The Problem:

Divvy is a Chicago based bike sharing company with 585 stations across the city. Divvy is a great way to get around Chicago, but getting from point A to point B smoothly relies on an available bike at your desired starting point, and an empty slot to check out your bike at your destination. If a customer looks to start their journey, and finds out there aren't any bikes available at the closest station, this will sour their experience as they now have to travel to another station or find another mode of transport. The same is true if a customer plans to end their trip at a particular location and finds there are not any empty slots available to dock their bike. They now have to find another location within their trip time window, and backtrack to their desired destination.

By analyzing Divvy trip data we can uncover insights on proper stocking requirements for all Divvy stations. These insights will be instrumental in helping Divvy's operations team stay in touch with customer needs. Properly stocked stations are essential for retaining customers and growing the Divvy brand.

The Data:

The 2017 Divvy bike sharing data is available on their website in two main parts; trip data, and bike station data. The trip data contains a unique trip id for each trip taken and contains information on the timing of trips, stations used, and what type of user took the ride. The two user types refer to a “customer” who purchased a 24 hour pass, and a “subscriber” who purchased an annual membership. For Subscribers there is gender and birth year information as well. In the station data the main piece of information that we’re interested in is the station capacity (how many docking stations are available at each station). There is also information on the location of all the bike stations and the date they went online as well.

The trip data for 2017 came in four csv files, one for each quarter. These were loaded into dataframes individually so that the shape and the column names could be looked at closer. Once the formats of the individual data frames were aligned, they were concatenated and then merged with station data. The station data for Q3 and Q4 of 2017 was selected to provide the most up to date station info.

Two merges were performed in total. One matching the station id in the station file to the starting station id in the trip file, and the second matching the station id in the station file to the ending station id in the trip file. This will allow us to view both the information for the starting and ending stations on the same line for each trip.

After the merge duplicate columns and rows were dropped, and data types were changed to appropriately represent numbers, categories, and date time objects. During the inspection of the data I noticed a fair number of riders whose birth years listed would make them over 100 years old. As people of this age are not likely to be riding a bike around Chicago it did not seem that all of that information was accurate. This may be caused from users entering incorrect birth years while signing up, but because of the inaccuracies and that this info is only available for subscribers, it will not be very helpful in grouping all riders by age. Because this project's main focus is on station usage and proper stocking requirements, the birth_year column was dropped from the data set as well.

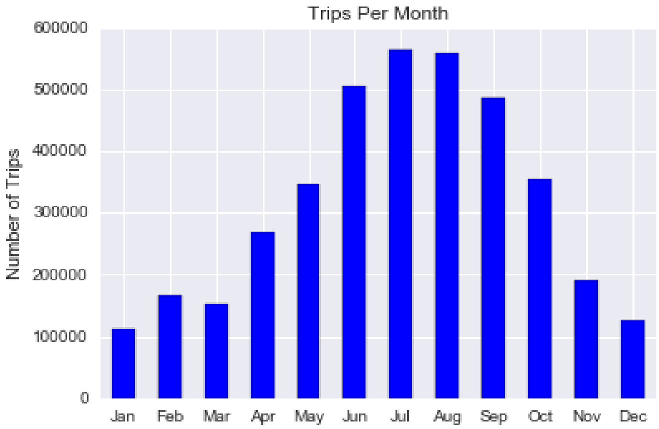
While checking for outliers in the data I found “Dependent” as value for user type. Customer and Subscriber were the only expected user types, but after grouping the data by user type, it was clear that there were a very small number of rides marked as “Dependent” in this category as well. Looking closer I noticed some those rides had gender information and some did not. Because the gender information was only tracked for subscribers, the dependents with that information were changed to “subscriber,” and those that didn't were changed to “customer” to avoid any confusion down the line.

Looking closer at the station capacity I found that the minimum capacity value was 0. As this threw up a red flag, I looked closer and found that two of the stations on the Q3 & Q4 stations file were listed with 0 capacity. This could have been because those stations were closed down in the second half of 2017, but because the trip data is available for all of 2017, it was important to find the correct capacity of those stations when the rides were completed. Our best option was to load in the Q1 & Q2 station file for 2017, and use that to find the correct capacities for those stations and update that information in the combined data frame.

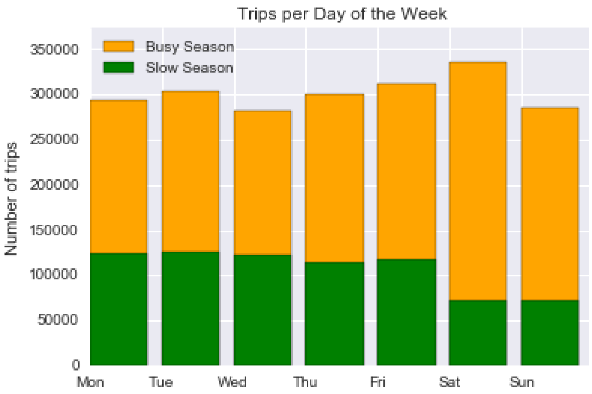
Continuing the search for any outliers, the trip durations were then inspected. The trip data from Divvy's website claimed that any trips shorter than one minute or longer than 24 hours were excluded. I found this to be the case and there weren't any trips included outside of those parameters.

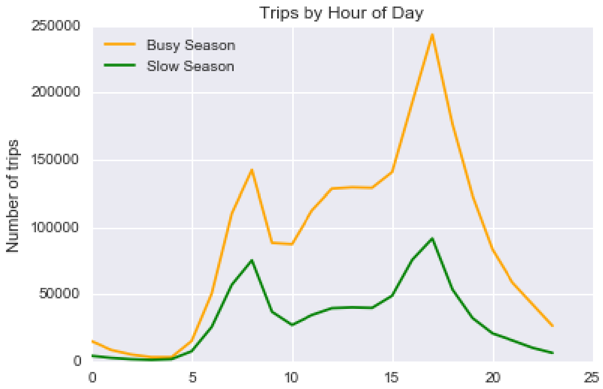
The story so far:

In order to make sure bike stations are properly stocked, I first looked at when Divvy riders are using the service. Looking at the number of bike trips taken in 2017 by month, it may not be surprising to see that the most trips are completed in the summer months, June through September. You'll notice that the slow season is found in the winter months, and the middle tier we see as April, May, and October mark the changing of the seasons.



The consistent number of rides in the winter months might suggest that not all of these trips are solely recreational rides on nice days around the city, but are part of the customer's daily routines or commutes. To explore this further, I took a look at a breakdown of trips by day of the week and different parts of the day. I found that overall Mondays and Tuesdays had the highest trip counts in 2017, and there's a spike in the number of trips around 8am and 5pm, prime commuting hours. Although those statistics reflect trip information for the year as a whole, it's important to see how those numbers fluctuate in busy and slow seasons. In the summer months (below to the left in orange) we can see that Friday and Saturday are the busiest days, while in the winter months (green) Tuesday is the busiest day. We also can see that Saturday and Sunday both have lower trip counts than every weekday during the winter. Despite Saturday being the busiest day in the high season, the average trip numbers per day were not found to be statistically significant for weekends compared to weekdays during that time.

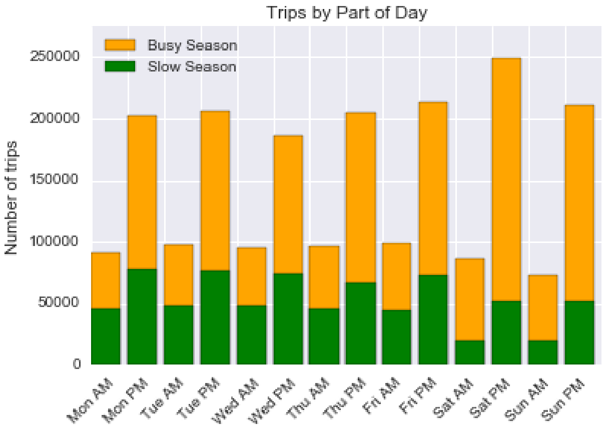




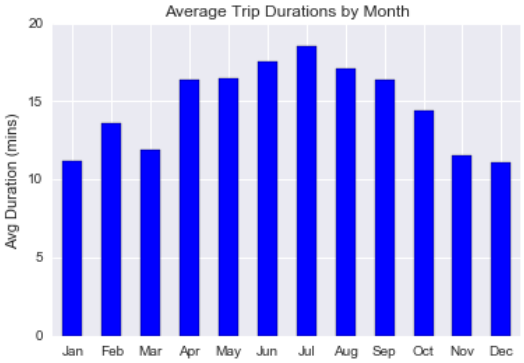
I took a closer look at the number of trips by hour of the day (above to the right) as well. You'll see that most of the trips occur around 8am and 5pm in the slow months as too.

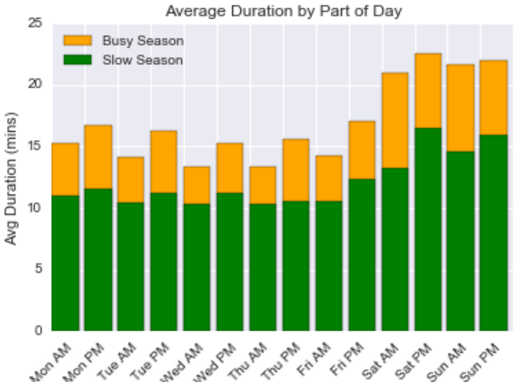
As 5pm was the highest trafficked hour I wanted to see if this could be attributed to it being a popular commuting hour. To test this I took a two sample t-test to determine if the average number of rides in the 5pm hour for weekdays was equal to the number of rides at the 5pm hour on the weekends. The test resulted in a very small p-value of 2.44e-15 letting us know that we could reject the hypothesis that the averages were the same. Using bootstrap replicates I determined the probability of a getting a higher average ride count at 5pm on the weekdays than the weekends was near 100%. The weekdays are a big contributor to the spike at 5pm.

Looking closer at the number of rides by hour, it seemed that the overall number of rides in the afternoon seemed to be higher than the morning hours. I took a look at the breakdown by day of the week to try to get more insight.
It turns out that in both the high and low seasons there are more trips in the PM hours than AM hours on any given day of the week.

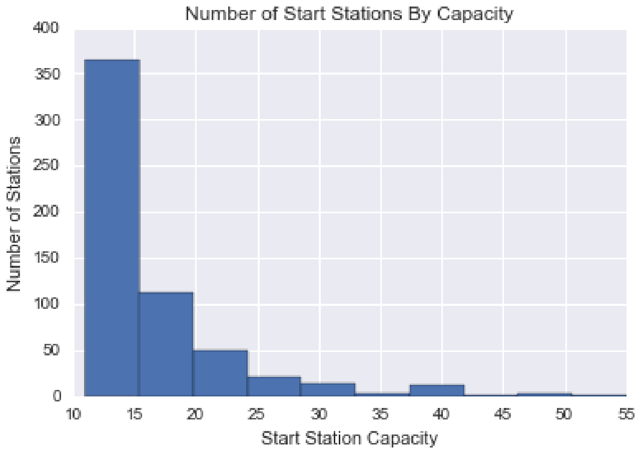


As the number of trips alone doesn't give us a full picture on how Divvy riders are using the service, I took a look at trip durations as well. I found that the average trip duration is 15.92 minutes and the majority of trips last between 5 and 15 mins. We see longer trip durations in the summer months than the winter, and longer trips on the weekends than weekdays. I also found that the longest trip durations seem to come at 1am and 2pm throughout the year (20.07 minutes and 19.56 minutes respectively), outside of the typical commuting hours. Trip durations during the prime commuting hours 8am and 5pm are 14.05 minutes on average.

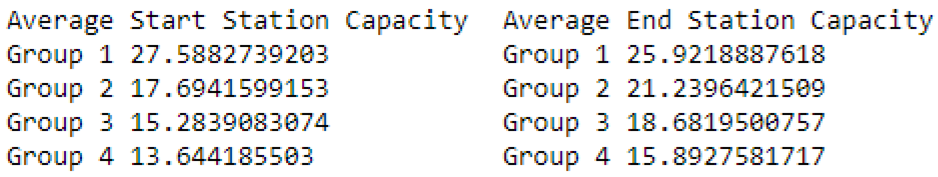




After taking some time to understand Divvy riders behavior, the next step was to take a closer look at the network of stations and their capacities. I found that most trips are coming from and ending at stations with capacities of 10-25. As the max capacity for a station in the city is 55, I needed to take a closer look at why it doesn't seem traffic is correlating with a higher station capacity. Looking at the distribution of capacities across all stations I found that an overwhelming majority of stations fall between a capacity of 10-15. This helps explain why much of the traffic from a high level perspective seems to be connected to lower capacity stations. Only 7 out of 585 stations have a capacity over 40.



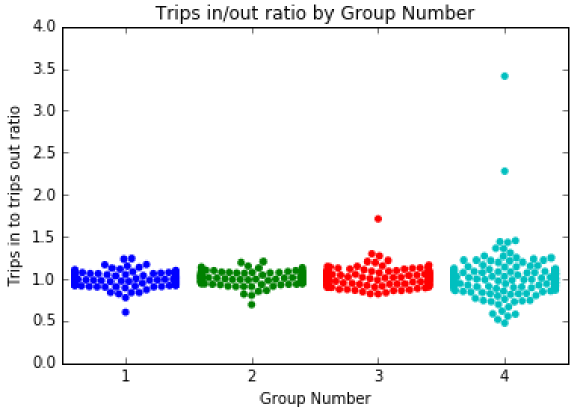
To look at this a different way I grouped all of the stations into 4 equal tiers based on their number of trips for the year. Highly trafficked stations were marked as group 1 and the lowest trafficked stations were marked as group 4. Looking at the average capacity for each group, the average capacities were higher in each group as you move up to the higher trafficked tiers. The average capacities can be found below:

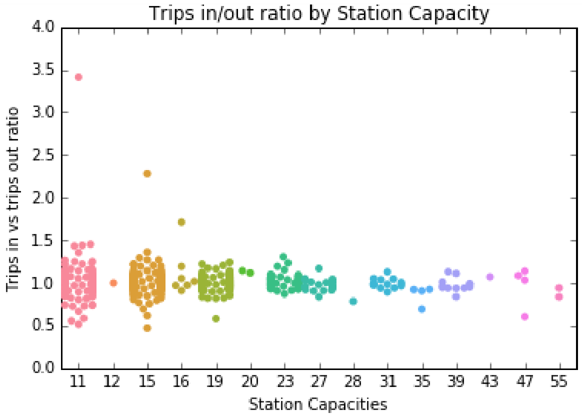


```
Average Start Station Capacity   Average End Station Capacity
Group 1 27.5882739203            Group 1 25.9218887618
Group 2 17.6941599153            Group 2 21.2396421509
Group 3 15.2839083074            Group 3 18.6819500757
Group 4 13.644185503             Group 4 15.8927581717
```

Next I wanted to investigate if there was a large change in which stations were used in busy and slow seasons. I found that 15 of the top 25 start stations and 15 of the top 25 end stations remained constant in busy to slow seasons. When I ran the same analysis for the top 50 start and end stations, I discovered 33 start stations and 31 end stations remained constant from busy to slow seasons. As about 60% of top stations remain constant in summer and winter months, it further enforces the theory that a large portion of Divvy trips are completed as a part of a daily routine that extends year round.

An important aspect I wanted to dive deeper into was the ratio of bikes that come in and out of each of the stations. To investigate this I calculated the in/out ratio for each station by dividing the total number of rides ending at each station by the total number of rides starting at each station for 2017. I found that neither number of rides, or station capacity had a significant impact on the average ratio as the average was very similar across all traffic group levels and station capacities. You will notice the variance does seem to be greater in the lowest traffic group and the smallest station capacities.





The in/out ratio was affected by the high or low season either. A two sample t-test did not provide sufficient evidence to reject the null hypothesis that the means were the same in both seasons.

Summary of Findings:

A summary of the insights that were uncovered regarding Divvy stations and Divvy customer behavior are provided below:

- The busiest time for Divvy in terms of number of trips is June through September. Their slowest times fall in January - March, and November - December.
- The number of trips are higher in the PM hours regardless of season or day of the week.
- The highest trafficked hour of the day is at 5pm, and the weekday trips at that time are consistently higher than weekend trips at that time. This lends itself to the belief that this spike is attributed to 5pm being a popular commuting hour.

- The peak number of rides in the morning is at 8am, and also aligns with a popular commuting hour on the weekdays.
- The number of trips are higher on the weekend days in the busy season while the number of trips are higher on the weekdays in the slow season.
- Although the busiest day of the busy season lies on a Saturday, there is not a significant difference in the average ride count on a weekend day compared to a weekday during that time.
- The average trip duration is 15.92 minutes, and this tends to be higher on weekend days than weekdays throughout the year. The trip durations tend to be longer in the busy season as well.
- 60% of the top starting and ending stations in 2017 remained constant in both busy and slow seasons.
- 62% of Divvy stations have capacities between 10-15, but the average station capacity is found to be higher for higher trafficked stations.
- The ratio of trips in to trips out of the station is not significantly affected by the number of rides going through the station or the station capacity.

Using the insights uncovered in the EDA a predictive model for stocking requirements can be built and applied to all stations throughout the city.