DSO 545: Statistical Computing and Data Visualization

Data Visualization for GoBike by Ford

May 8, 2017

Kannan Avalupet
Kimberly Norris
Bhavarthi Patel
Ye Yang
Xu Zhang

Executive Summary

Ford has recently introduced a bike sharing system within the San Francisco area. If they are to be successful and continue to grow this new venture, that is quite different from their core competencies of car manufacturing, they will need to understand how they have been doing thus far. By analyzing the user data from June 2017 through March 2018 we intend to answer 3 main questions for Ford. First, are certain weather conditions discouraging customers from renting bikes and is this an issue that needs to be mitigated? Secondly, are their certain locations that are more popular than others? This will help Ford understand if inventory/availability of bikes needs to be allocated among the different stations? And lastly, what type of user is the most frequent customer? Are they commuters or do they rent for leisure? This will help Ford better target their marketing to encourage more customers from the minority group.

After sourcing multiple datasets, cleaning the data, grouping it, and plotting it we have discovered some insights to answer the problem statements listed above. First off, there doesn't seem to be a strong swing in the duration of the rental when the weather changes, especially the subscribers usage seems agnostic to the weather. The total number of rentals does show a trend with the peak temperature for rentals being between 51 and 57 degrees Fahrenheit and a second smaller peak between 62 and 70 degrees Fahrenheit. These peaks are consistent with subscribers and general customers. By looking at type of weather, we could see the least favorable condition for people to rent bikes in was fog, for bother subscribers and general customers. Another interesting fact that came from this analysis to answer the first question, was that the average duration is much longer for general customers than for subscribers.

From the spider map we were able to understand the usage density between each stations and get an insight on the Highest footfall is for station 15 , with nearly 4970 outgoing trips to station 6 and with nearly 2703 incoming trips from station 81 alone.

Lastly, we have learned, by using heat maps, that the majority of people using the GoBike service are most likely commuters. The most popular start and end times are around 8-9am and 5-6pm Monday thru Friday. This also tells us that their commuters are using the bike for less than an hour to get to their destination.

The detailed analysis below will show what steps were taken to reach these conclusions and what next steps can be taken by Ford to improve their GoBike business model.

Introduction

We set out to learn what types of riders were using Ford's new GoBike bike rental service in the San Francisco area. To dig deeper into more than just the data that was provided on Ford's website, we wanted to add the element of weather to see how it impacted the number of rentals. Some of the insights that we focused on can easily be addressed by Ford to mitigate risk in this new business venture and encourage more customers by identifying what factors are drivers in increasing bike rentals.

This summary will walk thru the data that was used and where it was sourced from, what data cleaning need to be done in order to perform insightful analysis, summary of the analysis and potential next steps for Ford. This first step in data analysis and visualization will help highlight other data that could be analyzed to drive continuous improvement in Ford's business decisions in regard to their new GoBike model.

Data Description

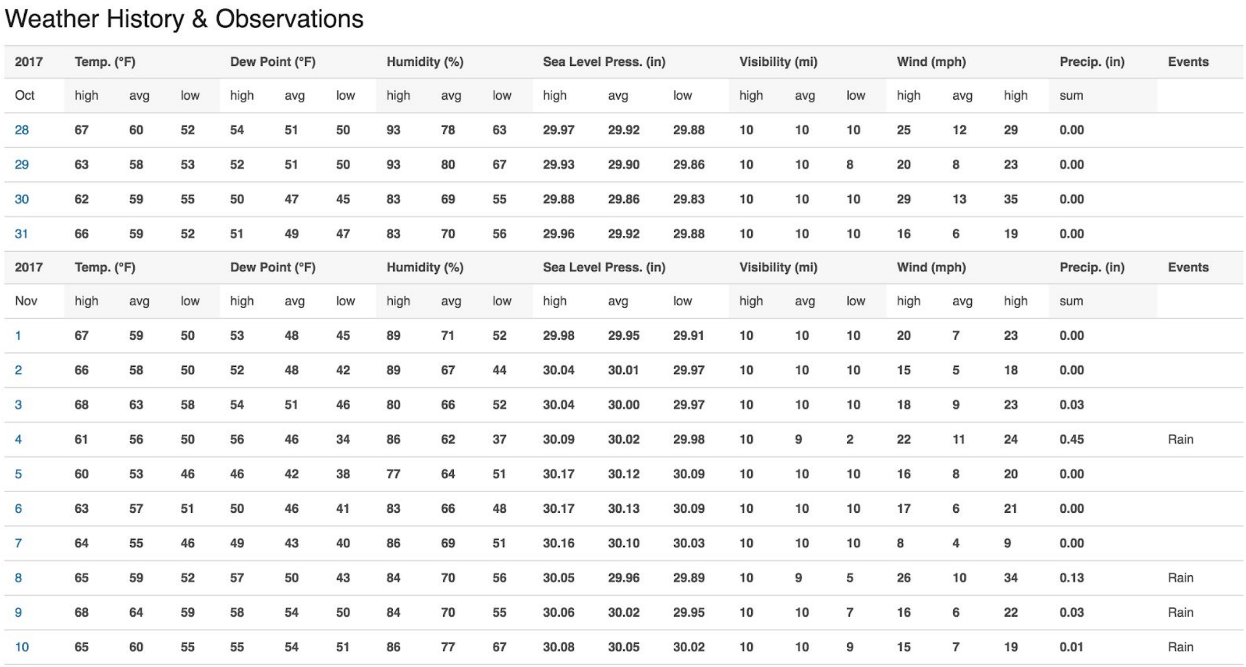
There were four datasets available to use on Ford’s GoBike website (https://www.fordgobike.com/system-data). The first dataset captured data from June 2017 thru December 2017. The other three datasets were for the individual first three months of 2018. These datasets included variables such as: duration, start time, end time, check out location, check in location, and some additional member information. All of these datasets were available for download and came in .csv formats.

The second dataset that was used came from the historical data available on the Weather Underground website (wunderground.com). We matched the custom historical data range to the date range captured in the GoBike datasets. In order for this data to be captured, we had to use the rvest package and web-scraping functionality in R.

Data Cleaning

All four the the Ford datasets needed to be combined into one data frame, but also required some tidying up of the data. The start and end times needed to be brought in as dates, for this we used the lubridate package. The original dataset also captured the times down to the millisecond, which was not necessary for this analysis so we trimmed that out of the date and time stamp.

The weather data required much more cleaning in order to be usable. Below is a screenshot of how the table was set up on the Weather Underground website.

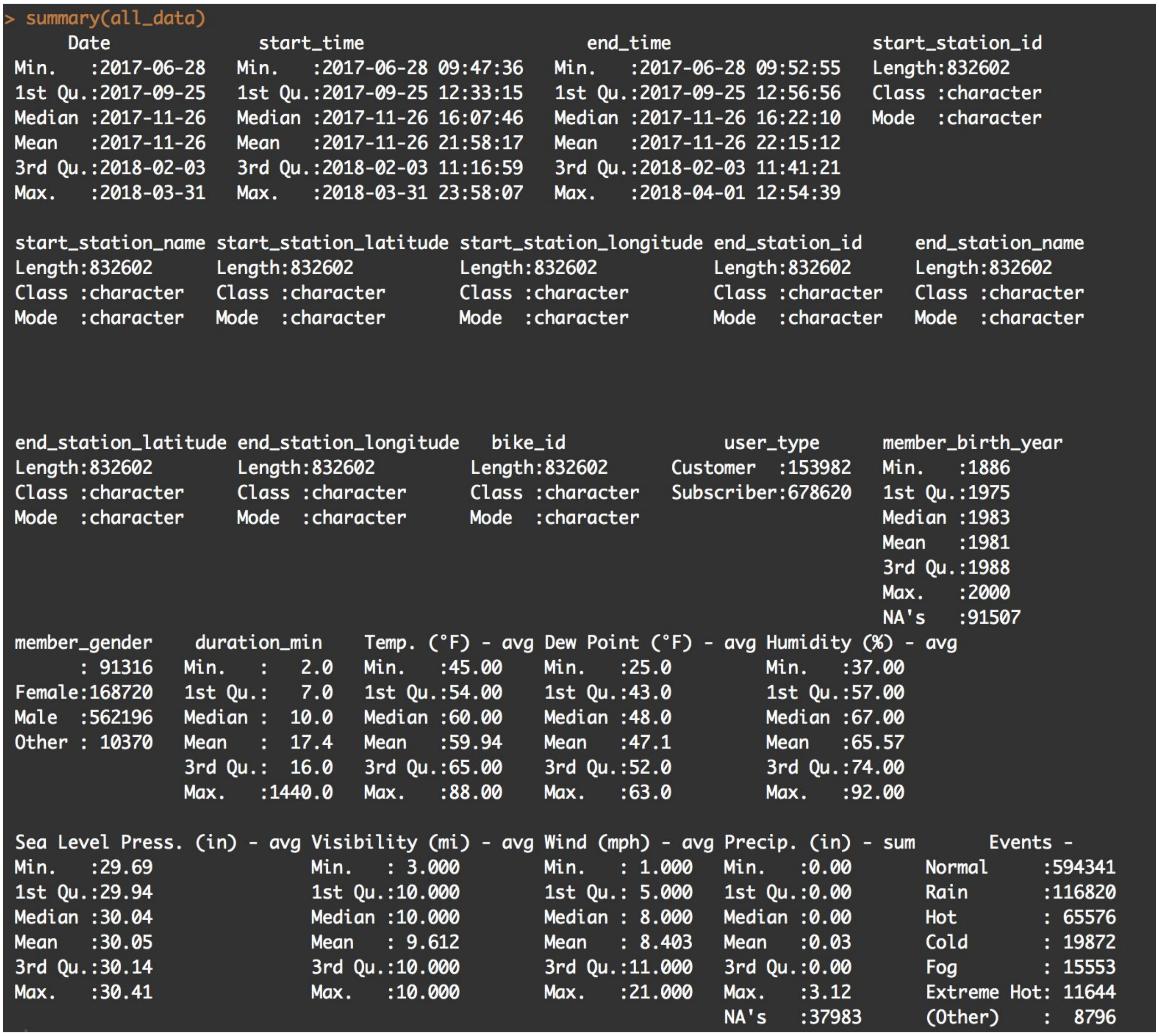


Weather History & Observations

2017	Temp. (°F)			Dew Point (°F)			Humidity (%)			Sea Level Press. (in)			Visibility (mi)			Wind (mph)			Precip. (in)	Events
Oct	high	avg	low	high	avg	low	high	avg	low	high	avg	low	high	avg	low	high	avg	high	sum	
28	67	60	52	54	51	50	93	78	63	29.97	29.92	29.88	10	10	10	25	12	29	0.00	
29	63	58	53	52	51	50	93	80	67	29.93	29.90	29.86	10	10	8	20	8	23	0.00	
30	62	59	55	50	47	45	83	69	55	29.88	29.86	29.83	10	10	10	29	13	35	0.00	
31	66	59	52	51	49	47	83	70	56	29.96	29.92	29.88	10	10	10	16	6	19	0.00	
2017	Temp. (°F)			Dew Point (°F)			Humidity (%)			Sea Level Press. (in)			Visibility (mi)			Wind (mph)			Precip. (in)	Events
Nov	high	avg	low	high	avg	low	high	avg	low	high	avg	low	high	avg	low	high	avg	high	sum	
1	67	59	50	53	48	45	89	71	52	29.98	29.95	29.91	10	10	10	20	7	23	0.00	
2	66	58	50	52	48	42	89	67	44	30.04	30.01	29.97	10	10	10	15	5	18	0.00	
3	68	63	58	54	51	46	80	66	52	30.04	30.00	29.97	10	10	10	18	9	23	0.03	
4	61	56	50	56	46	34	86	62	37	30.09	30.02	29.98	10	9	2	22	11	24	0.45	Rain
5	60	53	46	46	42	38	77	64	51	30.17	30.12	30.09	10	10	10	16	8	20	0.00	
6	63	57	51	50	46	41	83	66	48	30.17	30.13	30.09	10	10	10	17	6	21	0.00	
7	64	55	46	49	43	40	86	69	51	30.16	30.10	30.03	10	10	10	8	4	9	0.00	
8	65	59	52	57	50	43	84	70	56	30.05	29.96	29.89	10	9	5	26	10	34	0.13	Rain
9	68	64	59	58	54	50	84	70	55	30.06	30.02	29.95	10	10	7	16	6	22	0.03	Rain
10	65	60	55	55	54	51	86	77	67	30.08	30.05	30.02	10	10	9	15	7	19	0.01	Rain

You can see that the variable headers repeat for each month, and their are additional subheader rows. The other problem is the date was written as a single day of the month with the associated month only appearing in the subheader row.

After both datasets were clean we combined them into one dataset by merging them using the date.

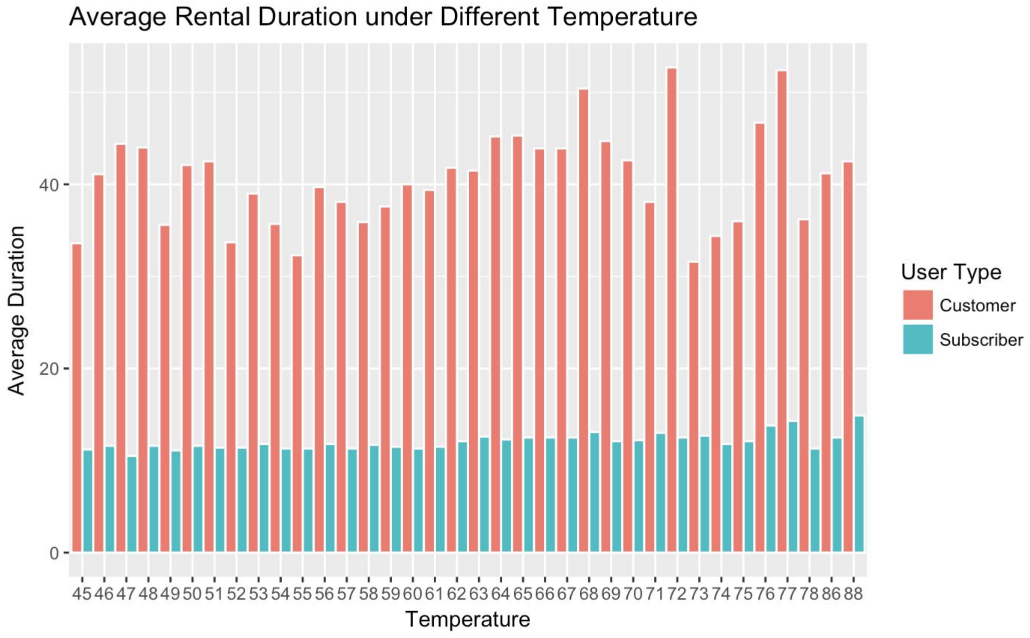


```
> summary(all_data)
      Date              start_time                     end_time                   start_station_id  
 Min.   :2017-06-28   Min.   :2017-06-28 09:47:36   Min.   :2017-06-28 09:52:55   Length:832602     
 1st Qu.:2017-09-25   1st Qu.:2017-09-25 12:33:15   1st Qu.:2017-09-25 12:56:56   Class :character  
 Median :2017-11-26   Median :2017-11-26 16:07:46   Median :2017-11-26 16:22:10   Mode  :character  
 Mean   :2017-11-26   Mean   :2017-11-26 21:58:17   Mean   :2017-11-26 22:15:12                     
 3rd Qu.:2018-02-03   3rd Qu.:2018-02-03 11:16:59   3rd Qu.:2018-02-03 11:41:21                     
 Max.   :2018-03-31   Max.   :2018-03-31 23:58:07   Max.   :2018-04-01 12:54:39                     

 start_station_name start_station_latitude start_station_longitude end_station_id     end_station_name  
 Length:832602      Length:832602          Length:832602           Length:832602      Length:832602     
 Class :character   Class :character       Class :character        Class :character   Class :character  
 Mode  :character   Mode  :character       Mode  :character        Mode  :character   Mode  :character  




 end_station_latitude end_station_longitude   bike_id               user_type      member_birth_year
 Length:832602        Length:832602         Length:832602      Customer  :153982   Min.   :1886     
 Class :character     Class :character      Class :character   Subscriber:678620   1st Qu.:1975     
 Mode  :character     Mode  :character      Mode  :character                       Median :1983     
                                                                                   Mean   :1981     
                                                                                   3rd Qu.:1988     
                                                                                   Max.   :2000     
                                                                                   NA's   :91507    
 member_gender    duration_min     Temp. (°F) - avg Dew Point (°F) - avg Humidity (%) - avg
       : 91316   Min.   :   2.0   Min.   :45.00    Min.   :25.0         Min.   :37.00      
 Female:168720   1st Qu.:   7.0   1st Qu.:54.00    1st Qu.:43.0         1st Qu.:57.00      
 Male  :562196   Median :  10.0   Median :60.00    Median :48.0         Median :67.00      
 Other : 10370   Mean   :  17.4   Mean   :59.94    Mean   :47.1         Mean   :65.57      
                 3rd Qu.:  16.0   3rd Qu.:65.00    3rd Qu.:52.0         3rd Qu.:74.00      
                 Max.   :1440.0   Max.   :88.00    Max.   :63.0         Max.   :92.00      

 Sea Level Press. (in) - avg Visibility (mi) - avg Wind (mph) - avg Precip. (in) - sum       Events -   
 Min.   :29.69               Min.   : 3.000        Min.   : 1.000   Min.   :0.00        Normal     :594341  
 1st Qu.:29.94               1st Qu.:10.000        1st Qu.: 5.000   1st Qu.:0.00        Rain       :116820  
 Median :30.04               Median :10.000        Median : 8.000   Median :0.00        Hot        : 65576  
 Mean   :30.05               Mean   : 9.612        Mean   : 8.403   Mean   :0.03        Cold       : 19872  
 3rd Qu.:30.14               3rd Qu.:10.000        3rd Qu.:11.000   3rd Qu.:0.00        Fog        : 15553  
 Max.   :30.41               Max.   :10.000        Max.   :21.000   Max.   :3.12        Extreme Hot: 11644  
                                                                    NA's   :37983       (Other)    :  8796  
```

Exploratory Data Analysis

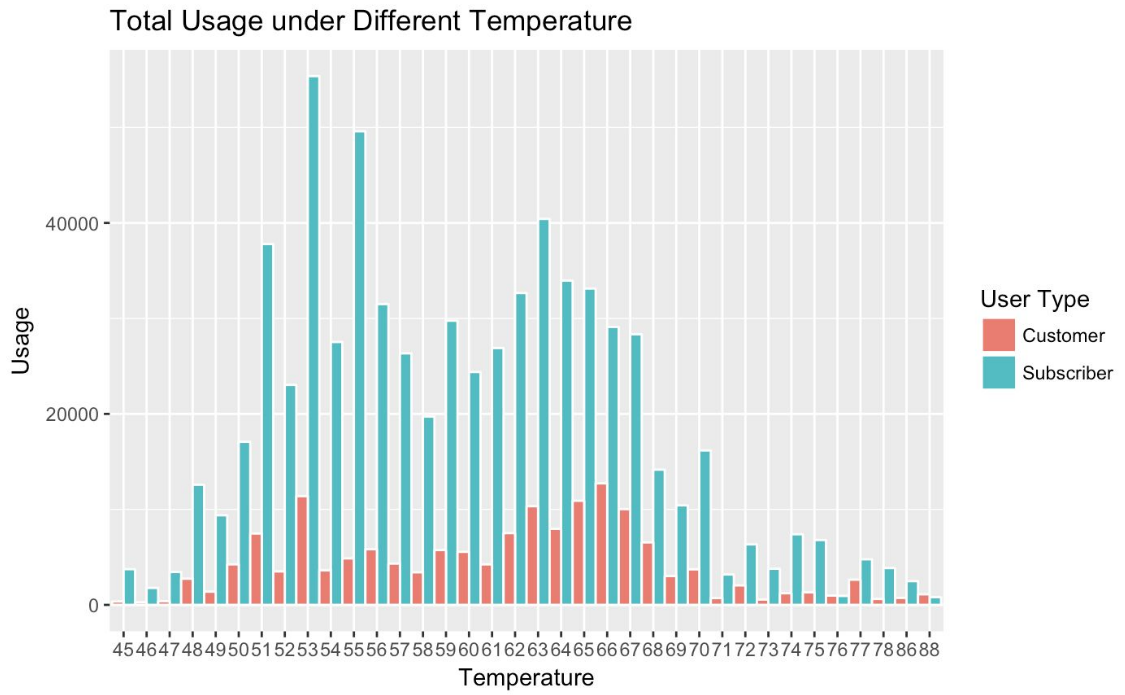
Here are some of the plots that were made throughout our analysis process.

Plot 1: First we built a bar plot to see the whether or not the average daily temperature influences ride duration among different user types.



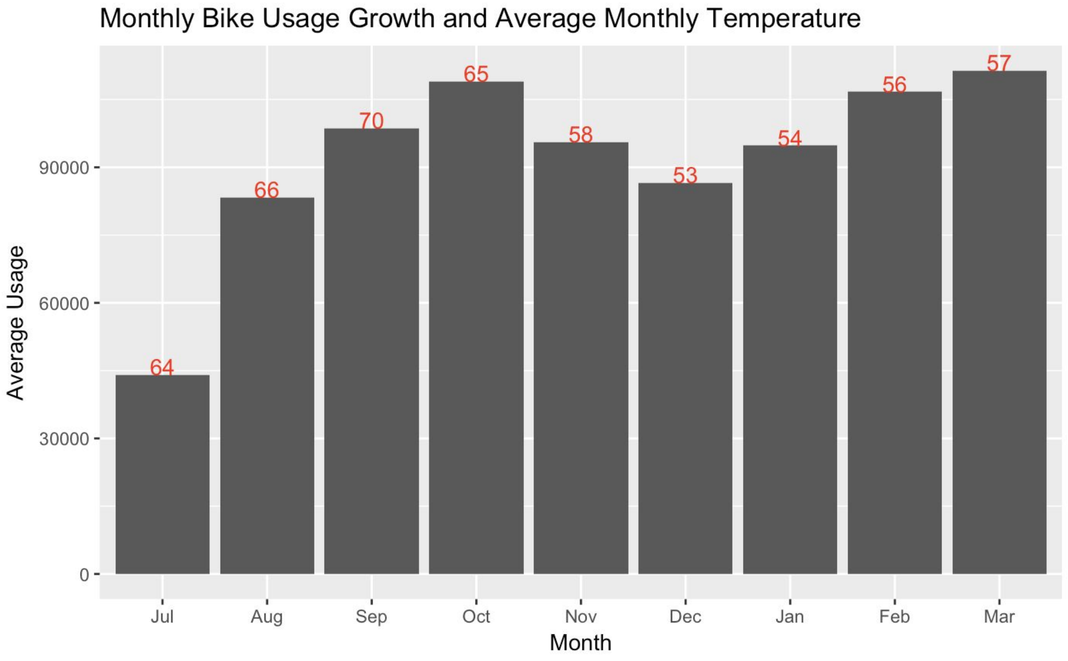
Subscribers and customers both show no significant difference on the duration regarding different temperatures. However, customers tend to rent bikes about three times longer than subscribers do.

Plot 2: Next, we looked at whether or not the temperature had an influence on the total number of bike rentals (usage).



When it comes to usage, subscribers tend to use the bike much more often than customers do. Also, we notice that within the specific ranges of temperature, both types of users tend to rent bikes more often. There are two peaks show in the graph, the highest being between 51 and 57 and the second between 62 and 70 degrees Fahrenheit.

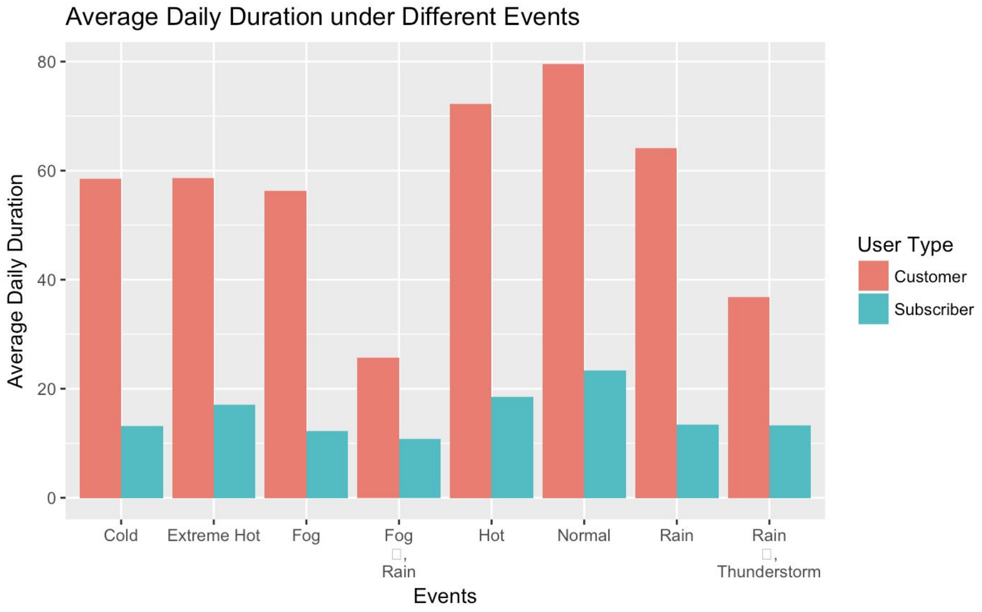
Plot 3: Next, we wanted to look at seasonality of temperatures.

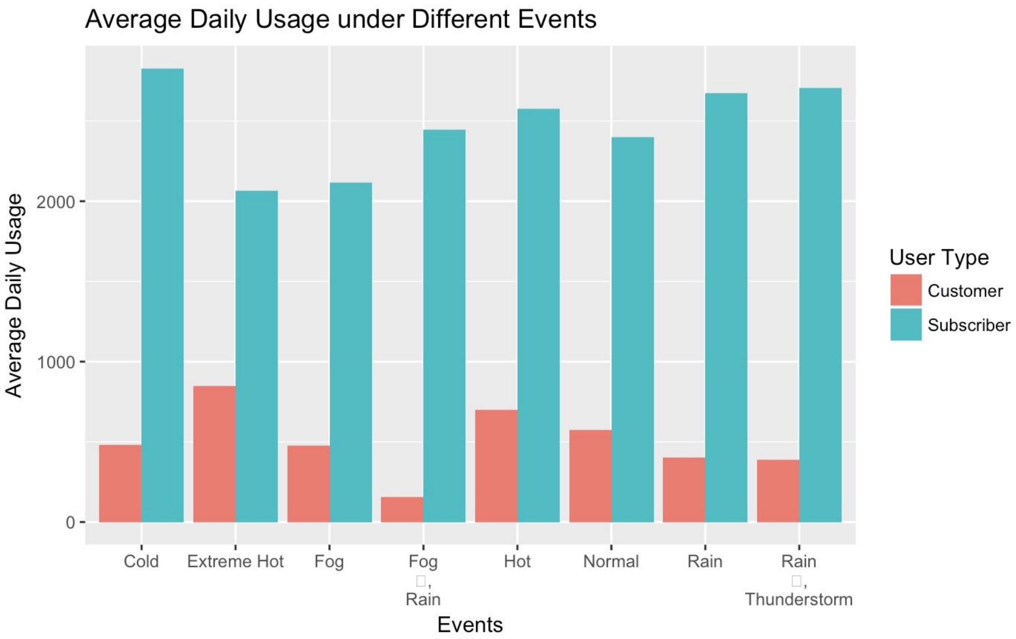


We can tell from the data set that during winter and summer, the usage is less than other seasons. However it is hard to make this conclusion because fordbike system officially launched in June 28, 2017. Two issues that make this data and its conclusions incomplete are: we do not have a full calendar year's worth of data, start-up growth could inaccurately portray bike usage during certain seasons.

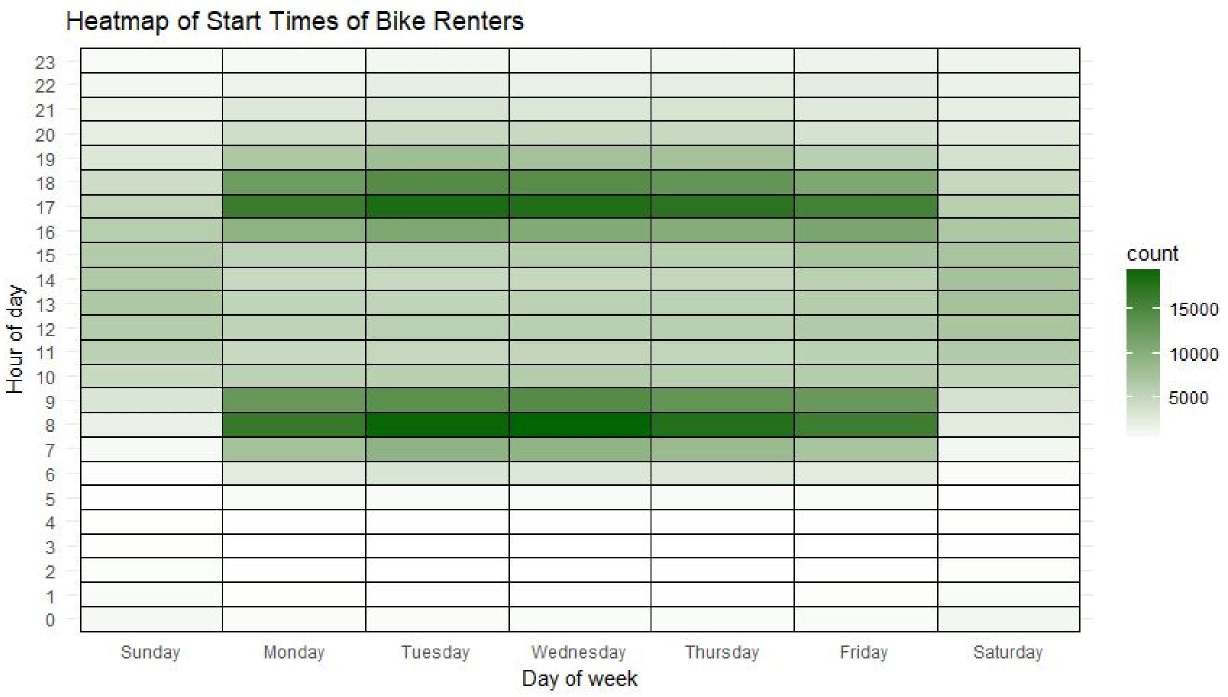
Plot 4 and 5: We know that certain temperatures do not always reflect the type of weather outside. For our next plots we looked at average duration and total usage, by user type again, but now by weather type instead of temperature.

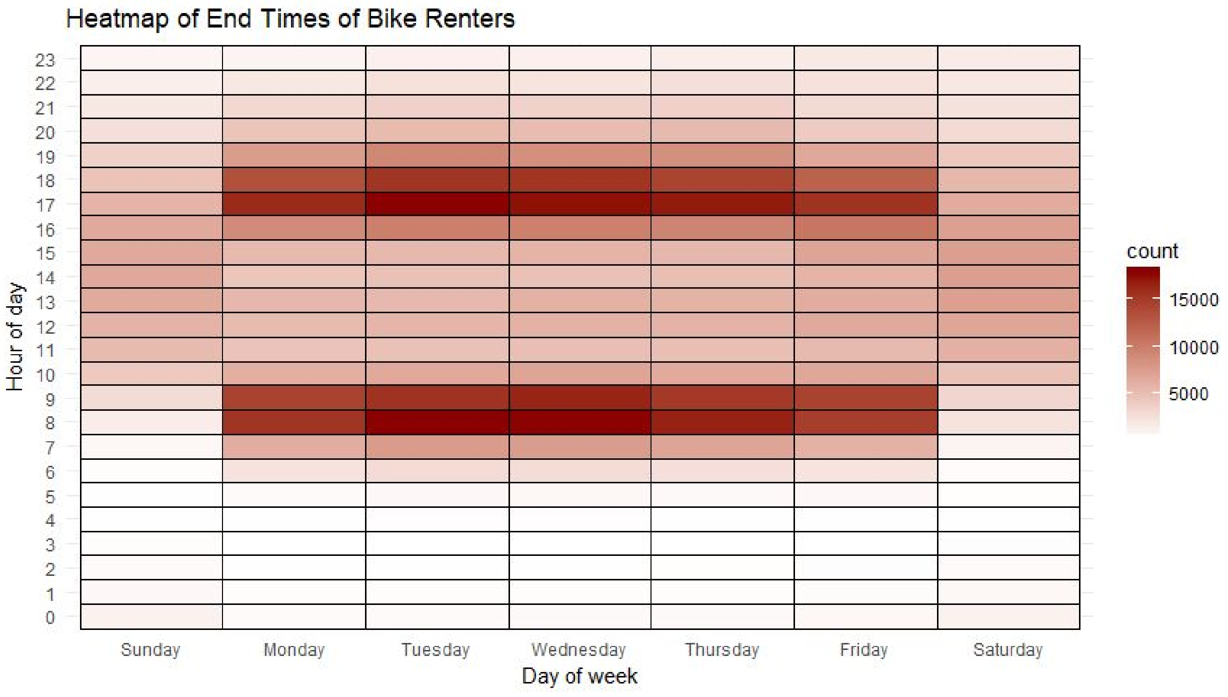
Some of the data was available in the original dataset, but we also created extra events like "Hot" and "Extremely Hot" based on the hot index and "Cold" based on the NWS Wind Chill Chart.





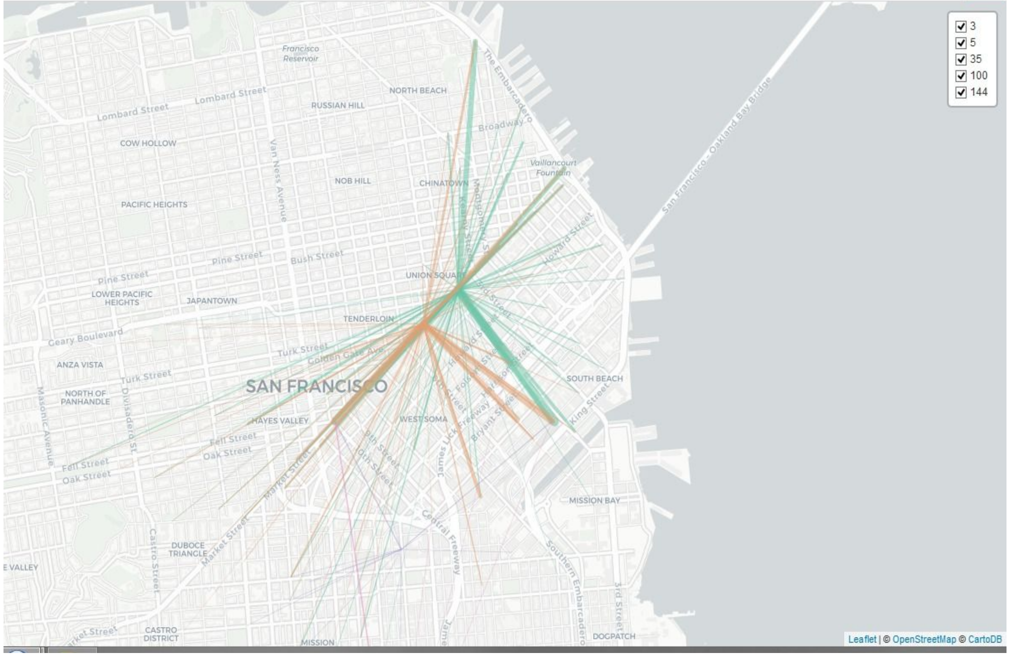
Plot 6 and 7: Next, we built heatmaps to understand when the most popular times for bike rentals were. This would provide indications of whether the majority of usage was coming from commuters or leisure riders. We built one plot representing the start times (in green) and one representing the end times (in red).





The heatmaps show that the peak times are Monday - Friday from 8/9 am and 5/6pm. This is a clear indication that the majority of riders are commuters and spend less than an hour on their commute via bicycle.

Plot 8: Lastly, we also built a spider map to show the trends of which stations and paths were most popular.



Conclusion

A lot was learned from doing this data visualization exercise, and a lot of action can be taken by Ford to improve their business. First, based off of the weather analysis, we saw that the lowest usage rates were when the temperature got over 80 degrees. Ford could partner with water distributors, such as Coke products (Dasani) or SmartWater, to install vending machines at the bike stations and if you purchase a water you can get a discounted bike rental. This could be a strong win win and make renting/riding a bike in warm weather a more worthwhile experience for their customers.

Another action Ford could take would be to offer discounted rates on weekends, since the majority of their customers seem to be commuters. This will encourage more business on the weekends. You could also offer additional weekend discounts to subscribers to encourage the general customers to become subscribers, which we know have higher usage rates.

Lastly, Ford can dig deeper into the popular routes that we highlighted here. By having more access to inventory at each station and whether or not customer decide not to rent because their aren't bikes available could be tied to this data to find any new trends. By more effectively allocating their bikes, Ford could also increase the number of customers their get in a day.