Harrisburg University of Science and Technology

Project Report

BIKE SHARE DATASET ANALYSIS

ANLY 500-90- O-2020/Late Fall –
Analytics I: Principles &Applications

Prof. Jonathan Wayne Korn

Team Members

Neeraj Gupta
Tripti Mishra

Table of Content

INTRODUCTION

A bike sharing system is a service provided by companies that allow users to rent bikes on a short-term basis. There are currently more than 500 bike sharing programs around the world. These services are usually aimed at reducing congestion, noise and air pollution by providing easy access to bikes for short-distance trips in metropolitan or urban areas. It is quite interesting to understand the number of users and their variability in using such services, the service location and availability of bikes and how do entities that oversee these systems manage the demand and supply in an efficient and cost-effective manner.

For our research project, we are using the bay area bike share dataset. This paper defines the problem statement, discusses the key attributes of the bay area bike share dataset and the approach we used to analyze the dataset to solve for the unbalanced station and non-uniformity of bikeusage between customers and subscribers.

PROBLEM STATEMENT

For our research purposes, we are interested in understanding the balance issues amongst the stations and bikes per the customers and subscribers of bay area bike share users. To elaborate, bike user's preference of renting a bike from a particular station varies over the year. This change in preference makes it difficult for the service providers to allocate enough bikes in that particular station and vice-a-versa. Due to this change in preference, service providers use previous year's records to allocate bikes in particular stations. It is important for the organizations to analyze the availability of bikes in order to increase the effectiveness and use of these services and thereby attract more users. The main issues we focus on resolving in this project are:

Unbalanced and underutilized Stations

Most bike sharing programs face the problem of variability in usage amongst stations where the trips quantity 'from' these stations is higher when compared to the number of trips 'to' the stations.

Non-uniformity in Bicycle usage

The stations that are in popular locations have high bike rentals when compared to the less popular station. Due to this non-uniformity in usage, some bikes are heavily used in popular stations and some are in 'like new' condition in less popular stations. This results in another problem where the regularly used bicycles get worn out and have to be bought to workshops to fix and repair.

The objective here is to analyze the information in this dataset and check whether there is a plausibility to recommend transfer of bicycles to balance the usages.

BAY AREA BIKE SHARE DATASET DESCRIPTION

In this project we are going to analyze bay area bike share dataset. Bay Area Bike Share started their operation in August 2013 in San Francisco Bay Area with 700 bicycles in 70 stations. The bike stations are available 24 X 7. We have selected the Year 2 Data and focus on trip dataset and station dataset.

Both datasets are available is .csv format and their structure are shown below

Trip dataset

Trip ID	Duration	Start Date	Start Station	Start Terminal	End Date	End Station	End Terminal	Bike #	Subscriber Type	Zip Code

Station dataset

station_id	name	lat	long	dockcount	landmark	installation

Customers and Subscriber

In the trip dataset there are two types of users: customers and subscriber. Customers use the bicycles for 3 day or less whereas the Subscribers are the individuals who have annual pass. When compared, subscribers use bikes greater number of times than the customers. The background information available is useful for the project as they indicate each and every trip along with the user information.

Relevant background information

The background information is obtained from the Bay Area Bike Share website dataset, i.e., the data extracted are from Sep 2014 to Aug 2015. This data is open source and contains multiple files. In the zip file, we use the 201508_station_data and 201508_trip_data. By analyzing the structure file, we could get some of the key points regarding the dataset and the information through the year. Bay Area Bike Share publishes the open-source data from September 2014 to August 2015. The data structure is shown below.

There are in total 354,152 rentals and by observing all the details we can identify the bicycle usage according to the particular station.

Trip ID	Duration	Start Date	Start Station	Start Terminal	End Date	End Station	End Terminal	Bike #	Subscriber Type	Zip Code
913460	765	8/31/2015 23:26	Harry Bridges Plaza (Ferry Building)	50	8/31/2015 23:39	San Francisco Caltrain (Townsend at 4th)	70	288	Subscriber	2139
913459	1036	8/31/2015 23:11	San Antonio Shopping Center	31	8/31/2015 23:28	Mountain View City Hall	27	35	Subscriber	95032
913455	307	8/31/2015 23:13	Post at Kearny	47	8/31/2015 23:18	2nd at South Park	64	468	Subscriber	94107
913454	409	8/31/2015 23:10	San Jose City Hall	10	8/31/2015 23:17	San Salvador at 1st	8	68	Subscriber	95113
913453	789	8/31/2015 23:09	Embarcadero at Folsom	51	8/31/2015 23:22	Embarcadero at	60	487	Customer	9069
913452	293	8/31/2015 23:07	Yerba Buena Center of the Arts (3rd @ Howard)	68	8/31/2015 23:12	San Francisco Caltrain (Townsend at 4th)	70	538	Subscriber	94118
913451	896	8/31/2015 23:07	Embarcadero at Folsom	51	8/31/2015 23:22	Embarcadero at	60	363	Customer	92562
913450	255	8/31/2015 22:16	Embarcadero at Sansome	60	8/31/2015 22:20	Steuart at Market	74	470	Subscriber	94111
913449	126	8/31/2015 22:12	Beale at Market	56	8/31/2015 22:15	Temporary Transbay	55	439	Subscriber	94130
913448	932	8/31/2015 21:57	Post at Kearny	47	8/31/2015 22:12	South Van Ness at	66	472	Subscriber	94702
913443	691	8/31/2015 21:49	Embarcadero at Sansome	60	8/31/2015 22:01	Market at Sansome	77	434	Subscriber	94109
913442	633	8/31/2015 21:44	Market at 10th	67	8/31/2015 21:54	San Francisco Caltrain	70	531	Subscriber	94107
913441	387	8/31/2015 21:39	Market at 4th	76	8/31/2015 21:46	Grant Avenue at	73	383	Subscriber	94104
913440	281	8/31/2015 21:31	Market at Sansome	77	8/31/2015 21:36	Broadway St at Battery	82	621	Subscriber	94107

Challenges

One of the main challenges we faced during the project is processing the overall data among the 354152-trip data. Before we do any analysis, we needed to process the data and we identified that the huge size of data made it very difficult for processing. A lot of time was spent in just cleaning the dataset, converting the format to time and date and allocating the right variable to get the exact weekday, month, day of the week, holiday and weekend specified for each data as well as segregating the time information into two different types to calculate hours and minutes so that appropriate duration of the bike rental can be calculated. With this huge dataset, we expected to acquire some other meaningful information that would help us calculate or predict number of rentals from a particular station. Using station data and tripdata was the only way to deal with anticipating the measure of Bike share. Subsequently, we needed to discover different indicators that could improve our underlying models. Finding the information for additional features was cumbersome. Moreover, combining this data into our working dataset likewise took a lot of time.

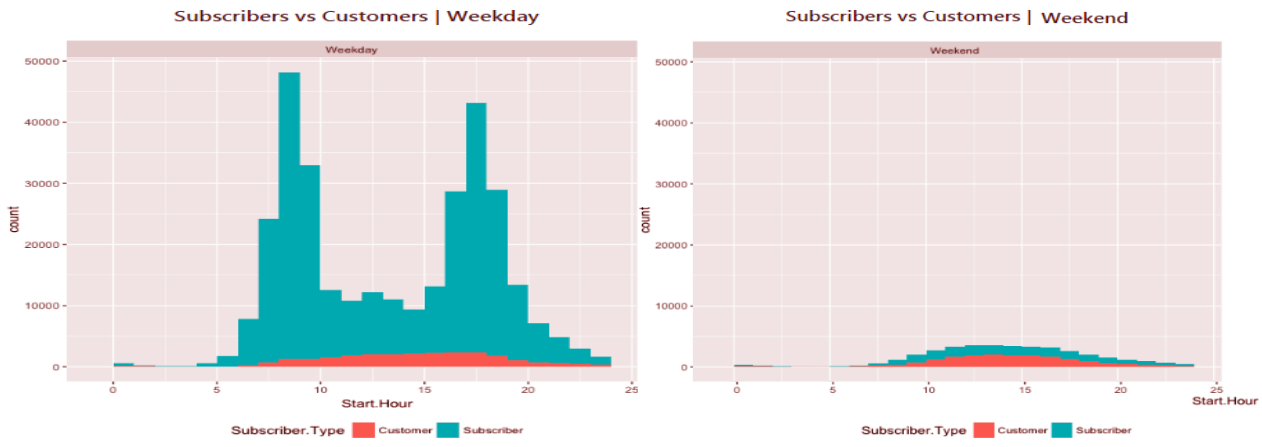
EXPLORATORY DATA ANALYSIS

The primary data is obtained from the bay area bike share data. This data was collected by a programmed machine in a periodic manner. Bike share information incorporated trip information that recorded each trip detail.In order to get a clear view, we developed scripts that would ascertain day by day elements to work on. We compare both the datasets to get clear ideas. Among the trips, there are 43935 customers and 310217 subscribers.

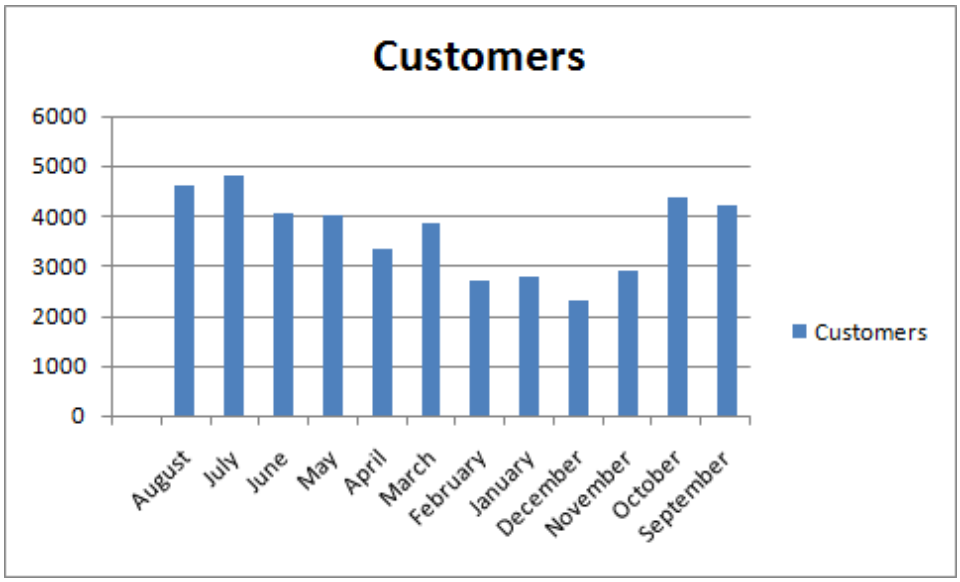
Year	Month	Customers	Subscribers	Total # of Trips
2015	August	4596	27308	31904
2015	July	4824	27645	32469
2015	June	4039	27868	31907
2015	May	3995	25547	29542
2015	April	3325	28036	31361
2015	March	3874	27752	31626

Time-based Usage

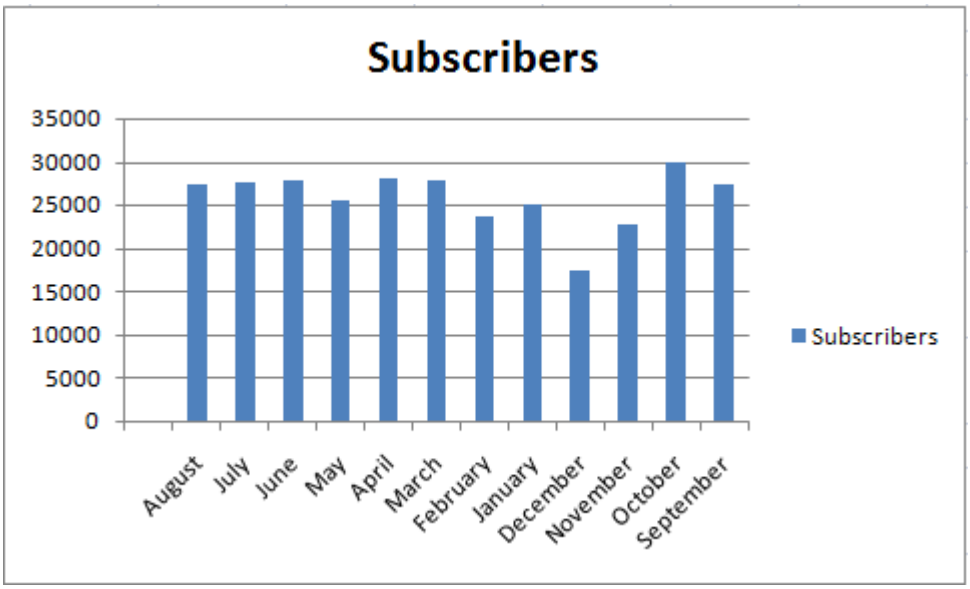
The graph below shows the variability in bike usage for customers and subscribers.



Month-wise rental by Customers



Month-wise rental by Subscribers



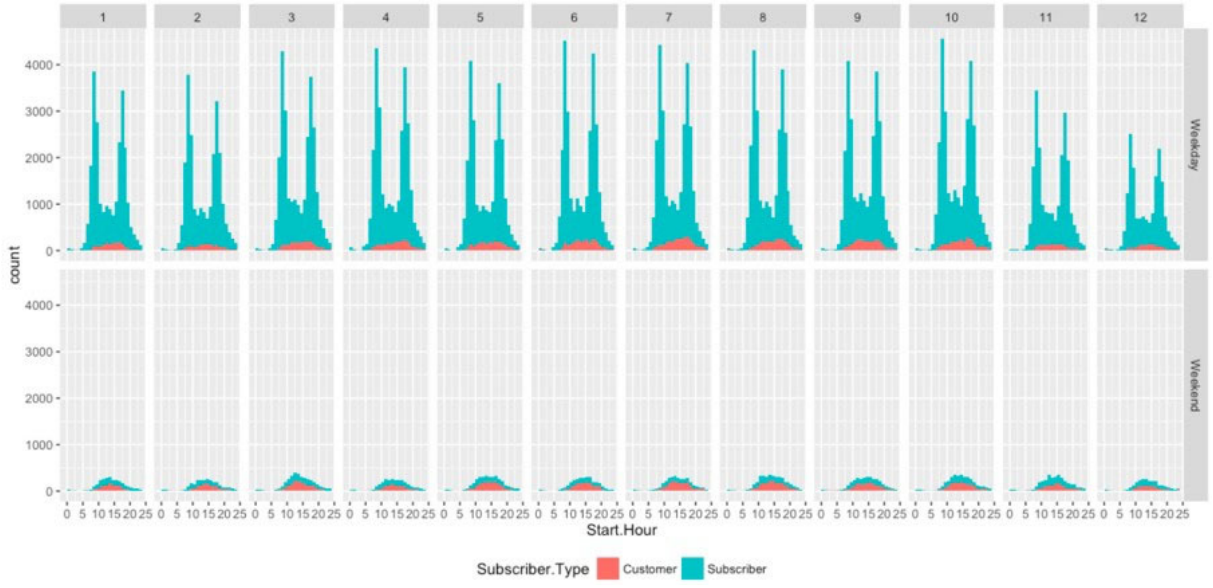
The above data describes the number of trips between subscriber and customer over the months. In December 2014 there is low number of customers and in the same December 2014 there is low number of subscribers.

TECHNICAL APPROACH: TOOLS AND ALGORITHMS

For doing any kind of analysis on this huge dataset, our approach was to first figure out what the usage pattern is for these bikes, during what day and time of the month as well as the year was this usage increasing and exactly which stations were beings used for these trips.

During weekdays, Subscribers use the service with peaks at 8AM and 6PM for commuting purposes. We also found that during weekends, Subscribers and Customers have a very similar usage pattern, suggesting that Subscribers are probably using the service for leisure purposes during weekends.

Month to month usage



The above graph shows that there is a seasonal pattern of the usage month over month. The smallest number of trips are recorded in December and the highest in June. It is interesting that people in October are also very active. This is probably caused by the fact that in Bay Area the weather allows to ride bicycle also in this month too and there are less people on vacations than in Summer months.

Stations

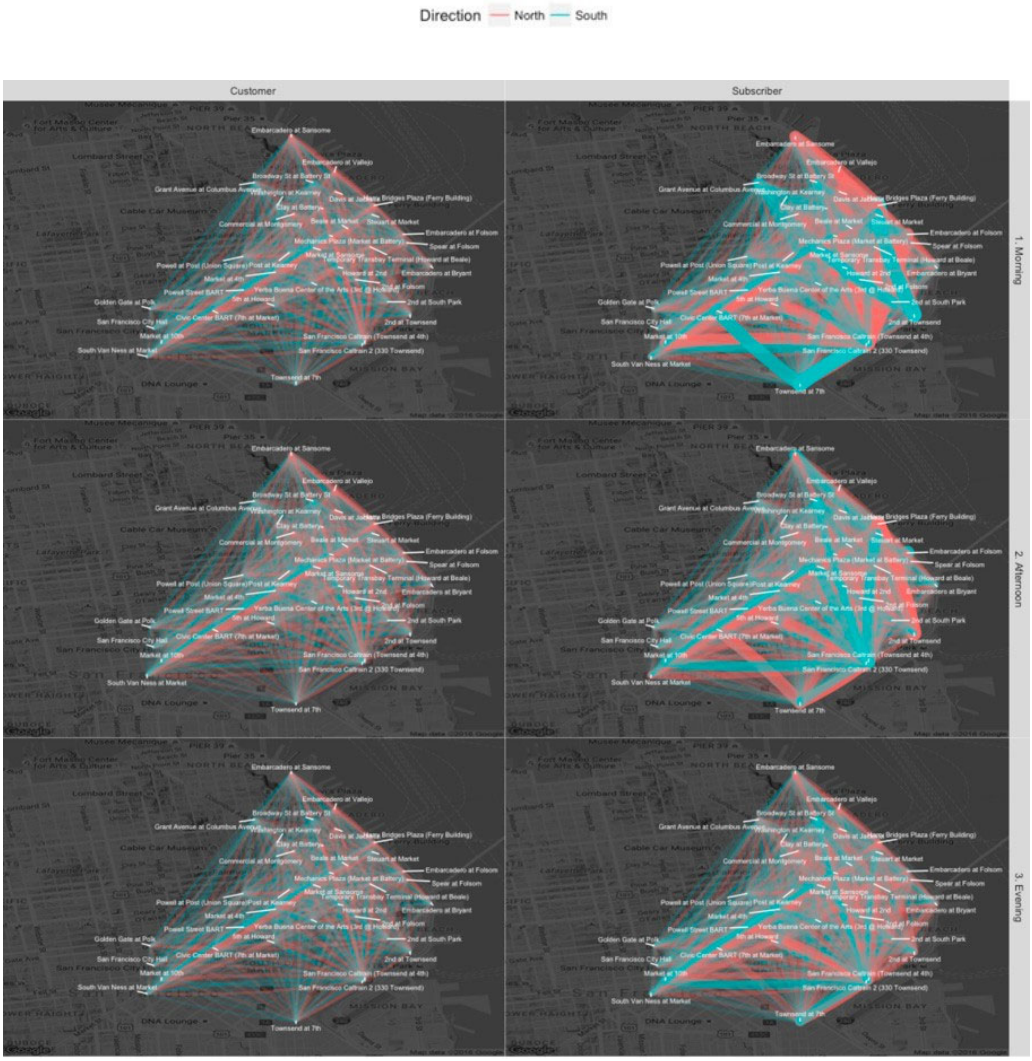
To better understand which stations had most traffic or were most popular among the riders and which route they preferred, we needed to analyze data geographically by using visualization techniques to plot the stations on a google map. We had to do perform quite a bit of research on exactly how we could do this in R, and after rigorous efforts using the google API-key we were finally able to plot these stations and routes on a map using the latitude and longitude measure in the station.csv dataset. Here are how 70 stations are spread in Bay Area (each white dot is an individual station):

Trips Direction

After plotting the stations on the maps, we ran some analysis code to determine the frequency and origin and destination of these trips. During this analysis, we found that 2.9% of trips end at the same station as they started. Out of those 7.15% are immediate changes, when a rider took a bicycle and gave it back in less than 2 minutes (e.g. decided to pick another bicycle for example).

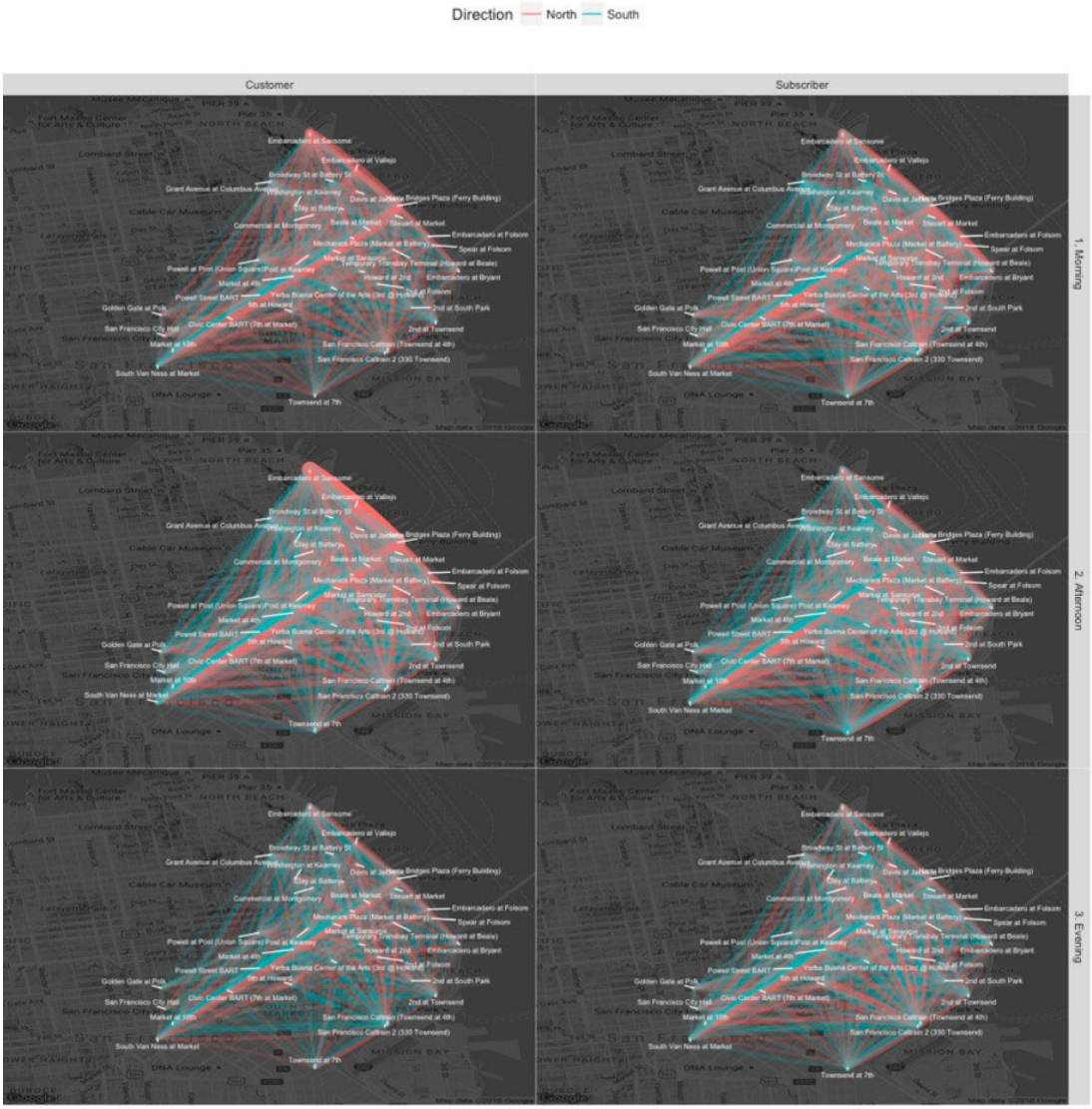
We then analyzed where these users were travelling using shared bicycles at different time of days on Weekdays and Weekends. We plotted maps on a general timeline of morning, afternoon and evening. On the maps below, we only show stations with high traffic were red line shows trips towards North and turquoise lines towards South.

San Francisco - Trips During Weekdays



Customers do not have route priorities depending on the time of the day. In mornings, many Subscribers travel to South towards Caltrain and Townsend 2nd and 7th St. There are also many Subscribers travelling from Caltrain towards Embarcadero. In afternoons, many Subscribers also travel from Downtown to Caltrain and from 2nd and Townsend to Ferry Building. In evenings Subscribers do not have such distinct routes apart from trips towards the south of Market St.

San Francisco - Trips During Weekends



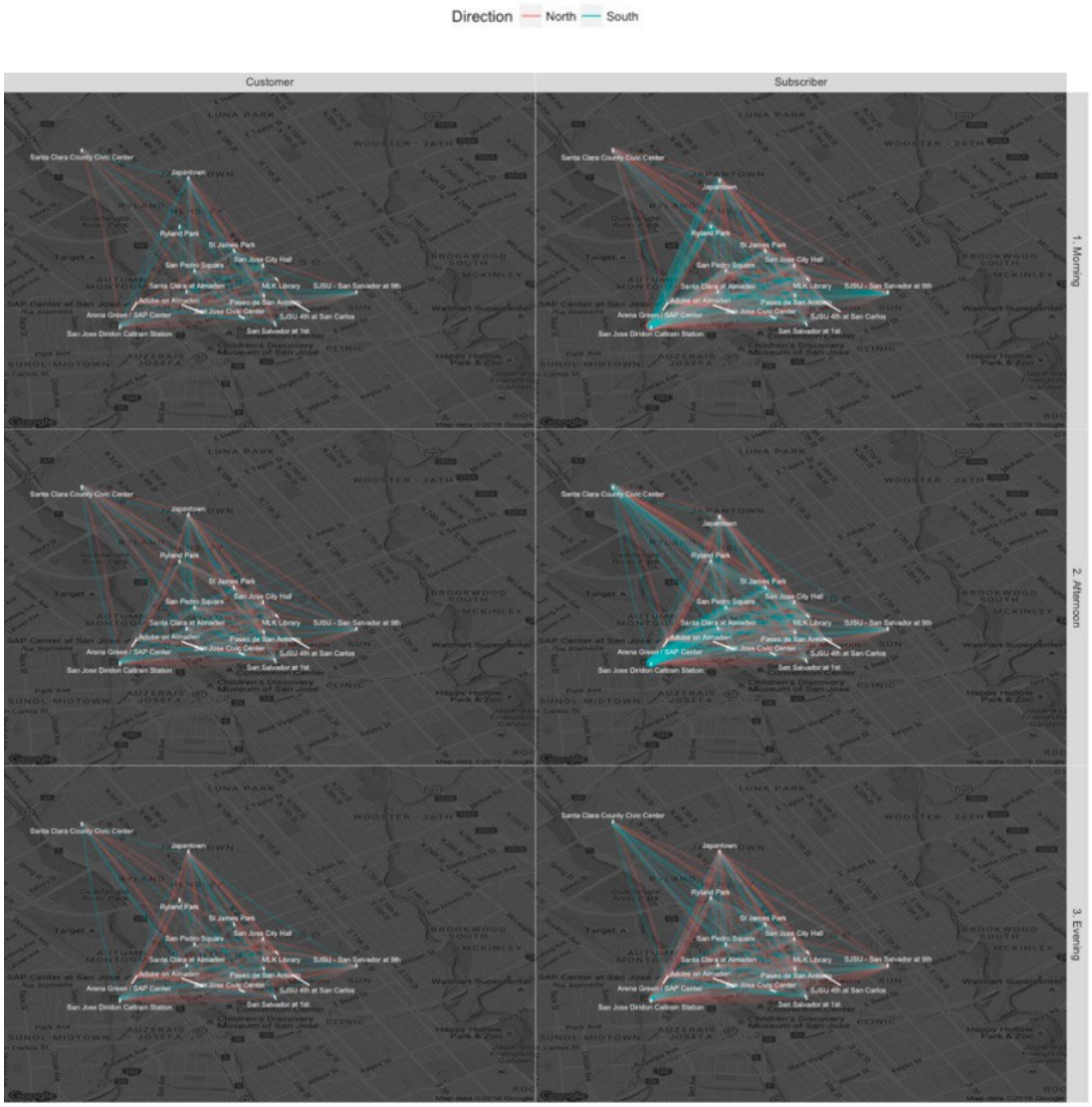
During weekends, the route preferences of Customers and Subscribers are similar (Market St and Embarcadero) providing an extra support for our hypothesis that Subscribers tend to use the service during weekends for leisure purposes.

Palo Alto - Trips During Weekdays

There is a clear pattern that Subscribers go in mornings to San Antonio Shopping Center from Caltrain Station and come back in afternoons. The Same in Mountain View - Subscribers go to Castro street in mornings and come back in afternoons.

San Jose - Trips During Weekdays

Many Subscribers go to San Jose Caltrain Station in mornings and come back in evenings.



FINDINGS

Underutilized Stations

From the analysis, we were able to find stations with high and less usage for bike rides.

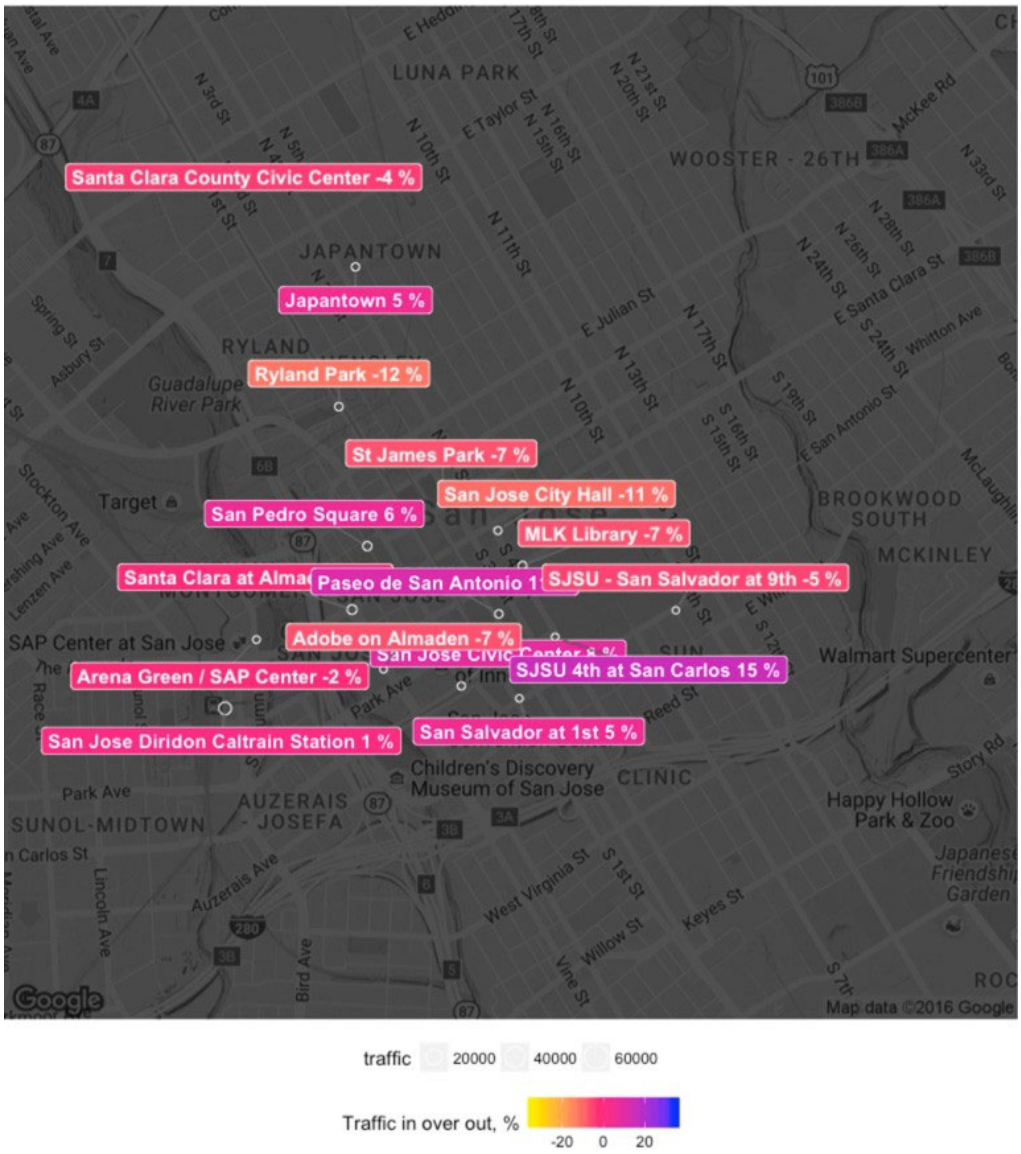
Blue/Purple are the stations which tend to have more bikes arriving than departing (up to 21%).

Yellow are those stations that tend to have more bikes departing than arriving (up to 32%).

Stations in San Francisco

Stations in Palo Alto, Redwood City, Mountain View

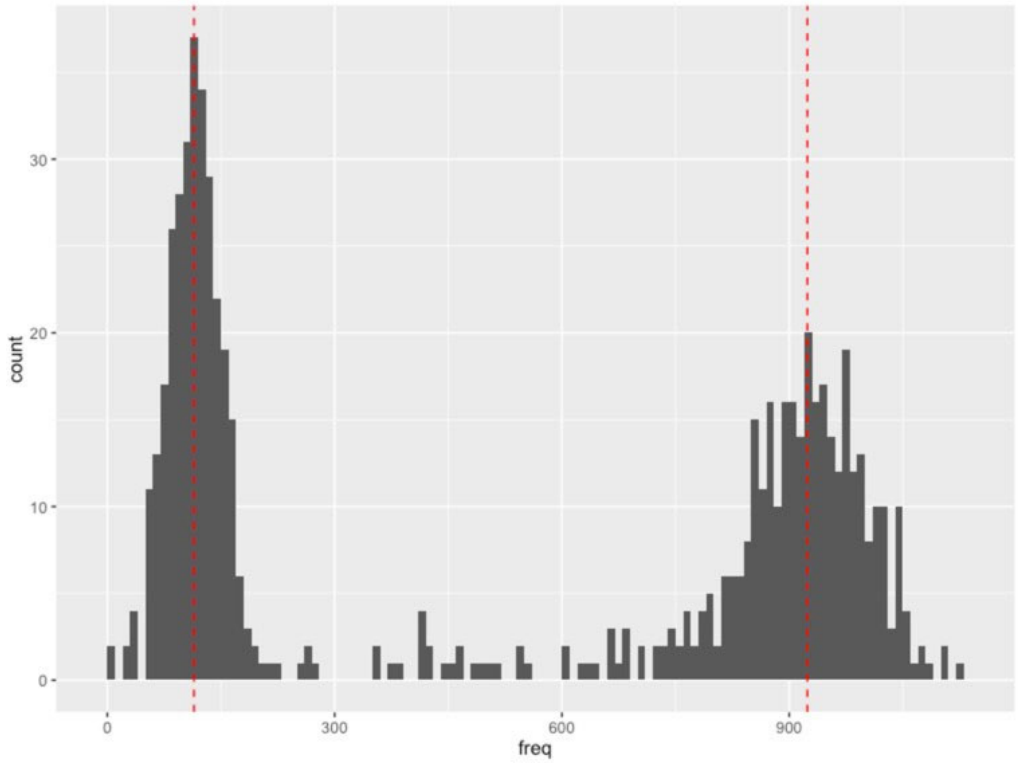
Stations in San Jose



traffic 20000 40000 60000

Traffic in over out, % -20 0 20

Bicycle Usage



Based on our analysis for usage of bikes on their bike ID numbers, half of the bicycles were used in average 114 times and another half 924 times. In the ideal case, if we can transfer bicycles and all bicycles are used equally often, then each bicycle would be used 530 times.

Recommendations

Based on our findings, we recommend the below transfers to make the distribution of bicycle usage a more uniform or normal rather than bimodal. To do it we believe that bicycles which were extensively used in areas with high traffic should be moved to stations with low traffic, while bicycles which are almost new should be moved from stations with low traffic to stations with high traffic. Moving bicycles is also a cost so we believe that the right way to do this transfer is to do it along the regularbicycle transfer caused by imbalanced stations usage.

Based on the trip's users did in the last day (based on the current dataset) we suggest transferring bicycles based on the following recommendations. These recommendations are balanced (the total number of bicycles to take off is equal to the total number to bring). The number of heavily-and used few times might be not balanced, but they are more priorities than an action order.

Even though bikes are moderately used in many US urban areas the usage from the bay area bikeshare have been increasing to huge amount. The accessibility thorough Bikeshare datasets gives aremarkable chance to examine urban cycling.

```
kable(recommendations[,c('Terminal','Station','Recommendation')])
```

Terminal	Station	Recommendation
10	San Jose City Hall	Bring 1 heavily used bikes
11	MLK Library	Bring 2 heavily used bikes
16	SJSU - San Salvador at 9th	Take off: 181
2	San Jose Diridon Caltrain Station	Take off: 213, 165, 663
22	Redwood City Caltrain Station	Bring 1 heavily used bikes
25	Stanford in Redwood City	Take off: 126, 196
26	Redwood City Medical Center	Bring 1 heavily used bikes
27	Mountain View City Hall	Take off: 35, 139
29	San Antonio Caltrain Station	Take off: 24
30	Evelyn Park and Ride	Bring 1 heavily used bikes
31	San Antonio Shopping Center	Bring 1 heavily used bikes
34	Palo Alto Caltrain Station	Bring 3 heavily used bikes
37	Cowper at University	Take off: 140, 230
39	Powell Street BART	Take off: 464, 423
4	Santa Clara at Almaden	Bring 2 heavily used bikes
41	Clay at Battery	Take off: 445
42	Davis at Jackson	Take off: 569, 86
45	Commercial at Montgomery	Bring 4 bikes used few times
46	Washington at Kearney	Take off: 395, 451, 290, 547
47	Post at Kearney	Take off: 523
48	Embarcadero at Vallejo	Take off: 491, 504
49	Spear at Folsom	Bring 8 bikes used few times
5	Adobe on Almaden	Take off: 714
50	Harry Bridges Plaza (Ferry Building)	Take off: 366, 609, 292, 419, 404, 583, 620
51	Embarcadero at Folsom	Bring 7 bikes used few times
55	Temporary Transbay Terminal (Howard at Beale)	Bring 2 bikes used few times
56	Beale at Market	Bring 13 bikes used few times
57	5th at Howard	Bring 11 bikes used few times
58	San Francisco City Hall	Bring 4 heavily used bikes
59	Golden Gate at Polk	Bring 3 heavily used bikes
6	San Pedro Square	Take off: 125, 163
60	Embarcadero at Sansome	Take off: 409
61	2nd at Townsend	Take off: 463, 66
62	2nd at Folsom	Bring 16 bikes used few times
63	Howard at 2nd	Bring 1 bikes used few times

64	2nd at South Park	Bring 4 bikes used few times
66	South Van Ness at Market	Bring 5 bikes used few times
67	Market at 10th	Bring 1 bikes used few times
68	Yerba Buena Center of the Arts (3rd @ Howard)	Bring 2 bikes used few times
69	San Francisco Caltrain 2 (330 Townsend)	Take off: 878, 334, 507, 16, 137, 278, 465, 602, 268, 516, 549, 214, 508, 526, 390, 222, 525, 614, 403, 594, 611, 353, 517
70	San Francisco Caltrain (Townsend at 4th)	Take off: 532, 29, 540, 310, 328, 327, 416, 67, 342, 556, 459, 500, 484, 158, 548, 372, 575, 597, 360, 422, 579, 531, 371, 432, 413, 709, 441, 427, 109, 274, 288, 538, 336, 619, 559, 495, 629, 635, 387, 535, 187, 637
71	Powell at Post (Union Square)	Bring 7 heavily used bikes
72	Civic Center BART (7th at Market)	Take off: 326, 370
73	Grant Avenue at Columbus Avenue	Bring 10 bikes used few times
74	Steuart at Market	Take off: 418
75	Mechanics Plaza (Market at Battery)	Take off: 512, 581, 462, 325
76	Market at 4th	Bring 8 bikes used few times
77	Market at Sansome	Take off: 322, 592, 361, 622, 189, 375, 434, 458, 563, 567, 510
8	San Salvador at 1st	Take off: 130
82	Broadway St at Battery St	Bring 3 bikes used few times
84	Ryland Park	Bring 2 heavily used bikes
9	Japantown	Bring 1 heavily used bikes

Limitation and Future Work

Some of the limitations associated with this dataset included the limited information and attributes availability. We tried to perform linear regression, but it was not very useful on this dataset and we did not feel like it added much value to our analysis. Probably a different type of dataset with more attribute information on the subscriber, age, gender, and occupation would provide more information on who the customers are. A time series model will also be valuable to add to the analysis with more years' data. There are many more transportation options coming up such as uber/lyft and scooter-sharing which will influence the count of bikes rent. There are still many important factors that would affect the count of bikes rent that have not been included into our models.

References

1. "OPEN DATA" retrieved from https://github.com/udacity/data-analyst/find/master

2. Shaheen. S, Christensen. M, Lima. I (2015) "Bay Area Bike Share Casual Users Survey Report: A Comparative Analysis of Existing and Potential Bike sharing Users".

3. Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg.