After deciding to look at last summer as my timeframe (June-August 2019), I decided to look at both basic demographics of each rental user and the busiest Citi Bike stations for both picking up and dropping off a rental bike.

To compile my data, I combined three months of Citi Bike csv’s into one data frame using jupyter notebook. I then cleaned and re-organized the data for to better suit my use. I created a new gender column which was able to change the binary indicators for each user to male, female and unknown. I also created an age column by subtracting the year (2019) from user birth year. I noticed some unusual ages, such as 131, so I eliminated all users aged 90 and over, thinking that was a logical age range. I also reformatted the trip duration column from seconds to minutes, as I thought that may be easier for the reader to distinguish. Lastly, in tableau, I created a new calculated field to determine the average rentals per day for each station.

In looking into the demographics of each rental, I examined user gender, age and consumer type (subscriber vs. non-subscriber). When looking at this data, some data gathering issues became apparent. I first noticed the unusual number of riders that were aged 50. I then noticed that the vast majority of those were also listed as unknown in the gender listing, and most unknown gender listings were also categorized as customer, or non-subscribers. This is when I began to think that is was a data gathering issue. I briefly thought about eliminating the data, however I felt it would be better to try to find the cause of the issue, which would help deliver better data analysis. This may be occurring when customers are entering in their data, perhaps not all fields need to be filled in and they are auto filled later. More research would be needed at the data entry point.

Then I began to look at the busiest stations in which bikes were both picked up and dropped off. This data, shown with average age, average trip duration and number of rentals for the time period will help with bike allocations efforts each night. When I was deep into this project, I thought that maybe even further breaking the data down into what days of the week were busiest but ran out of time to dig deeper. Another direction that I thought may be interesting I came about after I changed the start stations from the overall 3 months count to a daily rental count. I noticed roughly 20% of the stations are renting on average 2.5 bikes per day. I would like to try to find stations like this that have other similarly low volume stations close by and try to reduce some of the “low volume” stations, thereby keeping bikes more readily available in the higher volume stations.