School of Business
OPIM 5604 – Predictive Modeling
Quick Summary

Data Preprocessing

1. Prepare the data to conduct pattern discovery and predictive modeling analysis.
2. Data types – nominal, ordinal and continuous
3. Too many nominal variables can result in problems with data analysis. To the extent possible and when reasonable, try to frame the variables as continuous or at least ordinal.
4. Steps in data preprocessing
    a. Correct inconsistencies and obvious errors – for example, for 'gdp prediction' some may enter 0.05 and others may enter 5. Both represent 5% growth but were entered into the data inconsistently.
    b. Detect and deal with outliers
        i. Box and whiskers plots are helpful to find extreme values (very high or very low)
        ii. For numeric (i.e. continuous data) data that has extreme values, you should transform the data to make it look as 'normal' as possible. JMP provides a 'continuous fit' option to find the best transformation. This may significantly reduce the number of outliers in the data.
        iii. An excellent approach to detect outliers is to find data points that are outside of the 'structure' of the data set. For example, a height value of 6 feet may not be abnormal in of itself; but a height of 6 feet on an individual who is 5 years old is obviously an outlier. JMP provides options to detect such outliers. These are indicated by the Mahalanobis Distance values. Create them and use your judgment to keep/throw-away outliers
    c. Missing Values
        i. Number of options to deal with them
            1. Throw them out – danger is that it can result in severe loss of data
            2. Fill the missing values
                a. With average value of the variable
                b. Create a 'category' for missing data. For example, if there are missing values in 'gender' you can then create 'F', 'M', 'U', with 'U' representing 'unknown'. May be useful if you want to analyze missing values – may be there is something different about people who don't provide this information.
                c. Impute (means – fill in) missing values – JMP provides an option to impute the missing values (<u>continuous, numeric data only</u>) with the most probable values (based on analysis of the variable against other variables)
                d. Just leave it alone – predictive modeling techniques can deal with them.
                e. <u>Never impute (or fill in) variables you want to predict!</u>

d. Data reduction
    i. If the data set is too large and can create problems with analysis down the line
    ii. You can reduce the number of rows and/or number of variables (or columns)
    iii. Reducing the number of rows
        1. Sampling is a good technique to select only a subset of the data
    iv. Reducing the number of variables
        1. Logic – different variables essentially convey the same basic information. For example, 'bonus' and 'performance rating' both essentially capture information about how well an employee performed.
        2. Principal Components Analysis – For numeric (continuous data) variables – removes the overlap of information between variables. It does it by creating 'linear combinations' or 'principal components' that capture most of the variance of the variables.

5. Carefully document and justify all the data preprocessing steps you take to clean the data. Also, retain the original, untouched data set, along with the new and cleaned data set.