Assignment #3: Data Analysis and Regression (100 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.

(4) Outlier Identification

- Remove some observations from your sample in an attempt to produce better fitting models. These observations can be generically called 'outliers'. Note that an 'outlier' in a modeling context can be very different from the 'outlier' definition that you may have learned in an elementary statistics course. What is an outlier in a modeling context? Consider the terms outlier, leverage point, and influence point. Note that you might have individual observations that are different, and you might have groups of observations that are different. **Do not use your response variable to define an outlier.** Create a waterfall of your outlier definitions. Try to drop more than 100 observations, but less than 500 observations from your sample population.

Here is an example code snippet.

```
format outlier_def $30.;
if (x > 30) then do;
    outlier_def = '1. Definition 1';
    outlier_code = 1;
end;
else if (z < 0) then do;
    outlier_def = '2. Definition 2';
    outlier_code = 2;
end;
else do;
    outlier_def = '3. Definition 3';
    outlier_code = 0;
end;
```

Define your 'outlier' observations using a rational process. You can use EDA and model validation to identify rules to identify the population that you do not wish to include in your model. Report a table of counts for each of the outlier definitions using a PROC FREQ statement.

For modeling purposes you will be able to create a data set without outliers by including a DELETE statement in a SAS data step.

```
if (outlier_code > 0) then delete;
```

This outlier discussion should be its own section in your assignment report.

(5) Refit Multiple Regression Model without Outlier Observations

- Refit your multiple regression model. Did the fit improve? Compare and contrast the model fits. Use the UNPACK option in PROC REG to allow you to produce single plots for direct comparison.

This model comparison should be its own section in your assignment report.

(6) Model Comparison of Y versus log(Y)

- In this section we will fit two models using the same set of predictor variables, but the response variables will be SalePrice and log(SalePrice). You may use any set of predictor variables that you wish.

- Use the sample population from Part (3). This means that any observations that you might consider to be outliers in (4) and (5) are still in the sample population.

- How do we interpret these two models? How is the interpretation of the log(SalePrice) model different from the price model?

- Which model fits better? Did the transformation of the response to log(SalePrice) improve the model fit? In general when can a log transformation of the response variable improve the model fit? Should we consider any transformations to the predictors? If so, then fit one more model using any transformations that you find appropriate.

This model comparison should be its own section in your assignment report.

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 Assignment3_LastName.pdf.