Final Project Proposal

I. Group Info: Group 50 – Wenbo Hu, Martin Thai, Wenqian Zhao

II. Project description

A. Project Topic: This data visualization aims at visualizing and analyzing the Bay Wheel business operated by Lyft.

B. Dataset Description: This dataset contains Bay Wheels's trip data, which includes following attributes.

1. Trip.csv
    a) id
    b) duration
    c) start_date
    d) start_station_name
    e) start_station_id
    f) end_date
    g) end_station_name
    h) end_station_id
    i) bike_id
    j) subscribtion_type
    k) zip_code
2. Weather.csv
    a) max_temperature_f
    b) mean_temperature_f
    c) min_temperature_f

Although this is a real-time dataset collection which updates monthly on https://www.lyft.com/bikes/bay-wheels/system-data, we are planning to use the more comprehensive dataset of 2014 as it contains less missing values and extra information compared with the most recent ones; and the contributor also contains weather data that will boost our data analysis progress. Currently, we are thinking of analyzing the popular route decision by users (most popular Start Station - End Station combination), the trend of numbers of trips monthly etc.

C. Dataset link: https://www.kaggle.com/benhamner/sf-bay-area-bike-share

D. We chose this dataset to deal with the issue of global warming and answering governments' efforts in promoting sustainability. Bike sharing has been an economical, ecological, and popular choice in dense city areas where the distance of traveling is relatively short. As this service is

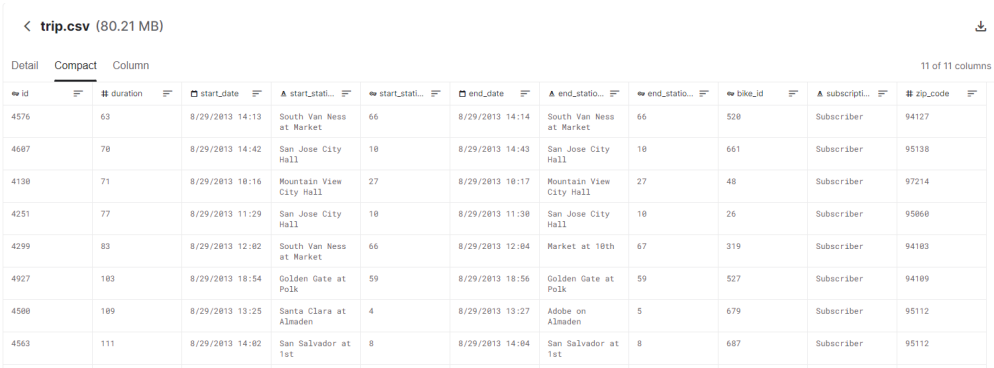
not yet popular in the La Jolla area, we would like to dig more into this topic and discover some facts from our visualizations.

III. Dataset of our choice

A. The URL of the dataset: https://www.kaggle.com/benhamner/sf-bay-area-bike-share

B. Dataset Image

1. trip.csv



trip.csv (80.21 MB)

Detail Compact Column 11 of 11 columns

id	duration	start_date	start_stati...	start_stati...	end_date	end_statio...	end_statio...	bike_id	subscripti...	zip_code
4576	63	8/29/2013 14:13	South Van Ness at Market	66	8/29/2013 14:14	South Van Ness at Market	66	520	Subscriber	94127
4607	70	8/29/2013 14:42	San Jose City Hall	10	8/29/2013 14:43	San Jose City Hall	10	661	Subscriber	95138
4130	71	8/29/2013 10:16	Mountain View City Hall	27	8/29/2013 10:17	Mountain View City Hall	27	48	Subscriber	97214
4251	77	8/29/2013 11:29	San Jose City Hall	10	8/29/2013 11:30	San Jose City Hall	10	26	Subscriber	95060
4299	83	8/29/2013 12:02	South Van Ness at Market	66	8/29/2013 12:04	Market at 10th	67	319	Subscriber	94103
4927	103	8/29/2013 18:54	Golden Gate at Polk	59	8/29/2013 18:56	Golden Gate at Polk	59	527	Subscriber	94109
4500	109	8/29/2013 13:25	Santa Clara at Almaden	4	8/29/2013 13:27	Adobe on Almaden	5	679	Subscriber	95112
4563	111	8/29/2013 14:02	San Salvador at 1st	8	8/29/2013 14:04	San Salvador at 1st	8	687	Subscriber	95112

2. weather.csv

weather.csv (438.06 kB)

Detail Compact Column

date	max_temp...	mean_tem...	min_tempe...
8/29/2013	74.0	68.0	61.0
8/30/2013	78.0	69.0	60.0
8/31/2013	71.0	64.0	57.0
9/1/2013	74.0	66.0	58.0
9/2/2013	75.0	69.0	62.0
9/3/2013	73.0	67.0	60.0
9/4/2013	74.0	68.0	61.0
9/5/2013	72.0	66.0	60.0
9/6/2013	85.0	71.0	56.0

IV. Tasks

A. Explore the trend of trips number in the past year monthly

1. Line plot
2. 1 key (month), 1 value (count)
3. Can be interactive with a sorting bottom

B. Explore the relation between the temperature of the day and time duration of the bike ride
    1. Scatter plot
    2. 0 key, 2 values
C. Explore the average duration of different types of riders (customer or subscriber)
    1. Stacked bar graph
    2. 2 keys, 1 value
D. Explore the linked relations between each stations, using the locations of bike stations by latitude and longitude
    1. Network + topographic map
    2. 2 keys, 1 value
E. Explore the most area of trips with choropleth
    1. Choropleth map
    2. Interactive with hover
    3. 1 key

V. Timeline and Work Distributions

A. Timeline
    1. Finding Dataset: Finished
    2. Drafting Project Proposal: Best before March 7th, depending on the actual progress. **No later than March 10th.**
    3. Data Cleaning: March 10th
    4. Code Implementation: Best before March 15th . **No later than March 16th.**
    5. Documentation: March 16th

B. Work Distributions
    1. We plan to have three separate data cleaning and preprocessing progress by each member to ensure that everyone's data is under the format in favor
    2. Each one will have at least one plot to code depending on the workload and actual time distribution
    3. Depending on the format of final documentation, the work distribution will be adaptively changed. However, the general proposal for that is each one will handle the part of responded plot, and we will aggregate our opinion to complete the final version