Capital Bikeshare Prediction: Data Investigation

Author: Matthew Cassi **Date**: October 4, 2017

Questions Asked

After cleaning up the data and joining all of it together the next step was to investigate trends in the number of rides and ride times.

Questions Asked:

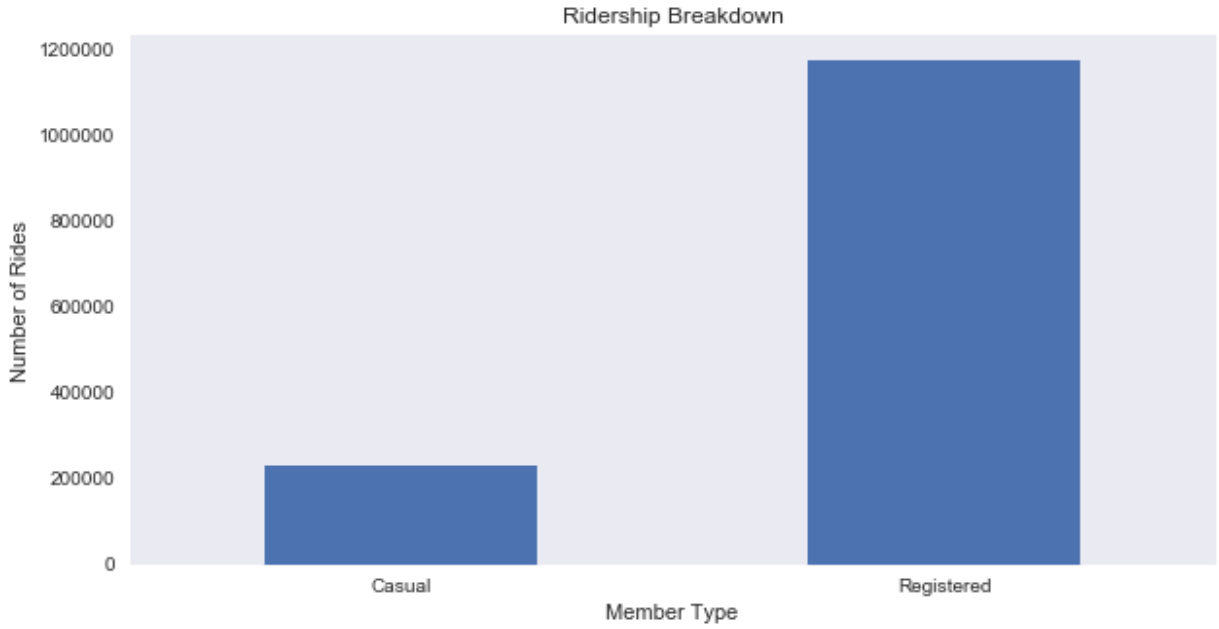
1. Do the number of rides change based on the season?
    - What about the number of rides based on Member Type?
2. Does the ride time differ based on the different seasons (Winter, Spring, etc.)?
    - Based on Temperature?
    - Based on Windspeed?
    - Based on Humidity?
    - Based on weather category (sunny, cloudy, rainy, snowy, etc.)?
3. Does the ride time differ based on the different Member Type (Casual rider vs. Registered rider)?
4. What is the average ride time on holidays vs. non-holidays?
    - Based on weekday?
    - Based on season?
    - Based on month?
    - How are these averages different for member type?
5. What are the most popular starting stations?
6. What are the most popular end stations?
7. What are the most popular start and end station combinations?

Answering these questions will provide further scenarios to investigate and insights as to what might cause ride times to increase between two bikeshare stations.

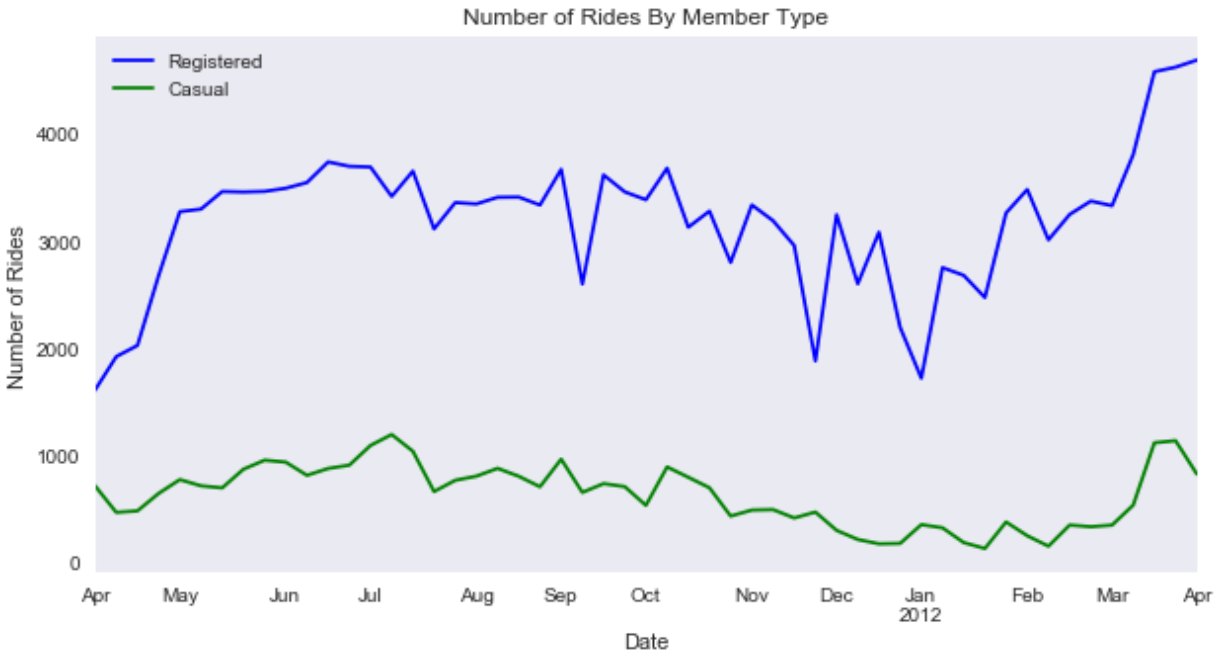
Number of Rides

The number of rides per day for an entire year can show if seasons cause an increase or descrease in ridership. Looking at the graph below on the left, the number of rides decreases when in the fall and winter seasons and increases during the spring and summer seasons. There is a lot of variability, which is difficult to read. The data were resampled based on the using weekly means and plotted. The trends remain the same but are easier to see.

There are 1.4 million rides included in the dataset. What is the breakdown of this by member type? Registered users greatly outnumber the amount of rides compared to casual riders, which is to be expected.



The trend should be fairly the same when looking at both member types (registered and casual) and number of rides on a daily basis. The graph below shows that the number of rides by registered and casual riders follow the same trends as the full dataset. Each have show a decrease in the fall and winter with an uptick in the spring and summer.



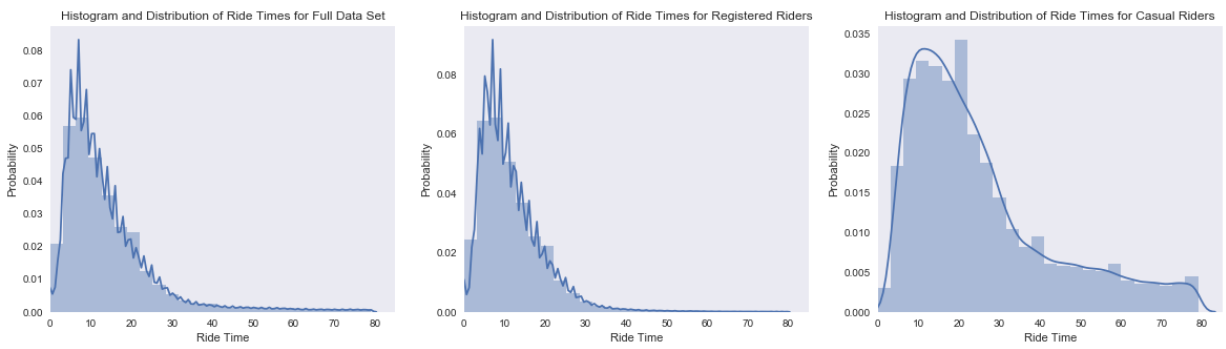
Based on the number of rides and the number of rides, one would expect that ride times would be greater in spring and summer compared to the fall and winter. This could be explained by riders deciding to ride to further locations because of more cooperative weather (warmer temperatures, rain vs. snow, etc). It can also be expected that the casual riders would be on bikes longer than registered riders.

Ride Times

The mean and median are both important to look at for the ride duration column of the dataset. The median is 11 minutes and the mean is ~13.9 minutes, which means the most rides are fairly quick.

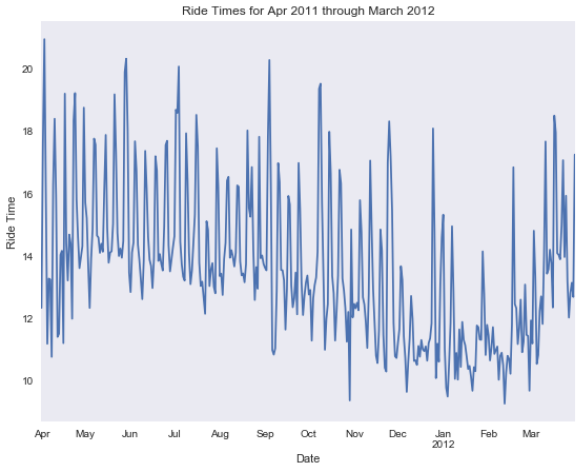
```
Mean: 13.8988544024
Median: 11.0
Minimum: 0.0
Maximum: 79.0
```

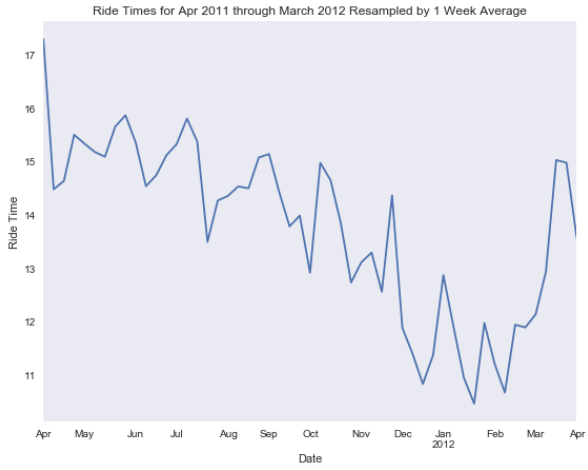
The distribution of the duration variable is right skewed meaning the most of the values occur to the left of the plot. The next question to ask is whether the distribution changes with regards to registered and casual riders.



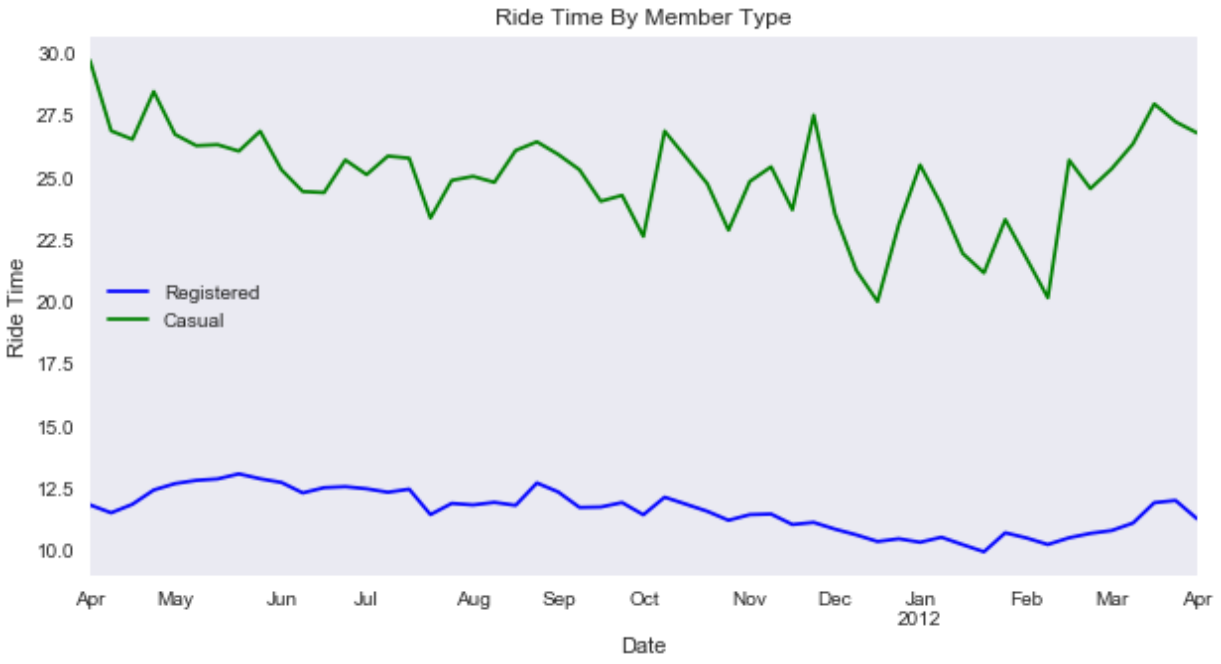
The distribution of the registered riders is very similar to the overall distribution. The casusal rider distribution is different from the other two distributions. Although the distribution is still right skewed, there is more variability in the data compared to the registered rider and overall distributions. In addition to the differences in distributions, the mean of each is different for each of the datasets (overall, registered, and casual).

Based on the analysis of the number of rides above, the ride time was expected to be shorter when there were less rides (in the fall/winter) and longer when there were more rides (in the spring/summer). The graph on the left shows the data based on daily averages whereas the graph on the right shows the resampled data based on 1 week increments. The average ride duration does decrease in the fall/winter and does increase in the spring/summer.





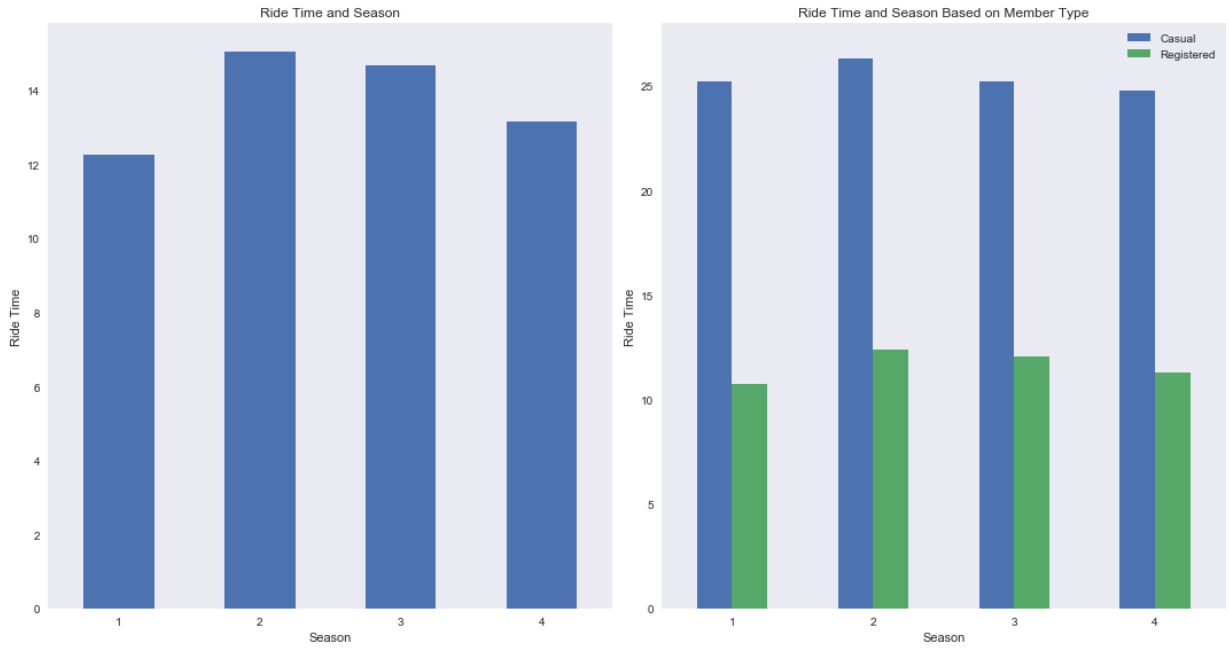
When looking at the breakdown of ride times for the two member types resampled over 1 week averages, the ride times also increase in the spring/summer and decrease in the fall and winter. However, it is not as drastic as the overall data.



Ride Times and Weather

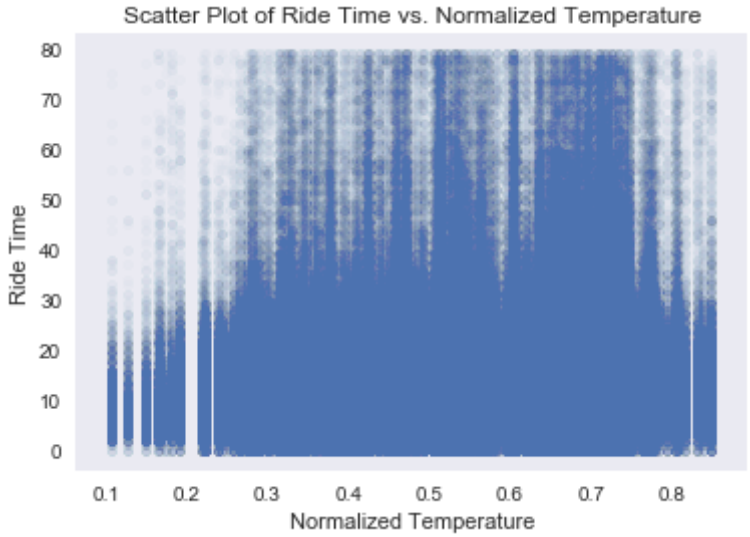
After looking at how ride times over the course of the full year, the mnext step is to examine how ride times are affected by weather. The variables included in the dataset are temperature, humidity, windspeed, and weather category (sunny, cloudy, rainy, etc).

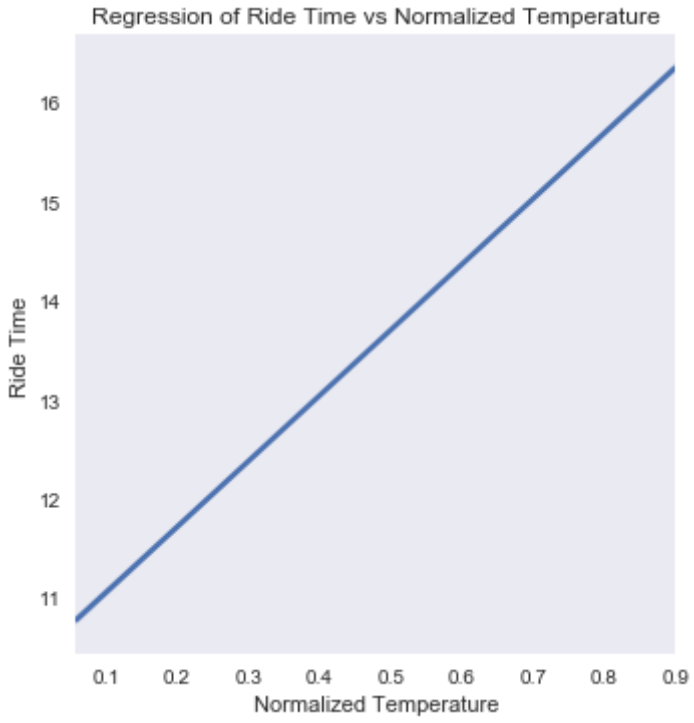
In the ride time analysis, the graphs showed that there was a change based on the season. The dataset contains a categorical variable for the season, which can be used to visualize the different means of each. The graph on the left shows the mean ride times for the different seasons for the entire dataset. The graph on the right shows the average ride times per season based on the member type. The graphs show that the average ride times in the fall and summer are greater than the ride times in the spring and winter. This is an interesting result based on the graphs in the previous section as it seemed as though the spring/summer values were greater than the fall/winter values.



The next step is to look at how the ride duration compares to the quantitative variables humidity, temperature, and wind speed. Below are the two graphs which show the scatter plot and regression plot for ride time vs. temperature. The scatter plot is hard to read due to the amount of points, so the regression plot was included. As the temperature increases, the ride time also increase, which is expected. Warmer temperatures should lead to longer rides.

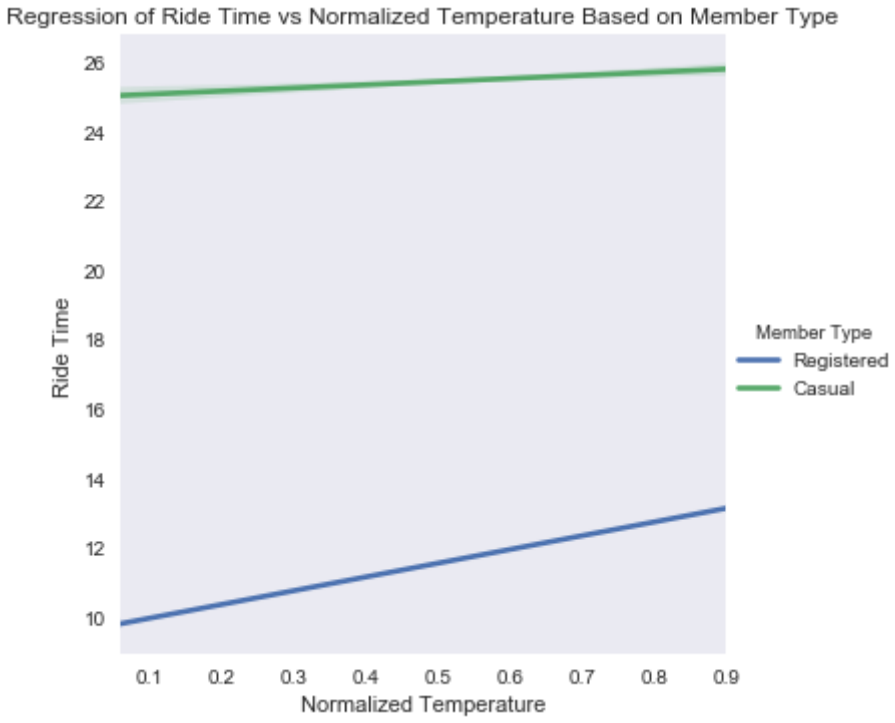
```
bikeshare.plot.scatter(x='temp', y='time_diff', alpha=0.02)
plt.grid(False)
plt.xlabel('Normalized Temperature')
plt.ylabel('Ride Time')
plt.title('Scatter Plot of Ride Time vs. Normalized Temperature')
plt.show()
```



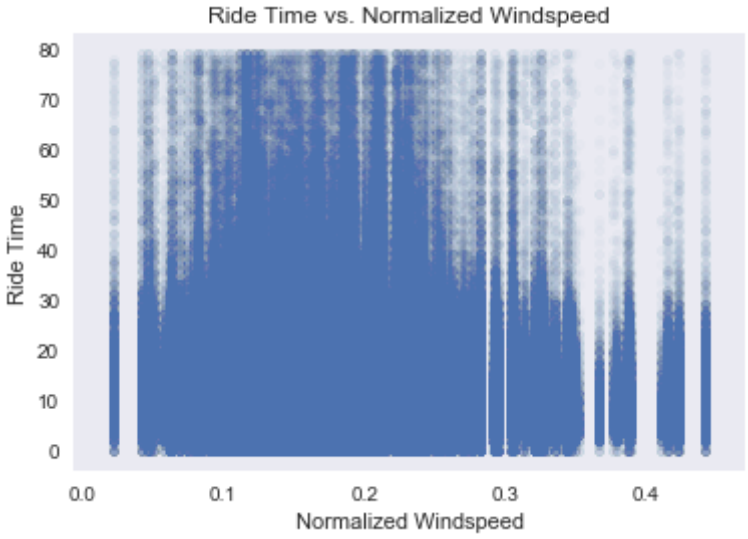


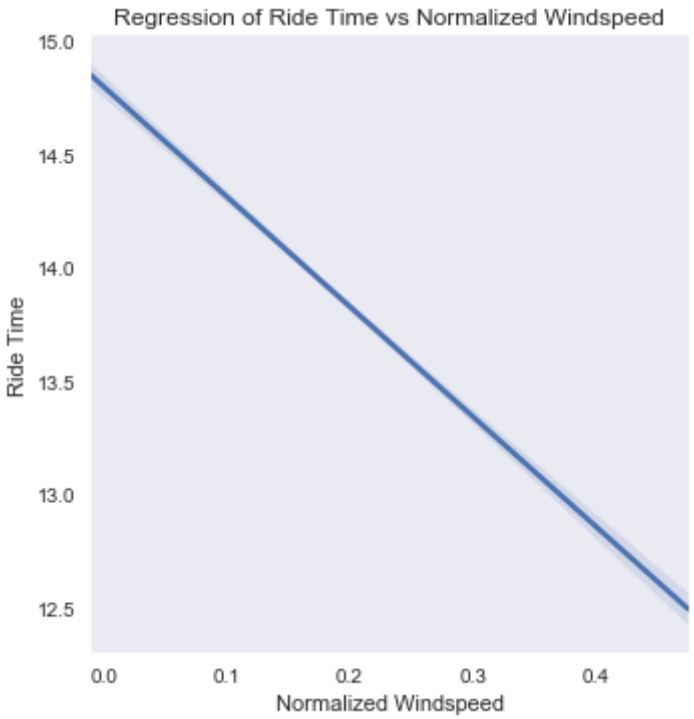
In addition to the overall dataset, the regression line is provided below for casual and registered riders. Both show the same results as the overall dataset, but the casual riders does not have the same increase. The rides for casual riders on increases a small amount when the temperature increases.

```
<matplotlib.figure.Figure at 0x117b212b0>
```



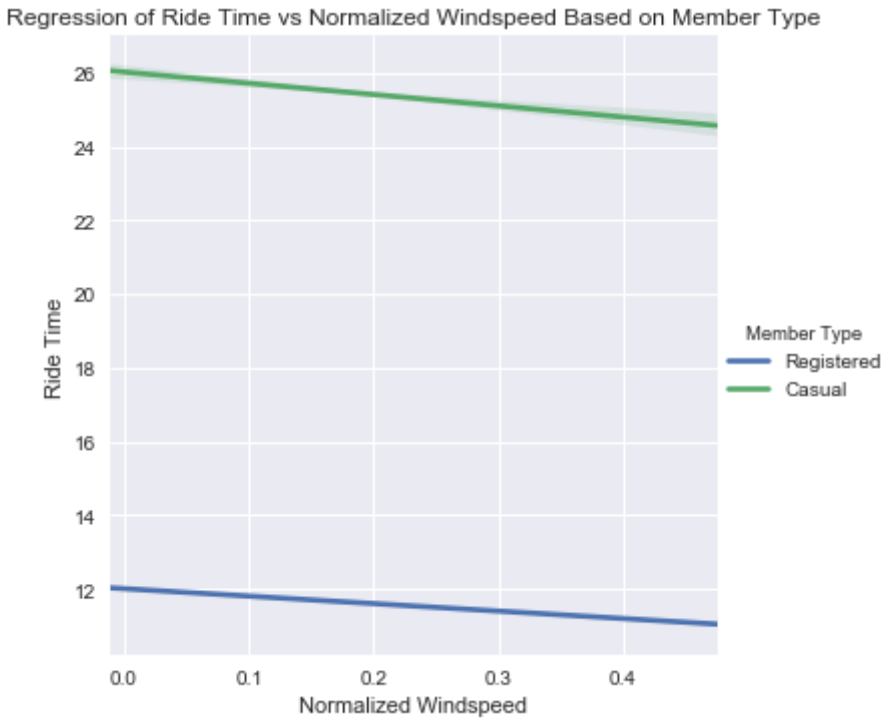
Windspeed was the next categorical variable to analyze. As with temperature, a scatter plot and a regression line were included. The scatter plot, although hard to read, seems to show that when the windspeed increases the ride time decreases. This is confirmed when looking that the regression line.





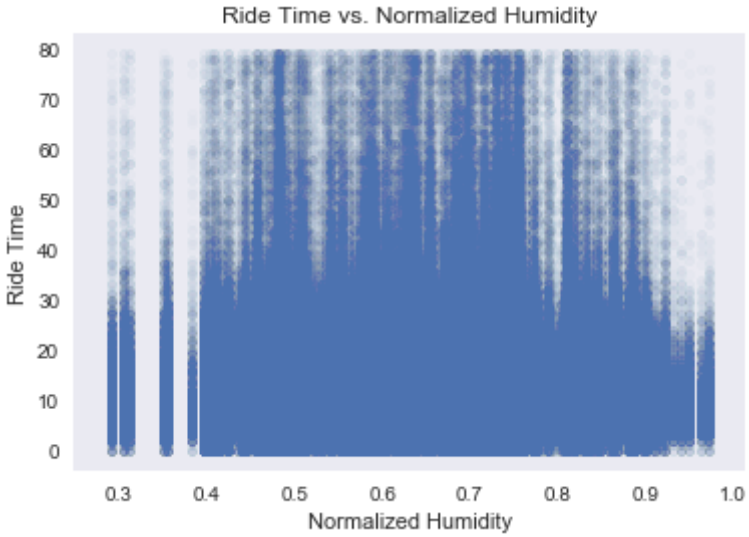
When the windspeed and ride times are broken down by casual and registed riders, the conclusions are the same in that the increase in windspeed causes a decrease in ride time.

```
<matplotlib.figure.Figure at 0x11e892518>
```

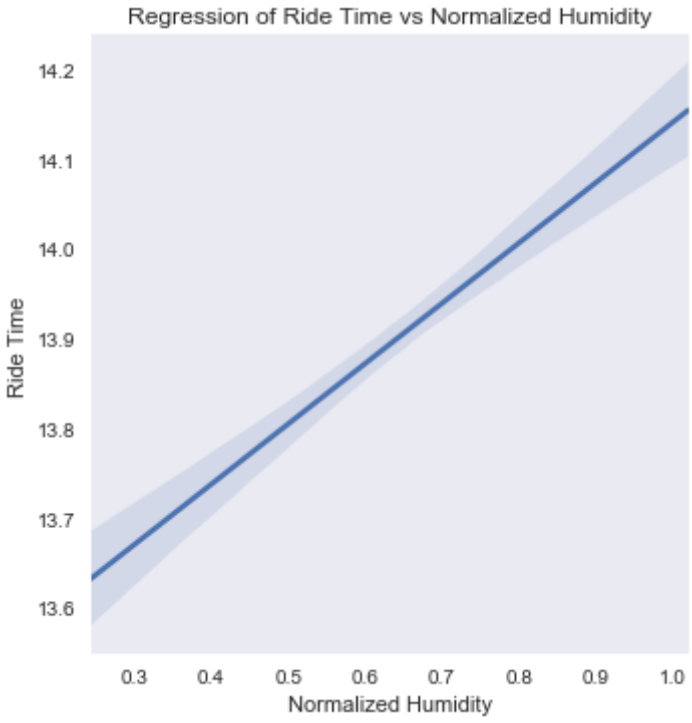


The final quantitative variable to analyze was the humidity variable and the ride times. Based on the scatter plot and the regression line for humidity, there seems to be an increase in ride time when the humidity increases. This did not seem correct as one would expect there to be a decrease. Increased humidity will increase the temperature feel so it was expected that there would be a decrease in ride time.

In [138]:

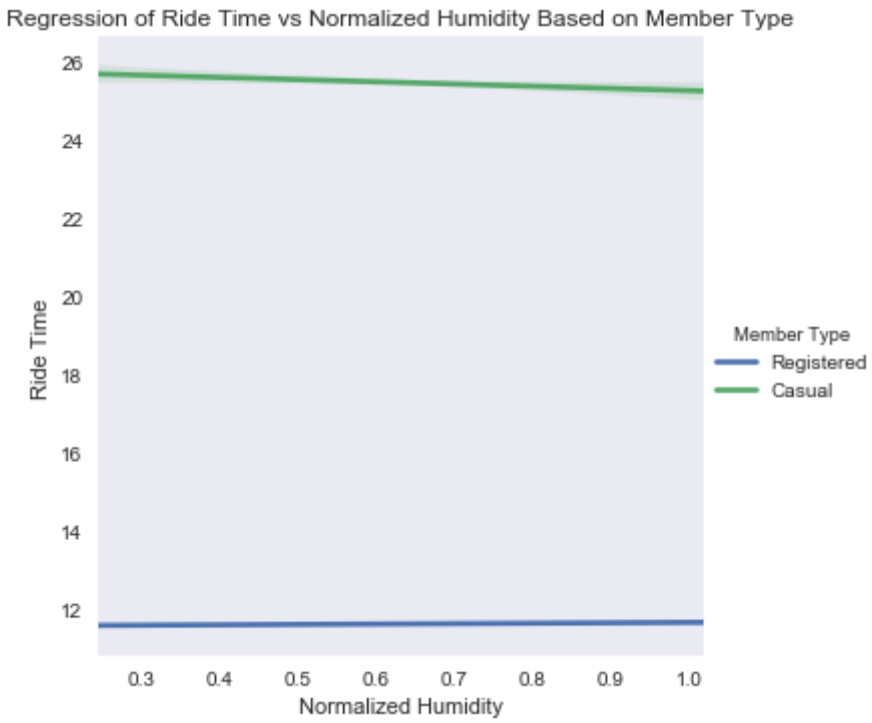


In [141]:



When ride time vs. humidity is broken down by member type, there does seem to be a decrease in the ride time, but only for casual riders. Registered riders do not seem to be affected by the humidity changes.

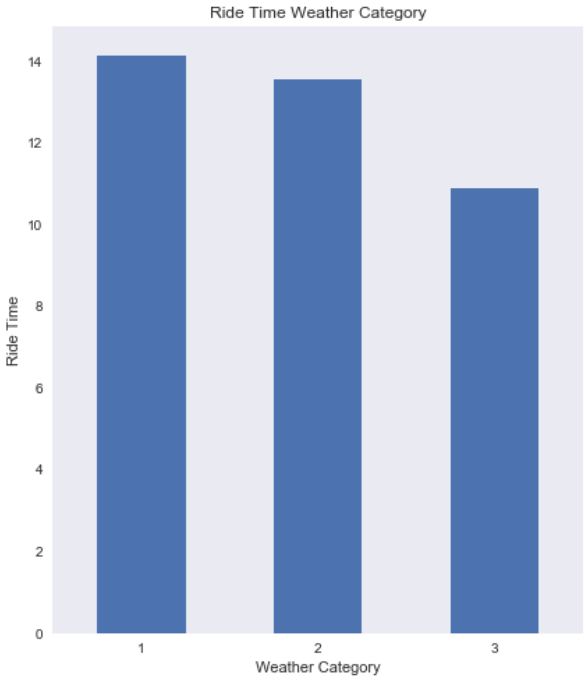
```
<matplotlib.figure.Figure at 0x117a3bcc0>
```

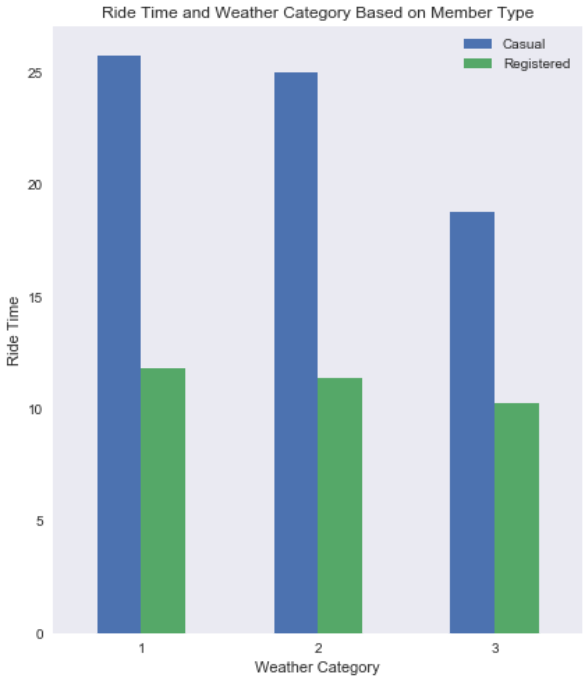


The dataset also contained a categorical variable for the weather during the day.

1. Clear, Few clouds, Partly cloudy, Partly cloudy
2. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

According to the graphs below, when the weather deteriorates, the avergage ride times decrease. This decrease also occurs when the data is separated by casual and registered riders. Casual riders have much longer ride times compared to registered riders, which has been consistent throughout the initial analysis.

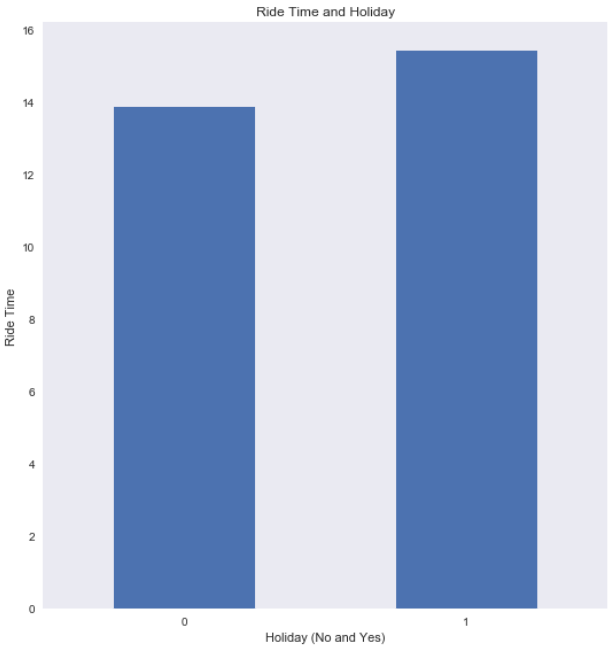


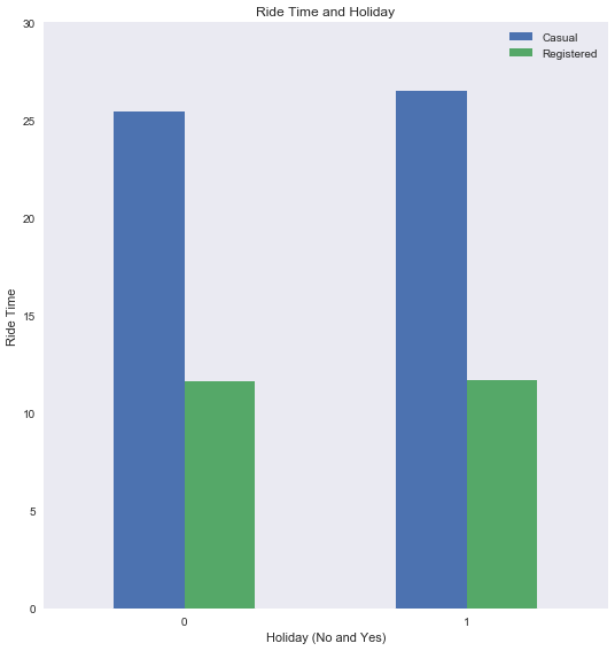


Ride Times and Calendar

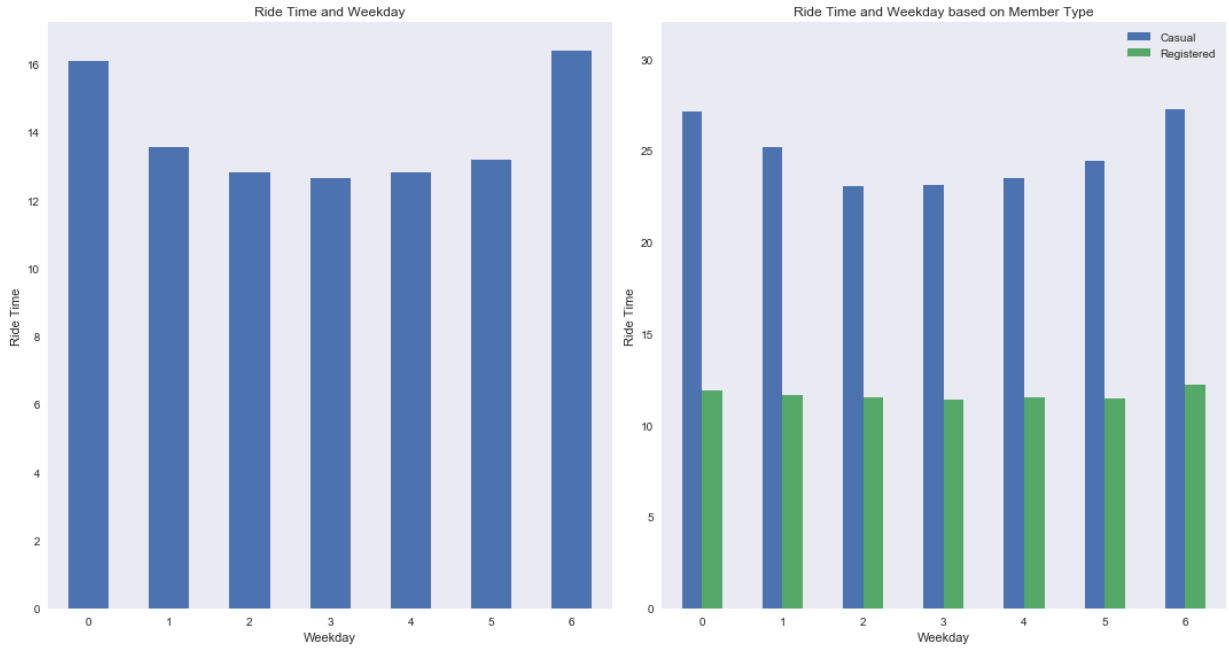
In addition to weather variables, the data also contains categorical variables regarding the day of the week, whether the day was a holiday, and whether it was a workday.

When looking at the holiday and the average ride times, there is a difference between when there is a holiday (1) and when there is not (0). The ride times increase and the comparison of member type show the same results.

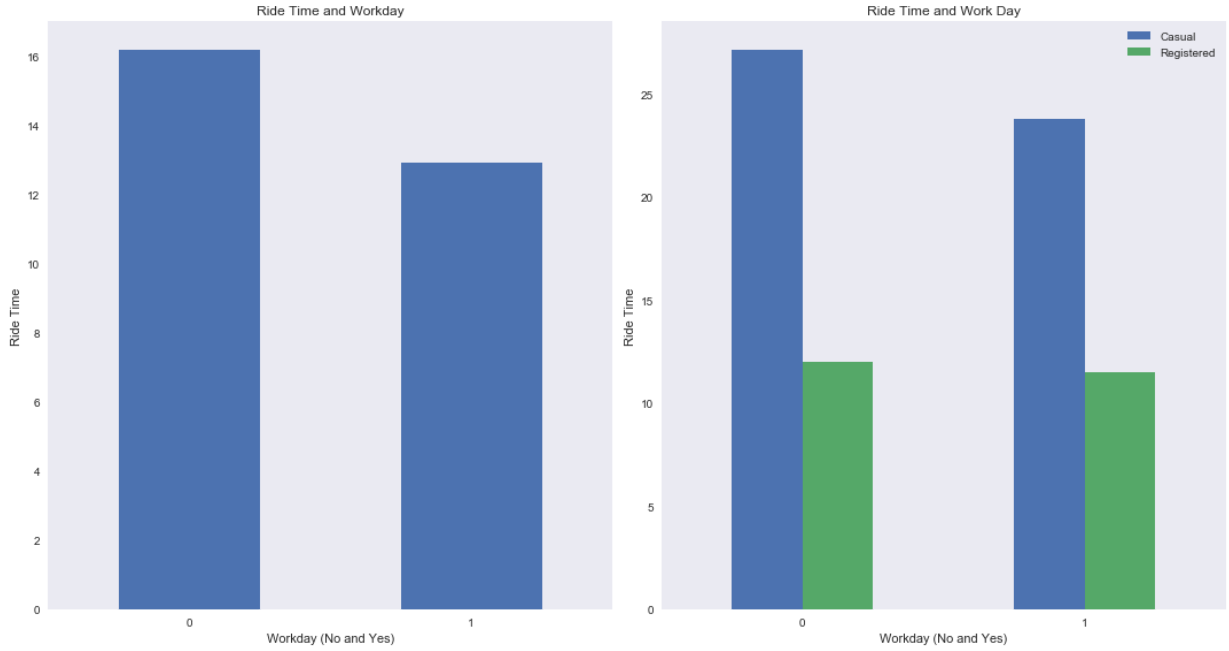




When determining the average ride times for the different days of the week for the overall dataset and the dataset broken down by member type, it shows the weekend rides are generally longer than weekday rides. However, the difference with registered riders is very minimal. The graph on the right also shows that the ride times for casual riders is much hgher than registered riders, which is consistent throughout.



The graphs of the average ride times and whether the day was a workday show that non-workday rides are longer than rides on workdays. The graph on the right, which breaks down the member type, shows that there is a larger difference in ride time for casual riders compared to registered riders.



Initial Conclusions

The biggest conclusion that can be drawn from the initial analysis is that casual riders are on the bikeshare bikes longer than registered riders. This is consistent throughout each of the different variables covered. This is to be expected as many casual riders are tourists and they seem to stop more to look at the different sites within Washington D.C..

Further investigation is needed (statistical analysis) to determine the effect of weather variables and the ride times. Just based on this analysis, warmer temperatures lead to longer rides while higher humidities and windspeeds cause a decrease in ride durationl.

In addition to the member type and weather, calendar events can also cause an increase or decrease in ride duration. Weekends, non-workdays, and holidays all increase the ride times for the overall data and for both member types.

Hypothesis to Investigate Further

Although the analysis above provides a good base for the understanding the dataset at hand, there are additional things to look at. The list below contains the additional analysis that should be completed to get an even better understanding of what variables change ride duration.

1. Analyze whether ride duration is affected based on multiple variables (like season and workday, season and holiday, etc.)
2. Determine the most common starting points, end points, and combinations of start/end points.
3. Determine if the number of bikes leaving and coming into a station is equal
    - This will show whether a station needs to have bikes transported to the station or not
4. Determine the most common areas (memorials, restaurants, neighborhoods)
5. Are the differences in means of ride times with categorical variables statistically significant?
6. Are the regression lines for the ride times and the quantitative weather variables statistically significant?