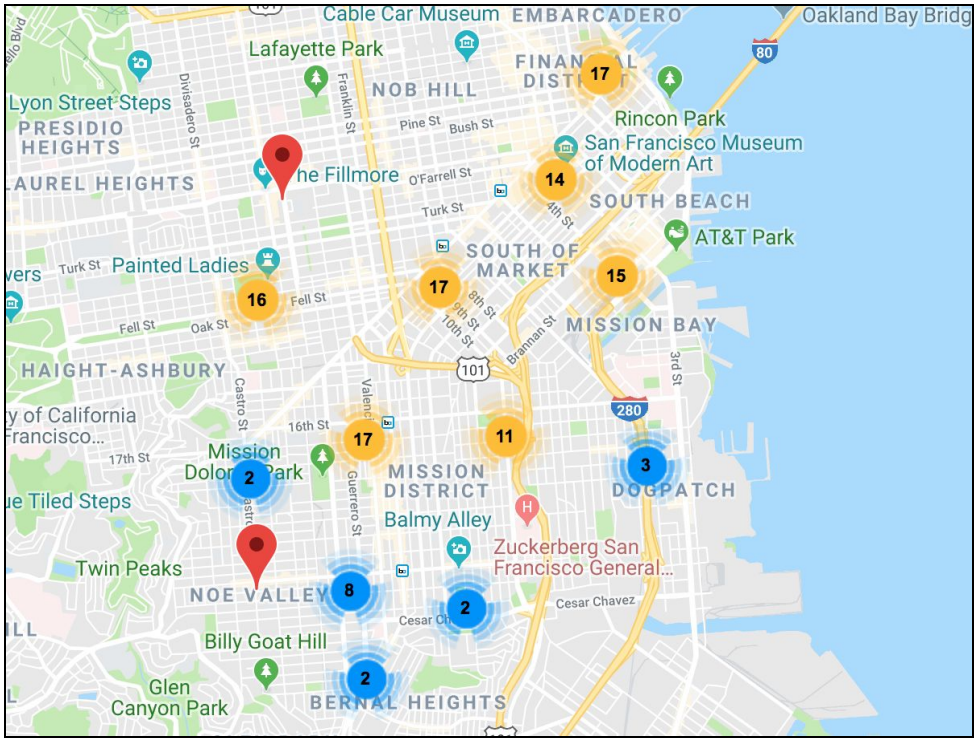



Figure 9: Google Maps marker clustering

By selecting a station from the drop-down list, or click on a marker, it will show only selected station and its top destinations on the map along with the list of trip information on the right column. Distributions of  the top 10 stations will be shown as lines drawing from selected station to given latitude and longitude points.


Figure 10: Markers with lines connected selected station to its top destinations with details on the right column

# Evaluation

Bike share is growing at an astounding clip across the U.S., with over 88 million trips made on a bike share system in the U.S. since 2010. According to our visualizations, in 2017 alone, riders took over 500 thousand trips, across 295 stations in the San Francisco Bay Area. We have seen the same patterns across different regions:

- The average trip duration is around 20 minutes.
- The peak hours are 8am and 5pm. (before and after working hours)
- The most number of trips were occurred in October.
- Almost 80% of the riders are men.
- About 80% of the trips made by subscribers.

In final visualizations, we walked through all analyses, visualizations and tools we used to help we learn about the data and answer questions:

1. **When** is the bike share system used? When is the peak time of the bike sharing system? Are there more riders on the weekdays or weekends?
2. **Where** are the stations? Where do people ride bike share?
3. **Which** are the stations that tend to be the most popular?
4. **Who** uses the bike sharing system? Female or male? Are there more customers or subscribers using the service?
5. **How** many trips in 2017? How many bike stations? How does the distribution of trips look like? How much is the bike share system used?
6. **What** is the average trip duration?

We believe our designs work well in many aspects. All visualizations are easy to understand in a few seconds with details-on-demand. Even though they are not perfect, they answer questions we asked about stations and members which help viewers gain more understanding of bike sharing services.

This is a list of ideas for the future improvements:

1. Making stations drop-down list easier to search. It could be replaced with an autocomplete text field.
2. Implementing brushing, for example, clicking on an element in a chart can highlight that group of records and reflect changes on the other charts.
3. Using Mapbox and Deck.gl could help us implement efficient map features with our large dataset.
4. Implement bar chart showing how many trips occurred during the period of time over the date range slider. This make viewers know which range doesn't have any data.
5. Utilizing space in the dashboard such as reducing the size of menu and number cards, so we could be able to see all information without scrolling.
6. Implementing histogram to see the distribution of rider ages.

# References


1. D3: Force-directed Graph Layout
   https://github.com/d3/d3-force
2. Google Maps: Arc between Markers
   https://jsfiddle.net/medmunds/sd10up9t/
3. Google Maps: Markers
   https://developers.google.com/maps/documentation/javascript/markers
4. Google Maps: Marker Animations
   https://developers.google.com/maps/documentation/javascript/examples/marker-animations
5. Google Maps: Marker Clustering
   https://developers.google.com/maps/documentation/javascript/marker-clustering
6. React: Declarative D3 transitions with React
   https://swizec.com/blog/declarative-d3-transitions-react/swizec/8323
7. React: Interactive Applications with React & D3
   https://medium.com/@Elijah_Meeks/interactive-applications-with-react-d3-f76f7b3ebc71
8. React: Range Slider
   https://github.com/react-component/slider
9. Tableau: Visual Analysis Best Practices
   https://www.tableau.com/sites/default/files/media/whitepaper_visual-analysis-guidebook_0.pdf