Lab One: Visualization and Data Preprocessing

You are to perform analysis of a data set: exploring the statistical summaries of the features, visualizing the attributes, and making conclusions from the visualizations and analysis. Follow the CRISP-DM framework in your analysis (you are not performing all of the CRISP-DM outline, only the portions relevant to understanding and visualization). This report is worth 20% of the final grade. Please upload a report (one per team) with all code used, visualizations, and text in a single document. The format of the document can be PDF, *.ipynb, or HTML. You can write the report in whatever format you like, but it is easiest to turn in the rendered iPython notebook.

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?). 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? Be specific.
    - [**10 points**] Give simple, appropriate statistics (range, mode, mean, median, variance, counts, etc.) for the most important attributes and describe what they mean or if you found something interesting. Note: You can also use data from other sources for comparison. Explain the significance of the statistics run and why they are meaningful.
    - [**15 points**] Visualize the most important attributes appropriately (at least 5 attributes). Important: Provide an interpretation for each chart. Explain for each attribute why the chosen visualization is appropriate.
    - [**15 points**] Explore 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.