The statistical analysis of the Washington DC bike share data set is an extension of our exploratory data analysis (EDA). During the EDA we noticed varying user counts for different years, seasons, temperatures, weather conditions and wind-speeds. It was claimed that user count increased over the two years, that temperature and wind-speed had significant effects on rider usage statistics, and that casual and registered riders had different usage patterns. Today we verify those claims statistically. The first step in the process is to refer back to BS_annual_hist.py where we get a first look at the total number of rides for different types of users over the course of 2011 and 2012. We find that the registered and casual users have dramatically different ride totals, but the question becomes: Do the two groups have different characteristics?.

We start answering this question by moving towards a revision of BS_day_of_week.py where we notice that total ride count is higher during weekdays, both overall and for registered users, meanwhile casual users have much higher ride counts during the weekend. This leads to the conclusion that registered users use the bike system as part of their daily routine, while casual users use the system for leisure. Collecting data regarding reasons for usage from both registered and casual riders would be a tremendous boon for the pursuit of understanding the two groups.

In BS_Statistical_Daily.py we do bootstrapping to test the null hypothesis that the seasons have similar means, which we easily reject, and observe plots of bootstrapped samples for mean and variance of total, registered, and casual users in the different seasons. We notice that all of the seasons and rider groups tend towards a simple bell curve, with the exception of mean and variance of casual rider count during the Spring. Both of these plots showcase positive skewness, which is certainly interesting. Thankfully knowledge of the region sheds light on the situation. The DC Metro area is well known for its Cherry Blossom Festival, which occurs annually during the end of March and beginning of April. This event brings in tourists from around the world, which would bring a sudden influx of casual users to the DC Bike-share system. The rest of BS_Statistical_Daily goes on to show the confidence intervals for different types of riders in different seasons, as well as interesting relative

standard deviation data (table included below). As excepted the variance is highest during Spring, possibly due to the cherry blossom festival. The other months have similar relative standard deviations for registered users, but percentages vary more dramatically for casual users. Perhaps they are more sensitive to temperature than registered users. We'll learn more in the correlation table.

Relative Standard Deviation Table

Season	Registered Users	Casual users
Spring	52.75%	115.42%
Summer	34.73%	66.50%
Fall	29.29%	49.42%
Winter	36.03%	82.92%

In BS_Statistical_Daily_Correlation we create a set of Correlation Matrices and Scatter-plots to observe the relationships between the independent variables (temperature, humidity, wind-speed) and the dependent variables (casual, registered, and count (abbreviated cnt in the data). Especially interesting is the Correlation Matrix for Fall 2011 & 2012. Here we learn that casual riders are more effected by temperature than registered riders, while registered riders are more effected by wind-speed than casual riders. That being, casual riders ride more when the temperature is higher, and registered riders use the service less when wind-speeds are higher. But we must be careful taking too much stock in these differences as the correlation coefficients are very low during the fall. When we look at the correlation matrix for Spring we notice that the correlation coefficient is much higher than in the fall. During the spring both casual and registered user count increases in accordance with temperature, and is barely affected by humidity and wind-speed.

In BS_Statistical_Hourly we divide the hourly data up into six different time frames, they are listed in the table below. Again registered users have the lowest relative standard deviation, which is especially low during the afternoon. Meanwhile the variance for casual users relative to their mean usage during the evenings is quite high.

Relative Standard Deviations for different types of users during different times of day

Time of Day	All users	Registered	Casual
Early (00:00 – 5:59)	121.52	120.34	157.82
Morning (06:00– 10:59)	81.85	87.51	118.83
Midday (11:00-14:59)	59.97	50.67	95.23
Afternoon(15:00-17:59)	58.65	32.25	88.35
Evening(18:00-20:59)	59.81	62.18	134.78
Night(21:00-23:59)	61.12	59.62	96.36

Finally, BS_Statistical_Hourly_Correlation.py generates a Correlation matrix for the independent and dependent variables in the hourly data alongside a scatter plot matrix for the same variables. It serves to reinforce the findings we have already discovered.