SMU Master’s Program

Data Mining Course - Summer 2018

Project 1: Data Visualization

Sarah, Sheri, Kyle, and Casey

**Grading Rubric**

**Business Understanding (10 points total).**

* ***Describe the purpose of the data set you selected*** 
  + ***(i.e., why was this data collected in the first place?).***

Wind turbine operators face a competitive marketplace for selling power. The historically low price of natural gas and the amazing success of solar and wind power has created thin margins. This has greatly benefited customers and the environment. Since the upfront costs are fixed, the operating cost of the wind farm can be the difference between making money and losing money in the long-term. In recent years there have been advances in technology and economies of scale that have made great improvements in the uptime of equipment. One such improvement is the flood of real-time sensor data measuring all sorts of parameters from the temperature of components to the humidity of the air. This data is collected to provide engineers with the data they need to troubleshoot equipment issues or analyze performance.

* **Describe how you would define and measure the outcomes from the dataset.** 
  + **That is, why is this data important and how do you know if you have mined useful knowledge from the dataset? How would you measure the effectiveness of a good prediction algorithm? Be specific.**

Data Understanding (**80 points total**)

• [**10 points**] Describe the meaning and type of data (scale, values, etc.) for each

attribute in the data file.

• [**15 points**] Verify data quality: Explain any missing values, duplicate data, and outliers.

Are those mistakes? How do you deal with these problems? Give justifications for your

methods.

• [**10 points**] Visualize appropriate statistics (*e.g.,* range, mode, mean, median, variance,

counts) for a subset of attributes. Describe anything meaningful you found from this or if

you found something potentially interesting. Note: You can also use data from other

sources for comparison. Explain why the statistics run are meaningful.

• [**15 points**] Visualize the most interesting attributes (at least 5 attributes, your opinion

on what is interesting). **Important: Interpret the implications for each**

**visualization**. Explain for each attribute why the chosen visualization is appropriate.

Page !17 of !39

• [**15 points**] Visualize relationships between attributes: Look at the attributes via scatter

plots, correlation, cross-tabulation, group-wise averages, etc. as appropriate. Explain

any interesting relationships.

• [**10 points**] Identify and explain interesting relationships between features and the class

you are trying to predict (i.e., relationships with variables and the target classification).

• [**5 points**] Are there other features that could be added to the data or created from

existing features? Which ones?

• Exceptional Work (**10 points total**)

• You have free reign to provide additional analyses.

• One idea: implement dimensionality reduction, then visualize and interpret the results.