Assignment #2: Regression Model Building (0 points)

Data: The data for this assignment is the Ames, Iowa housing data set. This data will be made available by your instructor.

Assignment Instructions:

In this assignment we will begin building regression models for the home sale price (the raw home sale price, not any transformation of the home sale price). We will begin by fitting these specific models.

(1) Define the Sample Population

- Define the appropriate sample population for your statistical problem. Hint: As it says in the two sentences one inch above this line, we are building regression models for the response variable SalePrice. Are all properties the same? Would we want to include an apartment building in the same sample as a single family residence? Would we want to include a warehouse or a shopping center in the same sample as a single family residence? Would we want to include condominiums in the same sample as a single family residence?

- Define your sample using 'drop conditions'. Create a waterfall for the drop conditions and include it in your report so that it is clear to any reader what you are excluding from the data set when defining your sample population.

(2) Simple Linear Regression Models

- In Assignment #1 you performed an initial exploratory data analysis of this data. Continue in this mindset and look at some exploratory views of the data to select what you believe are the two most promising predictor variables for predicting SalePrice. Note that simple linear regression models require a continuous predictor variable. Include this discussion in your report as its own section.

- Use these two predictor variables to fit two simple linear regression models. Use the automatically generated ODS output from SAS to rigorously assess the goodness-of-fit of each model. On what criteria are you assessing the model fit? Include each model in its own section of your report.

(3) Multiple Linear Regression Models

- Now combine your two simple linear regression models. Again, use the automatically generated ODS output from SAS to rigorously assess the goodness-of-fit of each model. Does this multiple linear regression model fit better than the simple linear regression models? Do more predictor variables always mean a better fit? On what criteria are you comparing the model fit?   Include the multiple linear regression model in your report as its own section.

Now let's consider a transformation of the response variable from the sale price to the natural logarithm of the sale price.

(4) Regression models for the transformed response log(SalePrice)

- Refit the models from (1) and (2) using log(SalePrice) as the response instead of SalePrice. Perform an analysis of goodness-of-fit and compare the models. Which transformed model fits the best? Do the transformed models fit better than the original models? Include each of these three models in their own section of your report. In the discussion of these models it is advantages to UNPACK the goodness-of-fit plots produced by PROC REG, place the relevant plots from the SalePrice model and the log(SalePrice) model next to each other and discuss the differences. It is important that you are displaying and discussing the differences in the goodness-of-fit between the two models.

Assignment Document:

All assignment reports should conform to the standards and style of the report template provided to you. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary results or information. The document should be submitted in pdf format. Name your file Assignment2_LastName.pdf.

Assignment Review:

The teaching assistant will:

- Check that you are writing a report that conforms to the spirit described in the assignment template.
- Check that you are including the appropriate output in the report.
- Note anything that is clearly wrong.

The teaching assistant will not comment on whether your discussion is correct or incorrect.