Additional EDA - Inferential Statistics

The code used for this process is broken down step by step at the link below:
https://github.com/dgeihsler15/Springboard/blob/master/Capstone1/Additional_EDA_inferential_statistics.ipynb

Initially in my EDA of Divvy's 2017 trip data, I looked into rider usage to better predict how the different Divvy stations should be stocked. Looking at the number of trips by month made it easy to separate out a busy and slow season, and from there I explored what the busiest days of the week and times of the day were in both seasons. Another focus was looking how at how the average trip durations changed throughout the week and throughout the year, and finally the different capacities for the divvy stations were studied as well.

In this portion of the EDA I wanted to dive deeper into my initial findings and learn more about ways to group Divvy stations to help make predictions for stocking requirements based on their similarities.

I started off this section by looking to see if there was any correlation between the latitude and longitude of the Divvy stations and the number of rides started from those stations. After creating a scatter plot and regression line for both the latitude and longitude of starting stations vs the number of rides, the pearson correlation coefficients were calculated as well. I found a slight negative correlation of .12 for latitude and number of rides, and slightly stronger correlation of .27 for the longitude and the number of rides. Both of these correlations however are weak, and by themselves, did not prove to be a significant factor in the number of rides coming through a station.

I also explored the correlation of the capacity of starting stations and number of rides started at those capacities, as well as the capacity of ending stations and number of rides ending at those capacities. After calculating the correlation coefficients, creating scatter plots, and adding in regression lines for starting and ending station capacities, I found a weak correlation between those variables as well. Looking back at the original EDA I found that there were many more low capacity stations than higher capacity ones. This could have been the reason the number of rides were slightly higher for the stations with lower capacity, so I decided to look at this another way.

After grouping Divvy stations into 4 tiers based on their total number of rides for 2017, I took a look a the average capacities for each of the tiers and found the average capacity got higher as you moved up each tier based on the number of rides. This was further proved by a two sample t tests showing a significant difference in average start and end capacities in group levels.

The next piece I wanted to get a better understanding of was how the number of trips into a station compared to the number of trips out of that station. I then found the in/out ratio for all of the rides in 2017 for each station. I looked to see how the number of rides affected the in/out ratio by comparing the ratios in the different group tiers, but the ratio was pretty consistent across all groups. I found the same while comparing the in/out ratio to different station capacities, the average ratio was similar across the different capacities, but the variance was higher for low capacity stations. Both of these findings were confirmed with a two sample t-test, as was a test to determine if there was a significant difference in the in/out ratio in the high and low seasons. The test again came back showing there was not a significant difference in ratio between seasons either.

The remaining hypothesis tests completed during this section were to confirm hypotheses made in the initial EDA. Please see those hypotheses below:

- The busiest hour of the day is attributed to typical commuting hours
- The number of trips are higher in the PM hours
- The number of trips are higher on the weekend days in the busy season
- The number of trips are higher on the weekdays in the slow season

- The average trip durations are higher on the weekend than weekdays
- The average trip durations are higher in the busy season

Each of these tests were completed using two sample t-tests to determine if there was a significant difference between the means in question . When a significant difference was found, I used bootstrap replicates to find a good estimate of the probability that the mean of one variable was greater than or less than the other.

Results:

The busiest hour of the day is attributed to typical commuting hours

During the initial EDA I found that 5pm was the busiest hour of the day throughout the year. To see if this could be contributed to 5pm being a prime commuting hour , I looked at the average number of rides during the 5pm hour on weekdays and compared that to the average number of 5pm rides on weekend days for the whole year. The results proved there was a significant difference in the means, and the number of rides on the weekdays at the 5pm hour were consistently higher than weekends.

The number of trips are higher in the PM hours

The initial visualizations were pretty convincing that this was the case, but the test results backed up the visualizations as well, and I found the average number of trips are higher in PM hours.

The number of trips are higher on the weekend days in the busy season

Despite finding that Saturday is the highest trafficked day in the high season in the initial exploration, the results of our test did not find significant evidence that there is a difference in the averages of weekdays and weekend days in the high season.

The number of trips are higher on the weekdays in the slow season

The t-test and bootstrap replicates confirmed the initial visualizations and found that the number of trips are higher on the weekdays in the slow season.

The average trip durations are higher on the weekend days than weekdays

The t-test and bootstrap replicates confirmed the initial visualizations and found that average trip durations are higher on weekend days than weekdays.

The average trip durations are higher in the busy season

The t-test and bootstrap replicates confirmed the initial visualizations, and found that average trip durations are higher in the busy season compared to the slow season.