Jeffrey Chao

IST 687 – Homework 6

Date Due: 02/25/2020

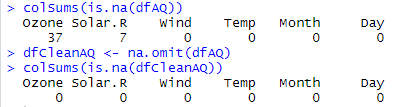
Date Submitted: 02/24/2020

**Step 1: Load the data**



**Step 2: Clean the data**

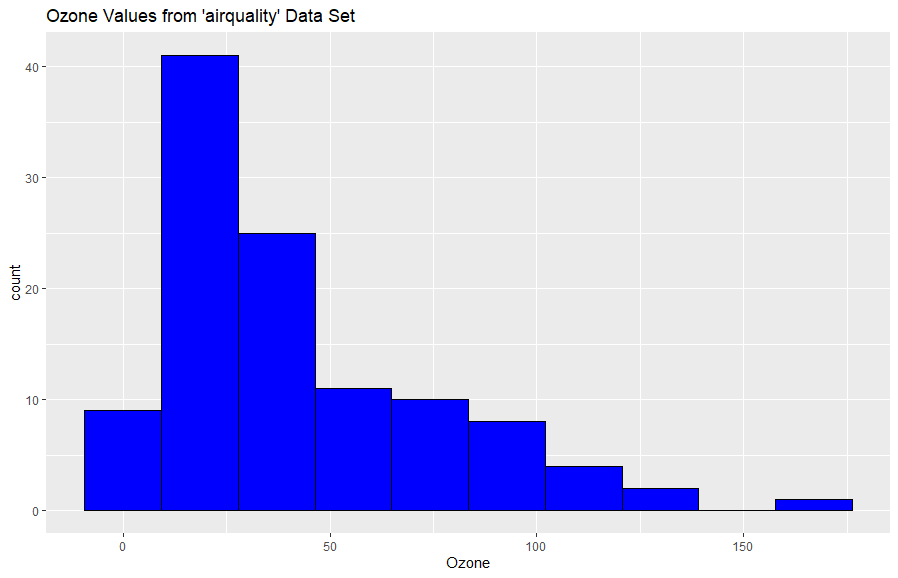




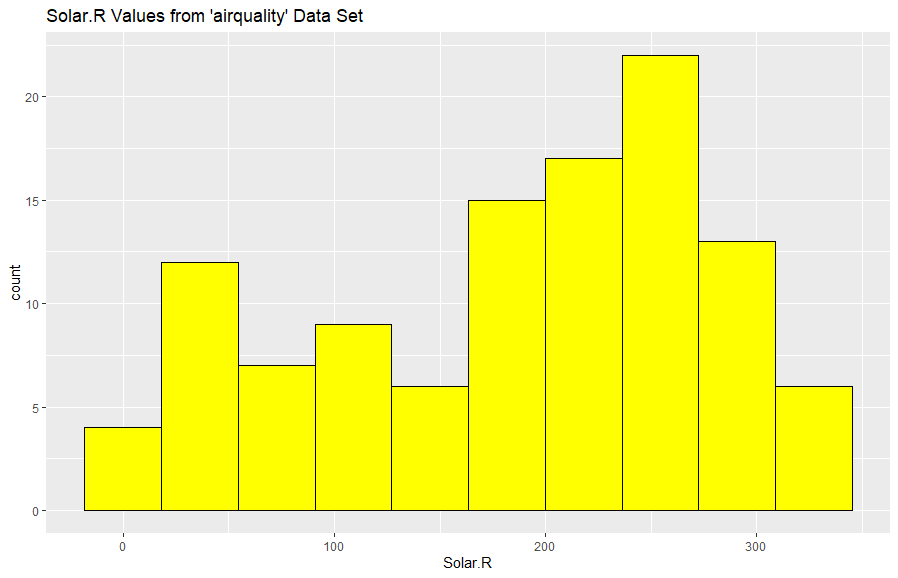
**Step 3: Understand the data distribution**

* *Histograms for each of the variables.*

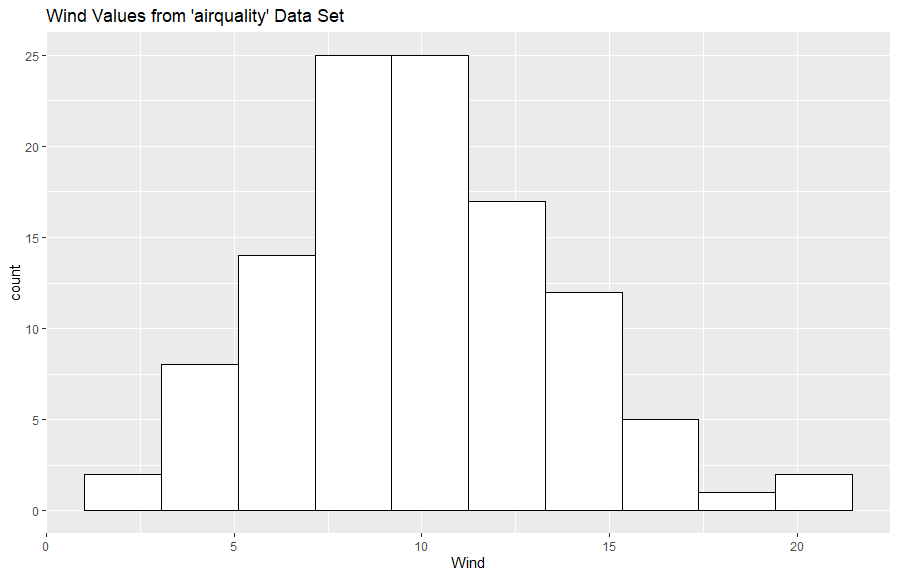




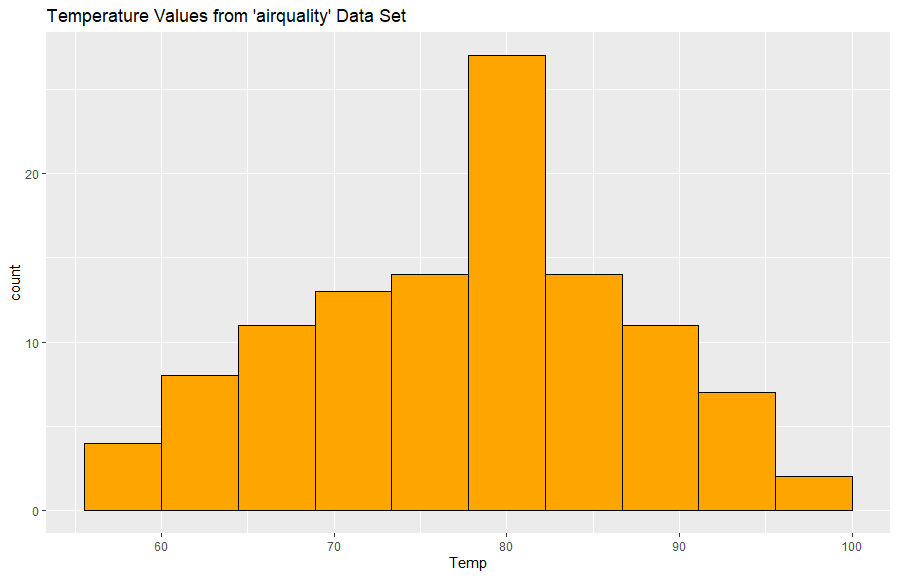




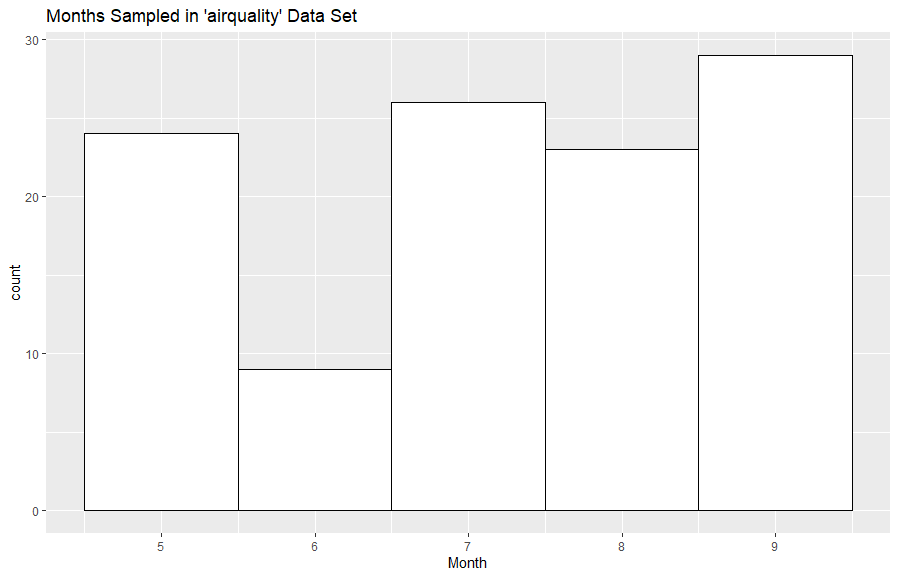




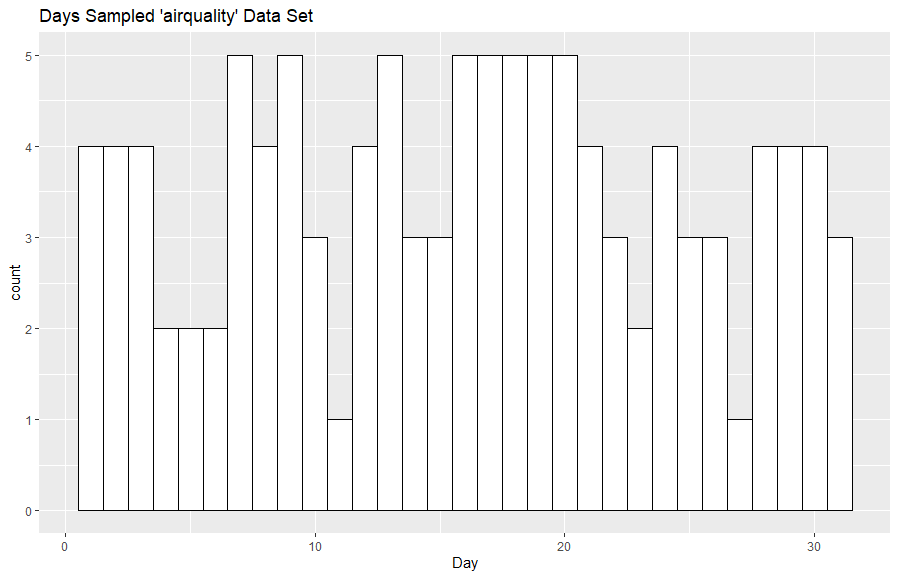






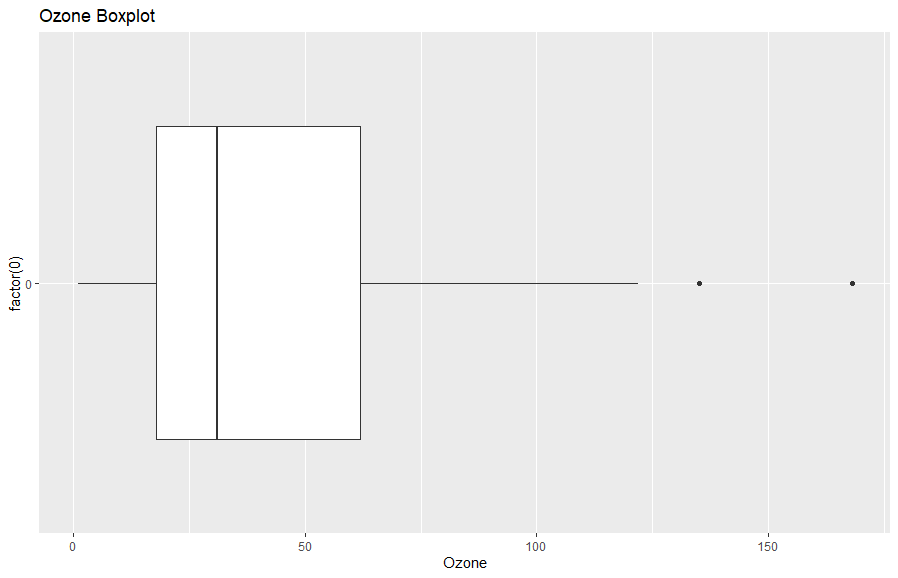






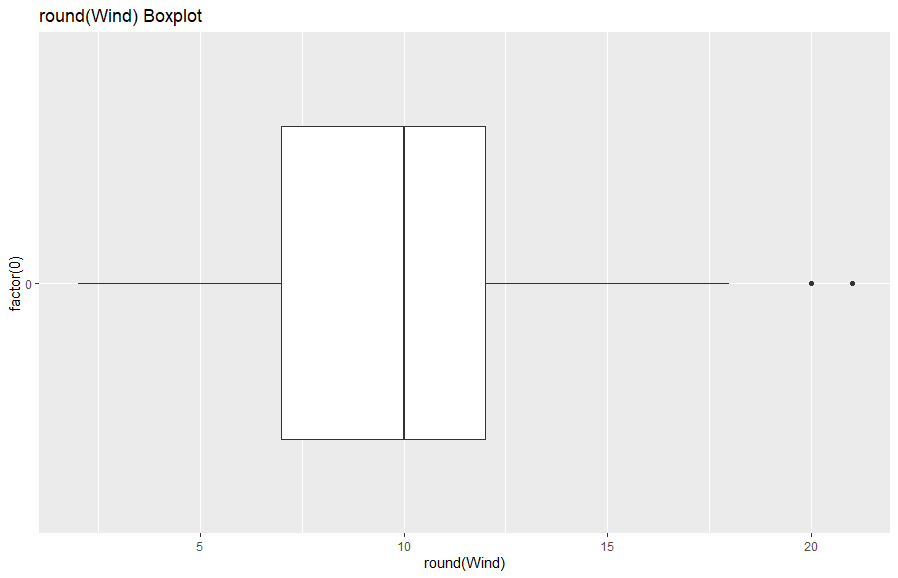
* *Boxplot for Ozone*





* *Boxplot for Wind values*

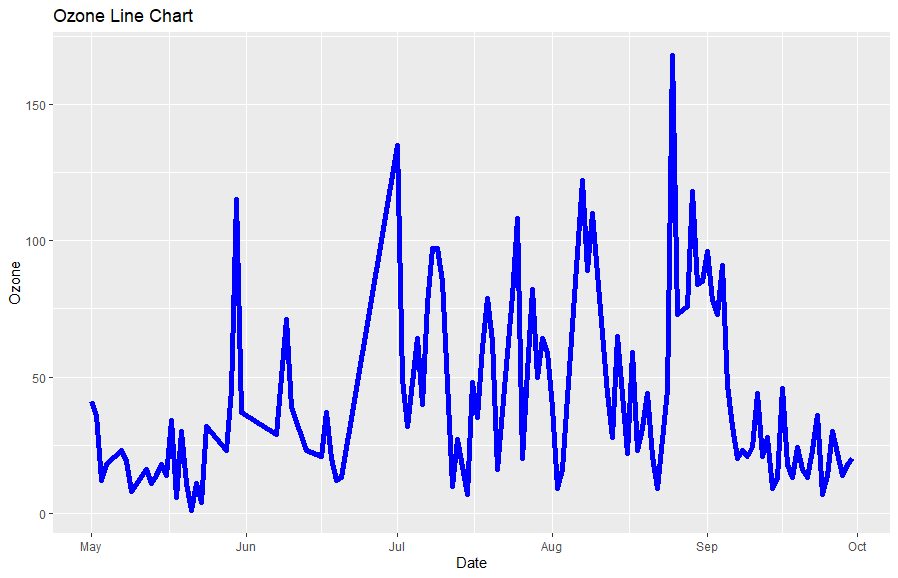




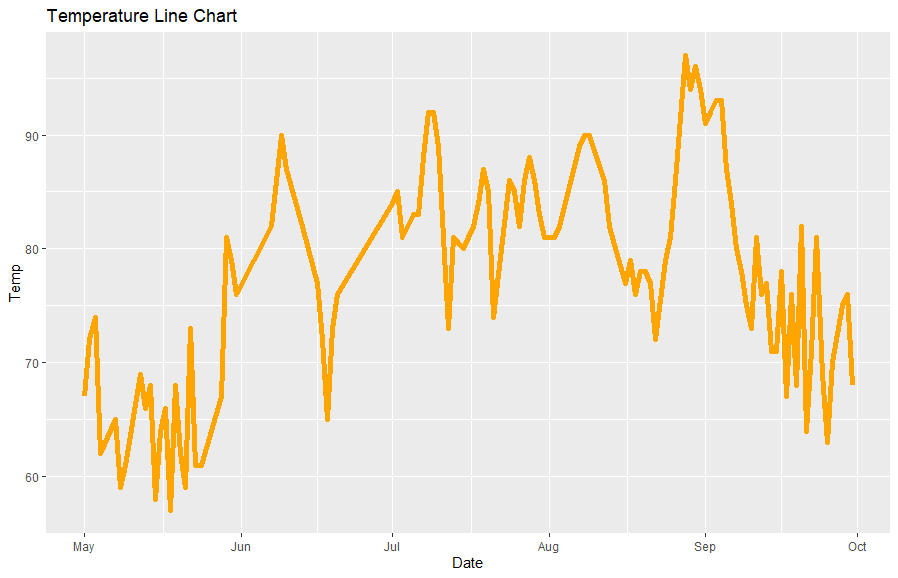
**Step 3: Explore how the data changes over time**



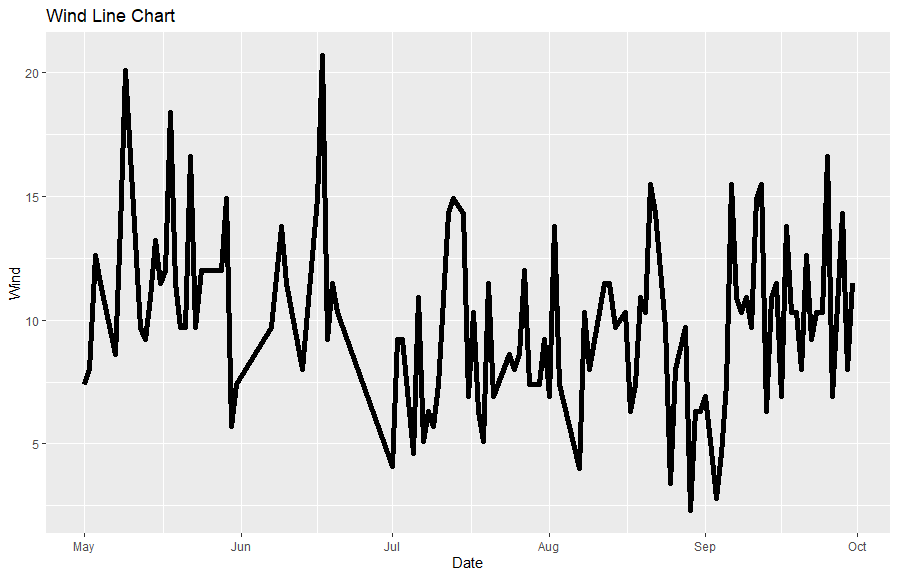




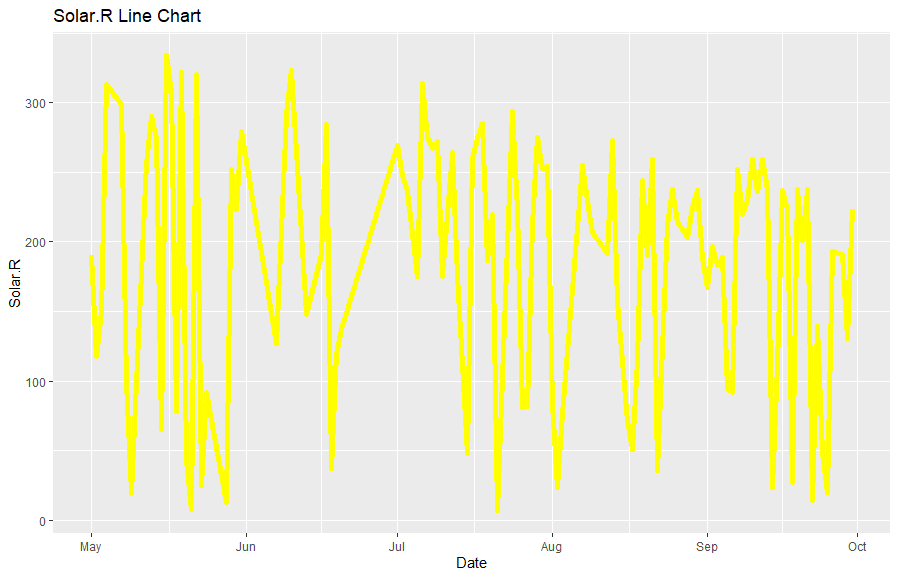


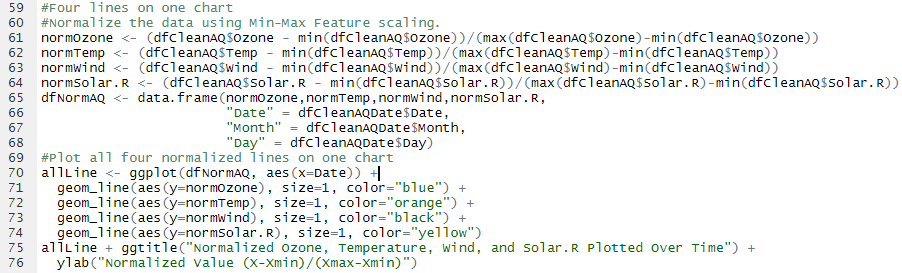


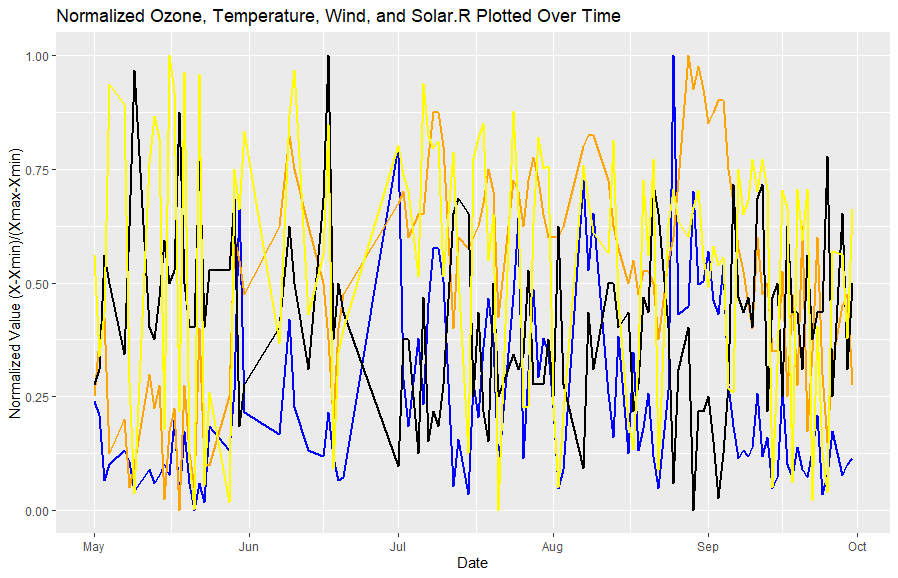




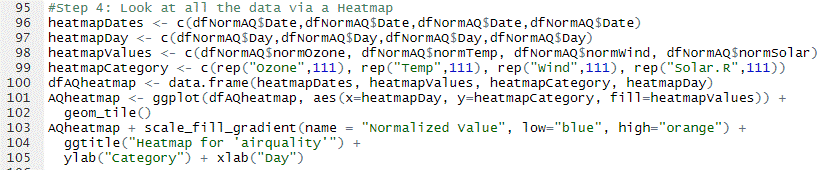


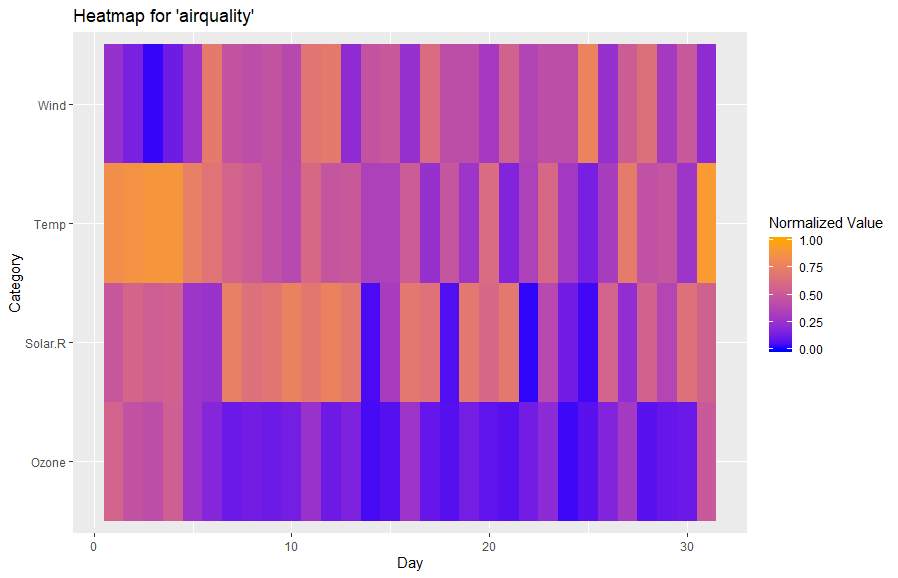






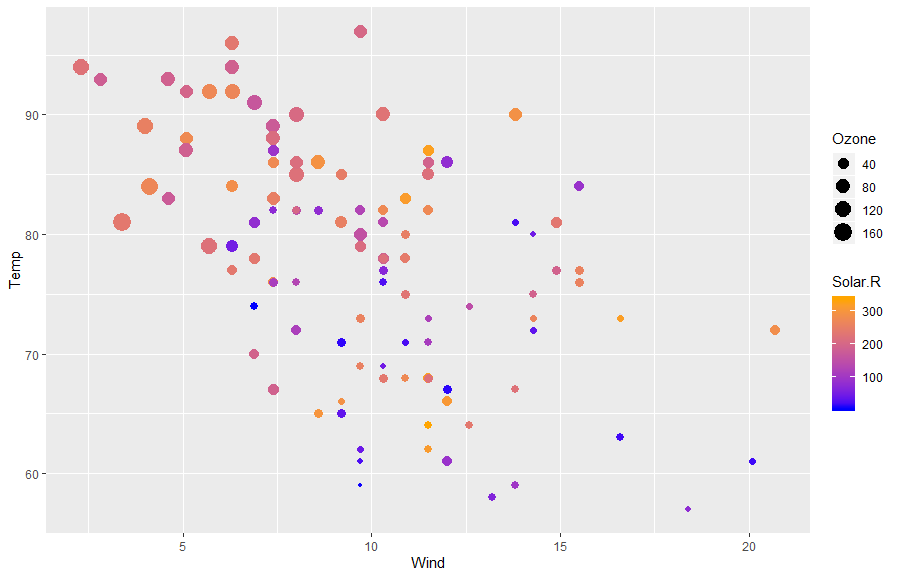
**Step 4: Look at all the data via a Heatmap**





**Step 5: Look at all the data via a scatter chart**





**Step 6: Final Analysis**

* *Do you see any patterns after exploring the data?*

1. Wind and Temp look like they are generally negatively correlated (as wind speed increases there is a decrease in temperature).
2. Temp and Ozone are generally positively correlated.
3. Wind and Temp seems to have normal distribution without any particular skew.
4. Solar.R looks like it has the least correlation with any of the other variables (Temp, Wind, Ozone, Month, or Day)
5. June had the highest count of N/A data in this data set.

* *What was the most useful visualization?*

For my personal preference, the scatter plot showing Temp, Wind, Ozone, and Solar.R all on one chart. By putting this information all together on a 2D grid, we can easily distinguish correlations and their relative strengths amongst the 4 non-time variables. Since this chart shows 4 different variables at once, instead of the typical 2, by doing combinations of these variables we can see six times as many correlations.