Project 1 - Final Report

Dataset description

1. **The dataset**

1) After careful consideration, we chose the dataset of bike renting of bay area, which can be found here, Bay Area Bike Share. This set consists of three parts. Firstly, the location of the stations, including longitude and altitude. (Figure 1) Secondly, the renting information of the bicycles, consisting more than 310 K lines with each line showing the start time, start location, end time, end location, and so on. (Figure 2) The final sheet of the set shows us the weather info around the year including 9 factors. (Figure 4) Here are parts of our data set.

station_id	name	lat	long	dockcount	landmark	installation
2	San Jose Diridon Caltrain Station	37.329732	-121.901782	27	San Jose	8/6/2013
3	San Jose Civic Center	37.330698	-121.888979	15	San Jose	8/5/2013
4	Santa Clara at Almaden	37.333988	-121.894902	11	San Jose	8/6/2013
5	Adobe on Almaden	37.331415	-121.8932	19	San Jose	8/5/2013
6	San Pedro Square	37.336721	-121.894074	15	San Jose	8/7/2013
7	Paseo de San Antonio	37.333798	-121.886943	15	San Jose	8/7/2013
8	San Salvador at 1st	37.330165	-121.885831	15	San Jose	8/5/2013
9	Japantown	37.348742	-121.894715	15	San Jose	8/5/2013
10	San Jose City Hall	37.337391	-121.886995	15	San Jose	8/6/2013

Figure 1 Station Info

Trip ID	Duration	Start Date	Start Station	Start Terminal	End Date	End Station	End Terminal	Bike #	Subscriber Type	Zip Code
913465	746	9/1/2015 0:10	San Francisco Caltrain 2 (330 Townsend)	69	9/1/2015 0:23	San Francisco City Hall	58	238	Subscriber	94107
913466	969	9/1/2015 0:15	Clay at Battery	41	9/1/2015 0:31	Washington at Kearny	46	16	Subscriber	94133
913467	233	9/1/2015 0:15	Davis at Jackson	42	9/1/2015 0:19	Commercial at Montgomery	45	534	Subscriber	94111
913468	213	9/1/2015 1:29	Clay at Battery	41	9/1/2015 1:32	Steuart at Market	74	312	Subscriber	94107
913469	574	9/1/2015 1:33	Steuart at Market	74	9/1/2015 1:42	San Francisco Caltrain 2 (330 Townsend)	69	279	Subscriber	94107
913470	623	9/1/2015 1:36	San Jose Diridon Caltrain Station	2	9/1/2015 1:47	Japantown	9	261	Subscriber	95112

Figure 2 Trip Info

Max TemperatureF
Mean TemperatureF
Min TemperatureF
Max Dew PointF
MeanDew PointF
Min DewpointF
Max Humidity
Mean Humidity
Min Humidity
Max Sea Level PressureIn
Mean Sea Level PressureIn
Min Sea Level PressureIn
Max VisibilityMiles
Mean VisibilityMiles
Min VisibilityMiles
Max Wind SpeedMPH
Mean Wind SpeedMPH
Max Gust SpeedMPH
PrecipitationIn
CloudCover
Events

Figure 3 Weather Variables

2) To help with the visualization, we also use a json map of the bay area, which is cut from the map of San Francisco.

2. The process of data

1) The route information

To show the frequency of the each route, we group the trip elements according the start station and end station. Then we summarize the count of different routes using Pivot Table form Excel. (we use color of the lines to indicate the frequency of different route) And to present the route in the map, we also map the location (latitude and longitude) of the station to the trip data. (Figure 4)

Start Station	End Station	count	start lat	start lon	end lat	end lon
South Van Ness at Market	South Van Ness at Market	136	37.774814	-122.418954	37.774814	-122.418954
Golden Gate at Polk	South Van Ness at Market	78	37.781332	-122.418603	37.774814	-122.418954
San Francisco City Hall	South Van Ness at Market	79	37.77865	-122.418235	37.774814	-122.418954
Market at 10th	South Van Ness at Market	74	37.776619	-122.417385	37.774814	-122.418954
Civic Center BART (7th at Market)	South Van Ness at Market	188	37.781039	-122.411748	37.774814	-122.418954
Cyril Magnin St at Ellis St	South Van Ness at Market	2	37.785908	-122.408891	37.774814	-122.418954
Powell at Post (Union Square)	South Van Ness at Market	90	37.788446	-122.408499	37.774814	-122.418954
Powell Street BART	South Van Ness at Market	401	37.783871	-122.408433	37.774814	-122.418954
Grant Avenue at Columbus Avenue	South Van Ness at Market	49	37.7979	-122.405942	37.774814	-122.418954

Figure 4 the route data for visualization

2) The weather data

After analyzing the dataset, we want to show how the weather would affect the renting behavior. There are three categories in the sheet, the maximum the minimum and the average. After calculation, the variance of the data is significantly small enough so that we just choose the average value. And the area is relatively small (the variance around the stations is not too high), so we just average the data of different stations to get the data of the whole area.

However, there are too many factors in the sheet, some of which are definitely of little use such as the sea level. And we can only show limited variables in a static chart. With the help of some examples, we think we can show 4 variables mostly. And the count is obviously a necessary one so we can only choose three factors from the weather sheet. To solve the problem, we use linear regression using MATLAB to figure out the priority of the factors. Below is the summary of the model (Figure 5), and we figure out that the temperature, visible miles, and wind speed are the most influential factors, which is just as we expected.

```
>> model = LinearModel.fit(test_we,SumofDuration)

model =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8

Estimated Coefficients:
                   Estimate         SE          tStat      pValue
                   ________    _________    _________    _______

    (Intercept)    -7.1602e+06    2.0897e+08    -0.034264    0.97269
    x1                -6946.8    6.0944e+05    -0.011399    0.99091
    x2                  26237    6.9539e+05      0.03773    0.96992
    x3              2.7133e+05    6.7602e+06     0.040137    0.96801
    x4                 -11905    3.4354e+05    -0.034655    0.97237
    x5                -5554.7    3.2102e+05    -0.017303     0.9862
    x6                 -17327    5.0153e+05    -0.034548    0.97246
    x7                 705.95         13922     0.050708    0.95959
    x8                  28720    1.2113e+06      0.02371     0.9811


Number of observations: 367, Error degrees of freedom: 358
Root Mean Squared Error: 1.37e+07
R-squared: 0.000142,  Adjusted R-Squared -0.0222
F-statistic vs. constant model: 0.00635, p-value = 1
```

Figure 5 Summary of the Linear Regression model

Data Mapping

[B. A description of the mapping from data to visual elements. Describe the scales you used, such as position, color, or shape. Mention any transformations you performed, such as log scales. (10 pts)]

1. **Route map**

We plotted the map of the bay area using coordinates from the json using the geo mercator scale. This scale was most suited as it gives a flat map. We then inserted SVG circle

and text elements to mark the start and end bike share stations from the station data csv file. We selected a color scale and displayed it near the map, indicating routes from least popular to most popular. After marking the stations on the map, we used the coordinates of the start and end stations to connect them using SVG line. Routes in deep red indicate to be the most popular choice, whereas in light red to be the least.

2. **Weather graph**

After finding the most influential weather factors affecting the bike sharing statistics, we decided to show a bubble chart representation for this data. Since we wanted to show four variables, namely, temperature, visible miles, wind speed and duration of biking, we represented two of these using X and Y axes (temperature and visible miles respectively). The color of the plotted circles represents a range of the wind speed indicated at the bottom of the graph and the size of the circle corresponds to the duration of biking (the greater the area of the circle, the more is the duration).

We used the linear scale for temperature, visible miles as the difference between the minimum and maximum values in their domain was not too big. We have used ordinal scale to map a range of values for the wind speed to distinct colors. For duration, the difference was very significant and hence we used the log scale. We realized that using log scale in this scenario wouldn't misrepresent the linear scale as the log scale is only being used for the area of the circle, and we wanted a comparative visual for the duration.

Visualization Description

The Bay area bike share utilizes solar powered bikes designed for an urban setting like the bay area, which helps the residents and visitors have an easy and affordable source of transportation for getting around the region. The system is available for use 24 hours a day, 365 days a year. These bikes are solar powered which makes them environment friendly.

A few interesting observations from the visualization are as follows:

- The bike share is heavily utilized and forms a dense network of routes overall.
- The start and end stations are located within a distance of 3 miles from each other.
- The less popular routes are more towards the interior area and the more popular routes are between the stations on the east border of the bay.
- Biking duration is more or less uniform across the 10 visible miles mark, in spite of varying temperature and wind speed.
- The bay area has visible miles equal to 10 for most time of the year.
- Biking frequency and duration greatly reduces when the number of visible miles goes below 7.
- Lower temperature and higher wind speed are generally resulting in less biking duration as expected.
- The biking duration does not vary drastically for a majority part of the year (indicates that the bikers have fixed routes that they travel on a regular basis).
- Duration of biking is more when the temperature goes above 50F.