Data Mining Fall 2016

Project 1: Data and Visualization

Assigned: 9/1/2016
Due: 9/22/2016 (via Canvas)
Points: 100

Please submit your report in **PDF format.**

Did you choose the right university?

Ranking universities is a difficult, political, and controversial practice. There are hundreds of different national and international university ranking systems, many of which disagree with each other. We will use a dataset compiled for kaggle.

https://www.kaggle.com/mylesoneill/world-university-rankings

The dataset contains three global university rankings from very different places and includes supplementary data about educational achievement and public and private expenditure.

Of course we are interested in SMU's ranking, but we are also interested in universities in Texas and how US universities perform compared to universities in other nations.

Write a report covering in detail all steps of the project. The results have to be reproducible using your report. Carefully describe every assumption and every step in your report. Also, mention any program/code/additional data that you are using for your analysis.

Follow the CRISP-DM framework

1. Business Understanding [10]

- Define the purpose of using data analytics/mining to achieve possible goals that university administrators, politicians (local and national), employers, students and parents may have. [3 points]
- What actionable results could your data mining project produce? [4 points]
- How would you define and measure effectiveness of this data mining project (you may choose to focus on a few of the identified goals)? What data would be needed and where would you get the data to judge the effectiveness of? [3 points]

2. Data Understanding [80]

- The meaning of the different variables can be found on the website. Describe the type of data (scale, values, etc.) for the variables in the data file(s). [10 points]
- Verify the data quality. Are there missing values? Duplicate data? Outliers? Are those mistakes? How do you deal with these problems? [10 points]
- Give simple appropriate statistics (e.g., range, mode, mean, median, variance, counts) for each variable and describe what they mean, especially, if you found something interesting. **Note:** You can also use data from other sources for comparison. [10 points]
- Visualize the most important variables appropriately (at least 5 attributes). **Important:** Provide an interpretation for each chart and explain for each variable why you chose the used visualization. Charts without explanation are useless! [15 points]
- What are the strength/weaknesses of SMU. Can you visualize these? [10 points]
- Explore relationships between variables with appropriate methods (a minimum of 5 relationships). Use, for example, scatter plots, correlation, cross-tabulation, group-wise averages. [20 points]
- The data contains only the country. If you can add the state or even the geolocation (GPS coordinate) of each university in the US, what could you do with this spatial data? Can you do this (exceptional work)? [5 points]

Exceptional Work [10 points]