Project #1. CollegeTuition

You work as a researcher at the US Dept. of Education and are requested to create a predictive model of college tuition based on a number of characteristics gathered from higher education institutions. The data collected from a sample of US Colleges has 1121 records and includes the following variables:

- tuition: College tuition ("out-of-state" rate for those with in-state discount).
- pcttop25: Percent of new students from the top 25% of high school class.
- sf_ratio: Student to faculty ratio.
- accrate: Fraction of applicants accepted for admission.
- graduat: Percent of students who graduate.
- pct_phd: Percent of faculty with Ph.D.'s.
- fulltime: Percent of undergraduates who are full time students.
- alumni: Percent of alumni who donate.
- num_enrl: Number of new students enrolled.
- public_private: Is the college a public or private institution? public=0, private=1
- fac_comp: Average faculty compensation.

The dataset is provided as a CSV file (hint: use the Var source node to read the file)

1. Explore the data to get some initial insights, if you think it is useful (your call)

2. Identify outliers and decide what to do with them

3. Missing data appears to be a problem with this data set. Prepare a copy of the data set, where the missing values are each replaced with their field means. Report on how this substitution has affected the fields (summary stats, etc.), if at all. What do you think of this method of dealing with missing values?

Use the dataset with missing data (not replaced) for items 4 to 6.

4. Provide a table describing the relationship of each explanatory variable with tuition (hint: use scatter plots). If the relationship is not linear, you can[1] make it so by transforming the predictor variable. Note: we did not cover this but I'm trying to make you be creative. For example, after looking at the scatter plot, you may find that Y does not have a linear relationship with X1, but you can make it linear by transforming X1(squaring X1 or taking the logarithm). How would you know this? Trial and error. For example square X1 and see how the scatterplot of Y and (X1)^2 looks like. Assuming you make your transformation (e.g. you squared X1), the new variable will be (X1trans), and the regression model will have the functional form, Y=b0+b1*X1trans+b2*X2+...+bk*Xk. This is still a linear regression as we have a linear combination of predictors. The only difference is that you have transformed some variables to improve the fit of your model.

[1] Not strictly required, but may help enhance your model. You decide.

5. Investigate the correlation among the predictor variables. Suggest a creative course of action (rather than simply omitting a variable) for dealing with any medium or strong correlations encountered (e.g. textbook, section 9.7; avoid any method linked to principal component analysis, as we have not covered it yet).

6. Use SPSS Modeler linear regression tool to investigate whether a linear relationship exists between tuition and the other variables. Investigate the differences in the models, if any, among these methods: enter, stepwise, backwards. Construct a table showing method, variables included, statistical tests on regression coefficients, goodness of fit metric(s), predictive accuracy metric on training and test data. Discuss. Which model do you prefer and why?

7. Use SPSS Modeler linear regression tool on the data set where the missing values are each replaced with their field means. Investigate the differences in the models, if any, among these methods: enter, stepwise, backwards. Construct a table showing method, variables included, statistical tests on regression coefficients, goodness of fit metric(s), predictive accuracy metric on training and test data. Discuss. Which model do you prefer and why?

8. Compare the best model in 6 and the best model in 7. Which model do you prefer and why?

9. For the final (chosen) model:

   a. Write out the estimated regression equation and explain the meaning of the coefficients
   b. Provide a full report of the chosen regression model and report its metrics (goodness of fit, predictive performance) and statistics on training and test data

Make sure you tweak your models to get the best performance. Use 70/30 partition in all cases

10. Decision tree classification: Using the public_private variable as categorical (flag), or deriving from it a flag variable, model the profile of a typical public and private college with a C5.0 decision tree algorithm, using all other variables as predictors (disregard tuition, given the typical difference in tuition between state and private institutions). Compute the confusion matrix and derive proper performance metrics.

Produce a report so that a reasonably intelligent readership can understand. The report should start with an executive summary of one page at the beginning, followed by your analysis.