Assignment #3: Data Analysis and Regression (50 points)

Data: The data for this assignment is the Ames, Iowa housing data set. These data have been made available as part of Assignment #1.

Assignment Instructions:

In this assignment we will begin building regression models for the home sale price (the raw home sale price SalePrice, 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 show 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. Produce the relevant diagnostic plots to assess the goodness-of-fit of each model. (Hint: Review the pdf note packets for how we should assess the goodness-of-fit. We are looking for two particular residual plots.) On what criteria are you assessing the model fit? Include each model in its own section of your report.

- Note that we always report the fitted model when we fit a linear or generalized linear model. This means that our report should contain a table with the coefficient estimates, t-values, p-values, etc.

(3) Multiple Linear Regression Models

- Now combine your two simple linear regression models into a multiple linear regression model. Again, produce the relevant diagnostic plots to assess the goodness-of-fit of each model 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) Neighborhood Accuracy

- Make a boxplot of the residuals by neighborhood. Which neighborhoods are better fit by the model? Do we have neighborhoods that are consistently overpredicted? Do we have neighborhoods that are consistently underpredicted?

- Compute the mean MAE and the mean price per square foot for each neighborhood. Plot them – Y=MAE and X=Price/SQFT. Is there a relationship between these two quantities?

- Group the neighborhoods by price per square foot. Create between 3 and 6 groups. Code a family of indicator variables to include in your multiple regression model. What is your base category? Refit your multiple regression model with your indicator variables. Compare the MAE of the original multiple regression model from (3) with your new multiple regression model. Which model fits better based on the MAE?

(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, but the models must include at least four continuous predictor variables and any discrete variables that may wish.

- 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_YourLastName.pdf**

The assignment pdf file and accompanying plain text files for Python programs should be included in a zip archive using standard zip compression with the name **Assignment3_YourLastName.zip**

Here is a reasonable section outline for this assignment report.

Section 1: Sample Definition

- Provide the waterfall of your drop conditions with counts.

Section 2: Simple Linear Regression Models

- Section 3.1: Model #1 (Name of Predictor Variable)
- Section 3.2: Model #2 (Name of Predictor Variable)

Section 3: Multiple Linear Regression Model – Model #3

Section 4: Neighborhood Accuracy

Section 5: SalePrice versus Log SalePrice as the Response

- Section 5.1: SalePrice Model
- Section 5.2: Log SalePrice Model
- Section 5.3: Comparison and Discussion of Model Fits