Week 2 Assignment
Due 6 pm Central, Thursday, June 7

Note: Late homework assignments are not accepted because the solutions are discussed at the start of every class on Thursday. Your solution to this assignment must be uploaded on eCampus as a PDF file. A common approach is to put the solution into a word document and then save that into a PDF file. Please only submit PDF files.

Assignment: Starting with this week, you are only expected to complete one of these parts. If you do both parts, you can double the points you receive for the assignment. In other words, you can look at this as two assignments. If you do one successfully you get 100 points for the assignment. If you have time to do both and you complete them successfully you can get 200 points. You are expected to complete at least one part successfully.

Data File: Diamondswmissing.xlsx

Part 1: Create a SAS EM project names "Week 2 Homework". In that project read this data file for this assignment. Import the data, ensuring that all attributes have the proper metadata described in the data dictionary. In this case, the target is 'Amount', and interval attributes.

Using the data dictionary for this file, identify outliers and replace them with missing. Impute all missing values using the 'Tree' method.

Use a 70/30 partition to build regression models for predicting 'Amount' using forward, stepwise and backward regression. Use the HP and Non-HP models, and select the best.

in regression model there is an option in sas miner for forward, backward

Using the best model, predict the amount for each case. Report the average predicted amount, the average observed amount, the minimum and maximum of these attributes.

Also print the top 15 observations from the final file containing the data plus, imputed values and predictions.

Part 2: Do the same assignment as Part 1 using Python.
However since Python does not impute using the 'Tree' method, impute using the average.

Use one-hot encoding for the nominal attributes, but do not scale the interval attributes. drop the original column in the evening

From these data, after imputation and one-hot encoding, fit a linear regression model for predicting 'Amount' . Use 70/30 cross validation (split) as you did in SAS EM.

Report a listing of your code, plus the same statistics you reported from Part 1.

Upload a pdf file containing your solutions to parts 1 and 2.

Part 1: SAS EM

1. A screen shot of your project window
2. A listing or screen shot of the min, max and mean of the predicted and actual amounts in your data.
3. A listing of the first 15 observations after imputation and prediction.

Part 2: Python

1. A copy of your python program
2. & 3. Same as part 1.

In order to receive full credit, please ensure all screen shots, code and tables are readable.