Austin Weather Data:

Statistical/Hypothesis Question

The goal of the project is to determine whether there are statistically significant differences between weather conditions between the year 2014 and 2016.

The data was available from the year 2013 to 2017. However, as seen from the code the complete datasets including 365 days was available only for 2104 and 2016. I wish there is a complete data se was available for the years 2013 and 2017. This would have given me more window to compare the data. However, I am little surprised to be able to detect statistically significant differences between the years 2014 and 2016.

In terms of the analysis, I don’t think anything is specially missing. The data was plotted to check for outliers, followed by histograms to visualize for any outliers. Further, in order to identify the differences, the data was plotted as PMF’s and CDF’s. Followed by the data was fitted to analytical distribution. Finally, the data was fit to linear regression and performed the hypothesis testing of difference in means. The testing clearly showed statistically significant differences between the means for the variables Average temperature, Average Dew Point Temperature and Average humidity percent. I am initially not expecting to see any major differences given the time frame of the data I am working is very narrow. However, from the dataset, I was clearly able to show the differences.

I wish there are couple more variables such as rainfall would have been really helpful to see between the years. The other information I felt missing was that the data is not completely available for all the years that are present in the dataset. Hence, I am restricted to the years 2014 and 2016 and not able to perform multiple comparisions.

The assumptions I made were that the data was completely normal for all the variables selected. However, I noticed in my normality plots that there were minor deviations at higher values of these variables. In addition, from the box plots there seemed to be some outliers. However, based on some readings about the weather data I made the assumptions that the outliers that are shown in the box plots may statistically significant but within the weather domain all those are real possible values and hence were not removed from the analysis.

The challenges I faced, with this dataset are predominantly in the initial clean-up of the data set. The data has some missing values. These missing values are different for different variables. Hence, when I remove the missing variables I am getting different length of a series for each variable and making it harder to plot. In addition I have variables where the missing values are denoted by “-“ or “T”. So I have to consider ways to specifically handle this type special characters. The analysis is shown in the where I figured out on how to handle this scenario where in the all the variables have same number of data points which eventually made easy for analysis and plotting.