Rhonald Reside

DS670 - Assignment 8

Midterm Draft

**Contribution**

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.”

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. I first thought working with .tar would be a nice contribution to the study. I have never worked with .tar, nor ever heard of it. Working with it would help other students who come across .tar an understanding of how to tackle it and convert it to a format that can be used on other tools and software. 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.

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 will use the JSON format and I plan to use logistic regression for my analysis. Statistically, the use of logistic regression to find the relationship between temperature, humidity and dew point. It will use binary logic to estimate the probability of a binary response based on our independent features, i.e humidity and dew point. I will use R Programming language to model the data and then fit it into the general linear model or glm() function. The library that has this is the ISLR library in R. I will use temperature as my Y-Variable, also known as my dependent variable. I label my variables as dew point and humidity; we would also know this as independent variables or our features.

**State-of-the-Art**

My competitor’s name is Mark F Lawrence. His article’s name is “The Relationship between Relative Humidity and the Dew point Temperature in the Mount Air.” His article focuses on a three things. First, he is interested in defining the relationship between Relative Humidity with Temperature, humidity and dew point. He discusses two formulas that he believes explain this relationship. In the first formula, he says relative humidity is the ratio of the actual water vapor pressure e to the equilibrium vapor pressure over a plane of water . It is the “saturation” vapor pressure. Here is his formula for this explanation:

RH =

Another way to understand this is it is the point at percentage of the air can hold before it cannot hold any more water and then condenses and in some places, turns into rain. This explains the sudden thunderstorms that occurs in Florida. It is that “real feel” weather you feel when standing in line at Disney in Florida where your shirt is full of perspiration in the summer months of June to August. That occurrence is relative humidity.

Lawrence also shows us another equation that describes the relationship between temperature, dew point, and pressure. He states it is the ratio of actual water vapor dry mass mixing ratio *w* to the equilibrium mixing ratio at the ambient temperature and pressure. He further explains it in the following formula:

RH =

He then explains the relationship between the variables through Linear regression for moist air. He states that relative humidity must be greater than 50 percent for it to become linear.

RH =

In response to his work, with the other variable I have available, i.e. wind direction, wind speed, and air pressure, I would be interested to know if these other variables have any kind of effect on relative humidity. As we work with the data, we will have error rates for each predictor and classification accuracy to help build a better picture of the relationship to temperature.

**Data**

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 ( x .

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 used excel to resave the JSON 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 discover any patterns or gaps in the data. I decided to use Tableau for this step. Even though the beauty of Tableau is that you blend your data from any format, the steps I took in excel to convert the data into a csv file made working with the data easier in Tableau. I was able to create a database in Tableau just for the weather data.

I find it easiest to visualize my data after it was loaded into Tableau. I created line graphs, heat maps and averages to help me build a story from the data. I discovered a few things after this step using Tableau. First, I found that there was a data gap in from June 8, 2014 to August 1, 2014. I could see this once I created a line graph with temperature, dew point and humidity. It was a very interesting visual find. I suspect that since this this data gap occurred throughout all the variable, there was most likely the possibility of data storage issue or they all ran out of battery at the same time.

The next discovery I found was the warmest day and the coldest 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 decided to create a heat map that represented the high temperature all the days. I also did the same for the cold and created heat map of the coldest temperature day

Next I loaded all the data in R. By doing so, 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. At this point it would not matter because visibility variable does not really affect relative humidity.

**Method**

I feel needed to apply complex statistic or machine learning to the project. R programming language made this step easier. I decided to first split my data into two data sets. I created a training data set and a test data set. This is an important step because you can run your statistics on your training data set. I hoped at this point that the statistics I have a very low error rate so that I can beat my competitive article. Just as a reminder, my competitive article has an error rate lower then 5%. The other point I would like to mention is that no matter how low your error rate is on your training data set, you expect your testing data set error rate to be slightly higher or at least very close to your training data set error rate. If your error rate for your testing data set is lower, then your training data set is not a good representation of your total data set.

I plan to use logistic regression as part of my analysis. Even though we can see there is a relationship with temperature, dew point and humidity when graphing in Tableau, I would like to see a statistical number that show this correlation. So when in R programming, I use the library ISLR for my statistical analysis. I plan to use temperature as my Y-Variable, also known as my dependent variable. I label my variables as dew point and humidity; we would also know this as independent variables.

From the ISLR library, I use the glm() function for the logistic regression model. I first use a binary response to tell me if my temperature is a hot day or a cold day. I use 65° as the decision variable. If the temperature is above 65°, then my algorithm will lable the temperature hot. If it is lower then 65°, then the temperature is labeled cold. By labeling my temperature inputs, the glm() function will create a confusion matrix. A confusion matrix is a table used to describe the classification performance of my model of my training data set. In this case, it shows us the accuracy of our classifier of 65°.

The generalized linear model us useful for predicting an outcome from a binary response from a data set. It is sometimes called discriminant function analysis because its assumptions are less restrictive. Our formula for the data set is

logistic\_model = glm ( Yresults ~ tmpm\_Fahrenheit + Dewptm + hum + Pressurem + wdir ,   
 data = data1 ,   
 family = binomial   
 )

In the formula above, we are calling our function a logistic\_model. Inside our glm() function, we have our independent variable and dependent variables, our data set, and binary response. Again, Yresults is our temperature. Our independent variable is the rest of our variables , , ,, and (i.e. temperature, dew point, humidity, pressure and wind direction.) Data is our training data set. Family equals binomial is our binary response to 65°.

We fit our data to model the glm() function. As an outout, we see the variable in the usual way and a receive a binomial error distribution. The estimators of the coefficient show us weight influences in a positive manner while displacement is negative. The goal is to calculate the predicted probability of temperature for the specified values of our predictors, i.e. dew point, humidity, pressure and wind direction.

I predict that there will be a strong correlation with humidity and dew point and weather. My hypothesis is that our output will agree with our competitive article. I think the glm() function will have lower error rate then our competitive article therefore making our model a better choice for relative humidity correlation. Our competitive article believes that you can lower your error rate when fitting the data without the help of computers. I believe otherwise, that it is very necessary to use computers to predict. Using computers will reduce your error.

My data download time was about 10 minute. This time reflected the amount of time to go Aarhus smart city data and download the data. Data upload time was 60 seconds. This time reflects the amount of time it took to open and load the data. Data Conversion took me about an hour. This was where I converted the file from JSON to CSV. I accomplished this part in Microsoft Excel. Data cleansing worked hand in hand with data conversion. In this process I cleaned and converted the data from multiple files. I was able to combine all the files into 1 csv file. I was lucky because it was somewhat of clean time stamp where all the variable were in sync.

My data exploration took about 3 days. This was where I spent most of my time in the process. I used Tableau, R Studio, and zeppelin. This was where I found the data gap, saw visual relation between my variable, find my hottest and coldest days, etc. Once I figured out what kind of story I wanted to build, I loaded my data to R studio to find some statistical values. Because I spent a fair amount of time in the data cleansing stage it was fairly quick to fit the data into the glm model. This process took about 20 second to run. Data visualization took some time in the Tableau phase. Since I am still learning Tableau, it took a little longer then I wanted to spend during this time to produce the graphs, data and maps. This process took me about 3 hours.

**Conclusion**

My work did outperform my competitors work because first on speed, he wanted to show that you can calculate without using a computer. My work is from computers and program. There is no way hand calculation can beat a computer and produce my high accuracy rate.

I agreed with my competitor that there was a direct relationship with temperature, humidity and dew point. We saw that in the very beginning with the use of Tableau. We confirmed using the glm() function in R. Our statistics showed the relationship.

We also saw that from the dataset, the warmest day was on August 2, 2014 at 80.6⁰ F, with average humidity of 37 and dew point of 11. We also found Coldest day was on March 11 at 26.6⁰ F, with average humidity of 84 and dew point of -4.

As I present my results to the class, I would like to explain what this all means. Its easy to see numbers and definitions, but if you are not able to comprehend or experience what it all means, my research is a failure. Therefore, I would first see if anyone really knows the difference between high relative humidity vs a lower relative humidity. I usually like to give a real world explanation so I would give the example of this upcoming summer and talk about vacation plans.

Having three kids excited to have me back from school, they are asking about going to Disney. I would ask if anyone has ever been to either Disney in California verses Disney in Florida. Reason I ask is because Florida has the extreme case of relative humidity. In this location, the air gets saturated daily with humidity that you have the instant thunderstorm. It is a location where if you just took a shower and got dressed to walk around the park, your shirt would most likely be wet from that humidity. It is sometimes unbearable to just go outside, let alone walk around an amusement park.

Disney in California does not have this kind of humid weather. They have that desert dry heat. You will not have that problem of sweaty shirts and unbearable heat. Your shirt will be fine all day in the 85˚ heat. California has a lower relative humidity because of this temperature, humidity, and dew point combination.

So where will I take my kids this summer? Almost guaranteed California. It is more pleasant to wait on line for that picture with Mickey Mouse. I think the two locations have a perfect explanation of real feel relative humidity. It is important to understand how relative humidity can affect us.