Public Bike Usage v. The Weather

MIDS-INFO-W18 Section 3
Project 2: Data Analysis
Presented: August 11th, 2016
By: Muying Chen, Shangjun (Jenny) Jiang, Bryan Chung

The Purpose:
The purpose of this project is to conduct an open preliminary review and analysis of accumulated data sourced directly from a local bicycle rental service called Bay Area Bike Share and find correlations between bike share activity and weather conditions using the data.

The Focus:
Our project will focus on four distinctive questions posed by our initial project proposal:

1. How does weather affect the frequency of bike share usage among its users?
2. How do seasons and seasonal patterns affect the frequency of bike share activity?
3. How does geographic location affect the frequency of bike share activity?
4. How does bike share activity compare among long term Subscribers vs. Customers?

The Context:
The Bay Area Bike Share, referred to now as BABS, is a green public self transportation system supporting a 700+ bicycle rental service, encouraging local users to participate in bike share usage. BABS maintains this bike “sharing” system over 70 bike share stations, the majority operating out of SF. BABS is especially popular with the SF locals and those who have a permanent resident in the Bay Area.

The Source:

- BABS Year 1: Aug 2013 - Aug 2014
- BABS Year 2: Sept 2014 - Aug 2015
- Dropbox Link: Combines All Files (alternative to the above)

The Raw Data Files
Raw data is accessible in .csv format. Four primary datasets each in three .csv files were identified to represent four unique aspects of the BABS data:

DATASET	REPRESENTATION	# FILES
station_data.csv	bike station attributes	3
status_data.csv	status of available bikes & open docks	3
trip_data.csv	trip activity among stations per user	3
weather_data.csv	daily weather figures reported per city	3

BABS dataset files

The Reconciliation

While it should be assumed that even a small amount of data 'scrubbing' is required for any dataset no matter what, the BABS datasets in particular became quite challenging in that each primary dataset came in three separate .csv files:

201402_station_data.csv	6 KB
201402_status_data.csv	607,814 KB
201402_trip_data.csv	16,816 KB
201402_weather_data.csv	79 KB
201408_station_data.csv	6 KB
201408_status_data.csv	637,733 KB
201408_trip_data.csv	20,164 KB
201408_weather_data.csv	79 KB
201508_station_data.csv	6 KB
201508_status_data.csv	1,061,760 ...
201508_trip_data.csv	42,005 KB
201508_weather_data.csv	155 KB

The initial task of reading in data files has now become a task of consolidating all three-part data files into a single dataset (as it should be), however we still ran into issues, almost identical ones, in fact. but with much the opposite reason: The two datasets we were most interested in reviewing were also the two with the least in common, datatype-wise.

We eventually were successful in pulling in data from these .csv files, and into a more singular and organized jupyter notebook window. However, incident proved to be the eye-opener that convinced us we needed to exercise much more review and control over the BABS datasets though we barely made any effort to review the data itself, feeling strongly that fixing control over this would fix the other problems too.

Dataset "Optimization"

A last point is what we ended up referring to as Dataset "Optimization". Mainly, we came to the following conclusion: All four datasets (station, status, trip, weather) held key data for their own dataset: (station hub activity, status changes, bike trips, and of course reported weather vales values per city). A good example was STATION having a date column but not in sequence

It was clear that some datasets would be easier to join if they shared the right column data, such as DateTime. Below are the results of our collective planning to remap new column data (taken from existing datasets) in order to "optimize" the joining process of combining data columns, a significant remapped column being "season" which is derived from datetime.month:

WEATHER
Date
Max_Temperature_F
Mean_Temperature_F
Min_TemperatureF
Max_Dew_Point_F
MeanDew_Point_F
Min_Dewpoint_F
Max_Humidity
Mean_Humidity
Min_Humidity
Max_Sea_Level_Pressure_In
Mean_Sea_Level_Pressure_In
Min_Sea_Level_Pressure_In
Max_Visibility_Miles
Mean_Visibility_Miles
Min_Visibility_Miles
Max_Wind_Speed_MPH
Mean_Wind_Speed_MPH
Max_Gust_Speed_MPH
Precipitation_In
Cloud_Cover
Events
Wind_Dir_Degrees
zip

STATION
station_id
name
lat
long
dockcount
landmark
installation

STATUS
station_id
bikes_available
docks_available
time

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

WEATHER
Date
Season *
Max_Temperature_F
Mean_Temperature_F
Min_TemperatureF
Max_Dew_Point_F
MeanDew_Point_F
Min_Dewpoint_F
Max_Humidity
Mean_Humidity
Min_Humidity
Max_Sea_Level_Pressure_In
Mean_Sea_Level_Pressure_In
Min_Sea_Level_Pressure_In
Max_Visibility_Miles
Mean_Visibility_Miles
Min_Visibility_Miles
Max_Wind_Speed_MPH
Mean_Wind_Speed_MPH
Max_Gust_Speed_MPH
Precipitation_In
Cloud_Cover
Events
Wind_Dir_Degrees
zip

* New shared data column

STATION
station_id
name
lat
long
dockcount
landmark
installation
Season *

STATUS
station_id
bikes_available
docks_available
time
Season *

TRIP
Trip ID
Duration
Start Date
Season *
Start Station
Start Terminal
End Date
End Station
End Terminal
Bike #
Subscription Type
Zip Code

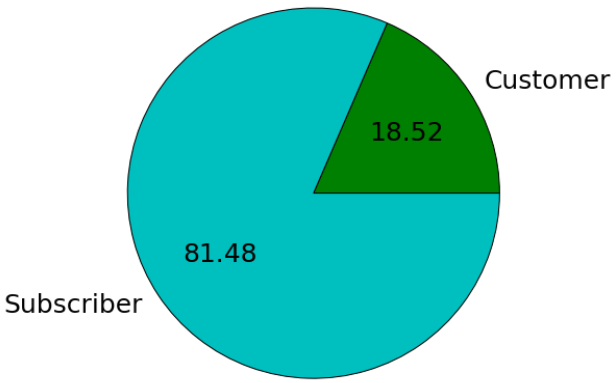
#The Plots

The plotting of key data points with simple data visualizations, such as summary tables, pie charts, and bar graphs, played a significant role in guiding our projected logical assumptions about what the data can tell us, as well as helping to frame more precise end analysis in our conclusions.

Plotting Outline

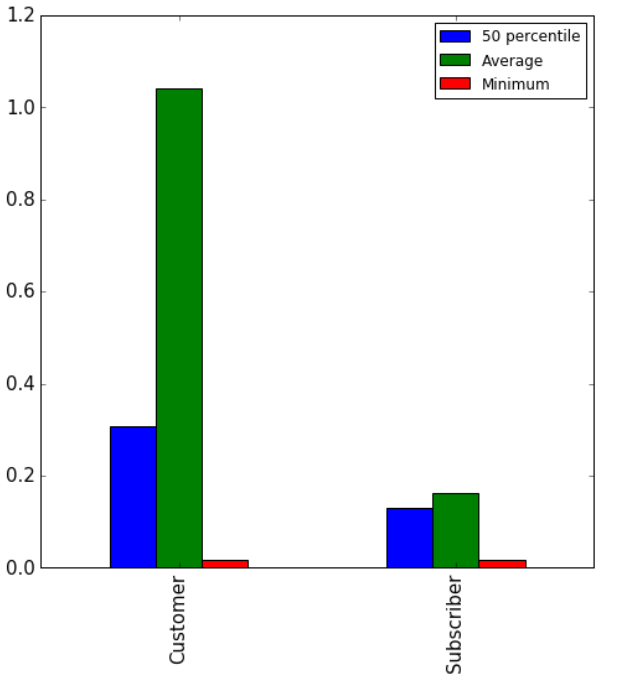
- How many people are using Bike Share across the Bay Area?

	trip_duration
count	313,705.00
mean	19.47
std	105.33
min	1.00
25%	5.77
50%	8.70
75%	12.90
max	12,037.27



Membership Total Count in Percentage 2014

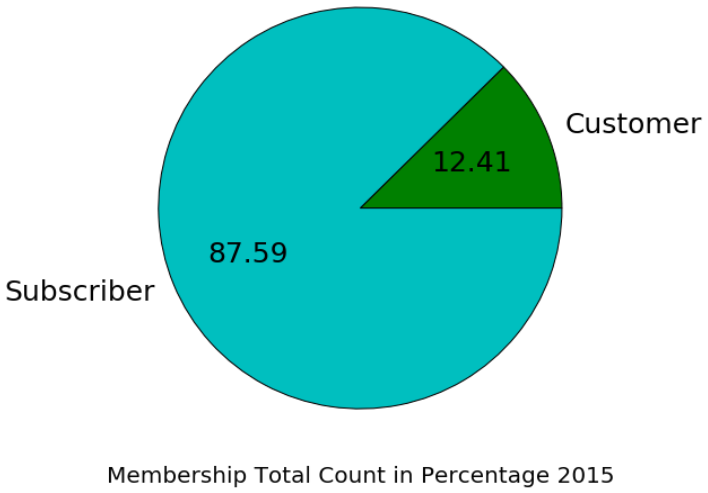
		trip_duration
membership		
Customer	**count**	58,105.00
	mean	62.56
	std	221.02
	min	1.02
	25%	10.87
	50%	18.43
	75%	37.83
	max	12,037.27
Subscriber	**count**	255,600.00
	mean	9.67
	std	44.64
	min	1.00
	25%	5.33
	50%	7.85
	75%	11.02
	max	11,941.33



User Duration of Trip Total Count 2014

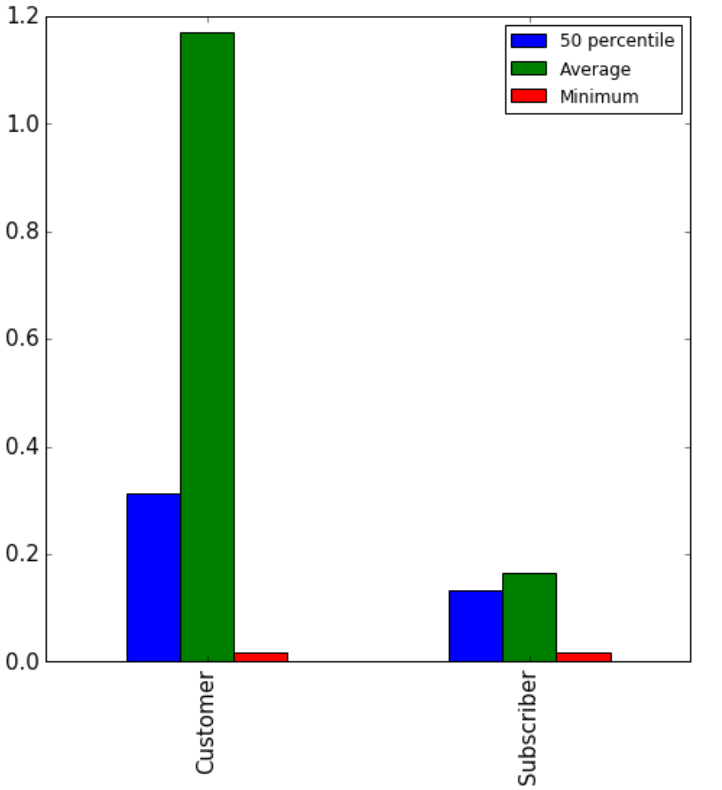
2015

	trip_duration
count	354,152.00
mean	17.43
std	500.28
min	1.00
25%	5.70
50%	8.52
75%	12.32
max	287,840.00



Membership Total Count in Percentage 2015

		trip_duration
membership		
Customer	**count**	43,935.00
	mean	70.24
	std	1,408.30
	min	1.00
	25%	11.12
	50%	18.85
	75%	40.38
	max	287,840.00
Subscriber	**count**	310,217.00
	mean	9.95
	std	66.28
	min	1.00
	25%	5.42
	50%	7.98
	75%	11.15
	max	30,876.50

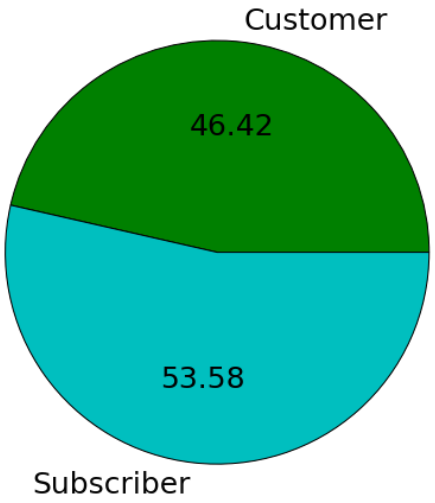


User Duration of Trip Total Count 2015

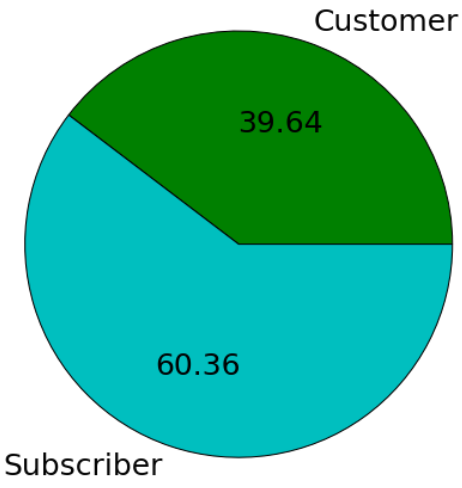
San Francisco 2014 and 2015

	trip_duration
count	3,636.00
mean	72.66
std	354.07
min	1.05
25%	5.56
50%	15.63
75%	44.95
max	12,037.27

	trip_duration
count	3,073.00
mean	70.50
std	492.46
min	1.10
25%	4.80
50%	14.38
75%	33.63
max	18,892.33



Trip Count in San Francisco in Percentage 2014

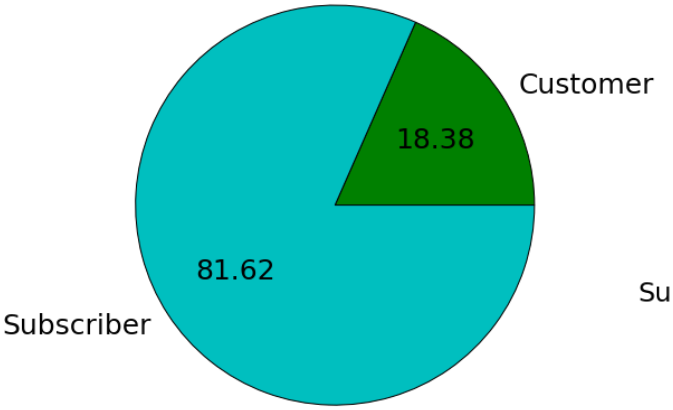


Trip Count in San Francisco in Percentage 2015

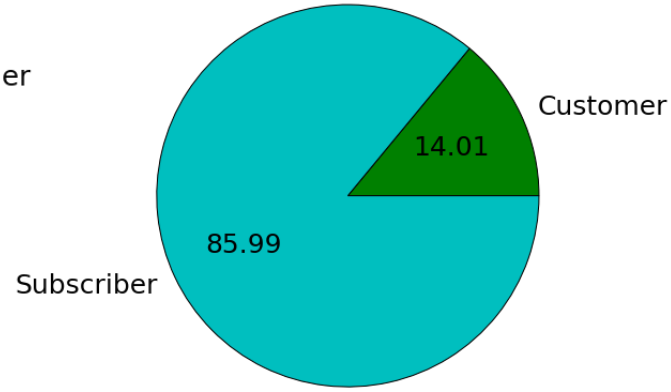
San Jose 2014 and 2015

	trip_duration
count	19,764.00
mean	22.66
std	163.44
min	1.02
25%	4.87
50%	7.57
75%	11.52
max	11,922.32

	trip_duration
count	17,956.00
mean	23.35
std	225.87
min	1.03
25%	5.20
50%	7.77
75%	11.05
max	10,932.32



Trip Count in San Jose in Percentage 2014

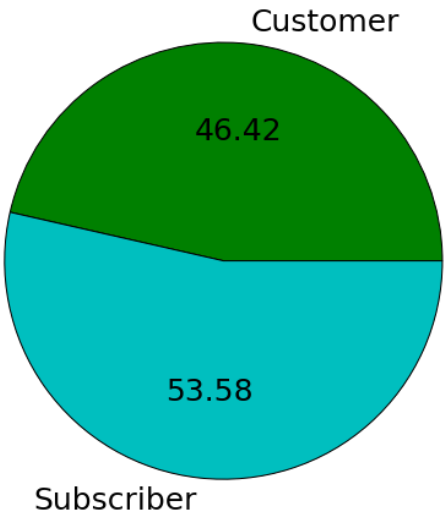


Trip Count in San Jose in Percentage 2015

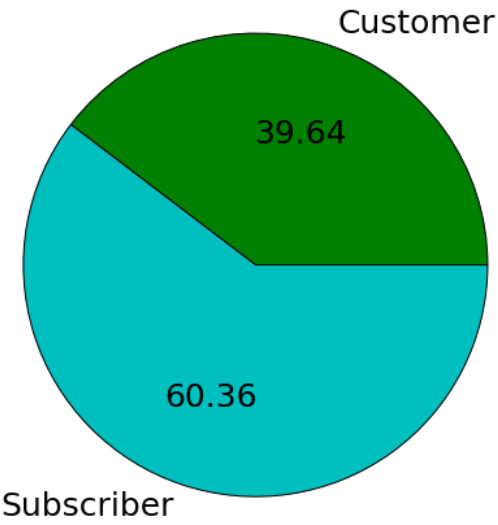
Palo Alto 2014 and 2015

	trip_duration
count	3,636.00
mean	72.66
std	354.07
min	1.05
25%	5.56
50%	15.63
75%	44.95
max	12,037.27

	trip_duration
count	3,073.00
mean	70.50
std	492.46
min	1.10
25%	4.80
50%	14.38
75%	33.63
max	18,892.33



Trip Count in Palo Alto in Percentage 2014

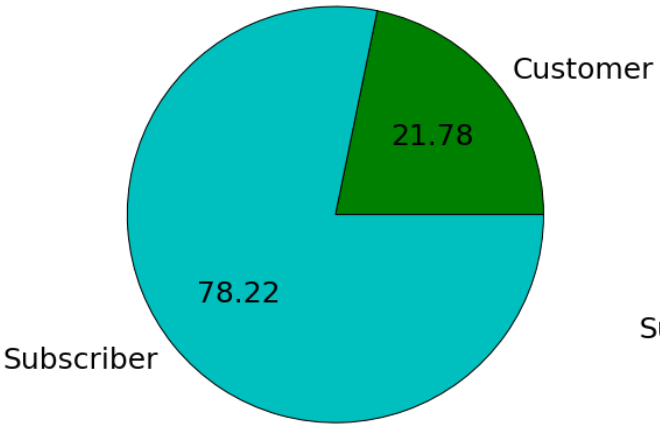


Trip Count in Palo Alto in Percentage 2015

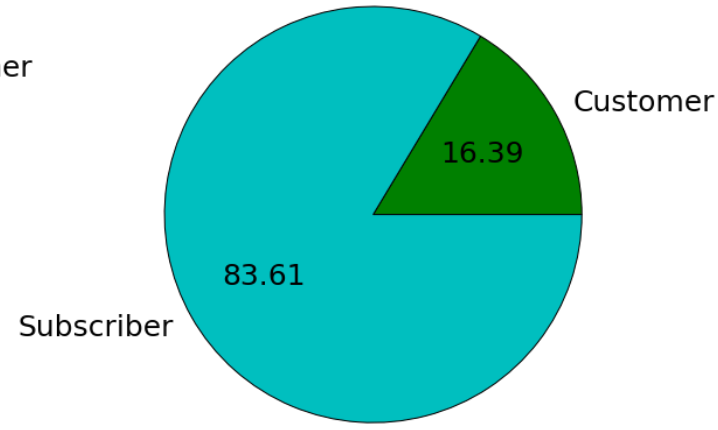
Redwood City 2014 and 2015

	trip_duration
count	1,391.00
mean	44.57
std	208.12
min	1.00
25%	3.82
50%	4.83
75%	8.72
max	3,831.90

	trip_duration
count	2,019.00
mean	38.13
std	350.76
min	1.13
25%	4.56
50%	10.35
75%	14.39
max	12,007.57



Trip Count in Redwood City in Percentage 2014

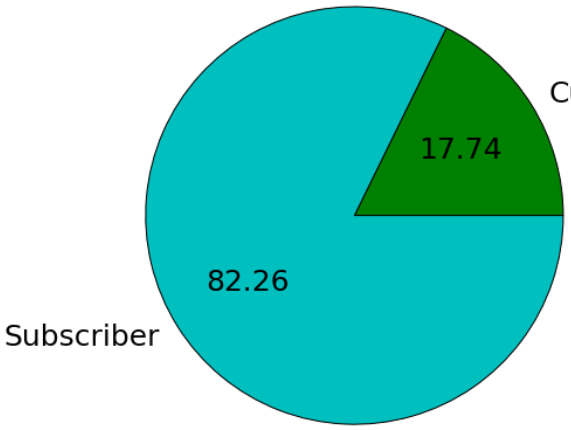


Trip Count in Redwood City in Percentage 2015

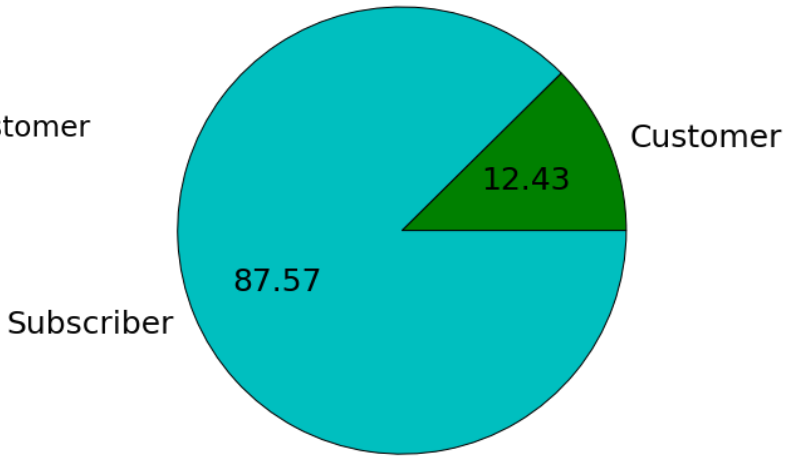
Mountain View 2014 and 2015

	trip_duration
count	1,391.00
mean	44.57
std	208.12
min	1.00
25%	3.82
50%	4.83
75%	8.72
max	3,831.90

	trip_duration
count	9,999.00
mean	23.83
std	338.98
min	1.02
25%	3.97
50%	4.90
75%	7.62
max	30,876.50



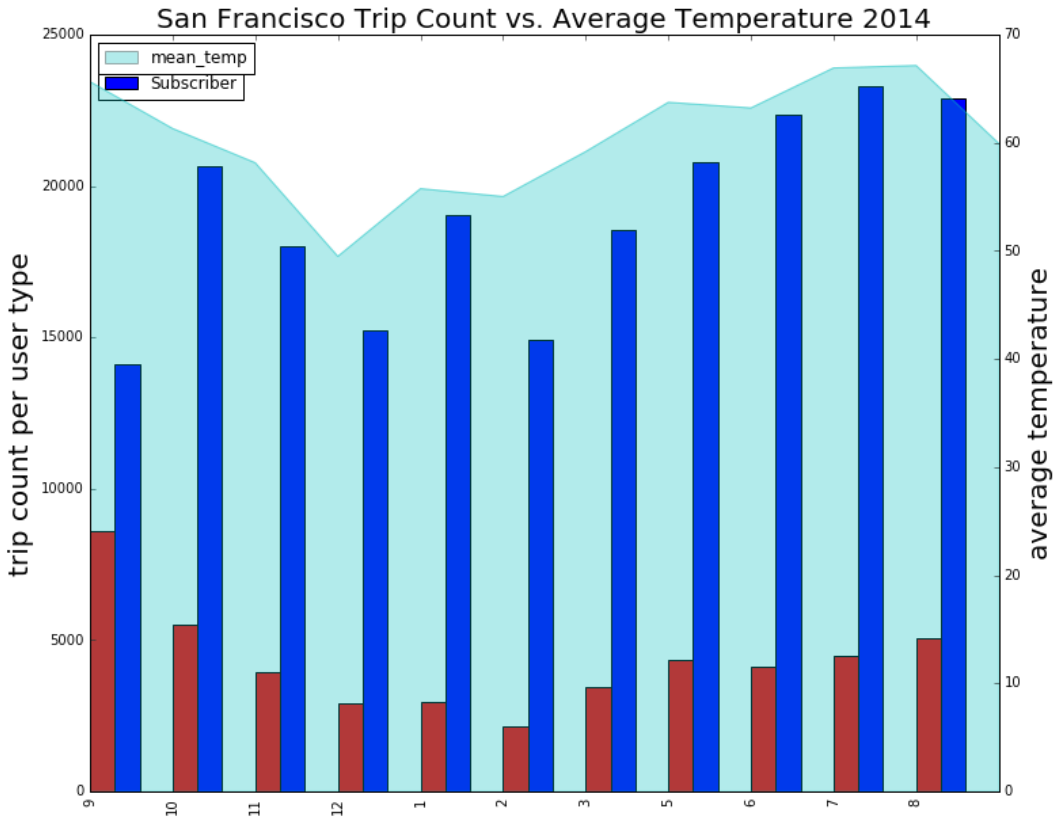
Trip Count in Mountain View in Percentage 2014

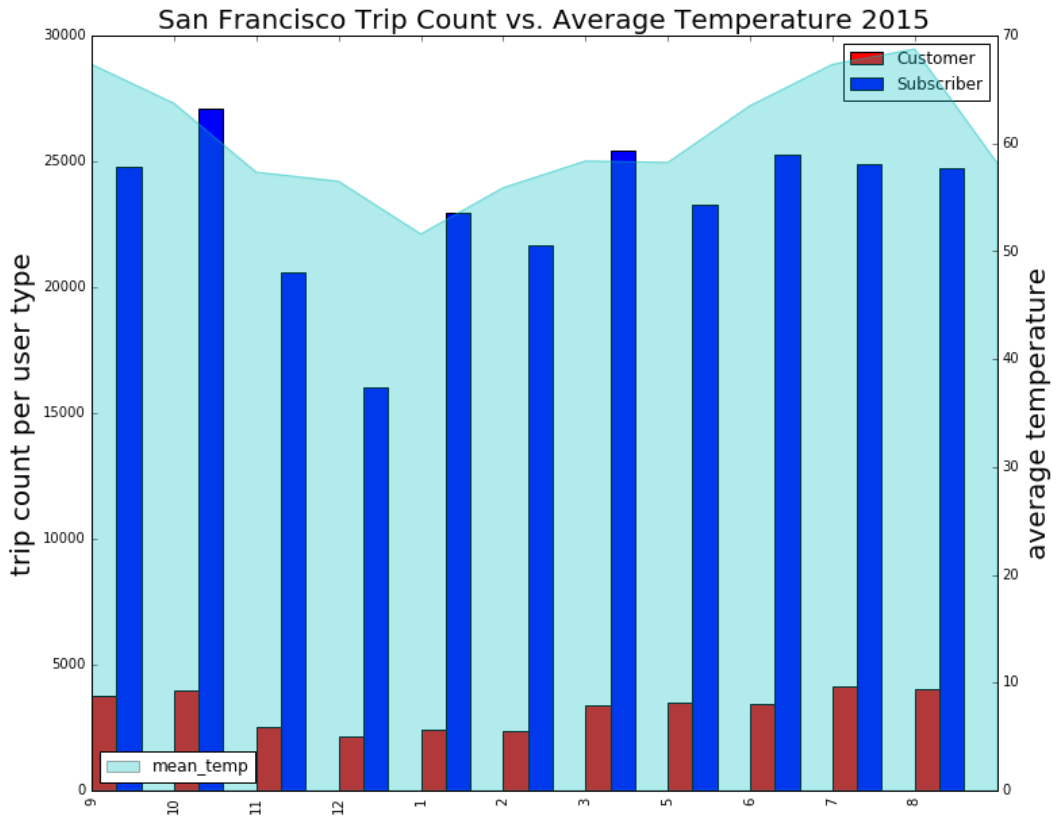


Trip Count in Mountain View in Percentage 2015

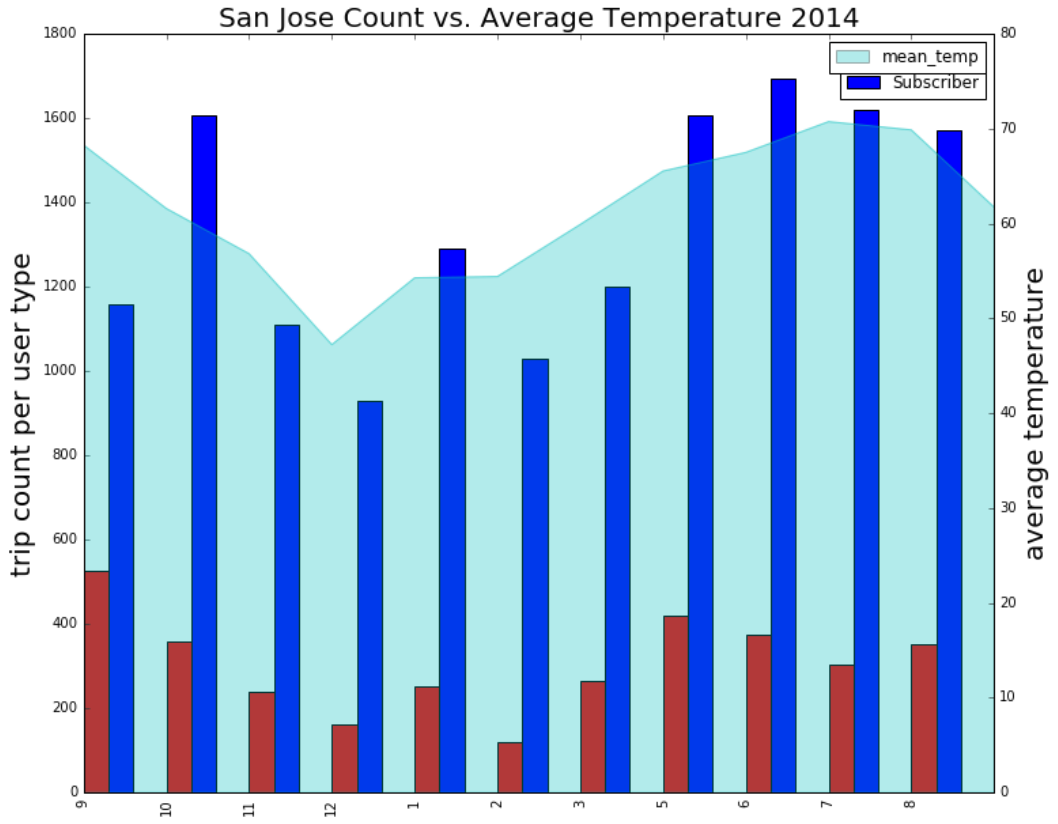
How does weather affect the number of trips for each user in each city?

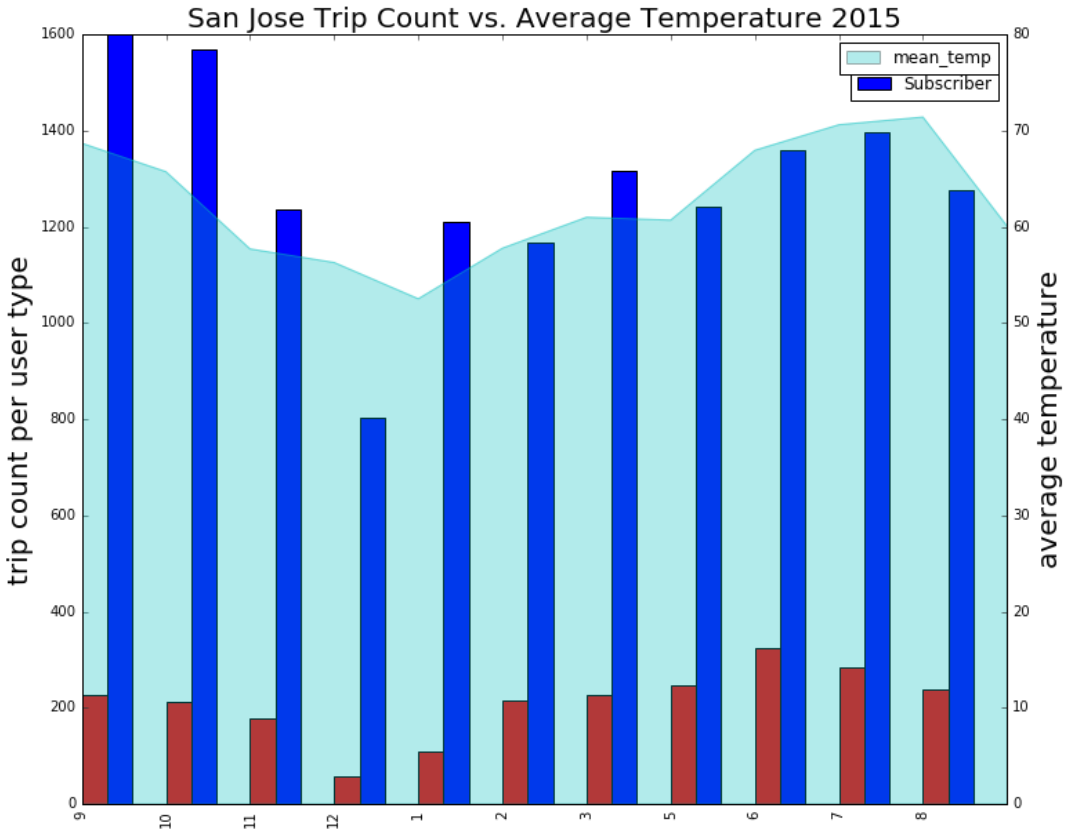
San Francisco 2014 and 2015



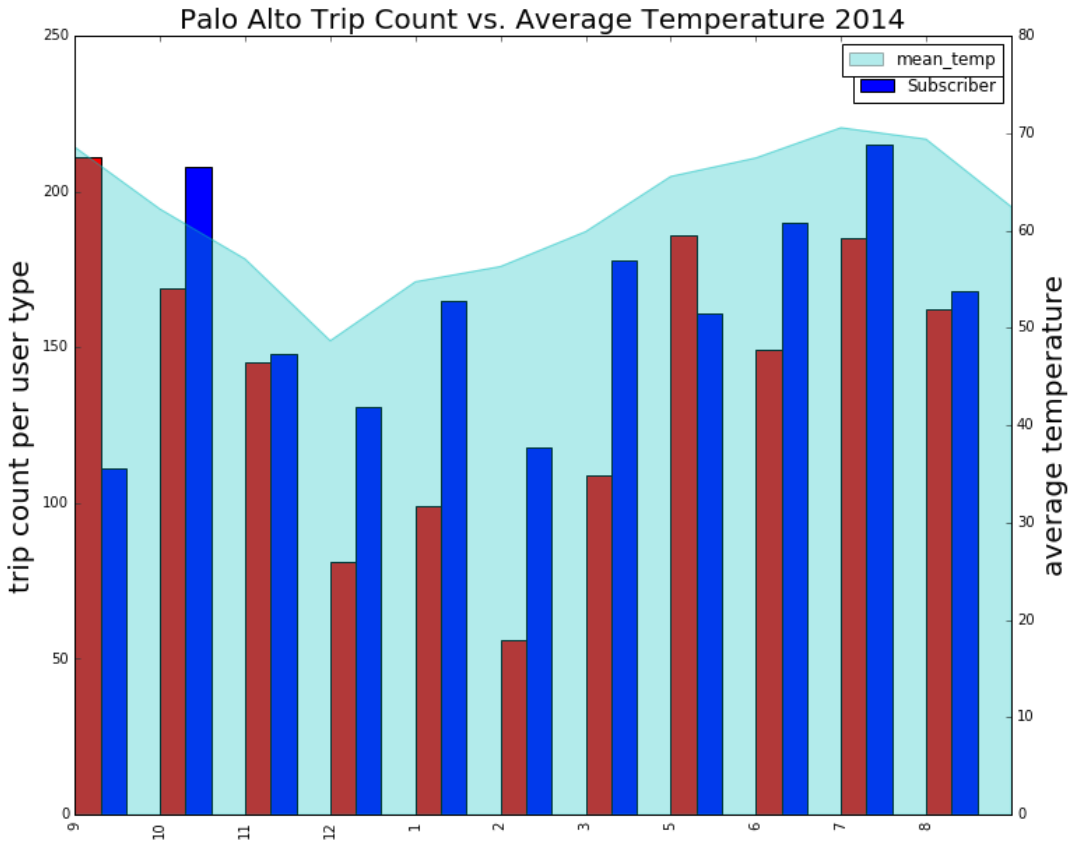


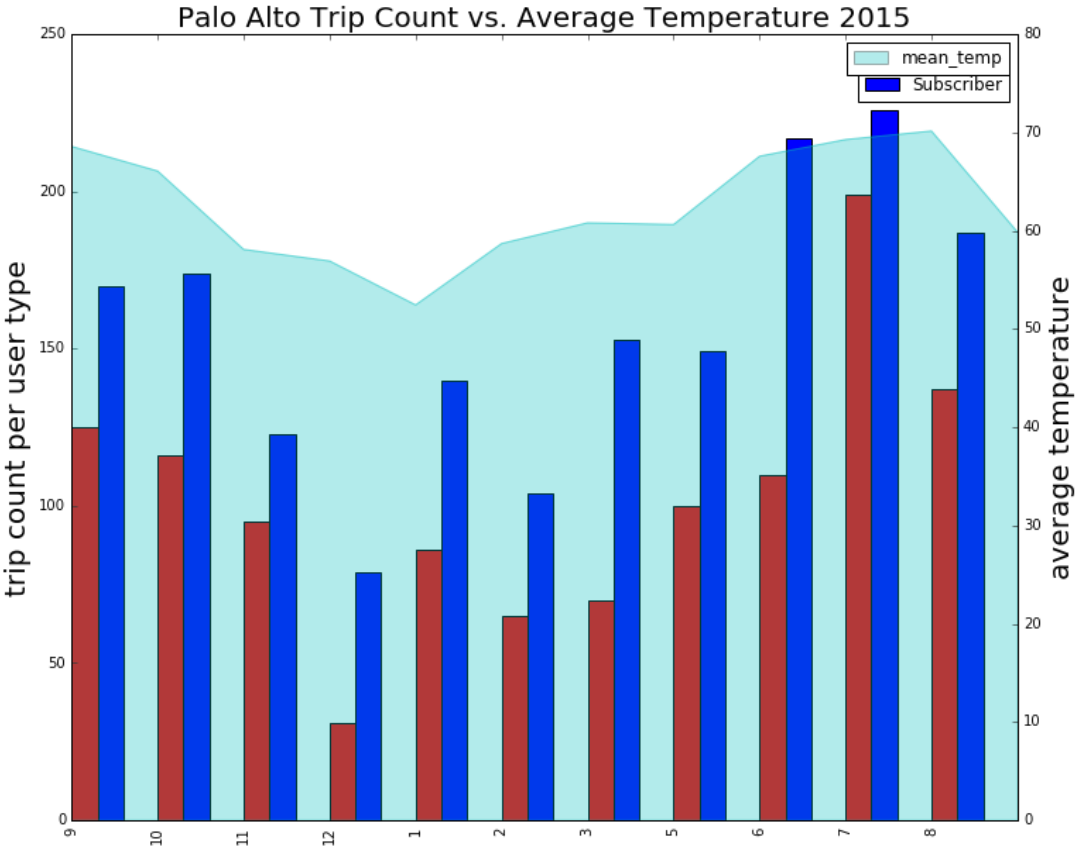
San Jose 2014 and 2015



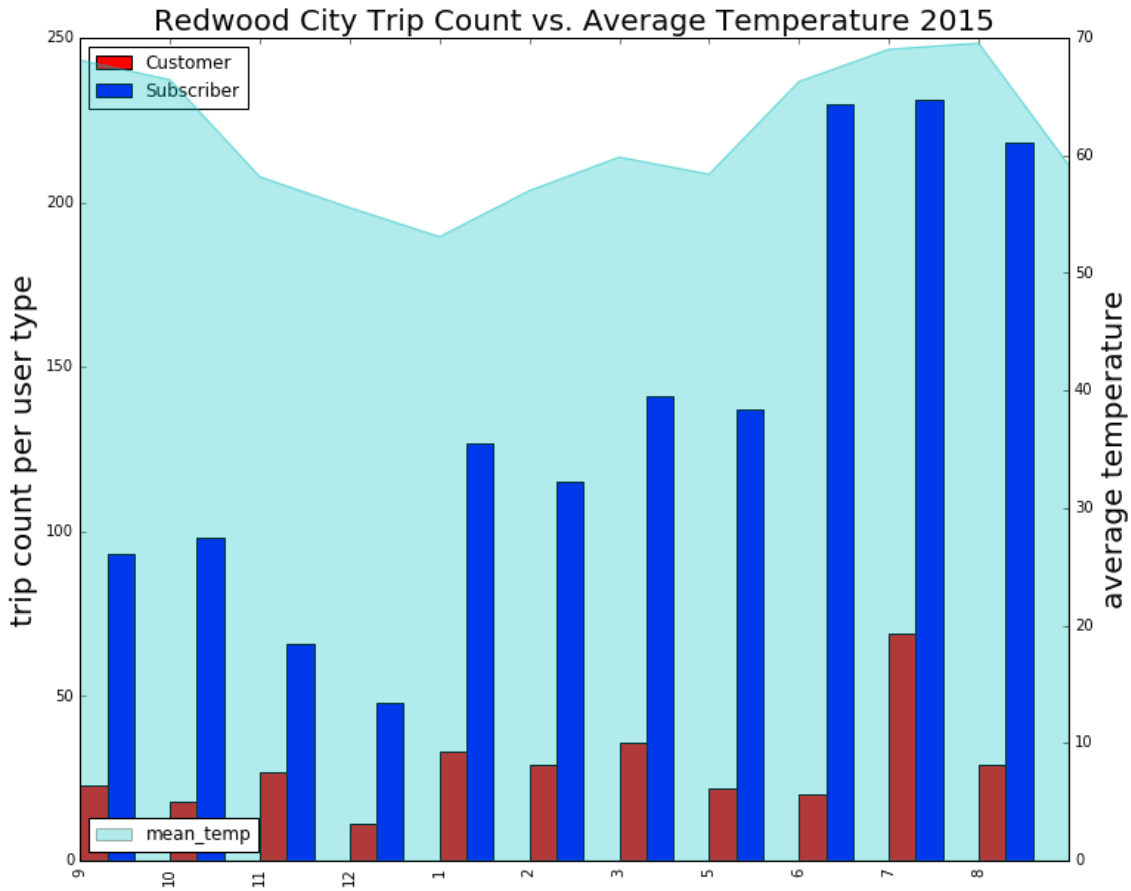


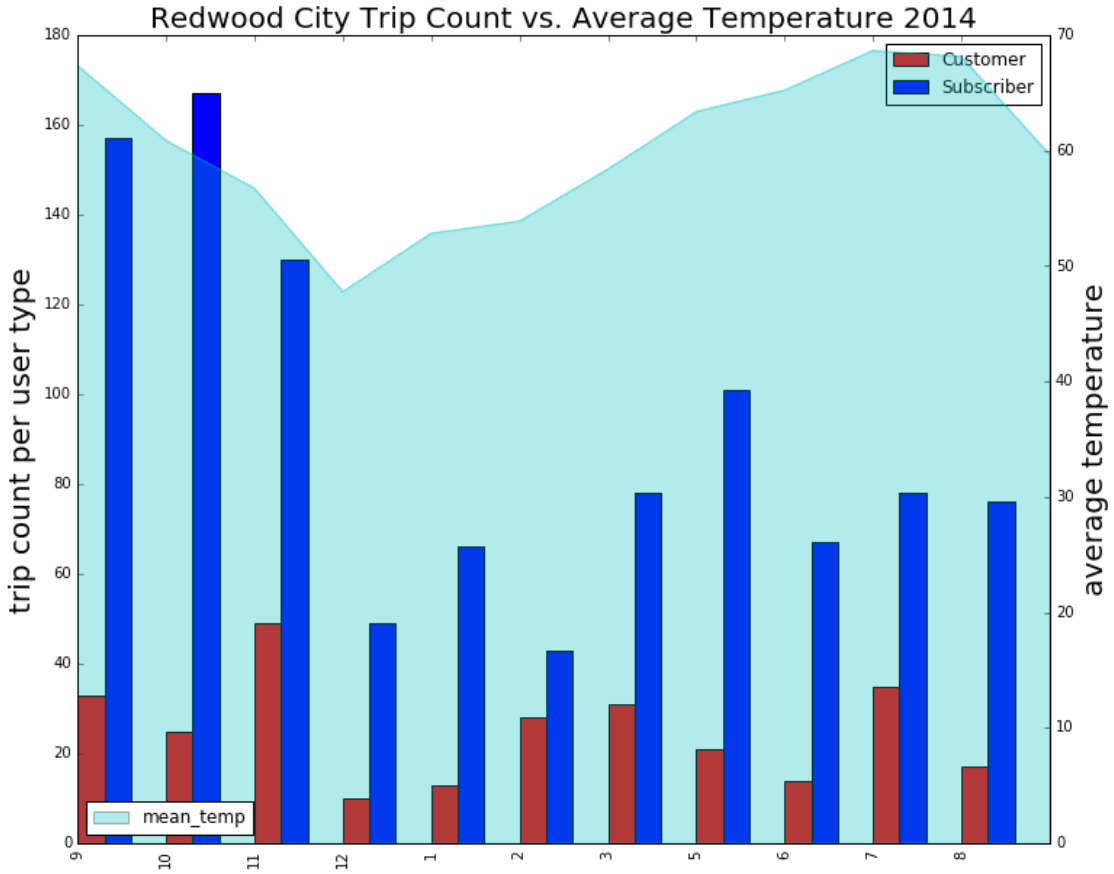
Palo Alto 2014 and 2015



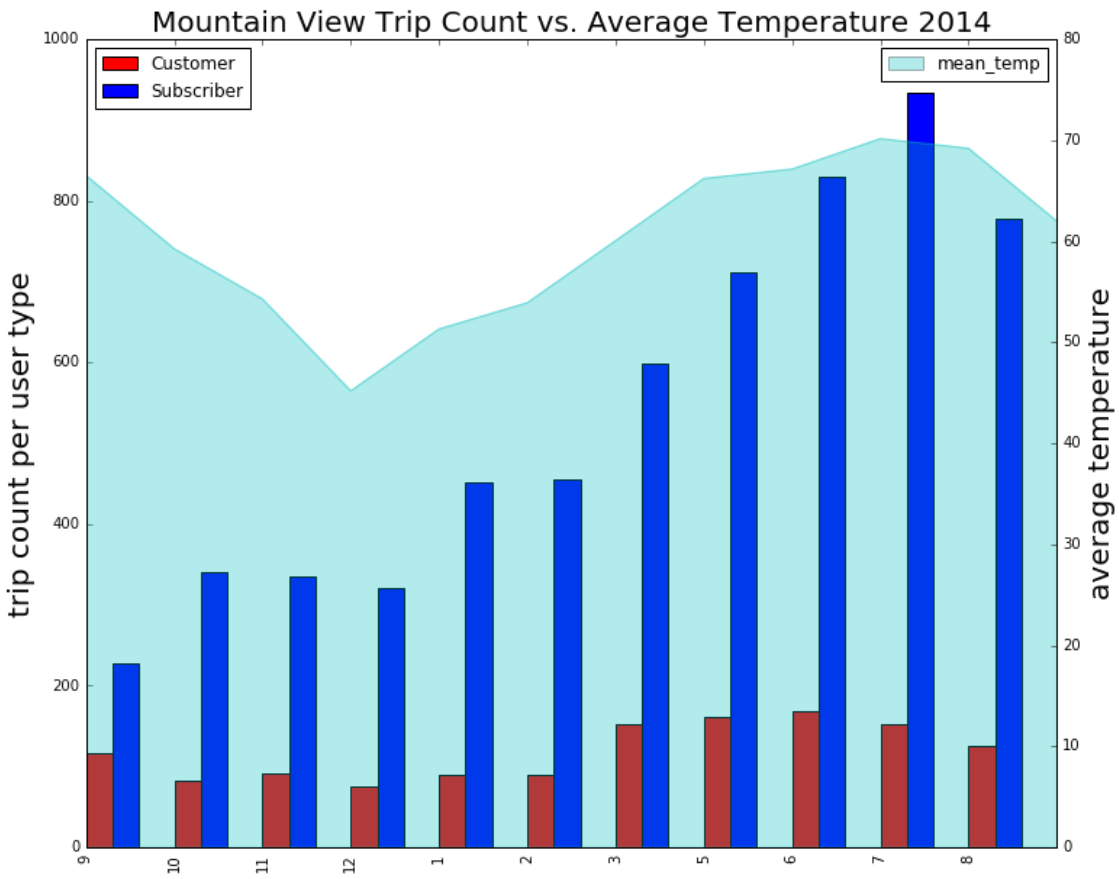


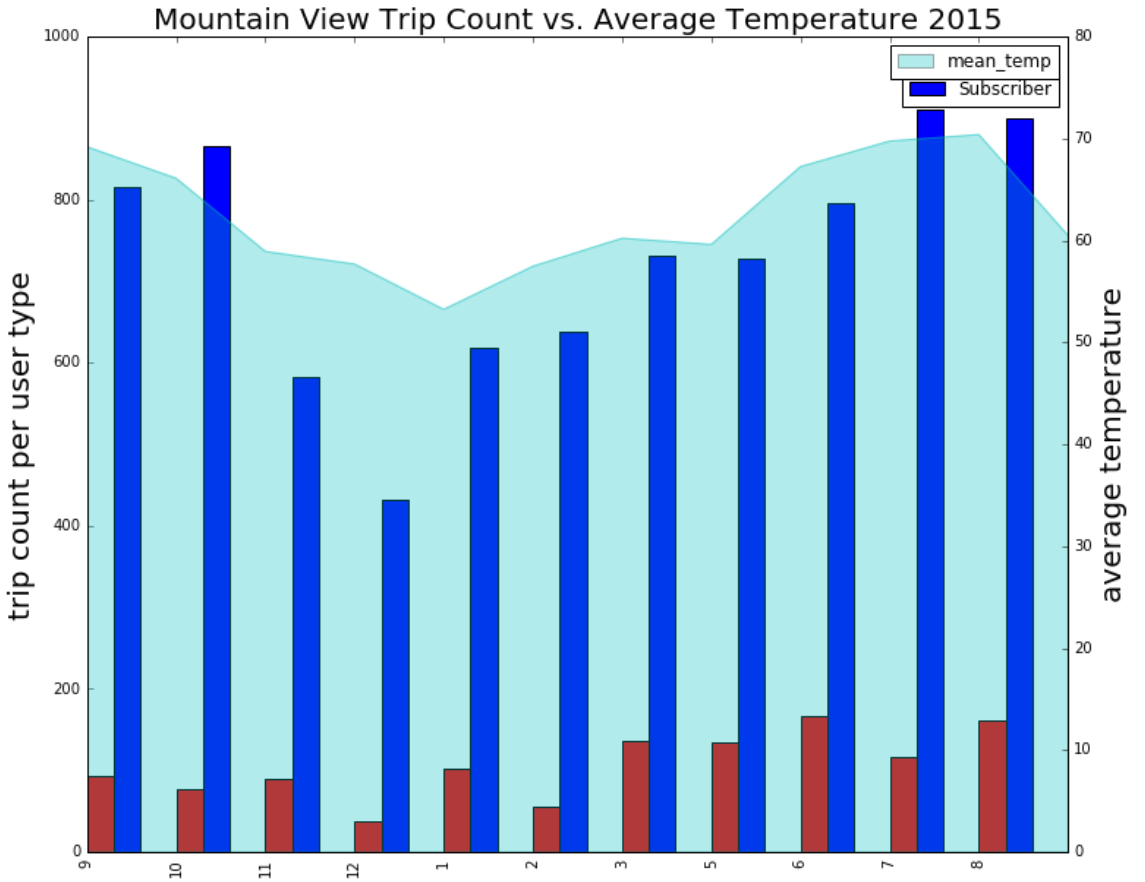
Redwood City 2014 and 2015





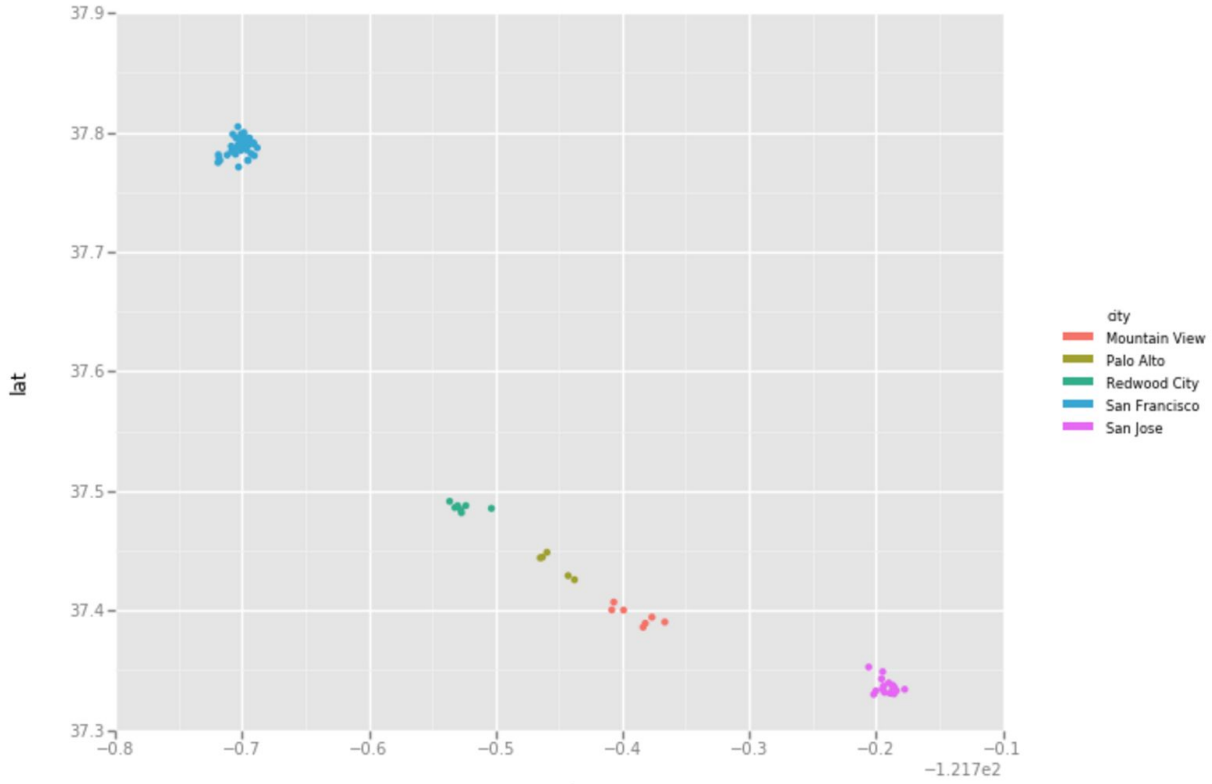
Mountain View 2014 and 2015



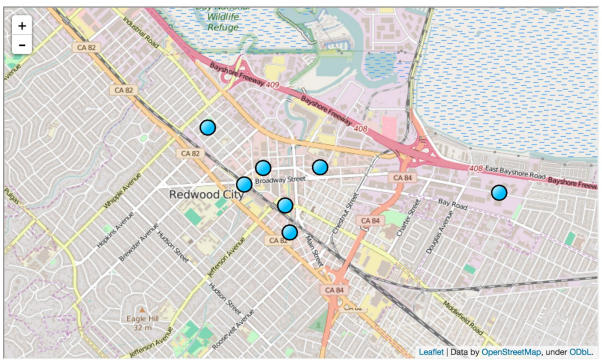


The Advanced Analysis

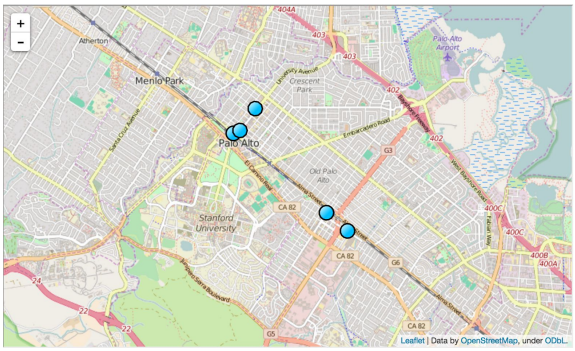
We first looked at the station data by city and we found they are very spread out by cities. So we decided to look at the dataset by each cities. We found a very cool library “folium” that can allow us to create a very detailed map and plot the bike rental places as points.



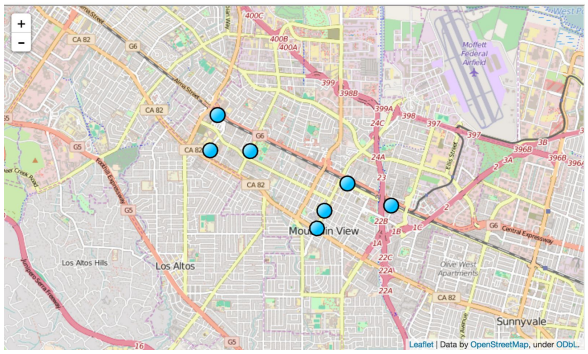
Maps



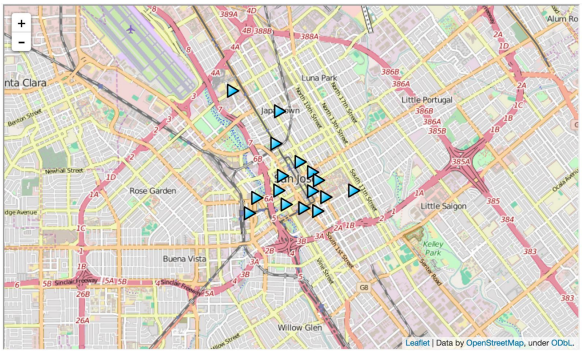
Redwood City



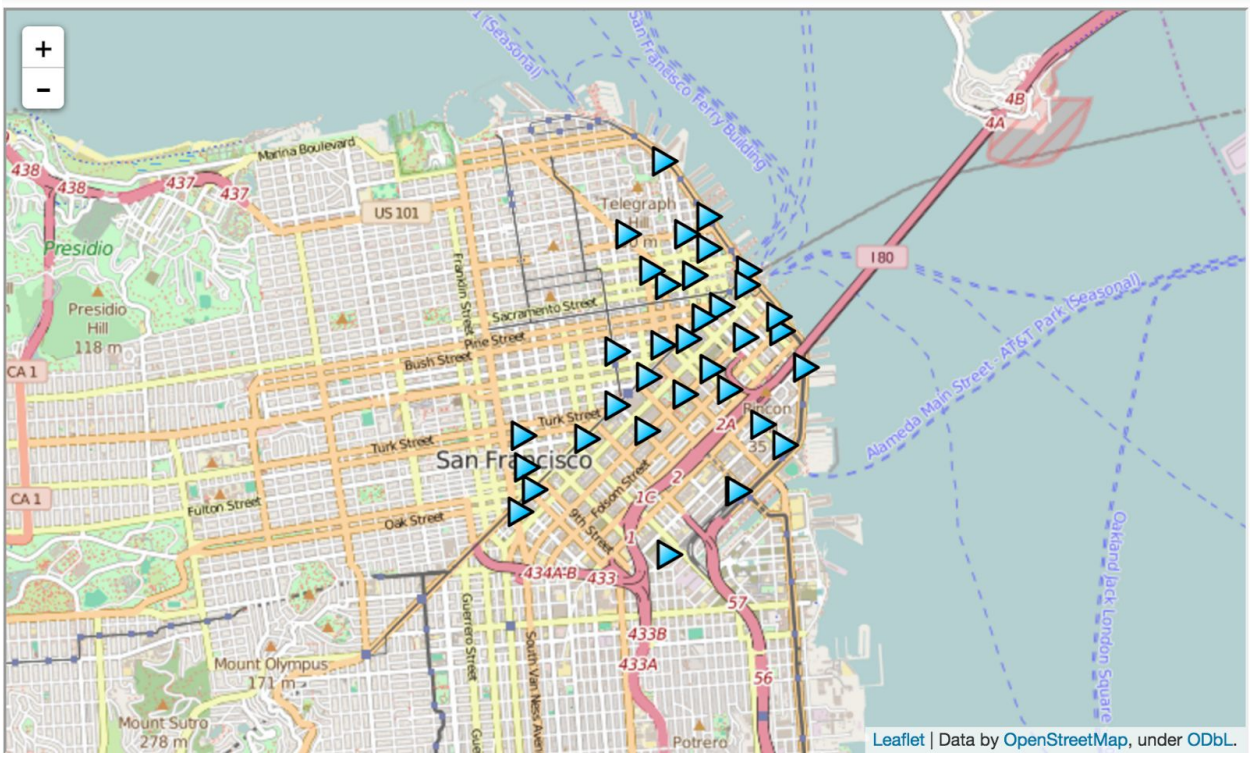
Palo Alto

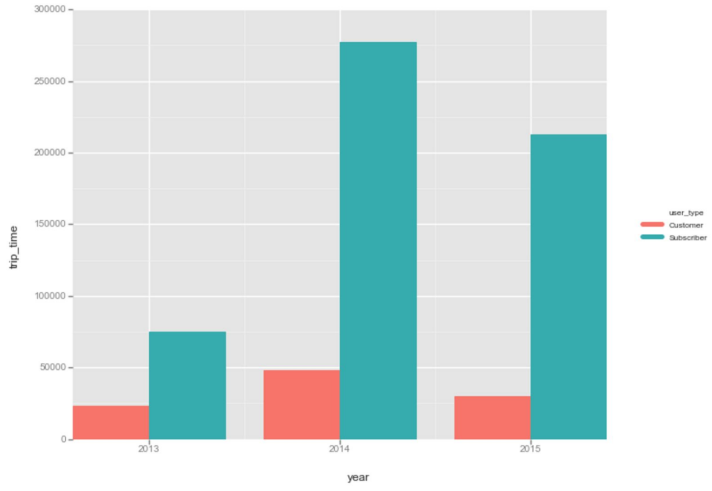


Mountain View



San Jose





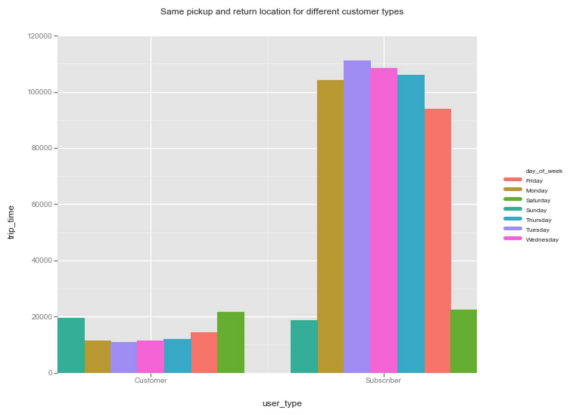
San Francisco

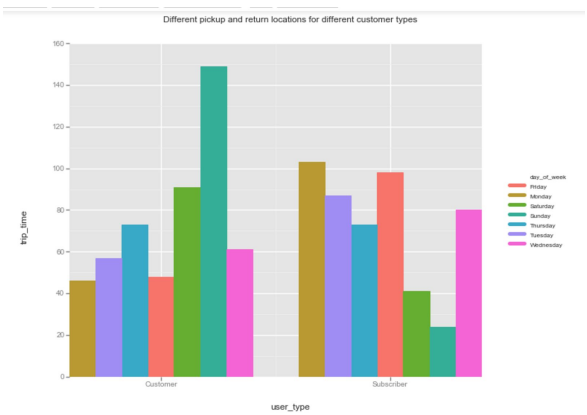
Same as our intuition, San Francisco has the most number of bike rental places. As we can see from these maps, the bike rental places are concentrated in the city center and tourism places. Especially for San Francisco, the majority of rental places are concentrated on the coastal locations. However, in other places, there are only limited locations.

Interestingly, when we look at the number of bike rentals based on the customer types, we can clearly see there are a lot of subscribers than customers. It is contrary to our original assumption that is there are more customers than subscribers.

To dig deeper into this issue, we looked at the start_city and the end_city. We divided the trip data into two types. One type is that the start city and the end city is the same; the other type is that the origin and the destination are different. Same as our assumption, the majority of the trip are started and ended in the same city.

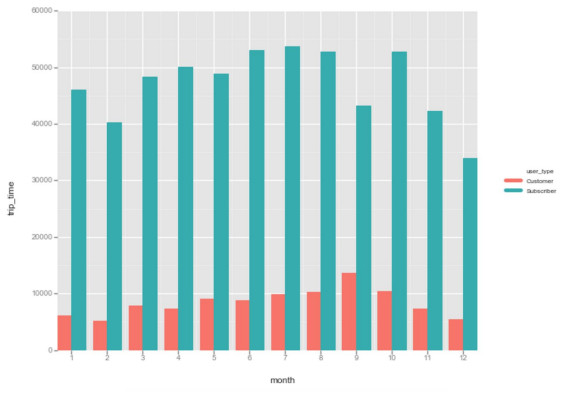
We have a very interesting findings when compared these two together. The majority of trip data that are returned in the same place are subscribers. However, when we look at the different return locations, majority of them are customers, especially on Saturday, as you can see from the plot on the right side.

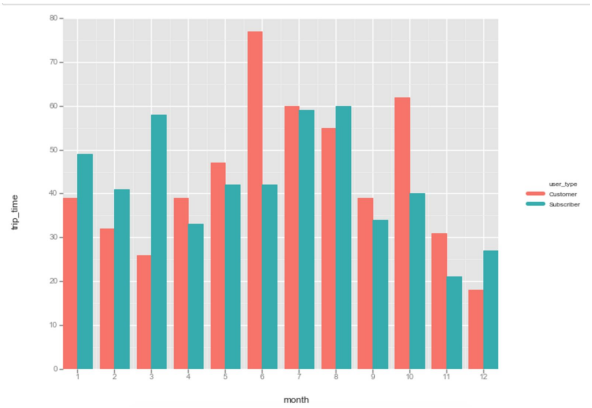




We also looked at the same dataset from a month perspective. The one on the right side is the same location dataset, the right is the different location dataset. The surprising thing about these two plots are the different peak time for customers.

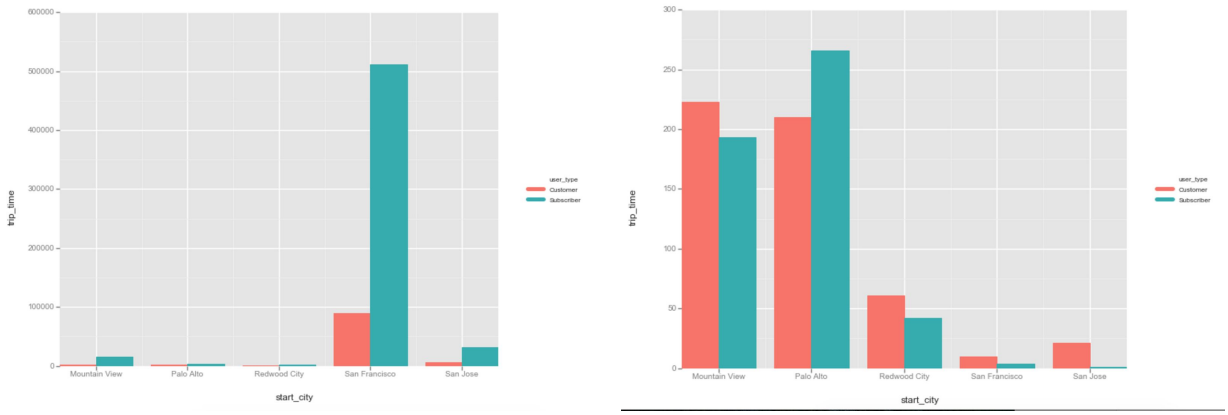
For the same location dataset, the most popular month for customers are September, but for the different location dataset, the most popular month switched to June. We suspect that a lot of tourists are visiting the Bay Area during June.



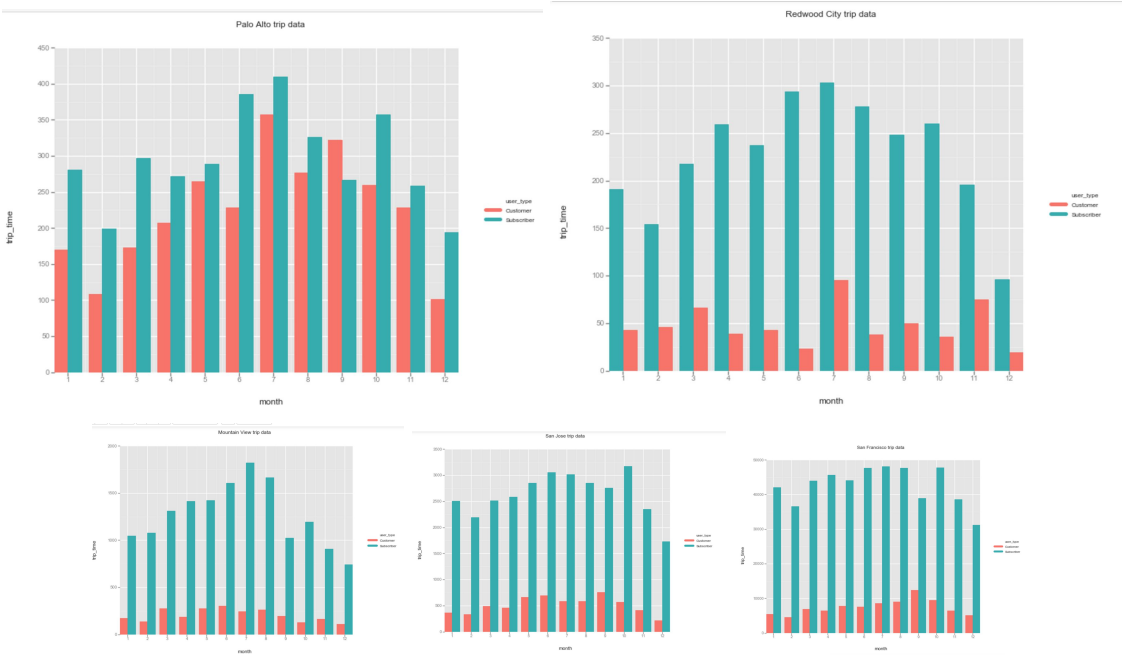


We also looked at the same dataset by different cities and by trip times. For same location returns, the majority are from San Francisco. Interestingly, for different locations, San Francisco is no the highest, Palo Alto and Mountain View are relatively high because the distance between Mountain View/ Palo Alto to San Francisco is quite far and takes time for people to bike from one place to another.

This also explains the rise of the customers in the different return locations. Majority of subscribers are local people and if they want to get around within the Bay Area, bike is probably not their first choice unless they only want to travel around the city. However, for customers (aka, tourists), they are not very familiar with the place so they are more likely to rent a bike and travel around different cities.



Then to further investigate the data by cities, we looked at them by months. Surprisingly, the number of subscribers and the customer in Palo Alto is almost the same, contrary to other cities which majority of the customers are dominated by subscribers. We think the reasoning behind the finding is that most people live in Palo Alto are relatively wealthy and have vehicles and their own bikes, so they don't need to rent a bike to get around. They may use their own bikes or just drive around.



We believe that still conforms with our assumption, but the thing we didn't expect is the huge number of subscriptions. We think this may due to the unawareness of the bike rental places and the inaccessibility of the rental places. As we can see from the maps, there are no bike rental places in the Golden Gate Bridge area.

The Conclusion

We wholeheartedly believe the data analysis that we have set forth has thoroughly addressed all for open questions with which we initially started this project. True, there is much data here to merely confirm the popularly believed assumptions of cause and effect. However there is also equally much if not more data present which proves just the opposite, and that even in the most commonplace kinds of data, so you might believe, there is a plethora of unique, bizarre, and unexpected insights.

Perhaps the Bay Area Bike Share service may need to increase the scale of their IT infrastructure, give just how valuable and insightful generative data output can offer, especially when it's free for the taking. On their website, BABS has stated their projections into the success of their endeavor: within two years, their fleet of 700+ bicycles will turn into an armada of 7,000!! We will remain curious to see just what that data tells us in the future.