Sam Gutentag

Capstone 1 Milestone Report

A. The Problem

The Bay Area Bike Share serves riders and users across five key cities in the San Francisco Bay Area, helping commuters get to and from work, tourists explore the city, and helps relieve congested roads. The program has been operating continuously since the end of August 2013.

- Goal 1 - Who are the important user groups?
- Goal 2 - Can usage patterns in these groups be identified?
- Goal 3 - What affects our Key Users Group Trends?
- Goal 4 - Can important service areas be identified?
- Goal 5 - How can the most active users be best served?
- Goal 6 - How can the lest active users be encouraged to become more active?

B. The Client

The direct client of this analysis is the Bay Area Bike Share Program. The dataset they provide follows the General Bikeshare Feed Specification[1] used by more than 65 other bike share programs across the United States and even a few around the world, making this analysis additionally applicable to several other programs.

C. The Data Set

The data is split into a set of four files, Station Data, Status Data, Trip Data, and Weather Data and is in total comprised of four sets of these files, each with different intervals of time.

Station Data is a record of each Station in the program, with the station ID, station name, station latitude and longitude coordinates, the total number of docks, and the installation date as the entries for each record. The Installation date is used as the date the station became active, the station was relocated, or had the number of docks altered.

Status Data is a record from every station in the system on one minute intervals consisting of entires of station ID, a date and time stamp, the number of bikes available, and the number of docks available.

Trip Data consists of entries with a trip identification number, trip duration in seconds, start date, start terminal, end date, end terminal, the bike identification number, user type, and user home zip code.

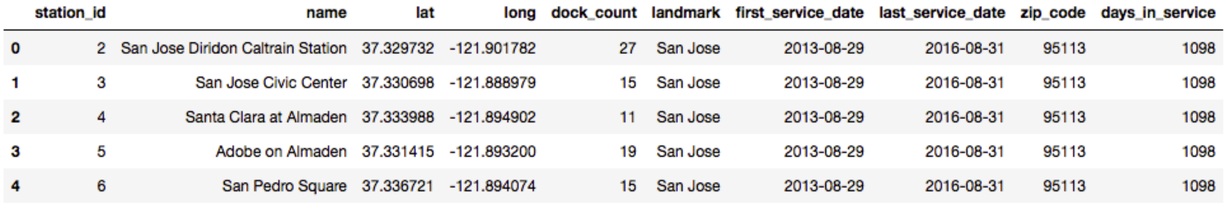
Weather Data is comprised of records for each day, in the key zip code for each of the five service areas and is comprised of date, temperature aggregations, wind aggregations, precipitation amounts, cloud cover, and notable events.

D. How the Data Set was Cleaned

The most important data set at the start is the Trip Data, as it is the record of the usage of the Bike Share Program as a whole. Each data set was cleaned with the intention to be compared to the trip data set, and additionally for comparison to the Status Data.

D1. Station Data

Files consist of multiple repeated records, in each reporting period the majority of the stations were not altered in any way. The first step was merging all four data files and removing duplicates. From there, Bay Area Bike Share provides a set of notes about special cases in which stations were renamed, relocated, expanded, closed, or opened. These notes were felt with specifically, removing records that show stations moving across the street or around a street corner as that is not important for this analysis. Stations that were moved more than a mile were given a second record with a matching Station ID but had the Last Service Day of the starting location and the First Service Day on the new record adjusted to match. Finally a Days in Service column was appended to each record for later popularity comparisons.



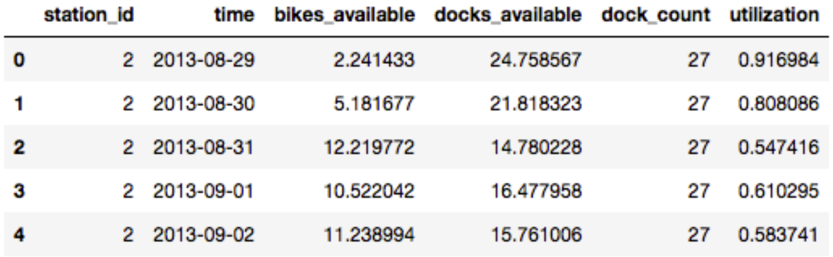
	station_id	name	lat	long	dock_count	landmark	first_service_date	last_service_date	zip_code	days_in_service
0	2	San Jose Diridon Caltrain Station	37.329732	-121.901782	27	San Jose	2013-08-29	2016-08-31	95113	1098
1	3	San Jose Civic Center	37.330698	-121.888979	15	San Jose	2013-08-29	2016-08-31	95113	1098
2	4	Santa Clara at Almaden	37.333988	-121.894902	11	San Jose	2013-08-29	2016-08-31	95113	1098
3	5	Adobe on Almaden	37.331415	-121.893200	19	San Jose	2013-08-29	2016-08-31	95113	1098
4	6	San Pedro Square	37.336721	-121.894074	15	San Jose	2013-08-29	2016-08-31	95113	1098

D2. Status Data

First correcting the date time formatting and then appending a column for total dock count by referencing the dock information in the Station Data, and then downsampled to 5 minute, 1 hour, and 1 day means. Before downsampling, some house keeping is needed to adjust records that take place in the Redwood City stations.

Stations 21, 22, 23, 24, 25, and 26 were closed on June 30, 2016, with stations 23, 24, 25, and 26 being relocated and getting new station ids 88, 89, 90, and 91 respectively. After these stations reopened, there was a delay in updating the station_id number and that is reflected in the status records. All Status records at each of these four stations have their station_ids corrected to the updated numbers.

Once this cleaning is taken care of, the dataset is downsampled to 5 minutes mean intervals, this provides a much lighter dataset for analysis. For each, a dock_count column was appended by joining the Station Data previously cleaned with special attention paid to adjustment for Station 73 expanding from 15 docks to 19 docks. This is important to double check the date of the record, as once the total docks are available, we can calculate utilization as the number of free docks divided by the total number of docks. A completely empty station is considered ‘fully utilized’.



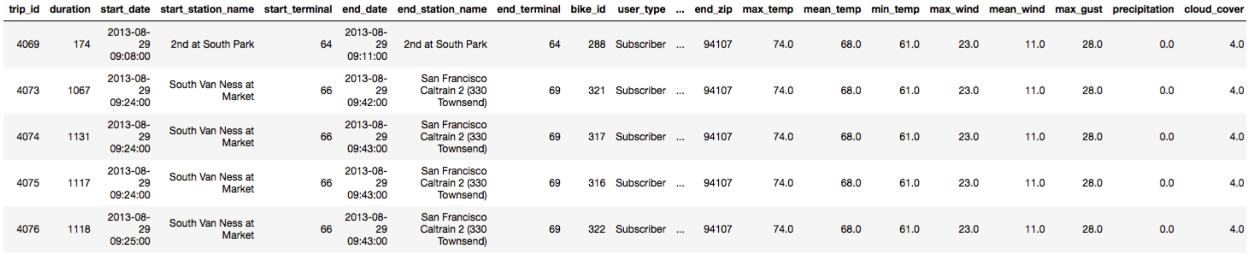
	station_id	time	bikes_available	docks_available	dock_count	utilization
0	2	2013-08-29	2.241433	24.758567	27	0.916984
1	2	2013-08-30	5.181677	21.818323	27	0.808086
2	2	2013-08-31	12.219772	14.780228	27	0.547416
3	2	2013-09-01	10.522042	16.477958	27	0.610295
4	2	2013-09-02	11.238994	15.761006	27	0.583741

D3. Trip Data

Trip Data is cleaned by first pruning outliers by trip duration, upwards of 97% of the nearly one million recorded trips last no more than one hour, those longer than one hour are discarded from this analysis. As with the Status Data, there are trips that report starting and ending at terminals 23, 24, 25 and 26 after they are relocated. For consistency, these records are updated to reflect the new station identification numbers in date ranges matching the Status Data and Station Data. For validation, graphical exploratory analysis is conducted to view trips starting and ending at each station near the opening and closing dates of the station for all stations.

To streamline latter analysis, Station Data was merge into each record to provide a start and end area, corresponding to the landmarks in the Station Data. In addition, key Weather Data was added to each record. Specifically, minimum temperature, maximum temperature, precipitation, cloud cover, and mean wind values were appended to each row.

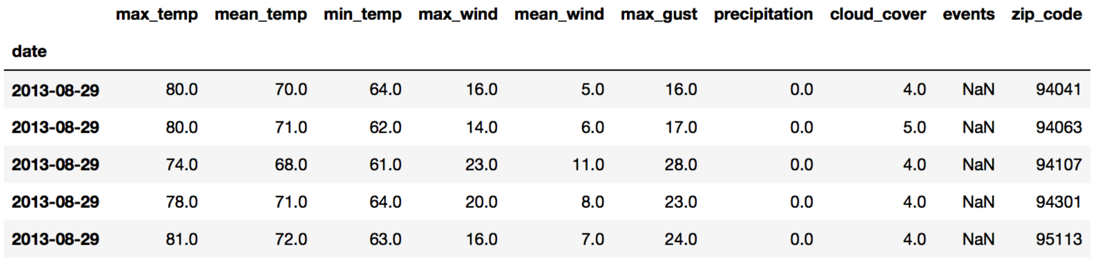
Finally data was saved to a single file for all data records, and in addition a file was saved of just Subscriber and Customer user type subsets.



trip_id	duration	start_date	start_station_name	start_terminal	end_date	end_station_name	end_terminal	bike_id	user_type	...	end_zip	max_temp	mean_temp	min_temp	max_wind	mean_wind	max_gust	precipitation	cloud_cover
4069	174	2013-08-29 09:08:00	2nd at South Park	64	2013-08-29 09:11:00	2nd at South Park	64	288	Subscriber	...	94107	74.0	68.0	61.0	23.0	11.0	28.0	0.0	4.0
4073	1067	2013-08-29 09:24:00	South Van Ness at Market	66	2013-08-29 09:42:00	San Francisco Caltrain 2 (330 Townsend)	69	321	Subscriber	...	94107	74.0	68.0	61.0	23.0	11.0	28.0	0.0	4.0
4074	1131	2013-08-29 09:24:00	South Van Ness at Market	66	2013-08-29 09:43:00	San Francisco Caltrain 2 (330 Townsend)	69	317	Subscriber	...	94107	74.0	68.0	61.0	23.0	11.0	28.0	0.0	4.0
4075	1117	2013-08-29 09:24:00	South Van Ness at Market	66	2013-08-29 09:43:00	San Francisco Caltrain 2 (330 Townsend)	69	316	Subscriber	...	94107	74.0	68.0	61.0	23.0	11.0	28.0	0.0	4.0
4076	1118	2013-08-29 09:25:00	South Van Ness at Market	66	2013-08-29 09:43:00	San Francisco Caltrain 2 (330 Townsend)	69	322	Subscriber	...	94107	74.0	68.0	61.0	23.0	11.0	28.0	0.0	4.0

D4. Weather Data

Weather records contain a lot of data points for each record, but cleaning is actually pretty straightforward. First the records are pruned to only include date, temperature, wind, precipitation, cloud cover, events, and zip code entries. Then using the date as the index, the data is sorted by date. Additionally, datasets for each service area are created for quick reference.



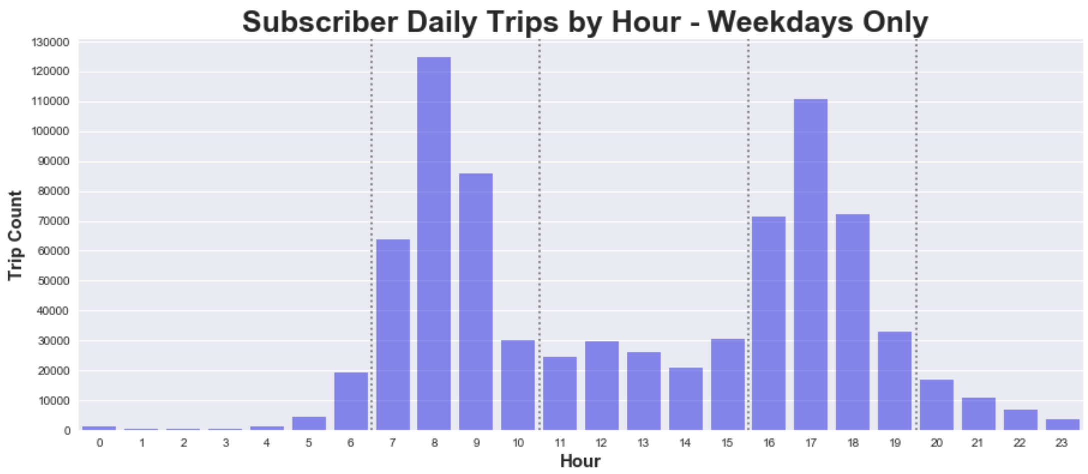
	max_temp	mean_temp	min_temp	max_wind	mean_wind	max_gust	precipitation	cloud_cover	events	zip_code
date										
2013-08-29	80.0	70.0	64.0	16.0	5.0	16.0	0.0	4.0	NaN	94041
2013-08-29	80.0	71.0	62.0	14.0	6.0	17.0	0.0	5.0	NaN	94063
2013-08-29	74.0	68.0	61.0	23.0	11.0	28.0	0.0	4.0	NaN	94107
2013-08-29	78.0	71.0	64.0	20.0	8.0	23.0	0.0	4.0	NaN	94301
2013-08-29	81.0	72.0	63.0	16.0	7.0	24.0	0.0	4.0	NaN	95113

E. Initial Findings

E1. Identify Most Important User Group

There are two key user groups that make use of the Bay Area Bike Share, Subscribers and Customers. Subscribers account for nearly 90% of trips taken. The most important group of users are a subset of Subscriber trips known as ‘Commuter Trips’.

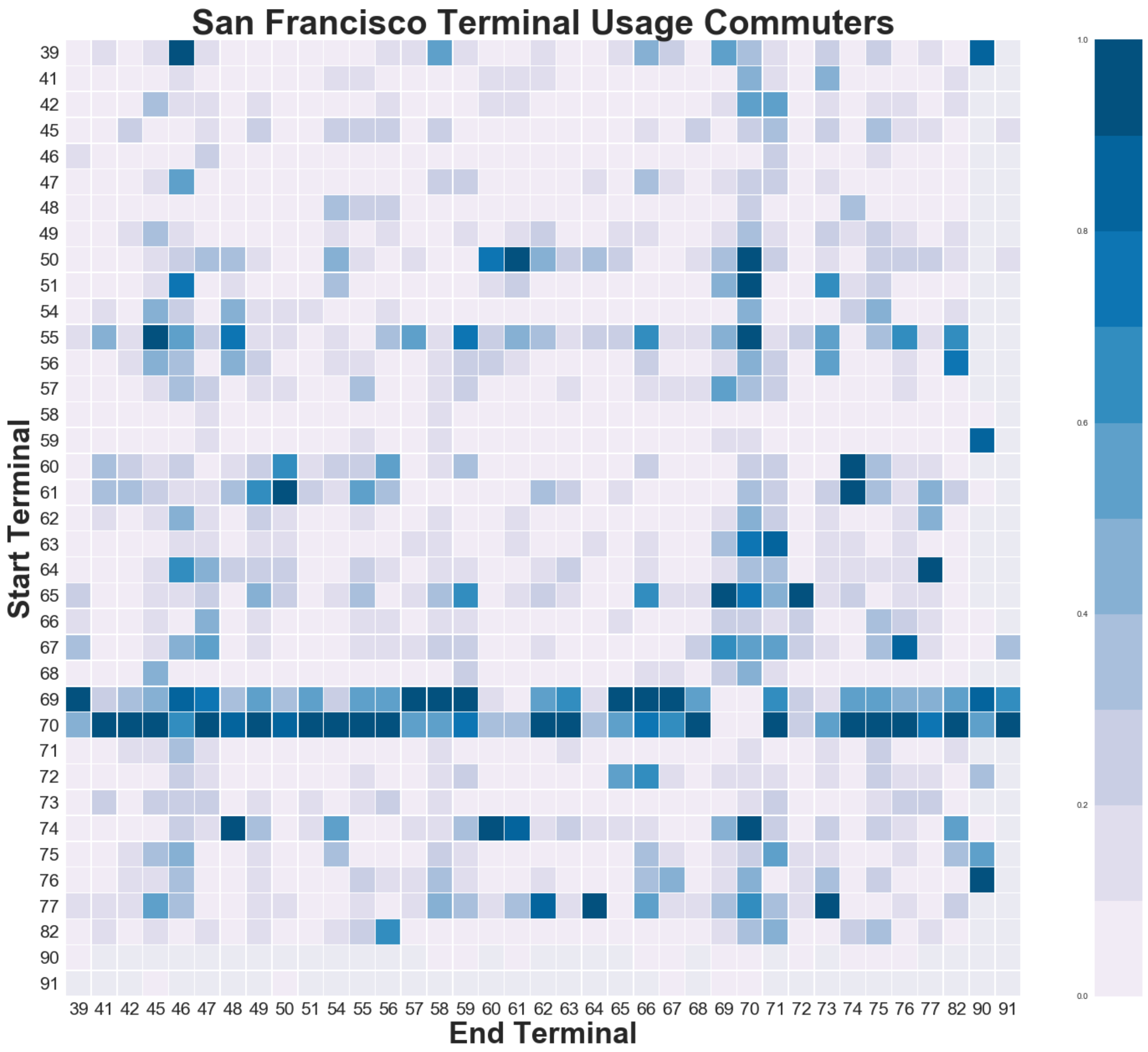
Commuter Trips are trips taken by Subscribers within San Francisco, during weekday commute hours 7am-11am and 4pm-8pm. This group of users account for over 75% of all trips.



Supporting this claim by looking at the most heavily used stations by Commuters, we find that the top ten stations for commuters starting and ending their trips are within a block of a mass transit terminal. Most notably, 25% of Commuters start their morning commute (and end their evening commute) at stations 69 or 70, which are located across the street from each other, just outside the final Caltrain Station in San Francisco, the local commuter rail system. In fact, all of the top ten morning and evening commuter bike stations are all within one block of Mass Transit systems.

	Top Morning Commute Start Terminal			Top Morning Commute End Terminal		
Rank	**Station ID**	**Terminal Name**	**Share**	**Station ID**	**Terminal Name**	**Share**
1	70	San Francisco Caltrain	14.58	70	San Francisco Caltrain	7.60
2	69	San Francisco Caltrain 2	11.64	65	Townsend at 7th	6.55
3	50	Harry Bridges Plaza Ferry Building	7.22	61	2nd at Townsend	6.54
4	55	Temporary Transbay Terminal	7.00	77	Market at Sansome	5.05
5	74	Steuart at Market	5.60	69	San Francisco Caltrain 2	4.28
6	61	2nd at Townsend	4.11	60	Embarcadero at Sansome	4.18
7	73	Grant Avenue at Columbus Avenue	3.82	55	Temporary Transbay Terminal	4.14
8	77	Market at Sansome	3.21	51	Embarcadero at Folsom	4.06
9	54	Embarcadero at Bryant	2.96	63	Howard at 2nd	4.02
10	67	Market and 10th	2.93	74	Steuart at Market	3.96

	Top Evening Commute Start Terminal			Top Evening Commute End Terminal		
Rank	Station ID	Terminal Name	Share	Station ID	Terminal Name	Share
1	70	San Francisco Caltrain	6.18	70	San Francisco Caltrain	17.29
2	65	Townsend at 7th	5.86	69	San Francisco Caltrain 2	12.09
3	61	2nd at Townsend	5.17	50	Harry Bridges Plaza Ferry Building	6.66
4	69	San Francisco Caltrain 2	4.68	55	Temporary Transbay Terminal	5.85
5	77	Market at Sansome	4.65	74	Steuart at Market	5.77
6	64	2nd at Southpark	4.36	77	Market at Sansome	4.31
7	60	Embarcadero at Sansome	4.16	61	2nd at Townsend	4.11
8	67	Market and 10th	4.00	65	Townsend at 7th	3.39
9	55	Temporary Transbay Terminal	3.97	39	Powell Street BART	3.05
10	74	Steuart at Market	3.97	60	Embarcadero at Sansome	2.95



E2. Data Story

Interesting patterns emerge when comparing the top Morning Commuter Stations to the top Evening Commuter Stations. Commuters can be split into two subgroups, riders commuting into the city, and those commuting out of the city.

Inbound Commuters are users who start their morning commute and end their evening commute at bike stations near Mass Transit Stations. As an example, this is a user who take the Caltrain from their home to the San Francisco Station and takes a bike from the Caltrain station to their office. Reversing the process at the end of their workday to return home.

Outbound Commuters follow the same process, only in reverse. Riding a bike from their home to a Mass Transit station to complete their commute.

E3. What Affects Core User Group

The nature of users being work day commuters keeps them using the program consistently. To access the affect of weather on these riders, we compute the difference in mean trips per day given several weather conditions, Hot Days, Cold Days, Rainy Days, and Windy Days.

Commuter trip numbers are meaningfully affected by hotter temperatures, with mean trip count **increasing 8.76%**, or average 62.68 more trips, on days with a Maximum Temperature above 75.86 degrees Fahrenheit

Commuter trip numbers are meaningfully affected by colder temperatures, with mean trip count **decreasing 14.70%**, or average 105.17 fewer trips, on days with a Minimum Temperature below 45.41 degrees Fahrenheit

Commuter trip numbers are meaningfully affected by Rainy weather, with mean trip count **decreasing 39.64%**, or average 282.48 fewer trips, on days with precipitation above 0.46 inches.

This is a substantial dip in Ridership, but it is important to also note that in across all Data Records, there are only 23 ‘Rainy Days’ across all 1099 days of recorded data.

Commuter trip numbers are meaningfully affected by Windy weather, with mean trip count **decreasing 7.58%**, or average 53.99 fewer trips, on days with a mean Wind Speed above 12mph.