Python Bridge Course Final Project
Analysis of San Francisco Public Bike Sharing Data

Authors: Shawn Kessler, Brian Schneider, Giles Ochs

Data Source:
The bulk our data was pulled from http://www.bayareabikeshare.com/open-data. We referenced data from 9-1-2014 to 8-31-2015, specifically (also known as the "YEAR 2 DATA"). When doing our analysis we ignored data outside of San Francisco. We've included most of the data in our zip file but the status csv file is too big and must be downloaded from the URL above or you can get it here: https://drive.google.com/open?id=0B9oI9y1i1FtZUU5TdXV2OUVkVkE (it's too big to view in Google Docs, so just download it after Google complains about not being able to display it).

Other data explored includes: the San Francisco Giants schedule for the '13 and '14 seasons. The US Postal Service data ended up not being useful but was explored. The data files are available in our zip file for easy reference.

The bike share data is separated into three main files: a station file which contains information about each bike station where a bike can be rented; a trip file which contains data about each trip, including what type of person rented the bike, where they started and finished, and their home zip code; and a status file that contains a minute by minute status of each station, including how many a timestamp and how many bikes for rent were available at that time.

There are two types of bike share users: subscribers, who ride bikes frequently and have long-term memberships, and customers, who ride bikes infrequently and have short-term memberships.

Data Cleansing and Merging:
The first hurdle with our data was that zip code entry did not appear to have much in the way of formatting constraints. The zip codes were "cleansed" in two ways. First, any zip code less than five digits long were removed. Second, any zip code that included a "-" were reduced to only included digits before the dash. This created bigger groups of data but a smaller number of groups.

Some initial manipulation of data was done to make column names more similar across the different files and easier to use. In addition, Shawn manually added zip codes to the stations' data file for all San Francisco stations. An effort was made to find a way to programatically go from the longitude and latitude coordinates offered in that file to a zip code but it ended up being too time consuming and updating the few rows by hand was a better use of time.

Analysis:

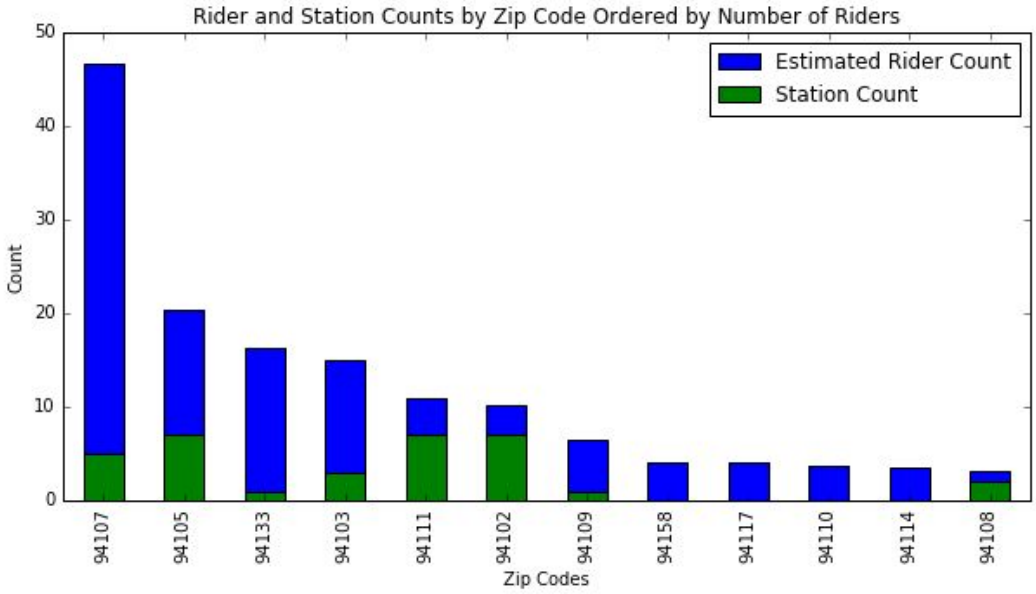
For our analysis we assumed the number of trips reported in the data was in some way proportional to the number of unique riders. None of the data has any information about unique visitors, each trip simply has a customer type and the customer's home zip code. Inasmuch as this assumption doesn't hold up, some of our analysis will not hold up either. It would have been great had this sort of information been available as it would have allowed us to make more accurate statements about the usage.

Questions posed and analyzed:

1. Which zip codes have few stations but many riders who live there?

   This question was asked with the intention of finding out which zip codes are likely good candidates for having bike rental stations added. The thought being if a person from an area is willing to use a bike in a different area then they would also likely be willing to use it in their own local area. This analysis in particular hinges on assumption #1, and we limited ourselves to the San Francisco bike stations because entering all zip codes by hand was anathema to time constraints -- and not really in the spirit of the project.

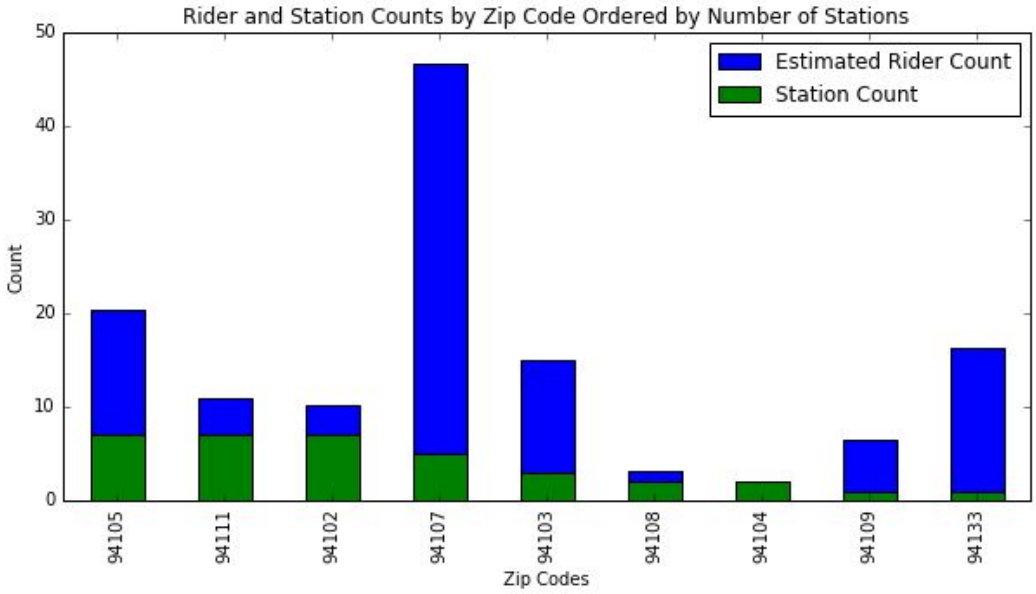
   We counted up the number of stations in each San Francisco zip code and the number of trips that were taken by individuals from all San Francisco zip codes. We then joined the two data sets together to see where there were many trips but no stations.



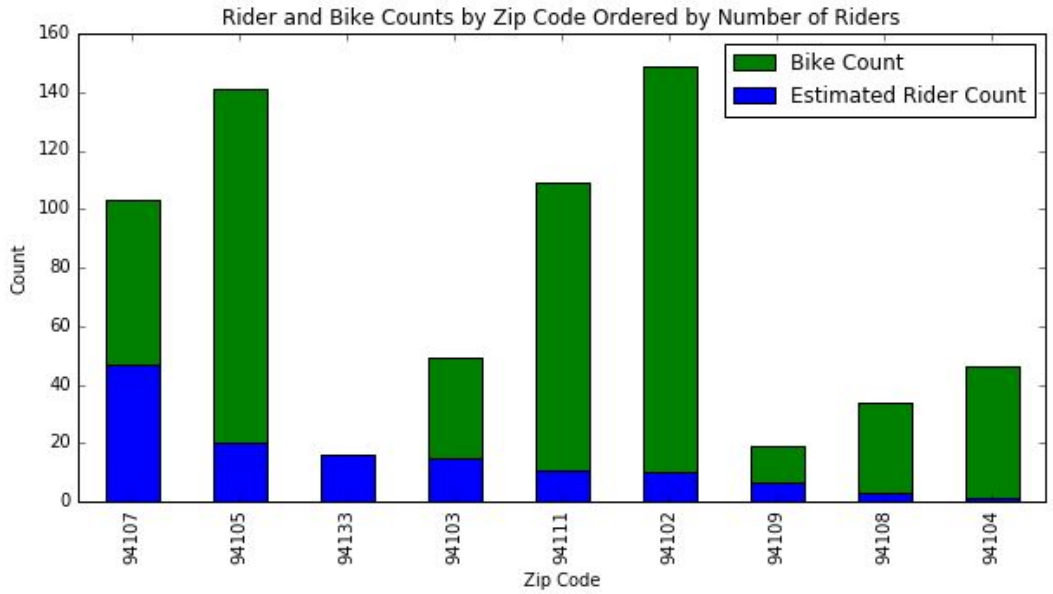
   The above figure shows that there are riders in four zip codes (94158, 94117, 94110, 94114) who use the bike share system but have no share stations in their home zip code. In addition, these four zip codes are adjacent to zip codes that already contain bike stations; this implies these riders are taking Muni a fairly short distance and then hopping on a bike. Such riders would potentially prefer to avoid Muni all together and ride directly to their destination from a station in their home zip code, if one existed.

There are other zip codes that have renters using bikes located in other zip codes but their trip totals were not as large and they weren't all adjacent to already existing zip codes with stations, so they would seem less ripe for new bike station placement.

Side note of interest: the zip code with the largest number of riders--by a very large margin--is not the the zip code with the most stations. With this in mind we also graphed the each zip code ordered by the number of stations in them:



This graph shows that the 94107 zip code has the most trips but not the most stations and that generally the two counts are not proportional. This might be an indicator that 94107 could use more stations. We will look at the number of bikes in each zip code to see if it accounts for the a smaller number of stations in an area with a lot of people who like to use the bike share system. It's important to remember that the "Trip Count" here is not the number of trips that originate or terminate in these zip codes but an estimate of the number of riders who live in these zip codes. We'll also need to explore those counts to see if this zip code sees trip usage proportional to the number of people who live in the zip code and use the bike share system.



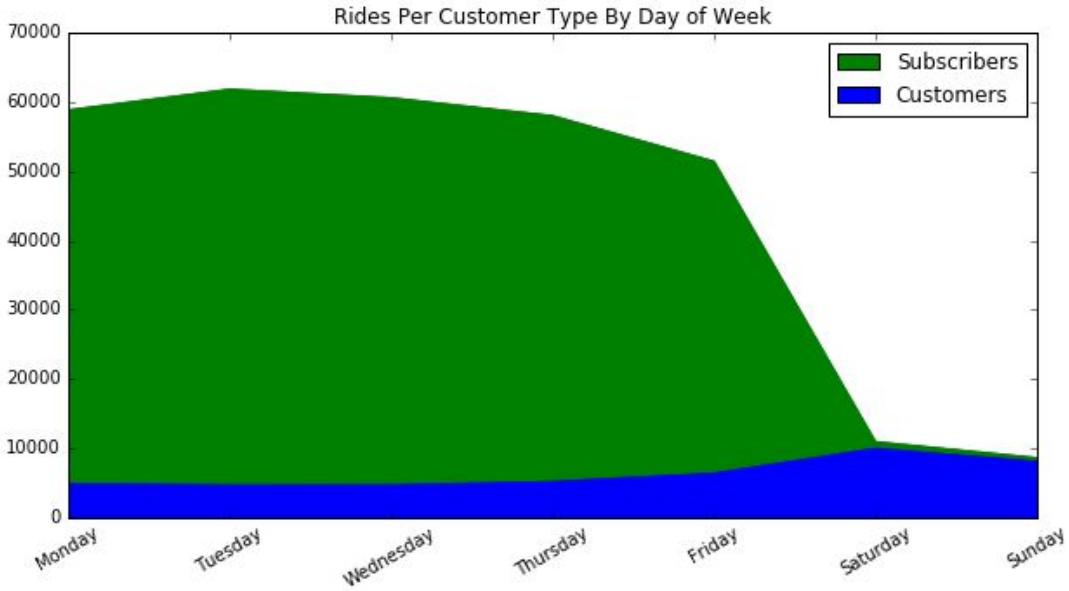
The graph above shows that the stations with the most riders don't have the most available bikes.

2. Can subscribers and customers teach us something meaningful about how people want to use the bike share service?

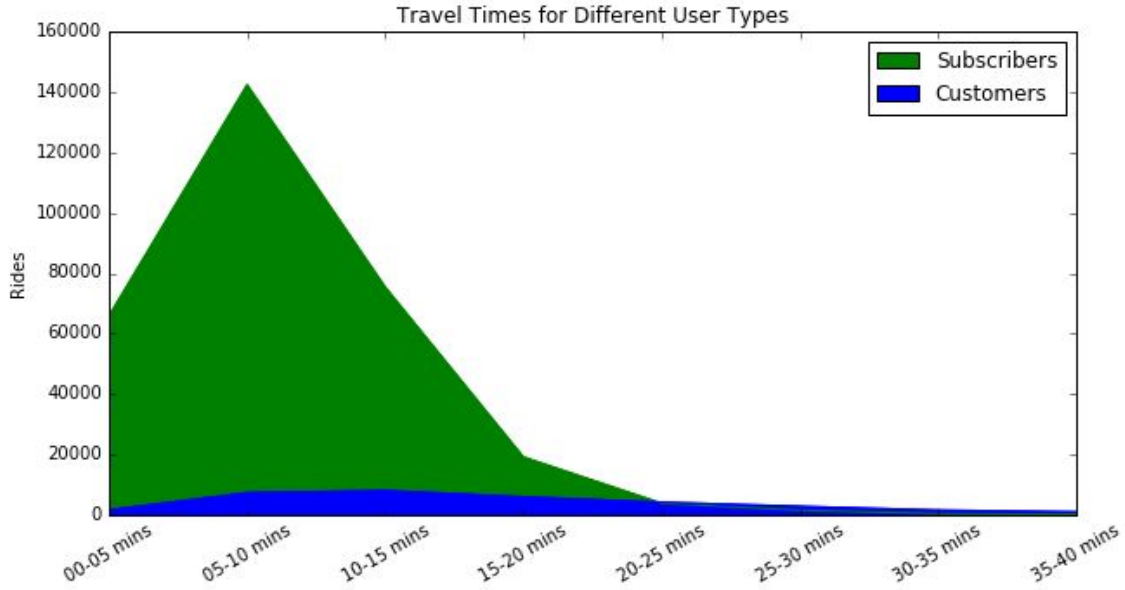
   Subscribers predominantly commute on weekdays, to and from work, most frequently at key transit stations: CalTrain and the Ferry Building. We see a slight dip in rides on Fridays, indicating that subscribers have different usage needs on said date.

   Customers, on other hand, ride more occasionally during the week and see an increase in volume over the weekend. The Ferry Building, in particular, is a major starting point for customers, indicating that on weekends, tourists, rather than commuters, have need for bicycle rentals.

   This leads us to the conclusion that the weekend is an opportune time to either fix bikes or increase business by renting bicycles to tourists. One opportunity we're particularly excited about is the prospect of supporting bike tours, to allow larger groups of tourists to explore the city.

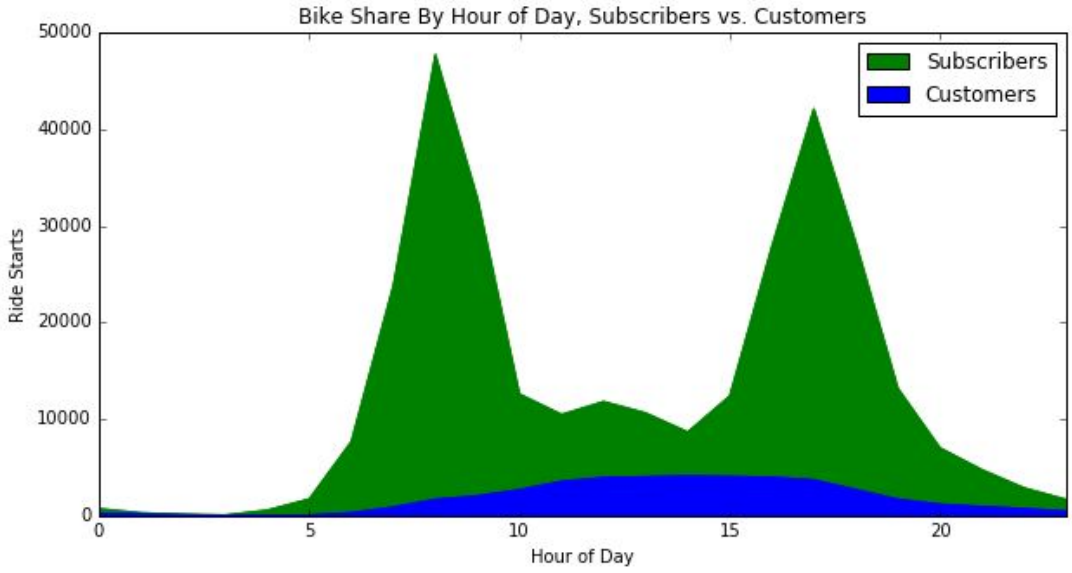


Subscribers and customers also have meaningfully different travel times. Subscribers, in particular, have very short route times, traveling from commuting hubs to the office. Customers, on the other hand, tend to take more time with each ride, indicating a less straightforward or more leisurely route. Initially we had considered attempting to better monetize bicyclists that ride over 30 minutes per ride, but that no longer seems like a worthwhile pursuit -- very few people participate for long rides.

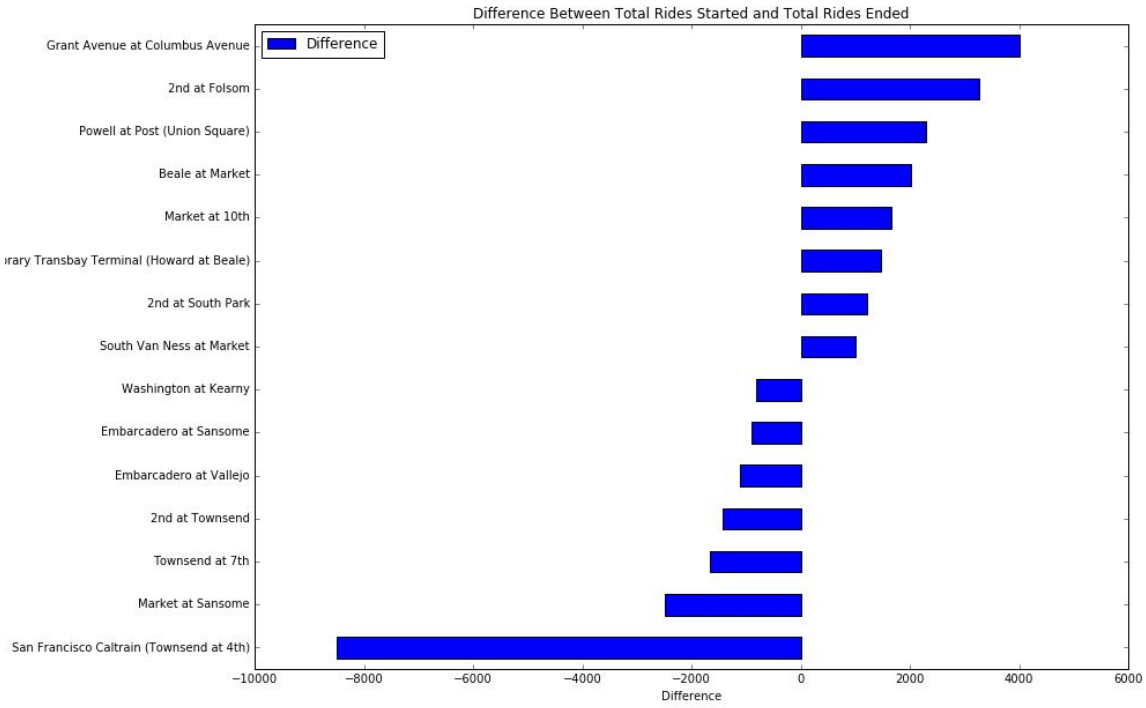


For subscribers, we see major usage spikes between 7:00 and 10:00 am, and 4:00 and 7:00 pm. This leads us to wonder how we can better use bicycles during work hours.

For customers, we see usage lift midday, with highest usage in the early afternoon.



3. How many rides start or end at given stations? Do some stations show themselves to have many rides either starting or ending at them?



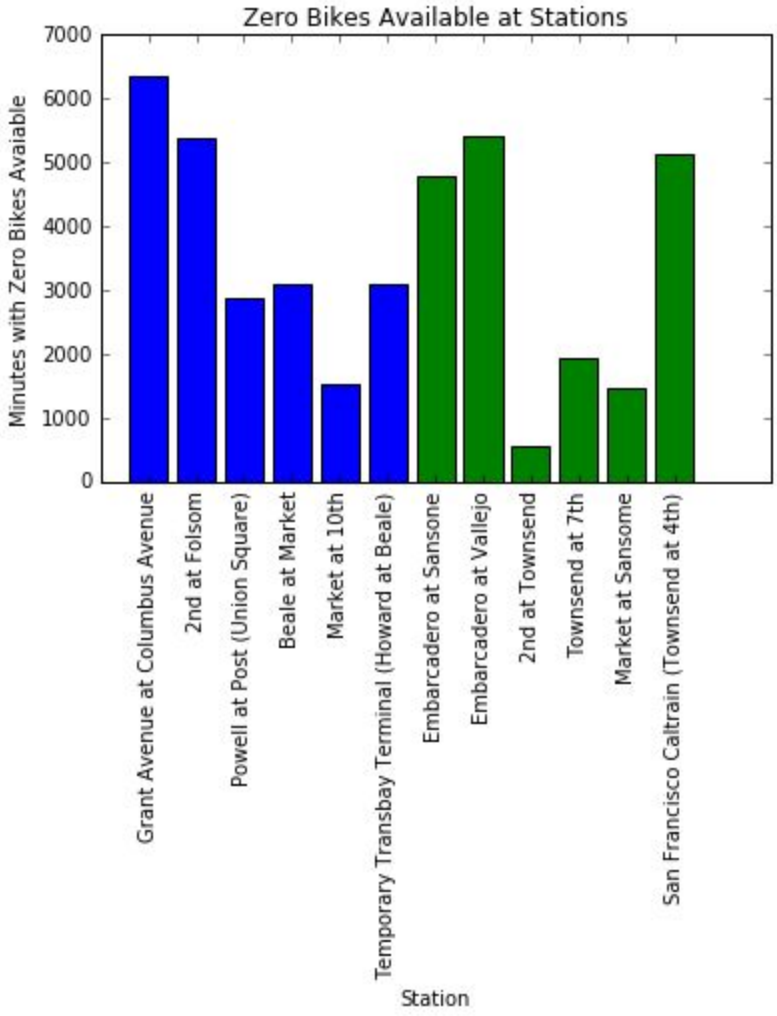
The above shows that there is significant one-directional traffic, either originating or terminating at certain stations. A positive difference implies that many more trips depart from a station then arrive there; a negative difference implies that many more trips arrive at a station then depart. Perhaps the most interesting example is the disproportionate number of rides terminating at the Caltrain Station. One explanation for this could be that people work commuters feel no need to rush to work in the morning (so they walk), but are rushed trying to make a specific train when they head home. In contrast, ~4000 more rides originated from the Grant Avenue station, which may be attributed to the station being at a high elevation. These findings raise questions of whether the company should be actively redistributing bikes to avoid pooling at certain stations, and shortages at others.

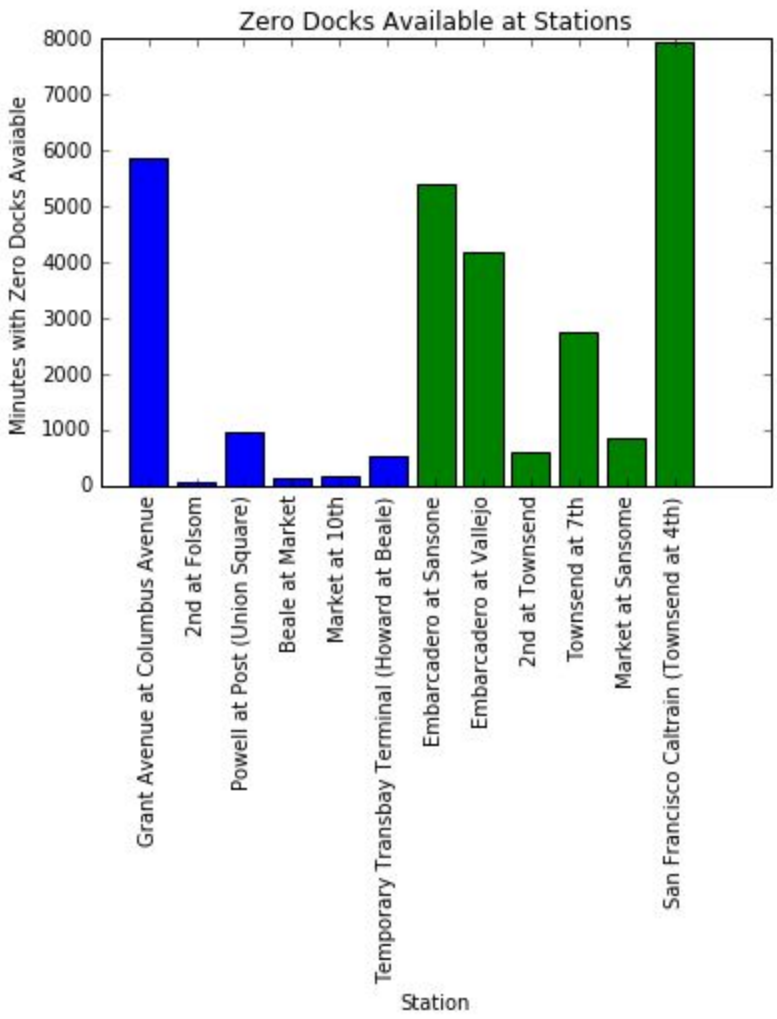
The chart above also begs the questions: (1) do stations with disproportionate outflow of bikes experience unusually high bike shortages? and (2) do stations with disproportionate inflow of bikes experience high dock shortages? The following two charts attempt to answer this.

The first chart below shows the occurrence (total minutes) of a station having no bikes available, for the stations plotted above with significant one-directional traffic. The second plot shows the occurrence (total minutes) of a station having no docks available, for the same stations. Regarding bike availability, while the stations with the highest depart-arrive difference have high levels of bike availability, so too do many of the

stations with significantly negative depart-arrive difference. By comparing this with the most common routes and station, it appears that station activity is a stronger driver of running out of bikes than is one-directional traffic.

In comparison, one-directional bike traffic did seem to have a more significant effect on dock availability. As shown in the second plot, stations with the largest negative depart-arrive difference (see: Caltrain) saw a significantly higher rate of dock unavailability than their counterparts where outflow was greater than inflow.





Summary:

1. Bike share is mostly a commuter tool for subscribers. Subscribers tend to ride Monday through Friday, and tend not to ride on weekends and holidays.
2. Usage is very predictable -- we know very well what our rider base will do in a day at any given time.
3. There's a major gap midday and overnight where bicycles are put to no use.
4. Bike share is mostly a leisure tool for customers. Customers tend to ride midday, with an increase in volume on weekends.
5. While subscribers tend to rent from commuter hubs, customers tend to rent from the Ferry Building.
6. Bike share could change their business and bike distribution from commuter stops on the weekdays to tourist destinations on weekends.
7. The Mission Bay, Dogpatch and North Beach neighborhoods are good potential expansions -- especially since highest volume of subscribers comes from Mission Bay

and Dogpatch. In addition, if the Golden State Warriors move to Mission Bay, there will be increased demand for transportation, especially on game nights. Bike share should capitalize.

8. Major transit points are where rides tend to start. Airports, other ferry terminals seem like possible bike share test locations.
9. Three stations (those w/ highest “no bicycles available” times) have lapses of ~40 mins/day, two of which are significantly uphill. These stations seem to be refilled less frequently.
10. One directional traffic trends is pronounced, but does not have a strong impact on bike and dock shortages. Instead, station ‘busy-ness’ is a much better predictor of if a station will have a bike or dock shortage.
11. Stations that have a significantly high ratio of inflow of bikes vs outflow (mainly those in really popular areas around Caltrain) experience a materially higher occurrence of dock unavailability. Thus, there is a strong argument for increasing the number of docks at those stations.