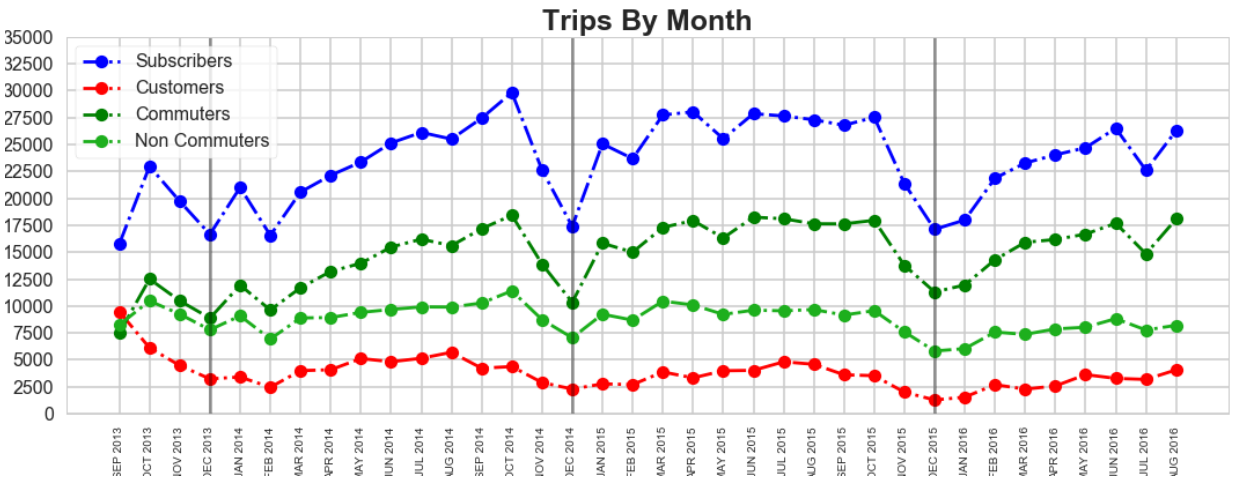
Bay for Bikers

Analyzing Bay Area Bike Share Usage Data

BY SAM GUTENTAG
JANUARY 12, 2018

Introduction

The Bay Area Bike Share Program is a network of rental bicycles distributed across San Francisco, San Jose, Mountain View, Redwood City, and Palo Alto California. Composed of 700 bikes and 74 stations, BABS accumulated nearly a million trips between September 2013 and September 2016. That is more than 24,000 trips per month from their subscription tier riders and an additional 3,300 trips to more casual riders, tourists, and other riders.



This level of ridership has proven to be very stable and this report aims to identify ridership patterns in these two groups of riders. We will provide insight in to where the program is best serving each user group and where potential improvements can be made to increase ridership or improve the system for all riders riders.

The Data

A. The Bay Area Bike Share Dataset

The dataset of 5 files for each of 4 date intervals: 2014_02, 2014_08, 2015_08, 2016_08

```
*_status_data.csv
*_trip_data.csv
*_station_data.csv
*_weather_data.csv
*_README.txt
```

The ‘status’ file set consists of records for each station on a minute by minute bases reporting the number of empty docks and parked bikes.

The ‘trips’ file set consists of records for each trip that is taken, reporting trip_id, start and end station_id, start and stop date, trip duration in seconds, bike_id, user type and the rider’s home zip code.

The ‘station’ file set consists of records for each active station in the program, reporting back the station id, station name, latitude and longitude coordinates, total number of docks, the city the given station is located within, and the date of installation.

The ‘weather’ file set consists of records of daily weather temperatures, precipitation, cloud cover, and several other weather metrics for each landmark location on a daily basis. This file is not needed as the DarkSky Weather Data provides much more detailed information.

The ‘README’ file sets consist of important notes provided by BABS with information about station relocation dates, the dates of station dock expansions, and the dates related to several stations being relocated but not having their station id numbers updated until later and several other notes. This file set is invaluable during data wrangling and cleaning of the data set as a way to make sure records reflect events as they occurred and corrects recording mistakes that would not be inferred by the data itself.

B. DarkSky Weather Data

DarkSky is a popular weather service aggregator and provides a substantial amount of information in the ‘Time Machine’ API that can be polled to gain hourly weather data at each station. A simple script allowed us to poll the API to find weather conditions for all stations in which a trip started from that station on each day. This accumulated to more than 66,000 daily records that are cleaned and attached to trip data in the wrangling steps described later.

C. Google Maps Elevation API

Google Maps needs no introduction, and the Elevation API end points provided elevation data for each of the stations in the bike share program. This information is also attached to each trip record and later used to calculate the elevation drop or rise between stations for each trip, spoiler alert, downhill trips are more frequent, and much shorter (read faster) than uphill trips.

Data Wrangling

The goal of the wrangling techniques used on this raw data set was to first correct station information based on the included README file notes, and then append by calculations or by join with other data sets columns to each file set that would be used multiple times over through the course of the analysis.

A. Station Data

The ‘station’ data is first concatenated to a single pandas data frame. First a ‘last_service_date’ column was appended to all rows with the date matching the last recorded trip end data from the ‘trips’ data set. Additionally, a ‘zip_code’ column was added to each row by referencing the README file to match original ‘landmark’ columns to the respective zip code.

The bulk of data wrangling for this data set is from direct notes included in the README files making manual adjustments to station first service and last service dates. When stations are relocated or expanded the ‘last_service_date’ column is updated and a duplicated row is appended with the ‘start_service_date’ set to a day after the ‘last_service_date’ in the original row. Finally a ‘days_in_service’ column is appended to each row that is the number of days between the first and last service data, this is used in later analysis to compare station popularity.

Additionally, the Google Maps API is polled for each station to append the elevation of each station in meters, and then meters are converted to feet.

B. DarkSky Data

Supplemental weather data was collected by polling the DarkSky Weather API for all daily and hourly records from each station on all days in which a trip was started from that station. This is a collection of 66,115 data records. The DarkSky data set provides weather records it provides, rain, wind, cloud cover, temperature and all other weather attributes fluctuate throughout the day so this added precision to trip information is useful when comparing the number of trips taken when it is hot, cold, windy, rainy, etc. Data is written to files with a station id appended to each row by region.

C. Trip Data

The ‘trip’ data is all concatenated into a single pandas data frame and pruned to only include trips that last at maximum one hour. The one hour cutoff is used to eliminate a small number of outliers where trips were noted to last several days, weeks, or in one case nearly a year. These are likely to be recording errors, and at worst case there are only 200 of the more than 998,000 total trips that are longer than hour long.

From here, and cross referencing the ‘README’ notes, we make adjustments to trips that start or end at stations that were not properly updated at the correct time. For example, several trips end at station 23 after the station had been relocated to station 88, but was not correctly reporting that it was in fact station 88. There are several of these mistake that are all adjusted accordingly. A ‘duration_minutes’ column is calculated from the original ‘duration’ column for easier analysis.

User zip codes are aggressively pruned and cleaned based on notes in the included in the README notes it shows that the collecting this data is done by riders at the time of each rental and does not well handle non-standard zip code formats (such as those from international customers) and is specifically noted as a non reliable data set. As such all zip codes are pruned to the first 5 numerical digits, and is ignored otherwise.

Additionally, a file is created is written for each year with trip data merged with Station and DarkSky data. This is done frequently in analysis, so a starter file already merged is handy. Columns for ‘elevation_change_meters’ and ‘elevation_change_feet’ are created by subtracting the elevations the end station from elevations at start station for each recorded trip.

D. Status Data

The ‘status’ data is first concatenated from four files to a single pandas data frame. This is a massive data set, at 3.2GB of uncompressed data the important part is first using the README notes to correct station_id numbers based on the updated values by date found in the

previously cleaned station data set. The status data set also includes all zero records for a station 87, a station that does not exists, and is removed.

Data from each station is resampled on 1 minute intervals to ensure fill in any missing values, a small subset of data is missing at a few stations in the early hours of the morning and forward filling in minutes that are missing data is reasonable.

Secondary files are created for each station with a 5 minute resampled mean. This is used when dealing with such a large data set to more rapidly iterate over the data and then only dig into minute by minute data when and where necessary.

A utilization column is happened to each data frame which is calculated as the number of docks available divided by the total dock count, that is, a fully utilized station is a time when the station is completely empty of bikes. Cleaned status data is stored in files by station id.

E. Bike Data

As an added bonus, the cleaned trips data set is used to create a bike data frame consisting of information for each bike. Records include trip counts, date of first and last recorded trip, total ride time, and days since last trip which is calculated as days between end record date in the trips data set and the date the bike was last parked. Other data including the age of the bike, the number of days in service, the number of days used and other metrics are collected. This information would be useful in identifying bikes that need servicing, replacing, or relocating to lower traffic areas. Bikes are a high cost item, and the longer they last the better the return on investment.

F. Supplemental Google Maps Elevation Data

Appending to Station Data, the Google Maps Elevation API is polled to collect elevation in meters and feet above sea level for each station. This data is also included when the Station Data is merged to each trip in the Trip Data. An 'elevation_change' column in meters as well as a column in feet is added to each trip record.

Analysis - Identifying Key Users and User Subsets

Bay Area Bike Share offers two service tiers to riders, Monthly Subscription Tier identified as Subscribers) and a more casual 'Pay as You Go' Tier (identified as Customers). Each group uses the program with different patterns and in significantly different volumes.

We will identify usage across hourly, daily, and monthly intervals, as well as identify important stations. We will look to identify subgroups of riders within this program tier to possibly identify new Service Tier Plans to expand the program to new riders and better serve current riders.

A. Rider Usage Patterns by Region

Working from the trips data, we can look at the breakdown of trips within each service region, and the share of trips taken by Subscribers and Customers. Note that for each region, the values here represent trips that start and end within that region, and that the 'Cross Region' row is all trips that start in one region and end in a different region.

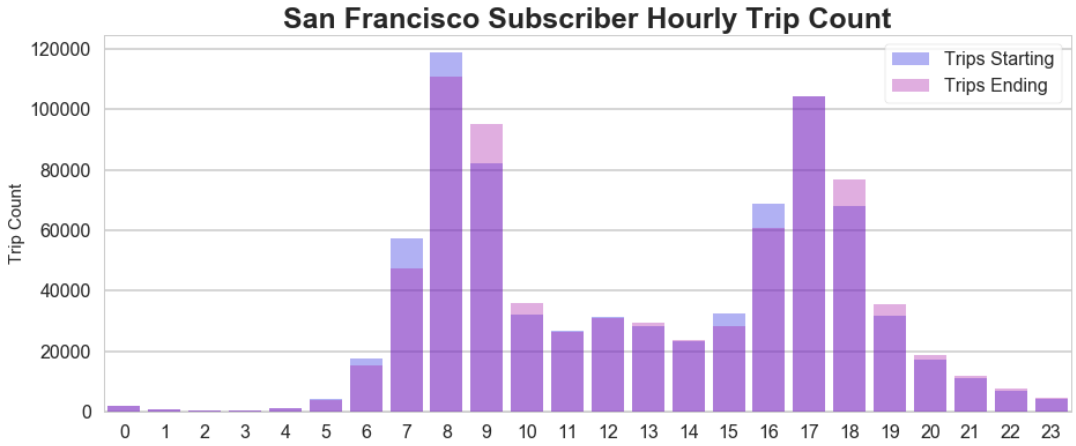
	All Users		Subscribers		Customers	
Region	**Trip Count**	**Trip Share**	**Region Trip Count**	**Region Trip Share**	**Region Trip Count**	**Region Trip Share**
All	983350	100.0000	846790	86.1128	136560	13.8872
San Francisco	891068	90.6155	771572	86.5896	119496	13.4104
San Jose	52781	5.3675	44117	83.5850	8664	16.4150
Mountain View	24646	2.5063	20933	84.8347	3713	15.0653
Palo Alto	9852	1.0019	5940	60.2923	3912	39.7077
Redwood City	5003	0.5088	4228	84.5093	775	15.4907
Cross Region	1420	0.1444	715	50.3521	705	49.6479

The vast majority of trips are taken within San Francisco, and only 0.144% of all trips traverse from one region to another. A consistent trend of trips being taken primarily by Subscribers is visible, Palo Alto stands out with a lower ratio of Subscriber to Customer ridership than all other regions but ridership is so low that this difference is not particularly noteworthy.

Subscription Members taking trips within San Francisco might be our key demographic of focus so far, accruing 90.6155% of all trips taken within the program and 86.5896% of all San Francisco trips, we can dig deeper into this subset deeper.

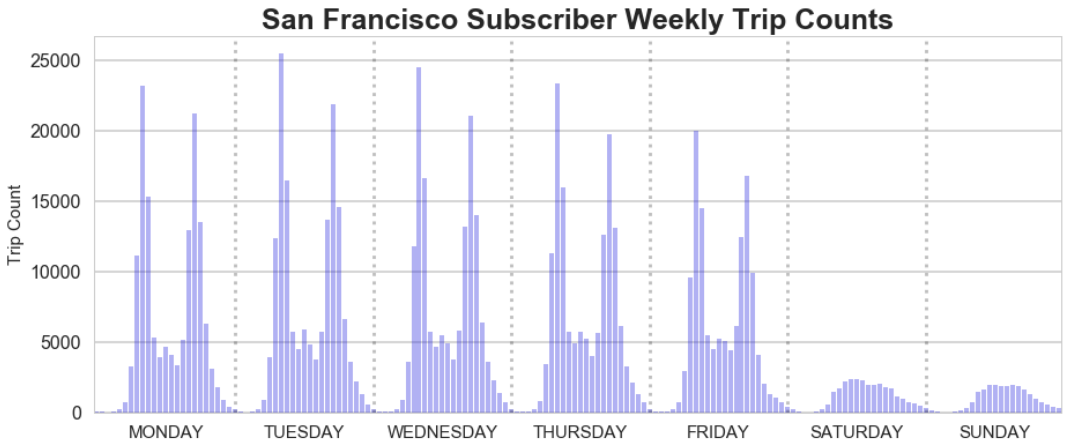
B.1 Subscriber Daily Usage Patterns

Significant peaks are immediately present in this subset of the data, with large peaks in ridership centered around 7am-10am and 4pm-7pm, commuter hours. In fact 33.6857% of Subscriber Trips in San Francisco take place in the morning commute window and an additional 30.8227% take place during the during the evening commute window.



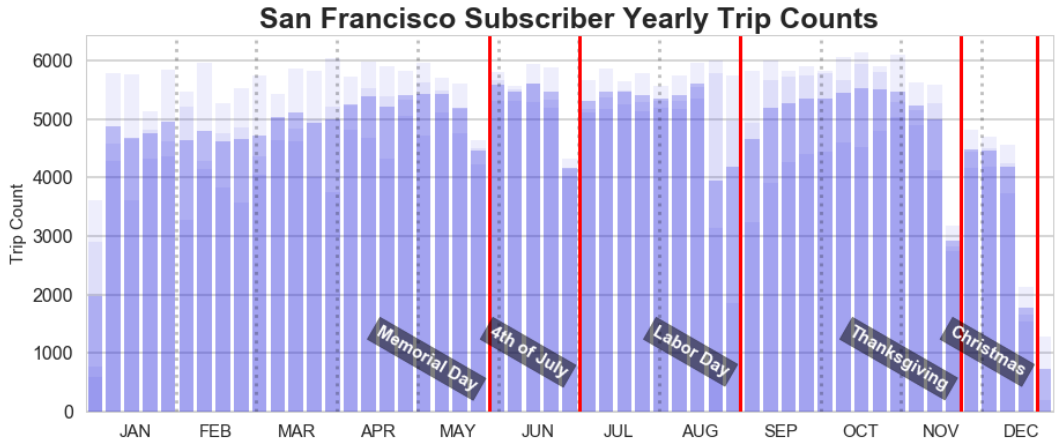
B.2 Subscriber Weekly Usage

Knowing that such a large chunk of trips are taken during what appear to be commuter hours, it is worthwhile to break down the data by day of the week. Again, to lackluster surprise, we see a heavy preference for San Francisco Subscriber trips to be taken during weekdays. In San Francisco, 96.9727% of trips start and 96.9683% of trips end on weekdays.



B.3 Subscriber Yearly Usage

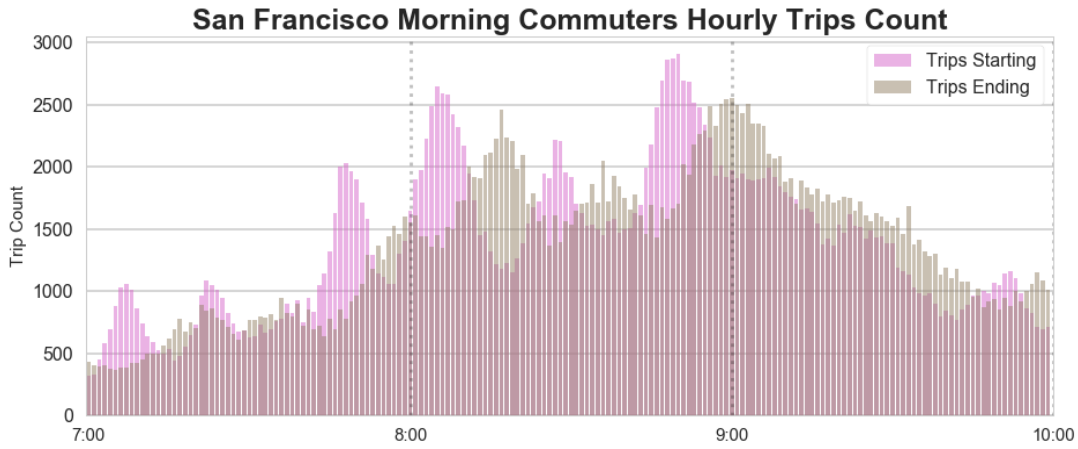
Zooming out one more time, we see that yearly Subscriber ridership are stable through out the year, only dipping around the holidays, holiday weekends, and in the winter months overall. This November and December dip might just be because of the holiday break, but weather may play a roll here as well, we will investigate that further later in this analysis.

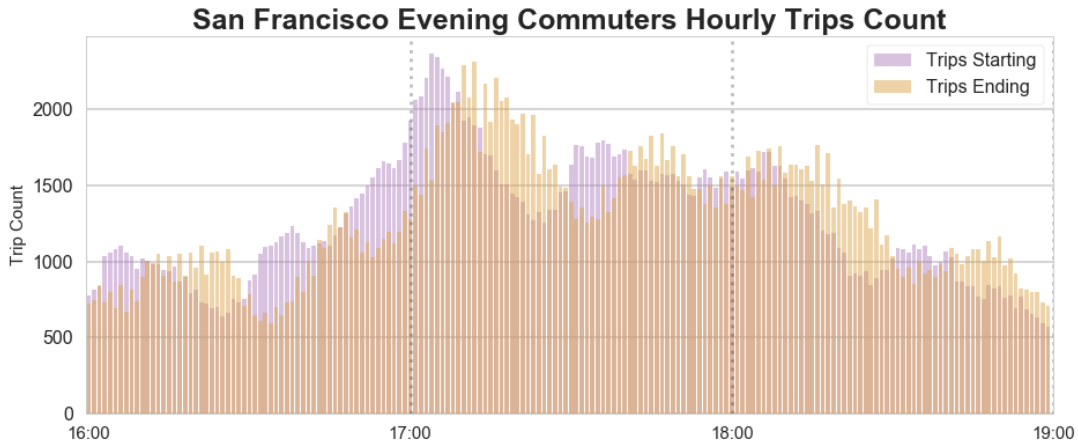


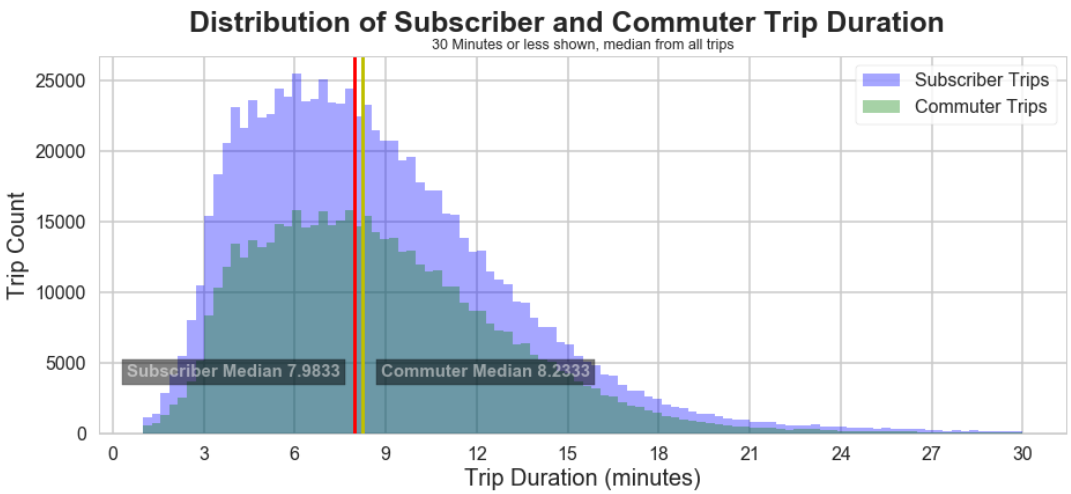
C.1 Commuters, The Missing Group

Looking at weekly trends, we can see clear trends in ridership on weekdays between 7am and 10am and between 4pm and 7pm. These windows are consistent throughout the entire year with the exception of holidays. This leads us to the narrative that the Commuter is a key user group to target for program growth and enhancement.

Zooming in on these Commuter windows, a minute by minute break down reveals the smaller peaks visible in these subsets of data. The offset of these counts also help illustrate the fact that many rides are 9 minutes or less in duration, a quick ride to work or home.

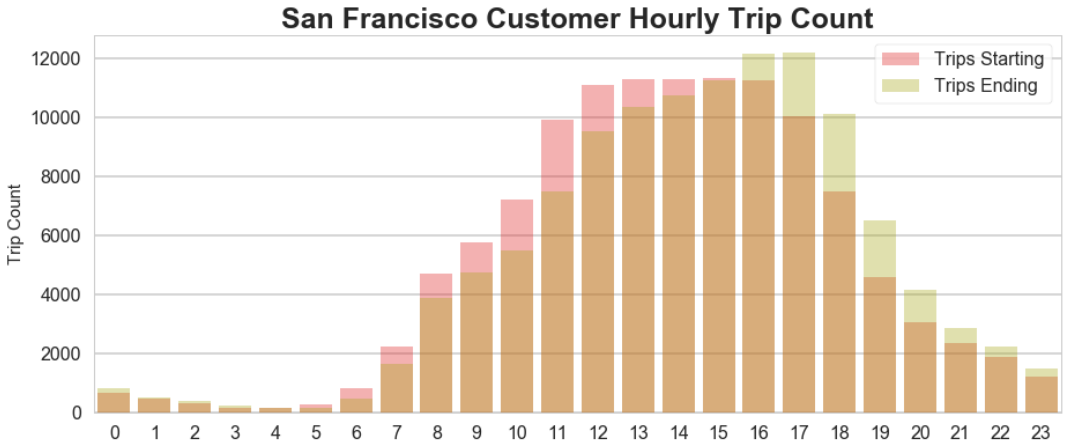






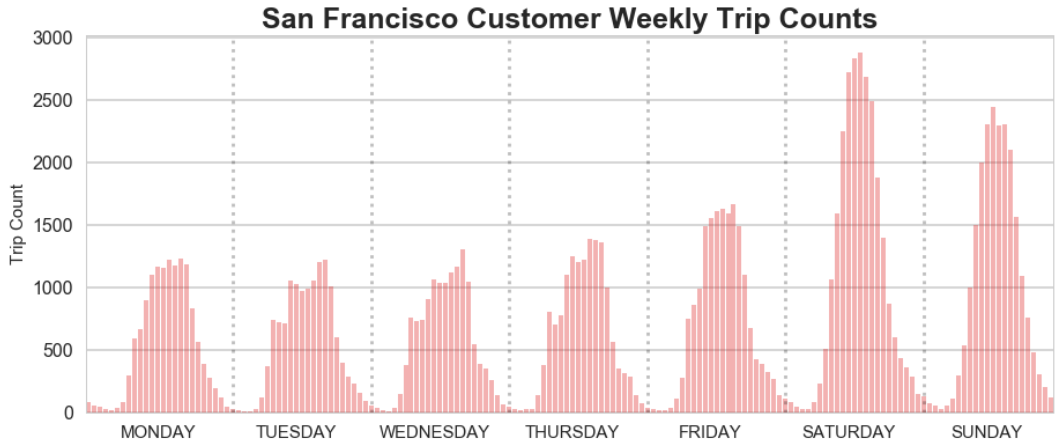
D.1 Customer Daily Usage

On an hourly basis, Customer usage is quite different from Subscribers, with a more normal distribution across the hour of day, Notable is the plateau of trips between 11am and 6pm, daylight hours.



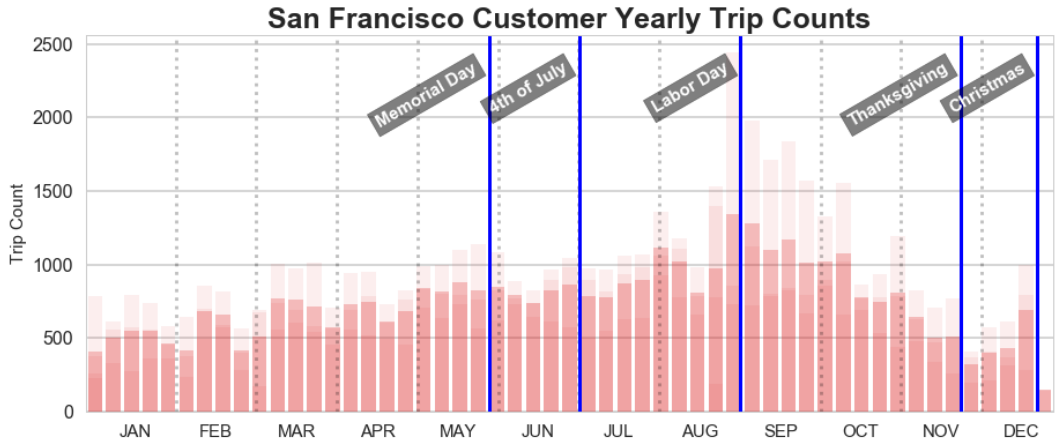
D.2 Customer Weekly Usage

On a weekly basis, Customer usage is also quite different from Subscriber usage, flipping that weekends are vastly more popular ride times than weekdays



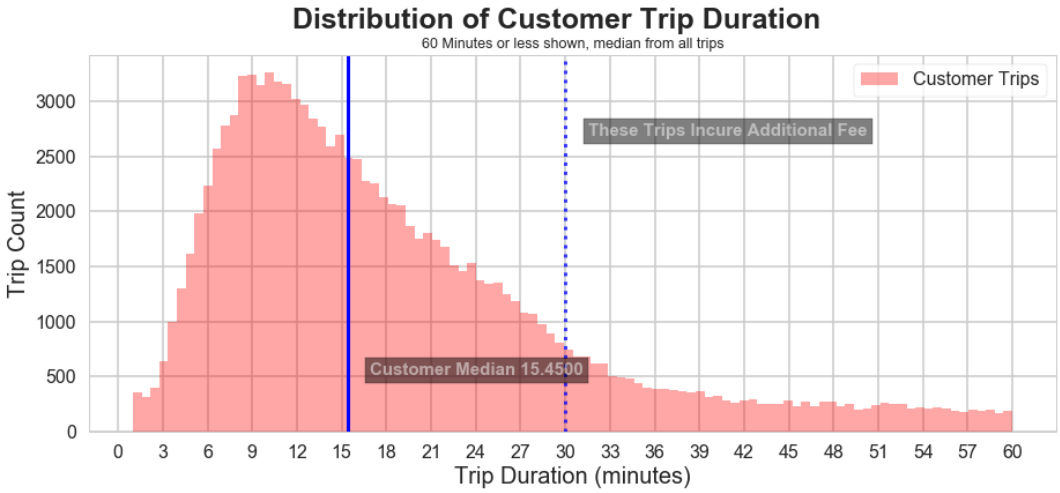
D.3 Customer Yearly Usage

Finally we zoom out to a yearly overview. Customers are flipped from Subscribers in that Holidays report much higher ridership counts than normal. Labor Day in particular, has the highest daily ridership for Customers every year. With one year seeing accumulating nearly 2500 trips, more than double the normal number.



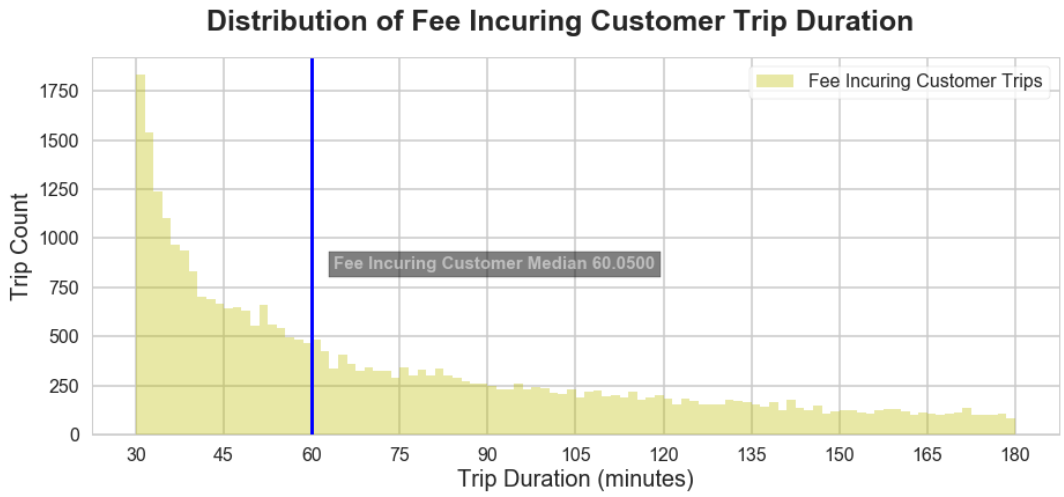
D.4 Customer Trip Duration

Customer Trips are often longer in duration than Subscriber and Commuter rides. However, we see a substantial number of trips are longer than the 30 minutes flat rate for rental period offered to causal riders. This is a possible region for Program Managers to work to ensure Customers are well enough informed of the pricing plans available to them.



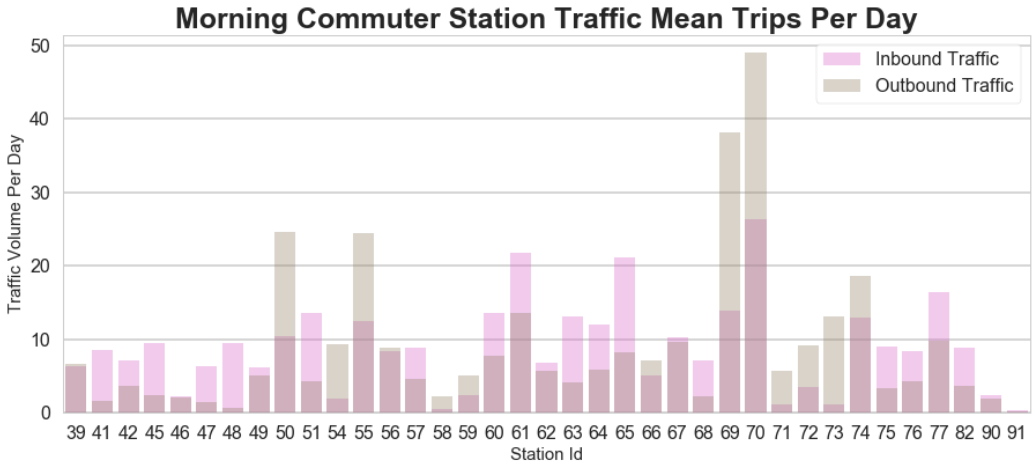
D.5 Customer Trips with Fees

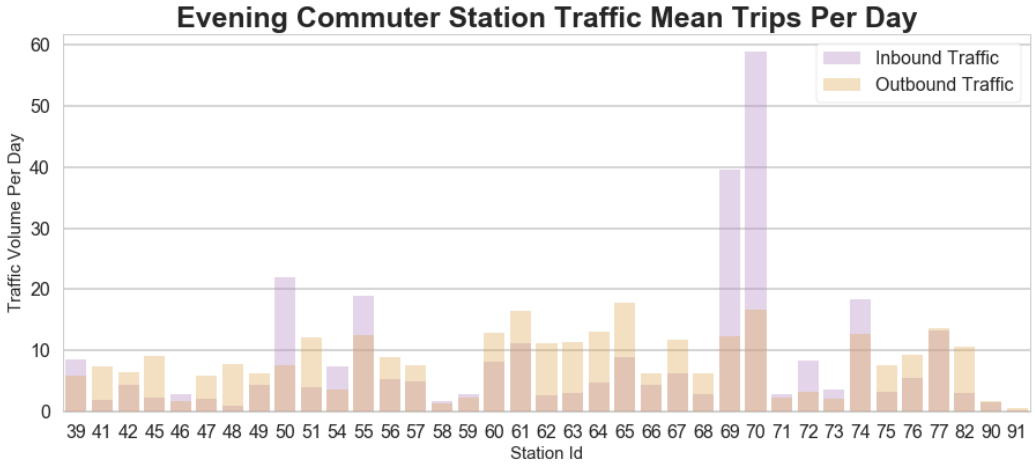
Of note, starting in 2017, after the dataset ends the Bay Area Bike Share program added a both a 1 hour and 3 hour rental option to Customer riders to abate this issue. Which given the distribution of fee incurring trips, is well placed.



Analysis - Identify Most Important Stations

Once we have focused in on a Core User Group, we will inspect ridership patterns to identify the most heavily trafficked stations. Taking a count of all Commuter trips inbound and outbound from each station in the morning and evening commute windows, several key stations become immediately apparent.





Stations 69 and 70 are bike traffic meccas, both during the morning and evening commute windows. In the Morning Commute window, these two stations are more heavily used as outgoing stations, that is more trips originate from these stations. The opposite is true in the Evening Commute window, with vastly more inbound traffic than outbound traffic on a daily basis.

As it turns out these two stations are directly across the street from one another, and are both located at the entrance/exit to the San Francisco Caltrian Station. This is a commuter rail line that originates in San Jose and terminates in San Francisco at the doorstep of these two stations.

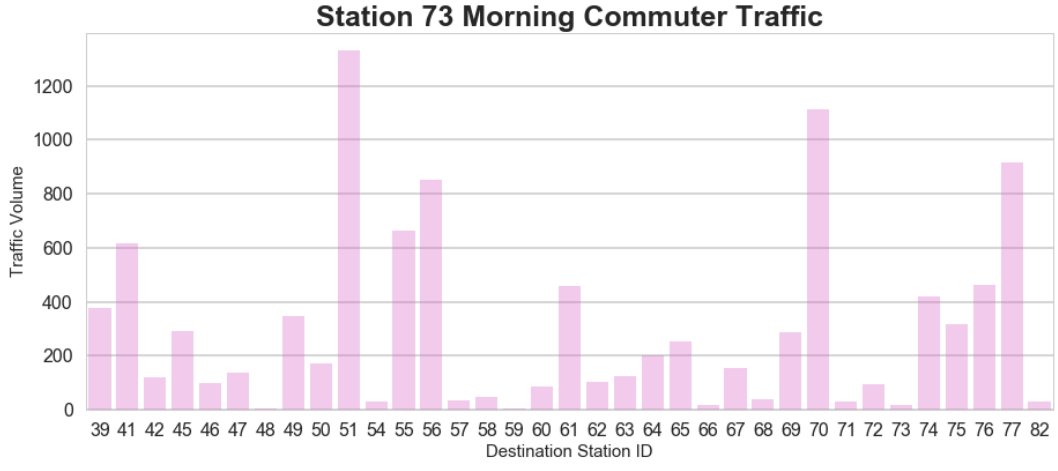
A. Morning Commuter Top Stations

Outbound Trips			Inbound Trips		
ID	**Station**	**Share**	**ID**	**Station**	**Share**
70	San Francisco Caltrain	14.978	**70**	San Francisco Caltrain	8.037
69	San Francisco Caltrain 2	11.691	**61**	2nd at Townsend	6.621
50	Harry Bridges Plaza Ferry Building	7.511	**65**	Townsend at 7th	6.427
55	Temporary Transbay Terminal	7.471	**77**	Market at Sansome	4.975
74	Steuart at Market	5.668	**69**	San Francisco Caltrain 2	4.206
61	2nd at Townsend	4.124	**51**	Embarcadero at Folsom	4.114
73	Grant Avenue at Columbus Avenue	3.980	**60**	Embarcadero at Sansome	4.141
77	Market at Sansome	2.971	**63**	Howard at 2nd	3.996
67	Market at 10th	2.940	**74**	Steuart at Market	3.918
54	Embarcadero at Bryant	2.865	**55**	Temporary Transbay Terminal	3.796

The top ten inbound and outbound morning commuter stations fit well in our commuter narrative. Green Stations above are all within 1 block of a Caltrain Station, Blue Stations are directly above BART stations, Red Stations are within one block of MUNI stations and Yellow Stations are along San Francisco's Embarcadero.

Caltrain is a commuter train rail, BART is the local subway system, MUNI the bus system, and the Embarcadero is a stretch of wide paved walking and biking paths along San Francisco's North Eastern Water front, which is nearly flat and also along a major MUNI light rail train line.

Trivia Fact, at Station 73 'Grant Avenue at Columbus Avenue' the majority of morning trips head towards, in depending order, Station 51 (Embarcadero at Folsom), Station 70 (San Francisco Caltrain), Station 77 (Market at Sansome), or Station 55 (Transbay Terminal). All logical destinations, but a second piece of information about these trips explains the station's popularity. All of these stations are at low elevations, and along direct downhill paths from Station 73, which is the highest elevation station in the entire Bay Area Bike Share System. Something a tired Morning Commuter keen to save pedals would use to their advantage.



B. Evening Commuter Top Stations

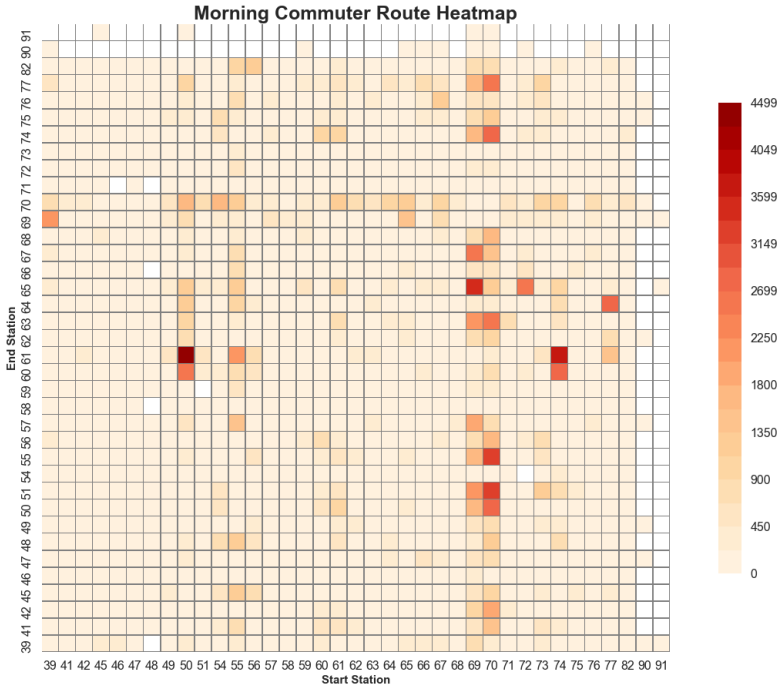
Outbound Trips			Inbound Trips		
ID	**Station**	**Share**	**ID**	**Station**	**Share**
65	Townsend at 7th	5.825	**70**	San Francisco Caltrain	19.316
70	San Francisco Caltrain	5.427	**69**	San Francisco Caltrain 2	12.999
61	2nd at Townsend	5.399	**50**	Harry Bridges Plaza Ferry Building	7.230
77	Market at Sansome	4.474	**55**	Temporary Transbay Terminal	6.190
64	2nd at South Park	4.294	**74**	Steuart at Market	6.034
60	Embarcadero at Sansome	4.228	**77**	Market at Sansome	4.350
74	Steuart at Market	4.147	**61**	2nd at Townsend	3.676
55	Temporary Transbay Terminal	4.107	**65**	Townsend at 7th	2.908
69	San Francisco Caltrain 2	4.012	**39**	Powell Street BART	2.820
51	Embarcadero at Folsom	3.941	**72**	Civic Center BART	2.723

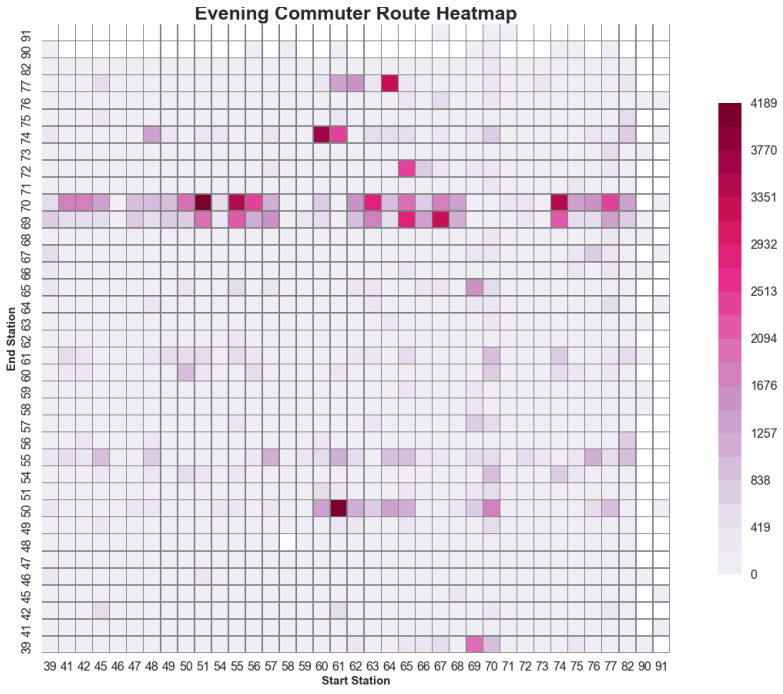
The top ten inbound and outbound evening commuter stations are largely the same stations and including the unique ones tell a similar story.

Another trend that is revealing itself, is that Morning Outbound and Evening Inbound trips are more concentrated at a few stations, with stations 69 and 70 accepting for 25-31% of trips in either group. Contrasting to the more even distribution of Morning Inbound and Evening Outbound stations.

Analysis - Route Popularity

Now that we have identified Important stations in both commuter time windows, we have to wonder, where are all of these riders coming from and going towards in each window? The San Francisco Caltrain stations 69 and 70, and additionally stations 61 and 65 each just one block away, stand out as a key nexus point for the Bike Share Program as a whole for both Morning Commuter start stations and Evening Commuter end stations.

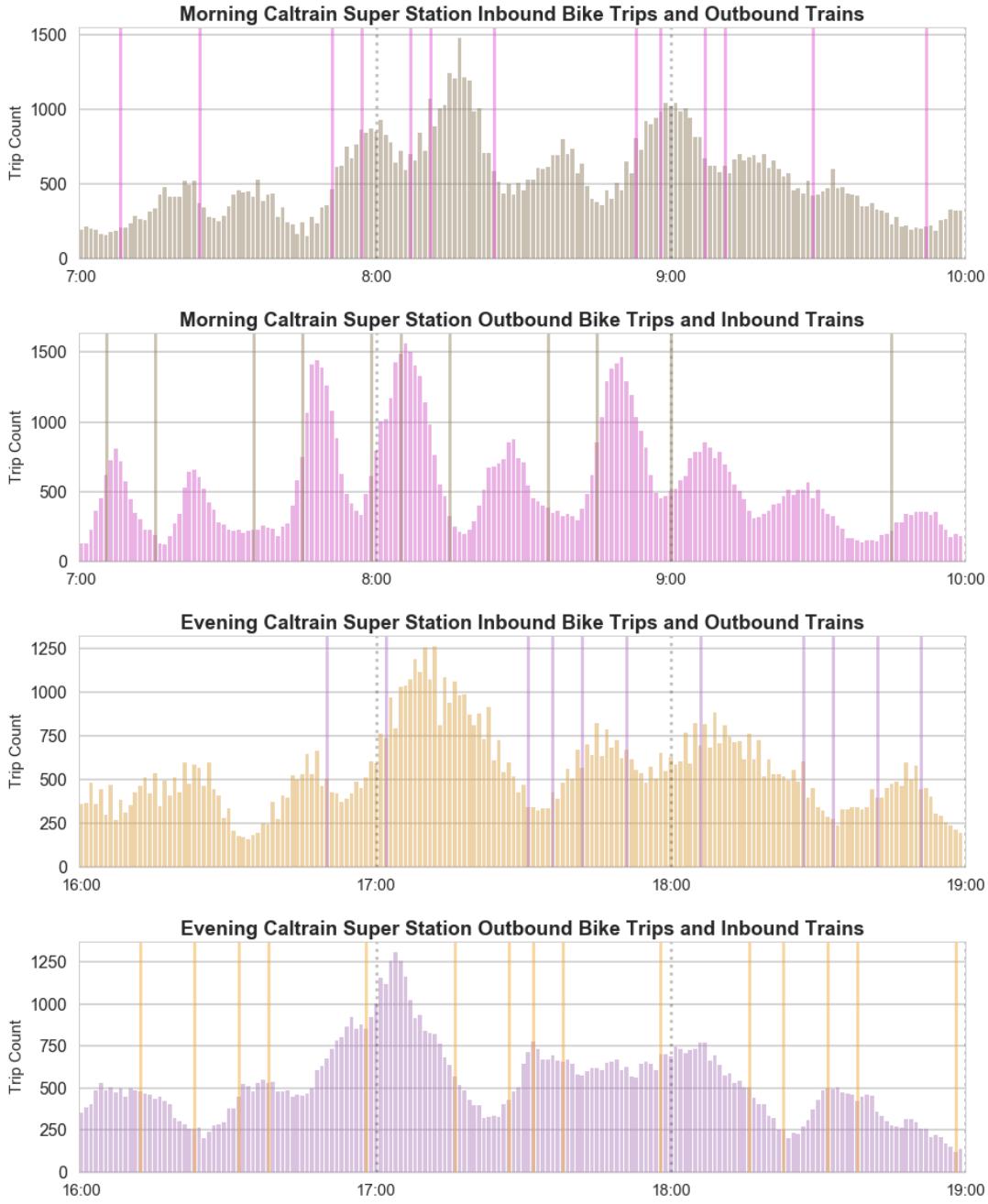




A. The Caltrain Super Station

With so many Commuter trips coming and going from stations 69 and 70, and given their direct proximity to one another, it is sensible to treat these stations as if they were one ‘Super Station’. Earlier in this analysis we looked at the minute by minute Commuter traffic in the evening and morning commuter windows. Lets inspect those again, but only using the Super Station, and given the proximity to the Caltrain Station, include an overlay of arrival and departure times for Caltrain trains into and out of the station.

Many Commuters are using the Bike Share to get to the Caltrain station before trains depart or start a bike ride shortly after a train arrives, showing that the Bike Share is a typical first or last mile effort in a rider’s daily commute.



Analysis - Weather Conditions that Impact Ridership

Ridership is going to be impacted by several different factors, such as time of day, temperature, windy, rain, availability of bikes, uphill or downhill routes, proximity to work, home, or public transit stations. We will look to determine the effect of these factors and how they help or hurt the ridership numbers.

Ideally we will find adjustable features to improve the system, such as adding docks or relocating bikes from underutilized stations to over utilized ones. For factors outside of program manager control, such as weather and geography, wether that be reducing the prohibitive factors, or taking advantage of low ridership time periods to better the system for upcoming dips or surges alike.

A.1 Recap: The Commuter Story

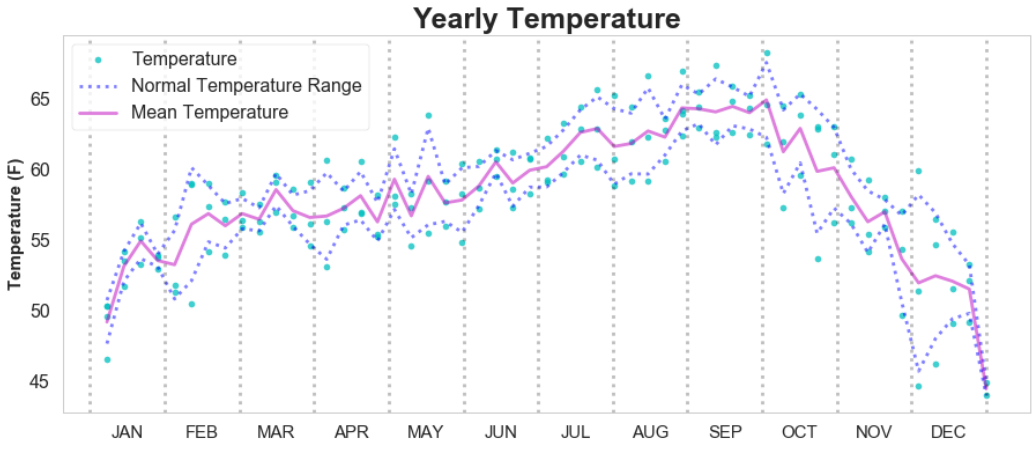
Lets recap, the first goal of this analysis is to identify the Core Users, and from there identify traffic patterns in these users. We have settled upon Commuters as our core user group and need to now identify factors that impact ridership, and how we can best abate or take advantage of dips and surges.

Two narratives of the 'Commuter' Riders could be:

Narrative A: Rider Tom boards the Caltrain near his home in San Jose, rides the Caltrain into San Francisco, disembarks at the Caltrain Station at 4th and Townsend, and then begins the last leg of his commute by renting a bike at the Caltrain super Stations, headed to work in the morning. Then taking this route in reverse in the evening to return home.

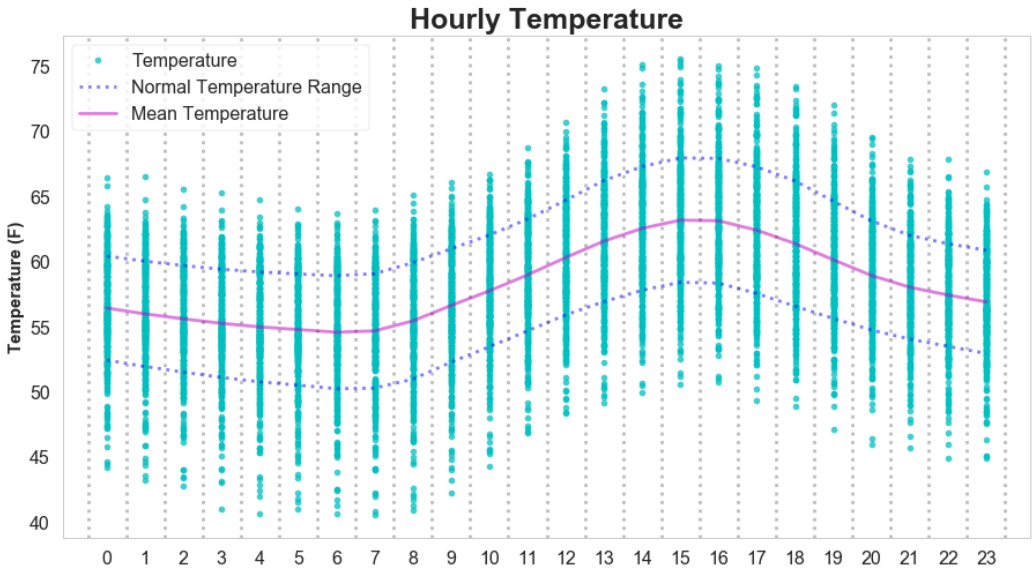
Narrative B: Rider Susan wakes up at her apartment in San Francisco and starts her morning commute by renting a bike just outside her home at Grant Avenue at Columbus Avenue. She then enjoys the downhill ride towards the Embarcadero or the Caltrain Super Station where she continues her commute on another form of public transportation.

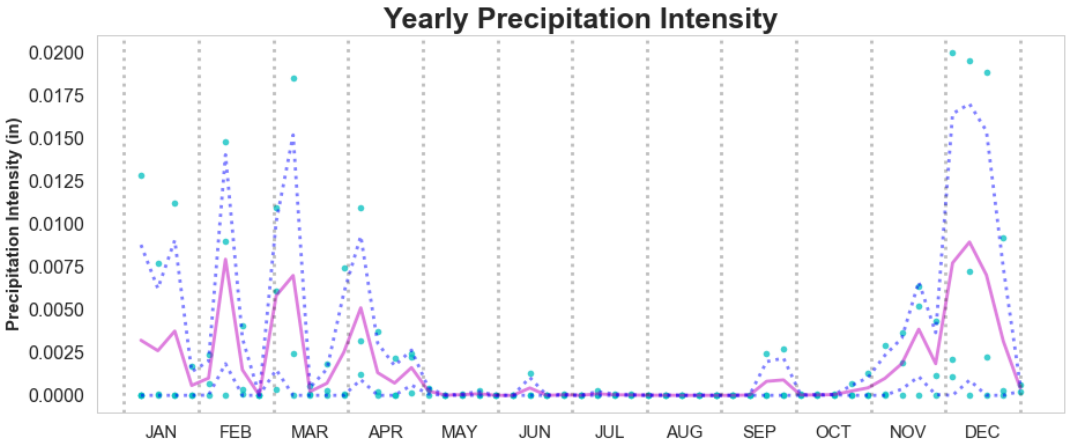
Commuters are on a fixed routine, Monday to Friday, Morning and Evening, the number of trips taken by commuters has seen few if any dips in ridership. This next analysis will focus on the impact of weather events on Commuter ridership at the Caltrain Super Station.



B.1 San Francisco Climate

San Francisco has a very temperate climate, it does not rain often, and temperatures year round are rarely outside of a comfortable 50 to 65 degrees Fahrenheit. Rain is rare, in the years included in the data the entire state of California has been in a drought, focused in the winter months it does still sometimes rain, but even that sparingly.





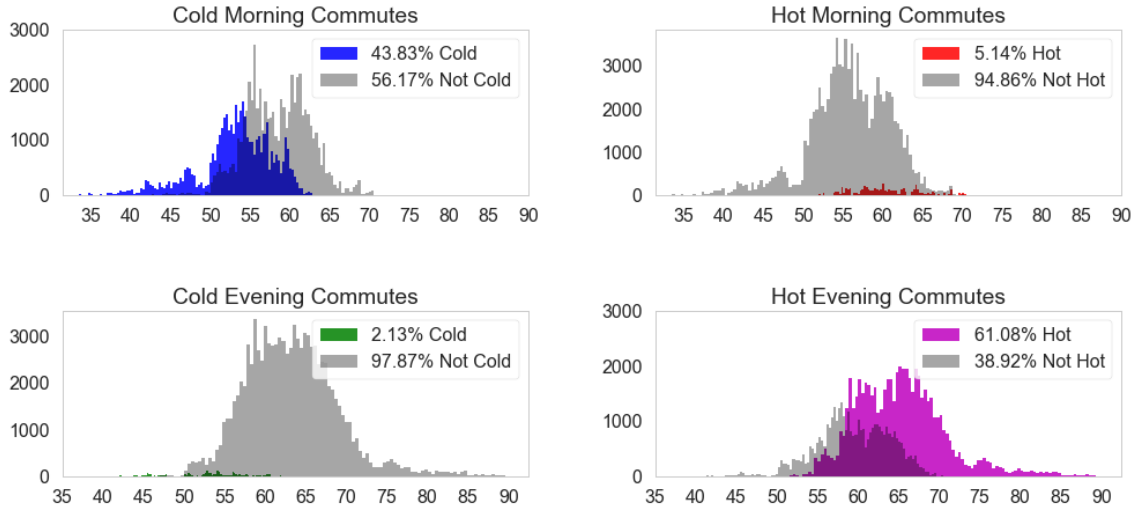
B.2 Classifying Weather Conditions

We are very specific in what days and trips classify as being abnormally hot, cold or rainy. Rainy Trips are trips started when the forecast reported was actively raining, the precipitation intensity (inches of rain per hour) was grater than zero, if the precipitation probability was greater than 50%, or if the weather forecast summary included the term ‘rain’.

Long story short, it does not rain too often in the Bay Area, with the DarkSky weather set covering a span of 26256 hours, only 2156 of these hours are classified as ‘Rainy’. Of the 25517 Morning Commuter trips, only 13559 (5.38%) and of the 237827 Evening Commuter trips, only 12947 (5.44%) of trips are classified as ‘Rainy’.
Temperatures are very stable throughout the year, so this analysis will classify trips based on the reported temperature relative to the average temperature for that day of the year.

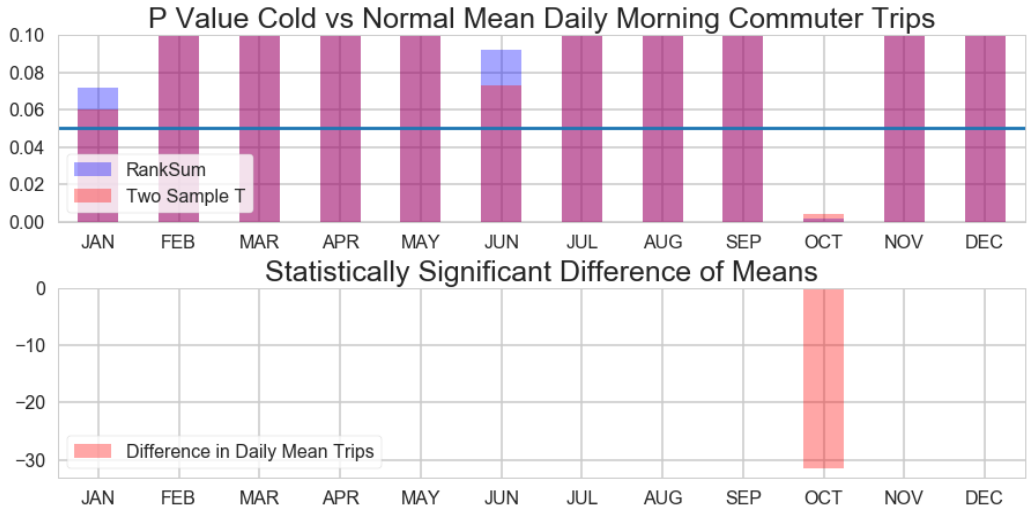
Trips are classified as ’Hot’ if the temperature at the time the trip started is more than a standard deviation above the mean for that day of the year or if the temperature is above 75 degrees. About 60% of Evening commutes are considered Hot Commutes while a negligible 5% of Morning commutes are classified as Hot Commutes.

Trips are classified as ’Cold’ if the temperature at the time the trip started is more than a standard deviation below the mean for that day of the year or if the temperature is below 40 degrees. About 42% of Morning commutes are considered Cold Commutes while a negligible 2% of Evening commutes are classified as Cold Commutes.



B.2 Is the 'Cold' a Factor in the Morning?

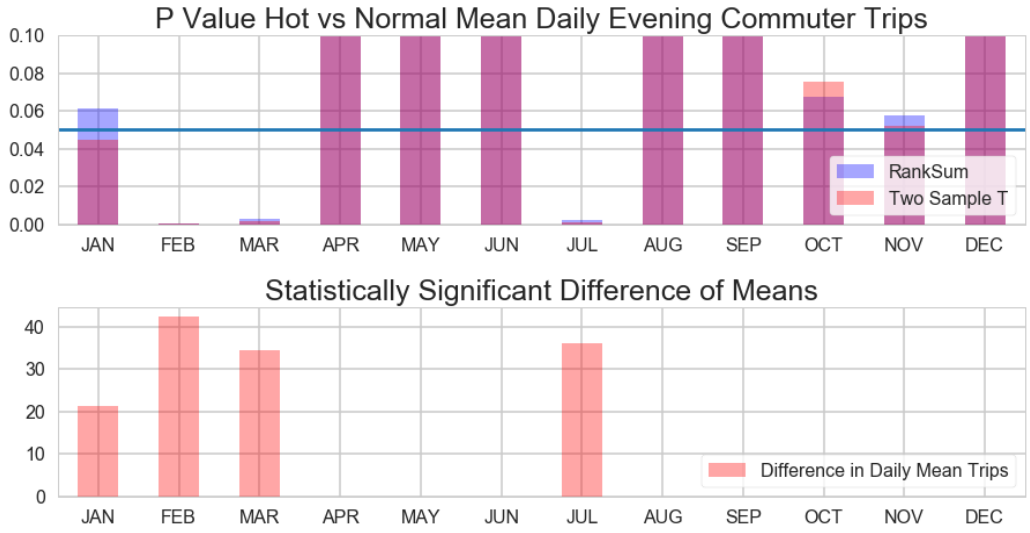
Very few Evening Commutes are considered cold, it is such a rare event that we can ignore these. Morning Commutes are split close to down the middle in cold and not cold trips. On a monthly basis, we investigate the statistical significance of the drop in number of trips on cold and not cold days using both a Two Sample T Test and the Wilcoxon Rank-Sum Statistic test.



October is the only month that sees a statistically significant dip in Morning Commuter ridership due to the cold, with each day seeing on average 30% fewer riders. This drop in ridership is likely accounted for by temperatures getting colder. Cold weather morning commutes in November and December are at similar numbers but with traffic reduced in these months due to other potential factors, the difference in mean trips per day is not statistically significant.

B.3 Is the 'Heat' a Factor in the Evening?

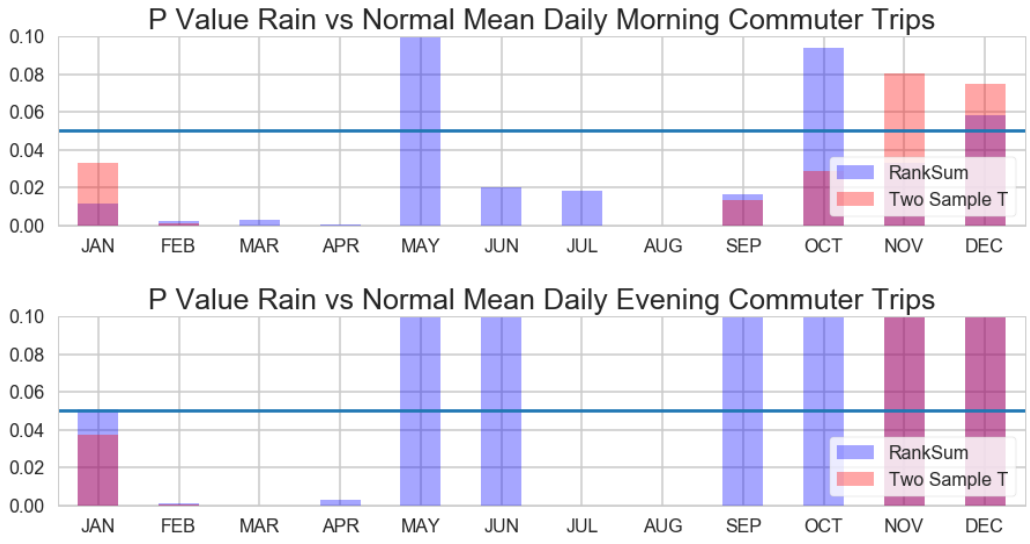
Very few Morning Commutes are considered hot, it is such a rare event that we can ignore these. Evening Commutes are split close to down the middle in hot and not hot trips. On a monthly basis, we investigate the statistical significance of the drop in number of trips on hot and not hot days using both a Two Sample T Test and the Wilcoxon Rank-Sum Statistic test.



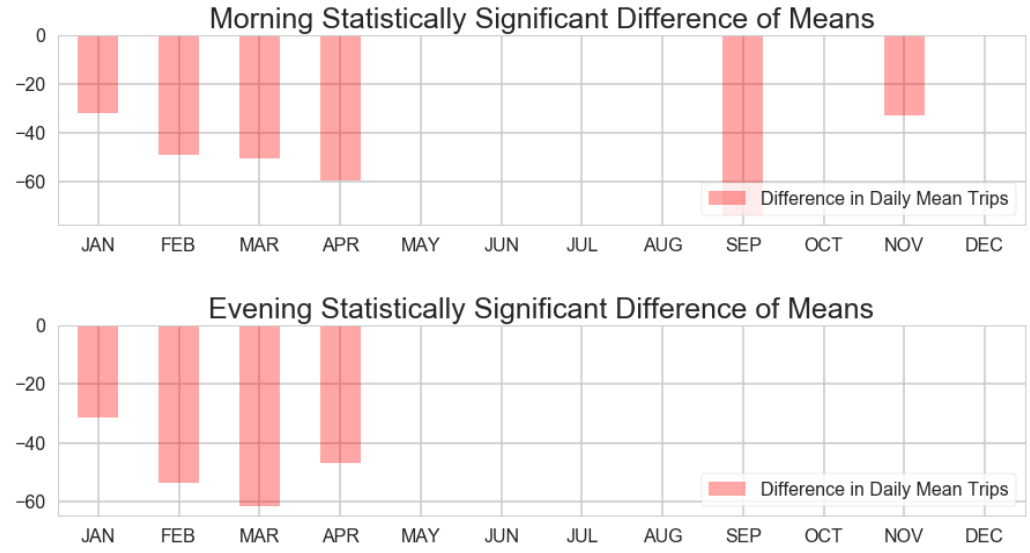
These results are a bit more interesting, it is evident that in January, February, March, and July hotter temperatures actually boost ridership! January, February, and March likely see the increased ridership when temperatures are warmer as a good change of pace from the colder and often rainier winter months.

C. No Riders on the Storm

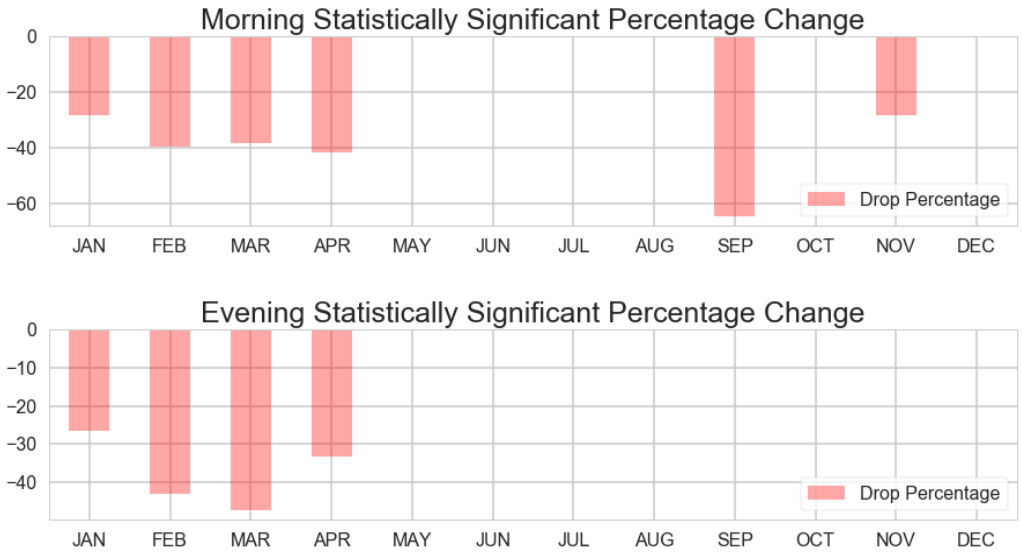
Now that we have identified the meaningful effects of temperature on ridership in each month (few as they may be) we can look at the impact of rain on morning and evening commutes. As previously noted, rain is not very common in the Bay Area, but when any amount of meaningful rain accumulates it is between late November and the end of March.



In both the Morning and Evening commute window, our p value tests show that there is only a statistically meaningful difference in ridership January through April, and in September for Morning commutes only. Special note, there have been no recorded rainy Morning commutes in August, nor any rainy Evening commutes recorded in July or August.



Between 20 and 60 fewer daily commuter trips take place when it is raining, looking at the change in percentage for each time period helps us see a bit deeper that rain reduces the number of rides by 20 to 50 percent as well. This is a big influence.



Findings Summary and Actionable Items

The key take away of this analysis is that the Missing User Group for the Bay Area Bike Share Program are Morning and Evening Commuters. These are users with a Subscription tier membership to the program paying a flat fee for unlimited rides less than 30 minutes. Additional fees are incurred for trips longer than 30 minutes, only 0.6231% of Subscriber trips are longer than 30 minutes.

The public data sets provided by the bike share program do not include any from of user identifying information that would be used to track specific user habits but with such a large proportion of trips occurring in the same time windows, with the same daily, weekly, and annual trends, a third 'Commuter' Tier membership might be worth investigating as a way to expand the program to more riders not ready to commit to 24 hour bike access.

Situated just below the current Subscription Tier price point, a new 'Commuter' membership option could be created. This tier would allow riders to use the Bike Share program on weekdays between 7am-10am and between 4pm-7pm only. This plan would also limit rides to just 20 minutes before incurring an additional fee instead of the current 30 minutes offered by the Subscription Tier plan. This 20 minute window is well above the median 7.9 minutes duration exhibited by current commuter rides.

This new tier option would attract potential users not ready to commit to paying for the unlimited plan and would also help program managers better concentrate the rides. This provides program managers with both an overnight and midday window in which bikes can be rebalanced through the system or maintenance. Eventually, these new Commuter Tier users could be pushed to a higher priced Subscriber Tier as they grow to depend on and enjoy using the bike share program.

Additional analysis, charts and supporting code can be found at gutentag.co/bayareabikeshare