**Rhonald Reside  
DS-670  
Assignment**

**Final Teleprompter Script**

**Slide 1**

Good evening everyone, if you do not know me by now, my name is Rhonald Reside. First of all, I would like to say if we are too look at this Venn Diagram, and understand what each piece means to be a Data Scientist. That is what it means for Statistic and Modeling, Coding with Performance on Big Data, and Communicating results for Decision-Making, I would say that I spent most of my time here, in the intersection of Statistic and Modeling and Coding with Performance on Big Data. As we saw in the previous presentation of weather, I wanted to share a different angle on weather. I would like to point out that our colleagues did an excellent job interpreting their findings. Here in the intersection, I know I spent at least 75% of my time with this research. It was important for me to know and really understand my data so that I can give you a compelling story. So lets move to the next slide.

Slide 2 - **Contribution of Competitor’s Article**

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 =

Slide 3 - **Description of Your Contribution**

I plan to use logistic regression as part of my analysis. I will use logistic regression to find the relationship between temperature, humidity and dew point. I used R Programming language to model the data and then fit it into the glm function. I will use a ISLR library in R. 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.

Slide 4 - **Data Source and Content**

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. 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.

The data files that the weather data set came in was JSON file format. Weather data set came with seven different types of files (i.e., 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 discovered that there was a data gap somewhere in the month of June. 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.

**Slide 5 – Your Method**

1. Data aggregation Download data
2. Clean and structure data. Converting JSON file to CSV
3. Visual analysis of data using Tableau and R programing. Graph the variables, i.e. temperature, dew point and humidity. Create heat maps of the cold temperature and warm temperature.
4. Use R programing to create statistics with the weather data. This can confirm the visual findings found from Tableau. We can confirm our warmest temperature, our coldest temperature, and finally the relationship between temperature, dew point and humidity.
5. Split the weather data set in R. I would be creating a training data set and testing data set. Splitting our data would be important when we cross validate and evaluate our glm model.
6. Fit our data into the glm() function.
   1. Binary response of 65. We classify our temperature if its hot or cold.
   2. Output from the glm() function will include:
      1. Error rate
      2. Z-value
      3. Weight of variable
7. Evaluate the model. Understand the output from the glm() function
8. Apply model to the test data set.
9. Conclusion

**Slide 6 – Quantitative Results 1**

Used Tableau to visualize data set. 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. 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.

We can visually see that there was a relationship between Temperature, Dew point and Humidity. From the graph, we discover a few other things. 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.

I 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 used 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. 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°.

**Slide 7 – Quantitative Results 2**

In Tableau, I created a heat map for hot and cold days. 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.

I used Zeppelin to further analyze my weather data. I aggregated the data using Zeppelin to create a subset of warmest and coldest day. One neat feature I was able to create was this drop box. It gave the ability to extract the temperature, dew point (average), and humidity (average). It confirmed the accuracy of Tableau and R programming.

**Slide 8 - Discussion: Comparison with Your Competitor**

My competitors work produced an eç rror rate of less than 5 percent. He explained and confirmed the Direct relationship between temperature, humidity and dew point. We see my error rate is .09 percent with my classification rate of 90.34 percent. The classification decision was based on a binary response of 65˚. This was my threshold. My logistic regression was able to predict a 90.3% accuracy rate on the test data.

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.

Slide 9 - **Performance on Big Data: Time Measurements**

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.

**Slide 10 – Conclusion**

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.