New York City vs. San Francisco Bicycle Sharing Systems

Yannie Lee, Hawina Bulcha, Andrew Lam
Python Bridge Course
Project 2
August 15, 2016

Introduction

The goal of this project is to compare bikeshare data from two cities, New York City (NYC) and San Francisco (SF), to understand how user behavior differs between the two locations. We will also be merging this information with weather data to better understand how weather factors influence ridership in each of the cities.

As a first step, we created two dataframes, one for each city. Each dataframe includes trip and weather information. We leveraged NYC bikeshare trip data and merged it with weather data collected from the government using date as a key. Since the weather data was collected through various weather stations, we decided to use the Central Park station data in our analysis. However, we used trip and weather data provided by the bikeshare organization in San Francisco, instead of merging in government weather data since the weather was readily available. To do this, we had to merge the trip data, which included station ID's, to the station data, which included station zip codes. We used Google to map these zip codes to actual city names (San Francisco, Mountain View, etc.) and then mapped those city names to the weather data.

After creating the dataframes for each city, we analyzed and mapped fields across the dataframes, standardizing the header names and contents for overlapping fields. This was particularly difficult for the weather data, since we had to understand weather metrics to accurately map the data and ensure we were doing an apples-to-apples comparison. The reason for standardizing the header names and contents is to enable us to leverage the same code when we analyze the data simply by changing the CSV files loaded.

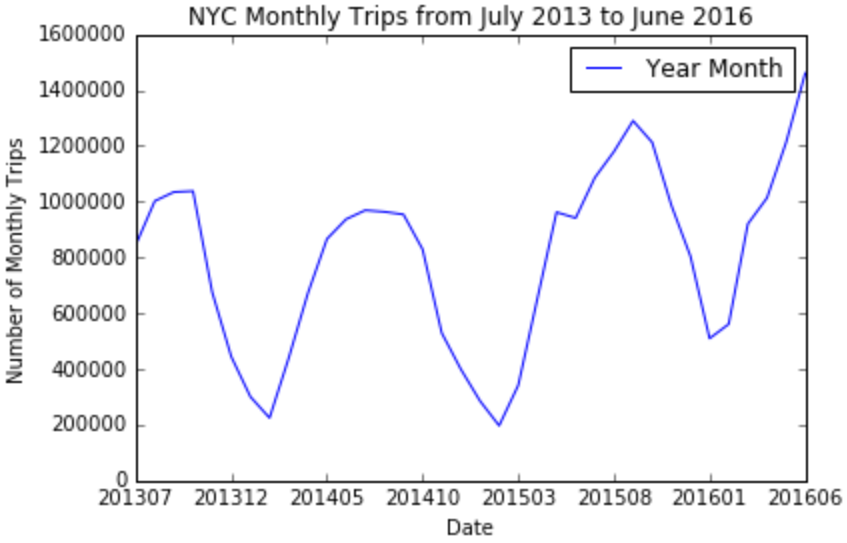
Here is a summary of the dataframe contents:

Field Name	NYC Source	SF Source
tripduration	NYC Trip Data	SF Trip Data
starttime	NYC Trip Data	SF Trip Data
stoptime	NYC Trip Data	SF Trip Data
start station id	NYC Trip Data	SF Trip Data
start station name	NYC Trip Data	SF Trip Data
end station id	NYC Trip Data	SF Trip Data
end station name	NYC Trip Data	SF Trip Data
bikeid	NYC Trip Data	SF Trip Data
usertype	NYC Trip Data	SF Trip Data
city	Added by Team	Station Data
weather key	date	date + zip
Precipitation	From Gov Weather File, PRCP	From weather file, Precipitation In

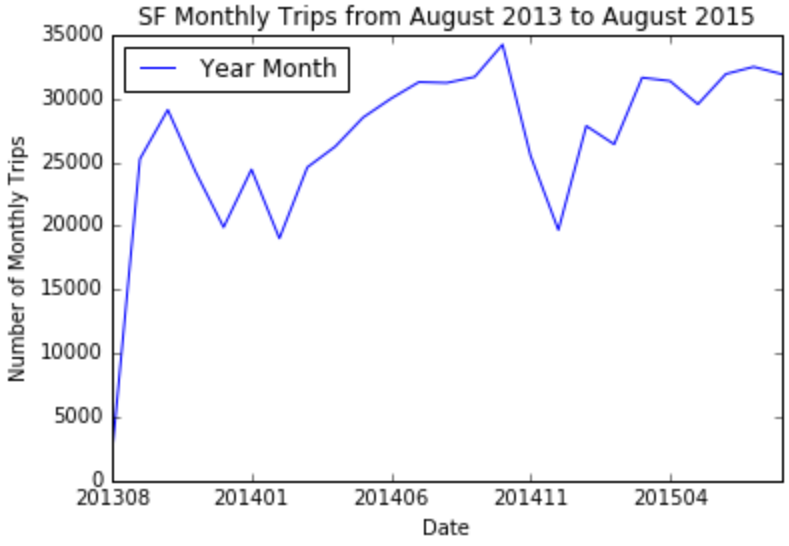
max temp	From Gov Weather File, TMAX	From weather file, Max TemperatureF
min temp	From Gov Weather File, TMIN	From weather File, Min TemperatureF
avg wind	From Gov Weather File, AWND	From weather File, Mean Wind Speed MPH

Overall Ridership

To begin our analysis, we wanted to explore the overall trend in ridership by city. Our hypothesis was that ridership would go down in winter months, which was confirmed by the New York data:



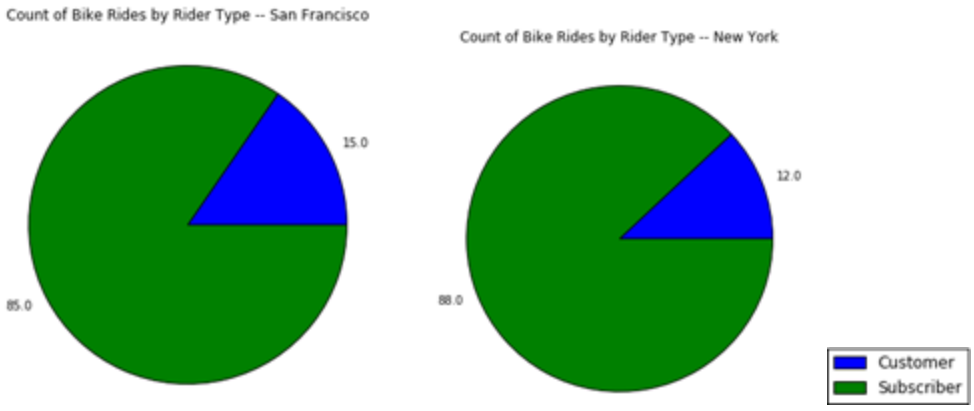
As you can see, ridership sharply falls as winter approaches with the troughs in the January - February timeframe. The peaks are generally in the summer timeframe with the exception of 2015 when ridership continues to rise until mid-fall.



The San Francisco data, however, is much more uneven. Although there are dips in ridership during the winter months, the graph does not display a smooth sine curve like the New York data, suggesting that the fluctuations in ridership cannot be fully explained by weather. This, in large part, makes sense since the variation in New York weather is much more severe than San Francisco.

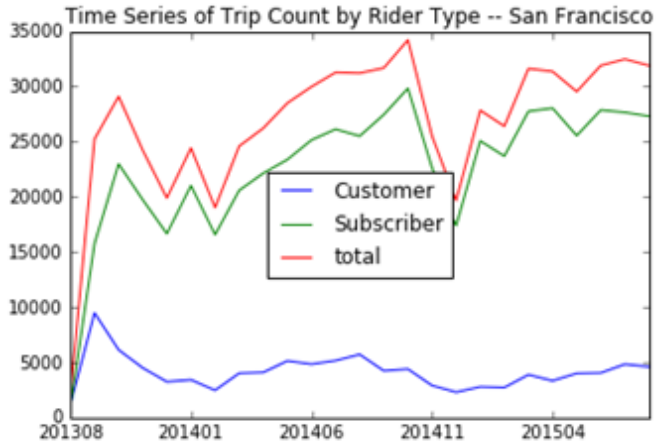
Rider Types

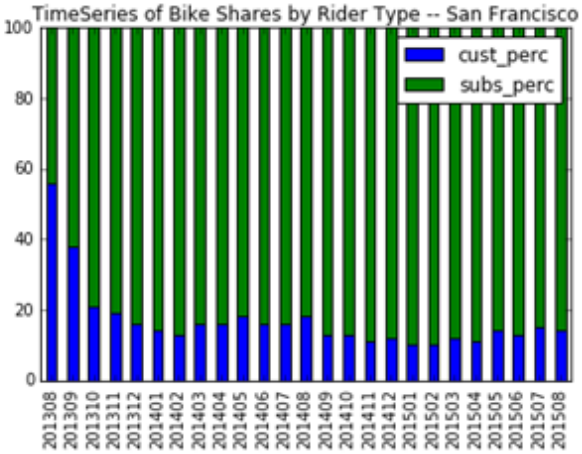
A second question to explore in the New York and San Francisco Bike Share Data is the distribution of Rider Types. Both datasets feature two types of riders – subscribers who have an annual pass and customers. Whereas the data is captured in the form of Bike Rides, without a unique reference to identify each customer or subscriber. So, the initial question of "How many Subscribers versus Customers use Bike share?" should be reframed as: What proportion of total count of trips taken overall did each Rider Type contribute to?



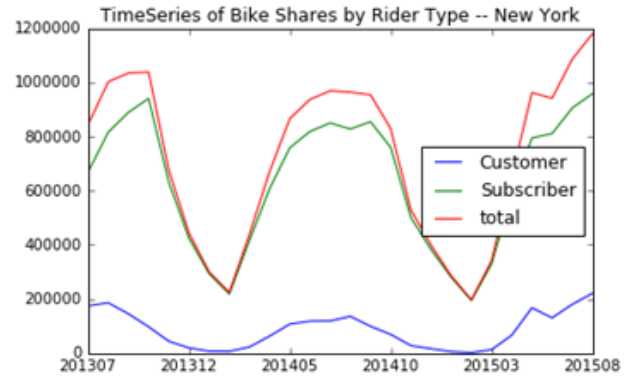
As we would expect, Subscribers take more Bike Trips compared to Customers in both cities. In NYC the proportion is 88% to 12% and in San Francisco, it is 85% to 15%.

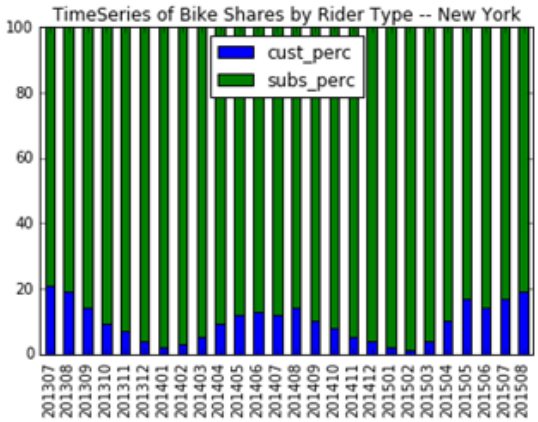
A second question to consider regarding Rider Types is how does the above trend of trips taken by customers to subscribers changes over time.





Looking at the trend over time for San Francisco reveals two facts. First, the line graph (left), shows that customer bike trips are slightly more sensitive to seasonal changes. Subscriber bike trips however follow a more irregular pattern. And as they comprise 85% of the observations recorded, the customer trend is buried in the initial overall ridership analysis. Second, the stacked bar graph to the right shows that in the first two months after the program was launched, the proportion of Customer bike rides was comparable to Subscriber bike rides, dropping down to its regular levels in the third month. As we are looking at trips and not rider population, this could mean either that the number of customers in the initial months was proportional to the number of Subscribers or that Customers took more frequent bike share trips overall in the initial stages.





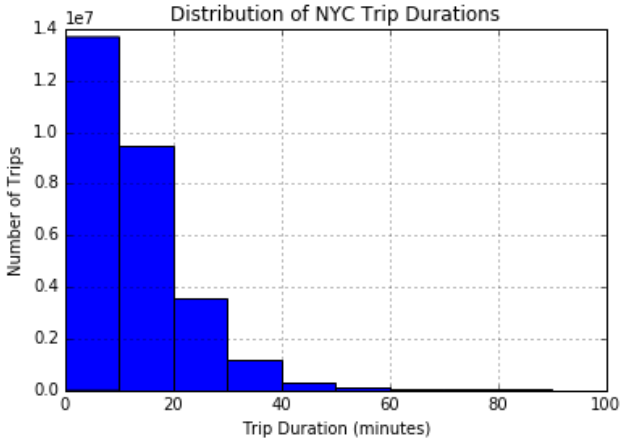
The above observations however do not hold for New York. New York Subscriber trip peaks and troughs are more pronounced, compared to their San Francisco counterparts. In addition, the customer trip count to subscriber trip count ratio in the first couple of months of the program are similar to their proportion in overall data set.

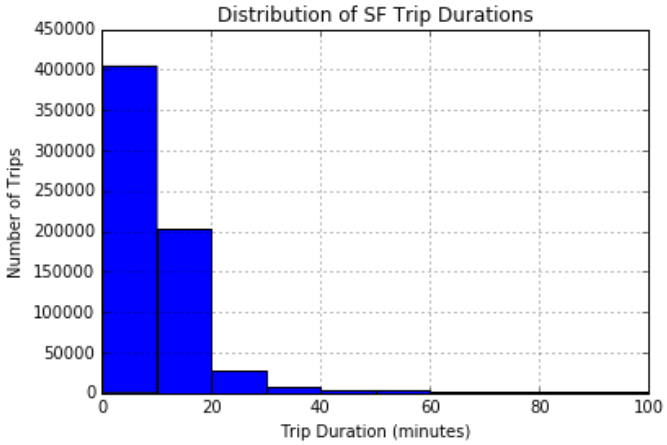
Trip Times

Next, we wanted to explore the distribution of trip times. The distributions for New York and San Francisco appear to be somewhat similar with fewer trips as trip duration increases. For both cities, the largest bin is in the 0-10 minute range. This might suggest that riders use the bike sharing system to cover distances that are too far to walk, but short to bike.

Furthermore, the histograms confirm that riders are sensitive to paying additional fees. In New York, additional fees are assessed for rides greater than 30 minutes if you have a Day Pass or 3-Day Pass. For riders with an annual membership, fees are assessed for rides greater than 45 minutes. Thus, there are very few trips above 45 minutes in duration.

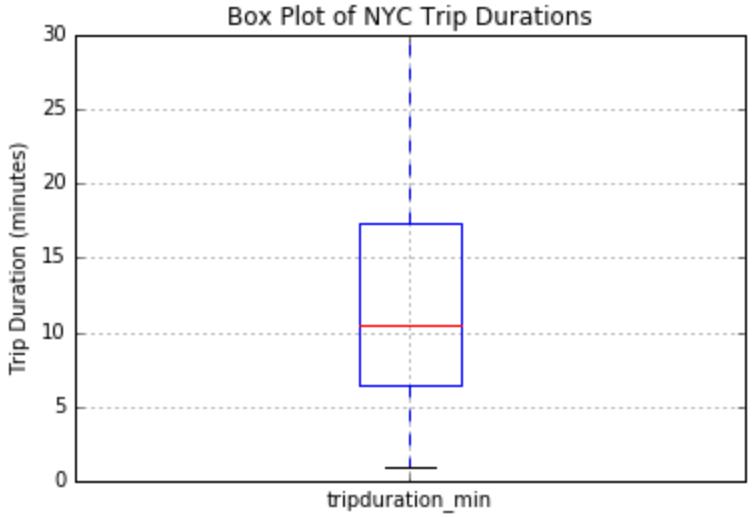
In San Francisco, the fee structure is somewhat similar with overtime fees assessed on trips greater than 30 minutes, regardless of your membership type. As you would expect, there is a sharp drop off in trips greater than 30 minutes.

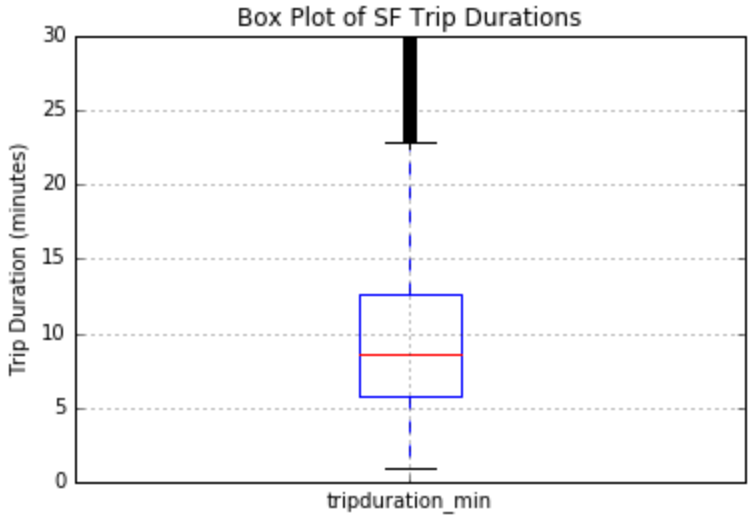




Note: The histograms were limited to 0-100 minutes for readability

In addition, we also created box plots to provide an understanding of the interquartile range. The New York data has an interquartile range of approximately 6 minutes to 17 minutes and a median just above 10 minutes. Similarly, the San Francisco data has an interquartile range of approximately 6 minutes to 13 minutes and a median of 9 minutes. This confirms what we learned from the histograms - that the vast majority of people take trips of less than 30 minutes.

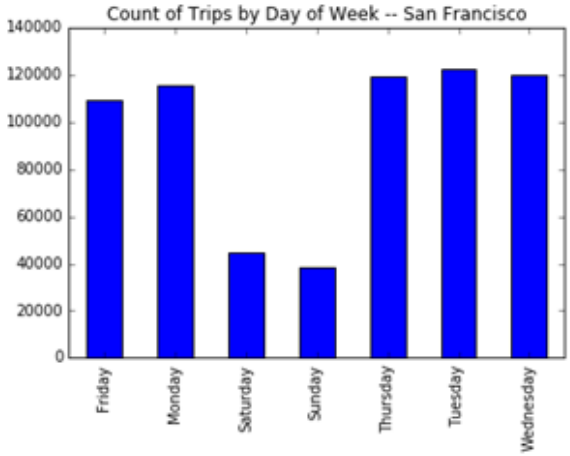


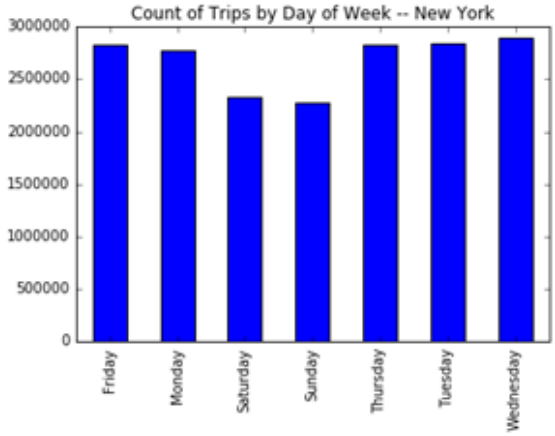


Note: The boxplots were limited to 30 minutes for readability

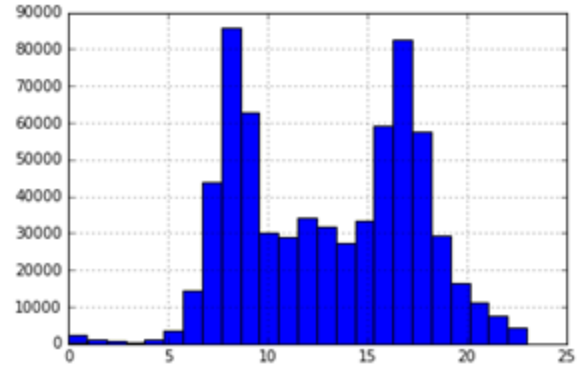
Time of Day Analysis

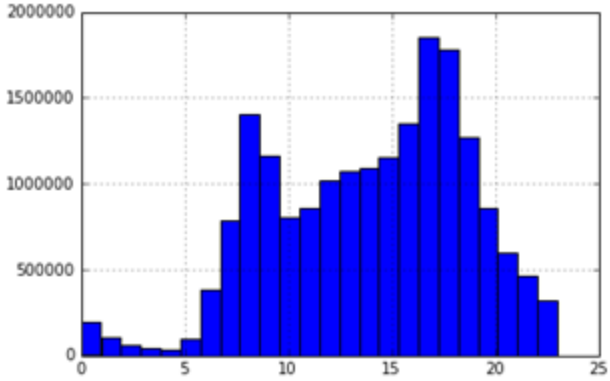
Building on the idea of variation from one month to the next, another point to consider is the change in count of trips when the unit observed is the time of the week or time of the day.





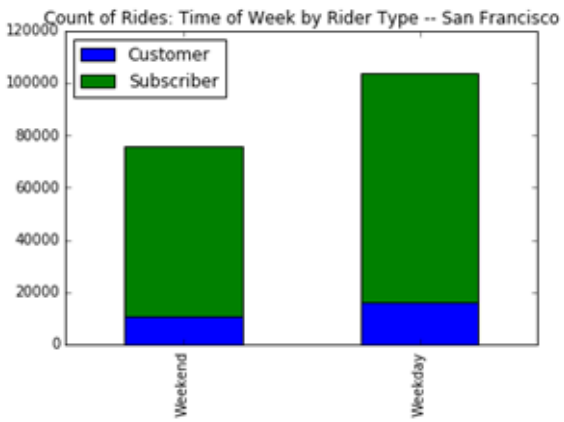
From the above figures, it is apparent that the count of bike trips stays relatively the same throughout the weekdays in both New York and San Francisco but drops on weekends. The drop is much exaggerated in San Francisco. This could be an indicator that commutes make up a higher proportion of bike trips in San Francisco compared to New York. The count of trip times by time of day analysis (below) supports this hypothesis.

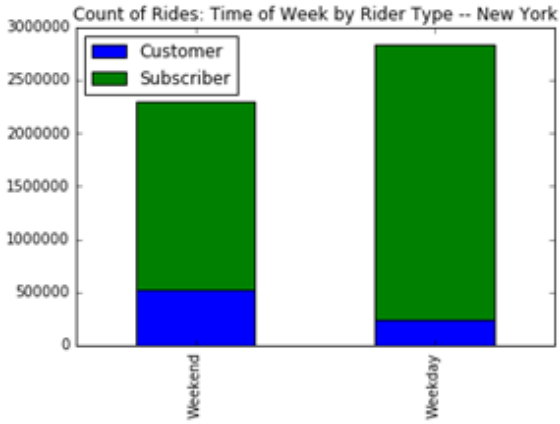




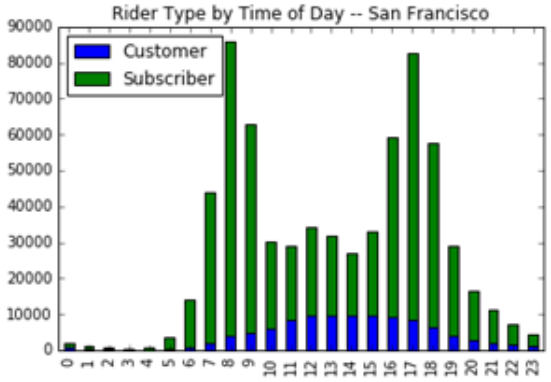
The San Francisco plot (left) is bimodal in nature, peaking markedly during commute times (5am-9am and 4pm-6pm). A similar trend is visible in New York, although much less pronounced. Commute times have a less profound effect on the total number of bike trips taken.

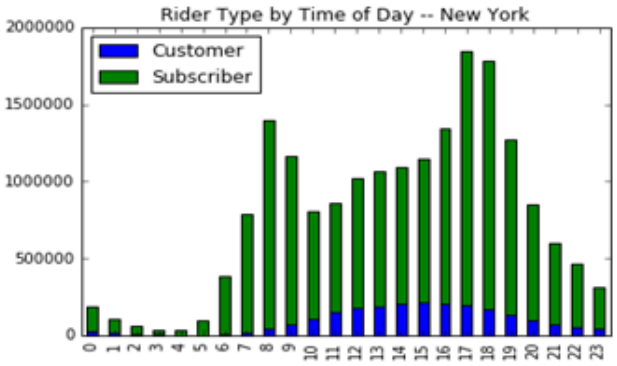
To add another layer of complexity, we ask how the proportion of bike trips taken over weekdays or weekends or at different times of the day vary according to the type of Bike Rider in the dataset.





From the above two graphs, there are more Customers coming out on a typical weekend compared to a weekday in New York. In San Francisco, the trend is reversed with fewer Customers coming out on a weekend. This explains why the dip when going from weekday to weekend is sharper in San Francisco to New York.



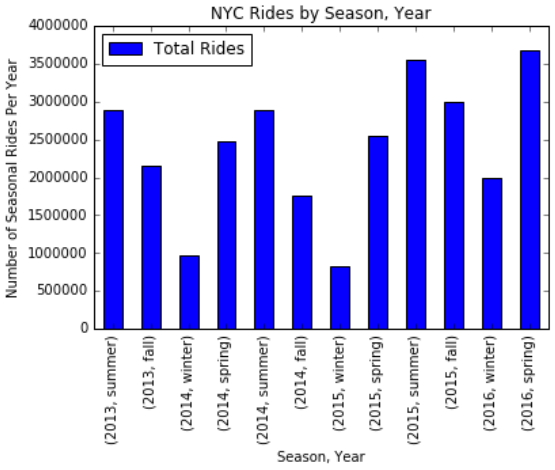


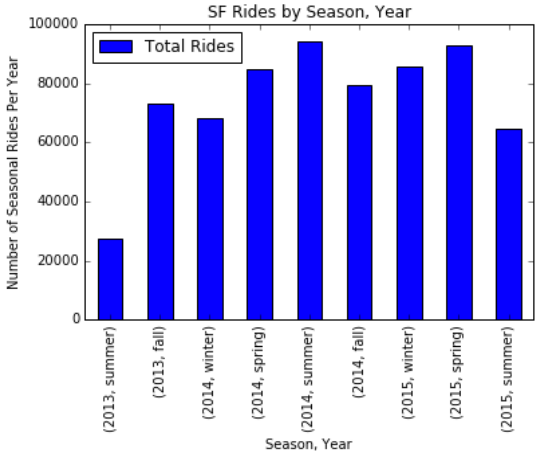
In addition, breaking out the time of day analysis by Type of Rider shows that Customer shows a regular bell-shaped curve in both cities, peaking in the middle of the day. The bimodal variation is all coming from Subscriber usage during peak commute times.

Impact of Seasons and Weather Metrics

We wanted to understand how weather impacted rider behavior, and how sensitive riders are to weather. To do this, we took a look at effect of seasons on ridership as well as rider sensitivity to three weather metrics: (1) temperature (minimum and maximum, degrees fahrenheit), (2) precipitation (inches), and (3) average wind speed (MPH).

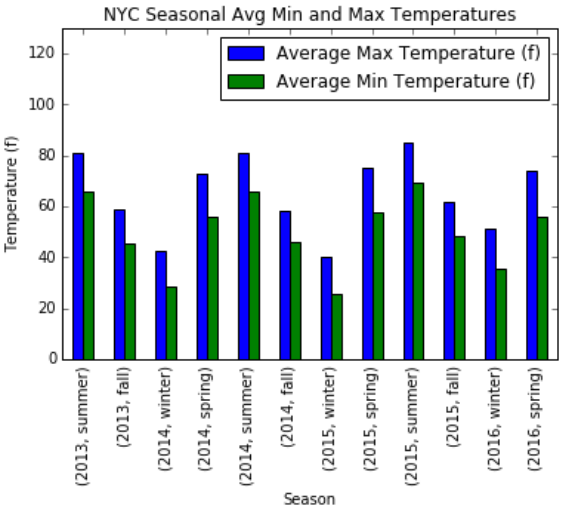
We first took a look at the effect of seasons on the total number of rides per season. We grouped every quarter consisting of 3 months into a season, starting with January through March as winter. Data for SF spanned from August 29, 2013 through August 31, 2015 while data for NYC spanned from July 2013 through June 2016. As such, it is important to keep in mind that data for summer 2013 and summer 2015 is incomplete in SF since summer 2013 only includes September data and summer 2015 only includes July and August data. Included below are two graphs:

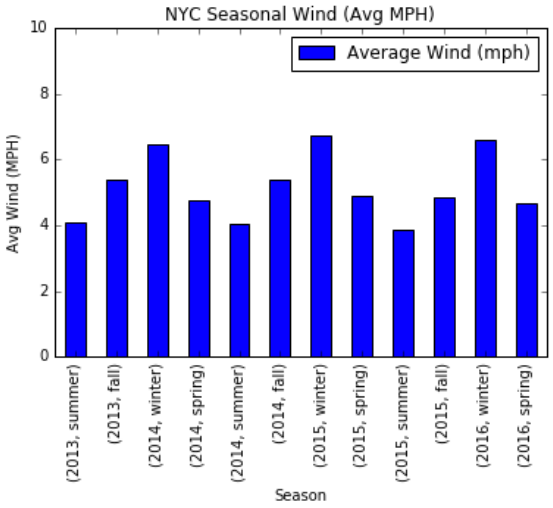


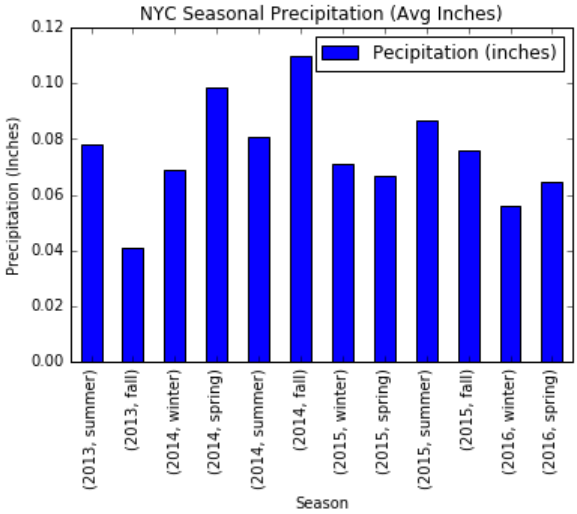


It is clear that NYC ridership is seasonal and heavily dependent on time of year, with the lowest number of rides in the winter and the highest number of rides in the summer. However, in SF, although data is missing for some data points, ridership doesn't seem to be seasonal. The number of winter rides is lower than the fall from 2013 to 2014, but there are more rides in the fall than winter in 2014 to 2015.

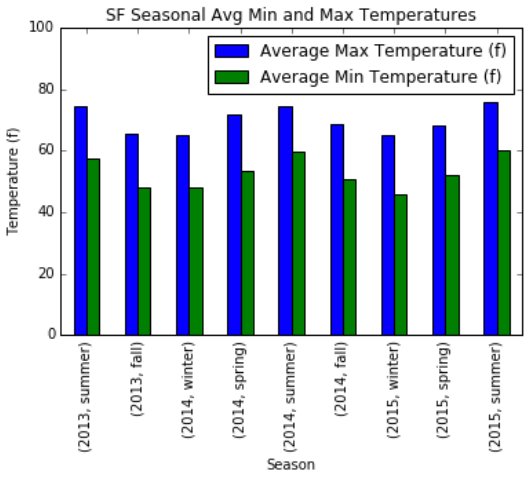
To dive deeper into drivers of rider behavior, we took a look at how specific weather metrics varied through the seasons for each city. We looked at minimum and maximum temperatures, average wind speed, and average precipitation by season for each city.

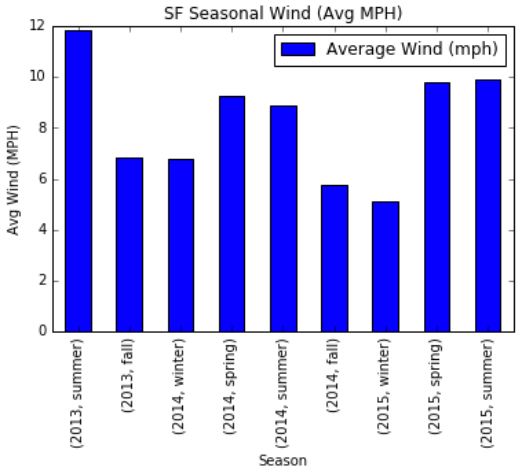


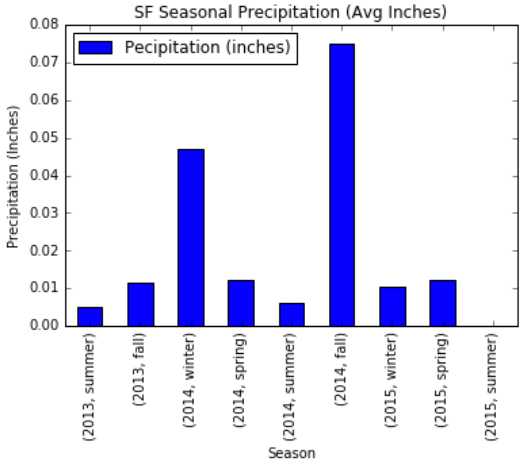




In New York City, temperature and wind speeds were seasonal while precipitation was not for the time period considered. Considering the NYC Rides by Season, Year graph in conjunction with the temperature and wind speed graphs, it seems that riders ride more often in warmer climates and prefer less wind. However, since there is no significant seasonal trend with precipitation, it is difficult to discern how it impacts rider behavior.





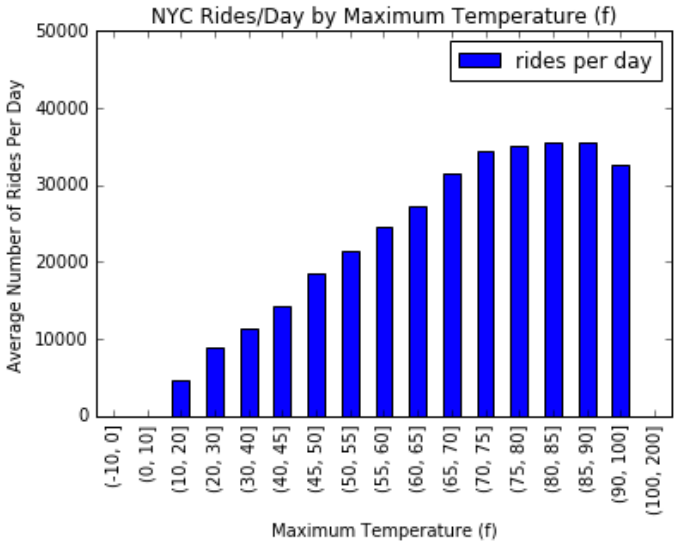


In SF, when considering the SF Rides by Season, Year graph in conjunction with the average precipitation graphs, it is clear that there are less rides in seasons where there is more rain. Precipitation spiked in the winter and fall of 2014, which is when SF experienced lower number of rides per season. However, unlike NYC, it was difficult to observe any trends for temperature and wind speed.

We weren't satisfied with the analysis and conclusions we were able to draw, however, and wanted to get a better sense of how specific changes in weather metrics would impact ridership. For example, how would a 10 degree fahrenheit increase in minimum temperature impact ridership? To answer these questions, we looked at a different ridership metric, average number of rides per day, so we could better understand how ridership changes based on specific metrics.

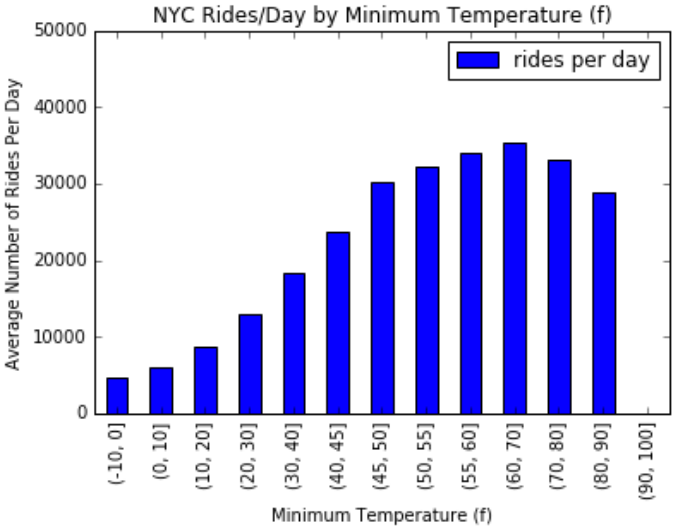
Temperature

In NYC, people seemed to prefer warmer weather, but weather that was not too hot. When considering the maximum temperature of the day, rides per day seemed to increase until it reached 70 degrees. Rides per day were pretty flat from 70-90 degrees, and then decreased when it was too hot, anything above 90 degrees. We saw a similar upward and then downward trend for minimum temperatures.



max_temp	count_nonzero	rides per day
(-10, 0]	0	NaN
(0, 10]	0	NaN
(10, 20]	7	4560.142857
(20, 30]	39	8760.051282
(30, 40]	126	11435.761905
(40, 45]	67	14166.686567
(45, 50]	60	18430.666667
(50, 55]	77	21442.246753
(55, 60]	81	24483.876543
(60, 65]	92	27220.315217
(65, 70]	82	31531.000000
(70, 75]	100	34344.300000
(75, 80]	97	34990.896907
(80, 85]	139	35627.122302
(85, 90]	99	35462.212121
(90, 100]	26	32591.500000
(100, 200]	0	NaN

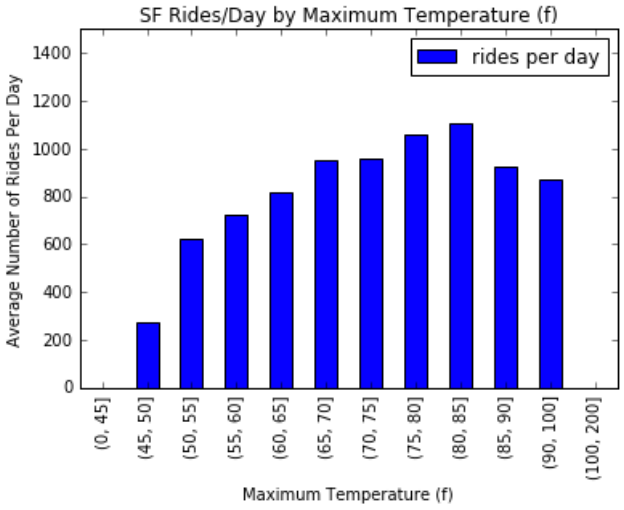
count_nonzero values is the sample size of each bucket



min_temp	count_nonzero	rides per day
(-10, 0]	1	4653.000000
(0, 10]	18	6129.055556
(10, 20]	55	8752.618182
(20, 30]	115	12935.391304
(30, 40]	172	18240.802326
(40, 45]	101	23713.603960
(45, 50]	88	30150.852273
(50, 55]	106	32256.216981
(55, 60]	88	34007.500000
(60, 70]	241	35374.547718
(70, 80]	103	33106.514563
(80, 90]	4	28906.750000
(90, 100]	0	NaN

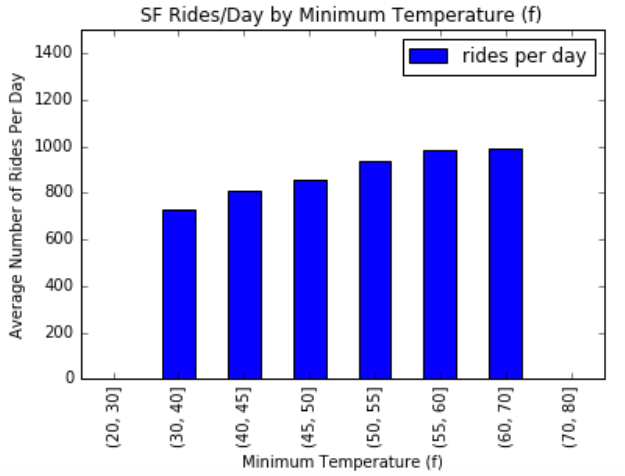
count_nonzero values is the sample size of each bucket

Although the trends in SF were not as strong, possibly due to a smaller sample size and a lack of data for temperatures that were outliers for the SF climate, the rides per day trends mirrored those in NYC. There were increasingly more rides per day in SF up to maximum temperatures of 80 degrees and then a decline thereafter. The minimum temperatures showed only an increasing trend, but never a downward trend, possibly because the climate in NYC gets warmer than SF.



max_temp	count_nonzero	rides per day
(0, 45]	0	NaN
(45, 50]	1	274.000000
(50, 55]	14	618.928571
(55, 60]	57	723.263158
(60, 65]	162	814.320988
(65, 70]	177	949.163842
(70, 75]	194	957.463918
(75, 80]	80	1060.350000
(80, 85]	29	1108.206897
(85, 90]	12	921.250000
(90, 100]	7	871.857143
(100, 200]	0	NaN

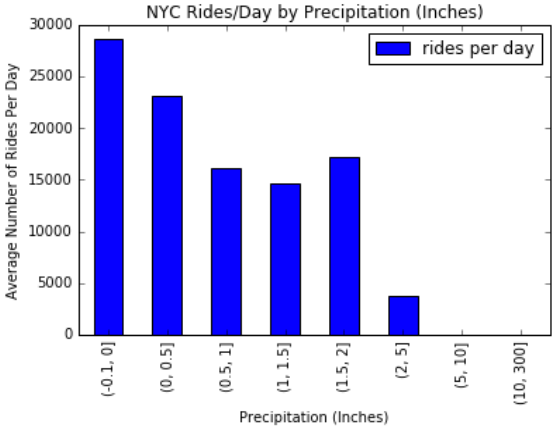
count_nonzero values is the sample size of each bucket



min_temp	count_nonzero	rides per day
(20, 30]	0	NaN
(30, 40]	33	726.666667
(40, 45]	80	808.850000
(45, 50]	153	856.150327
(50, 55]	220	937.077273
(55, 60]	153	986.183007
(60, 70]	94	991.882979
(70, 80]	0	NaN

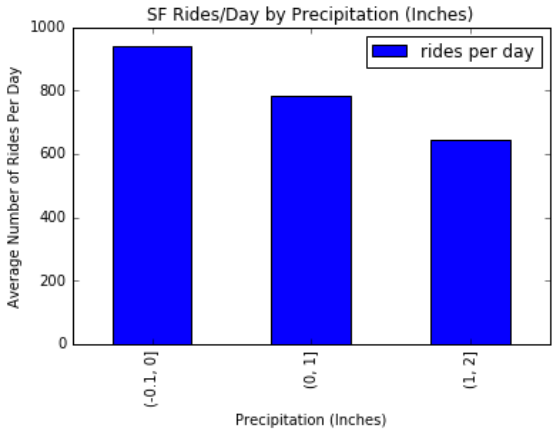
count_nonzero values is the sample size of each bucket

Precipitation



Precipitation	count_nonzero	rides per day
(-0.1, 0]	740	28642.191892
(0, 0.5]	274	23098.065693
(0.5, 1]	40	16178.575000
(1, 1.5]	21	14704.714286
(1.5, 2]	14	17212.928571
(2, 5]	3	3791.666667
(5, 10]	0	NaN
(10, 300]	0	NaN

count_nonzero values is the sample size of each bucket. Last 2 buckets don't have significant n sizes and should be weighted lightly.

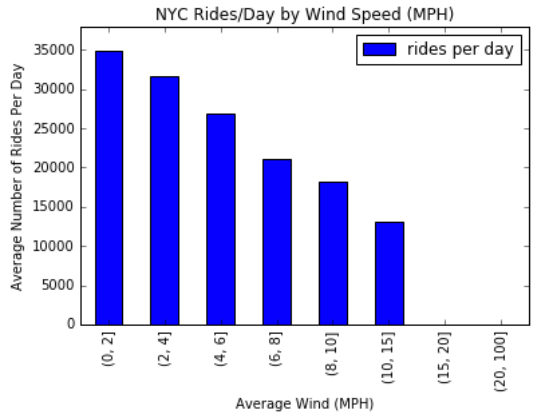


Precipitation	count_nonzero	rides per day
(-0.1, 0]	622	939.250804
(0, 1]	107	782.233645
(1, 2]	3	646.333333

count_nonzero values is the sample size of each bucket. Last bucket don't have significant n sizes and should be weighted lightly.

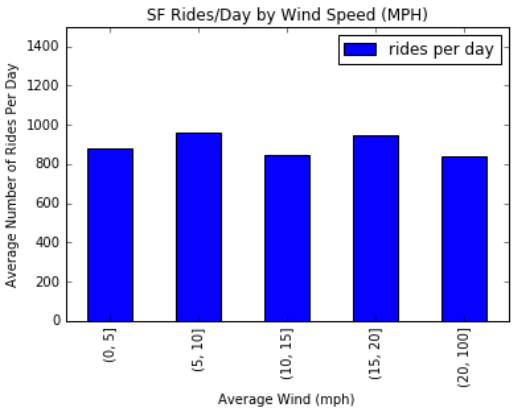
Both NYC and SF Riders are sensitive to precipitation, with decreasing rides per day as there is more precipitation per day. Looking at the data in this way allowed us to tease this insight out in NYC, where we were unable to draw this conclusion when looking at general seasonal weather patterns.

Average Wind Speed



avg_wind	count_nonzero	rides per day
(0, 2]	44	34949.477273
(2, 4]	315	31590.085714
(4, 6]	407	26930.417690
(6, 8]	177	21137.152542
(8, 10]	90	18231.577778
(10, 15]	51	13156.294118
(15, 20]	0	NaN
(20, 100]	0	NaN

count_nonzero values is the sample size of each bucket



avg_wind	count_nonzero	rides per day
(0, 5]	173	877.184971
(5, 10]	375	959.698667
(10, 15]	166	846.753012
(15, 20]	17	945.705882
(20, 100]	2	840.500000

count_nonzero values is the sample size of each bucket

Average wind speed is the only weather metric that riders from NYC and SF seem to react differently to. New Yorkers are more sensitive to wind speed, as we see rides per day declining significantly as average wind speed increases. However, San Franciscans seem to be immune to wind speeds, with ridership staying consistent no matter how windy it is. However, there is a smaller n-size for SF data, so the data for Average Winds of more than 15 MPH should be taken lightly.

Conclusion

In conducting this project, we set out to understand the similarities and differences between the bike share systems in New York and San Francisco. The historical data shows a sharp increase in the number of monthly trips for New York, but fairly flat results for San Francisco. The increase in New York could be a result of rising popularity or the addition of new stations, but it is difficult to draw a definitive conclusion without doing further analysis. Based

on the histograms and boxplots of trip duration, we also saw that the majority of people take trips less than 30 minutes. We suspect that this is largely a reflection of the pricing structures in both cities, which charge additional fees for rides greater than 30 minutes. In analyzing the distribution of Rider Types, we observed that the vast majority of trips were taken by Subscribers as opposed to Customers. In San Francisco, we observed that trips taken by Commuters were proportionally equivalent during the first two months of the program and later dropped. However, this trend did not occur in New York. Trips also peak during commute times, largely due to the behavior of Subscribers. We also saw that people from both cities reacted with weather in a similar fashion, preferring to take rides when the weather was temperate and when there was less precipitation. The only difference in rider preferences seemed to be average wind speed - New Yorkers were more sensitive to wind while San Franciscans were not sensitive at all.

Appendix

The code can be found in the folder associated with this report and is structured as follows:

1. Data Merge
2. Overall Ridership
3. Rider Types
4. Trip Times
5. Time of Day Analysis
6. Impact of Weather (broken out by cities)

***

Prompt

<u>The Report</u>

The report will be 8+ pages (including appropriately sized figures) and will be a report on what you found out from the data. This should focus on telling stories and explaining the narrative of the exploration and challenges associated with that. The report should not include any code - all code should be included in a sub-folder in either plain python files or in jupyter notebooks.

For the report, any graph, table, or figure should be annotated with why it is included. This is really to enforce just slapping graphs in your report that have no meaning.

***

GITHUB: https://github.com/yannie30/MIDSPythonBridgeProject2-BikeShareData
CSV Files: https://drive.google.com/drive/folders/0B0WHLqoqLz_tdHQwbGNsTWMtS2s

Outline

1. Introduction (Yannie)
2. Overall Ridership (Andrew)
3. Rider Types (Hawina)
4. Trip Times (Andrew)
5. Time of Day Analysis (Hawina)
6. Impact of Weather (Yannie)
7. Conclusion (All contribute main points from your section)