Rhonald Reside
DS670 - Assignment 2

[format: Word, JSON, PDF] Data Set Loading and Data Summary

The data set I was assigned was the weather data. The source of the data set is from CityPulse. The website contains a collection of smart city data sets from Aarhus, Denmark. Aarhus is the second largest city in Denmark, located east coast of the Jutland peninsula. The population of Aarhus in 2016 was about 1.378 million. By utilizing smart city data collection of Aarhus, Data Scientist are able "find effective and sustainable solutions to the challenges faced by many cities today. "

On the CityPulse Smart City website, the weather data set is available in two types of format. One format is in .tar format and the other format is in JSON format. The data is separated by two sets of dates. The first range of the weather data set is February 2014 to June of 2014. The second set of the weather data set is in August 2014 to September 2014.

I was first interested in working with the .tar format file. Since this is my first exposure to .tar, I was really interested in seeing its data structure and content. At first I tried to research what programs or tools can read this .tar file. Unfortunately, I was not successful to open the .tar file. From my research, it seemed to be just a proprietary file only readable to certain machines. As much as I wanted to work with .tar, I had to abandon this format.

The next set of files that the weather data set came in was JSON file format. Weather data set came with seven different types of variable. It came in with Dew Point, which was in degrees Celsius. Humidity was the next variable which came in percentage. Pressure was the next variable which came in the measurement of mBar. Temperature was the next variable which was measured in degrees Celsius. Wind direction was the next set of variables which was came in the measurement of degrees. The next variable is wind speed, which was measured in kilometers per hour. The final variable was visibility.

I decided to use excel to first look at the six files of variable. I figured it was easier for me to clean and structure the data in excel. Once you opened each file, you discovered that each measurement was also time stamped. I spent about 3 hours to clean up and turn the six files in a data frame. My goal was to combine all the six variables into one csv file. This way I can use several tools to study the data so that I can come up with a good solid story. While I was creating the file in excel, it was simple for me to convert the Celsius into Fahrenheit. I used the formula Temperature in Celsius multiplied by nine fifths and adding 32 ($T_{(°C)}$ x $\frac{9}{5} + 32$).

Once I had all my variable converted into a data frame, I discovered that the time stamp for each variable was perfectly aligned. This made it very easy and clean to find a measure at a specific time. In other words, if I wanted a measurement on February 2, 2014 at 1:00 AM, I had an exact measurement of temperature, dew point, humidity, pressure, wind direction and wind speed and visibility. I was able to use excel to resave the file as a csv. I knew once I got to this point that it would be easy to move the data into different environment to study and find any patterns that would help my story for the data.

My next step after my conversion from CSV to JSON, was to understand my data set and see if I can find any patterns or gaps in the data. I decided my favorite tool to use for this analysis is Tableau. I discovered a few things once I uploaded into Tableau. First, I found that there was a data gap in from June 8, 2014 to August 1, 2014. I suspect that since this this data gap occurred throughout all the variable, there was most likely the possibility of data storage issue. The next discovery I found was the warmest day in the data set. I found that August 2, 2014 was the warmest day measured at 80.60° F with a humidity of 37 and a dew point of 11. The coldest day occurred March 11, 2014 with a humidity of 84 and a dew point of -4. I found one of the variables to be incomplete. The visibility variable only ranged from February 2014 to June 2014. I therefore am deciding not to use this variable.

I think with what I have, I can probably find a relationship between temperature, dew point and humidity. I can probably find a story to see how each of this variable make up relative humidity. One thing I would like to add, is that during my discovery phase with Tableau, I was able to visualize a correlation between each variable. As the temperature spiked in some instances, the humidity and dew point also spiked. I am deciding too since relative humidity does not really concern pressure, wind direction and wind speed, I most likely not use these variables in my story.

One thing to add is that I also brought the weather data set to R Programming. After doing this, my analysis confirmed my discovery in Tableau. My data gap in June 2014 was accurate. The warmest day and coldest day were also confirmed in my analysis in R. I also graphed my data in R. This gave me the opportunity to see the strength of Tableau Verses R. Tableau's graphing tools is cleaner and easier to read.