Assignment #1: Exploratory Data Analysis for Simple Linear Regression

Points: 50 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 are going to get right at it and start writing some code to produce exploratory graphics for simple linear regression (SLR). We do not need to understand SLR to produce and interpret these graphics. We will understand linear regression more and more over the next two weeks, and this assignment should help aid in that future understanding. A starter script will be provided to help you complete this assignment.

With the Ames data set we are presented the problem of building a predictive model for the typical residential single family home. In this problem SalePrice is the response variable, the Y variable, and the features of the property are the possible predictor variables. The standard modeling questions in a regression problem are: (1) Which predictor variables are important?, and (2) What is the shape of the relationship between the predictor variable and the response variable? Immediately we should notice that we are interested in the relationships between the predictor variables and the response variables! This focus is fundamental to modern predictive modeling on larger data sets. 1970's small data statistical modeling typically performed by greenhorn students is frequently distracted by focusing on the relationships between the predictor variables. You can waste time on that if you only have ten predictor variables, but can you do that if you have a mere fifty predictor variables? Do we know how many pairwise comparisons exist with only fifty predictor variables? Type choose(50,2) into R and find out.

Let's reiterate our primary questions of importance for this assignment.

(1) Which predictor variables are important?
(2) What is the shape of the relationship between the predictor variable and the response variable?
(3) Do we have the 'right' response variable? Do we need to transform the response variable?

In this assignment we will investigate these three questions using traditional statistical modeling. Let's get our analysis started. Note that we will complete our analysis in pieces. Those pieces will generate the output and the data insights that we need to understand our modeling problem and answer the questions above. However, we will need to organize those pieces into a coherent statistical report. A coherent statistical report is not diary of your statistical analysis, nor is it an output dump, nor is it step-by-step output from the assignment. A coherent statistical report presents the statistical problem and your results in an intelligent manner.

Use the starter script for Assignment #1 to help you get started with this assignment. Note that we will focus on only six predictor variables: 'TotalSqftCalc','TotalBathCalc','QualityIndex', 'TotRmsAbvGrd','OverallQual','OverallCond'. We will investigate the relationship between these six predictor variables and the response variables SalePrice, and log(SalePrice).

(1) Pearson Correlation Analysis

- Use the corrplot package to produce a correlation plot for SalePrice and log(SalePrice). The sample code will show you how to produce this plot.

(2) Scatterplots and Scatterplot Smoothers

- Produce scatterplots between each predictor variable and **SalePrice**. Overlay the loess scatterplot smoother and the fitted SLR model. The sample code will show you how to produce this plot.

- Produce scatterplots between each predictor variable and **log(SalePrice)**. Overlay the loess scatterplot smoother and the fitted SLR model. The sample code will show you how to produce this plot.

- Each of these response variables will produce six plots. Organize them in your report in a 3 row by 2 column panel for presentation.

(3) Discrete or Continuous? Should we treat a discrete predictor variable as a continuous predictor variable?

In regression modeling have to make a variety of modeling decisions. One of those decisions is when to treat an inherently discrete predictor variable as a continuous variable. This decision is subjective, but it should be informed by the number of different values that the predictor variable takes, and the shape of the relationship between these values and the response variable.

- Let’s look at the three predictor variables TotRmsAbvGrd, OverallQual, OverallCond. We already have the scatter plots from (2). Let’s make a boxplot for each variable. The boxplot is also a smoother. Make a 3 row by 2 column grid with the scatterplot from (2) in the left column and the boxplot in the right column. The middle line in the box plot is the median value. The trend in the medians shows the shape of the relationship.

Do we think that it is appropriate to treat these discrete variables as continuous predictor variables?

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 Assignment1_LastName.pdf.

Here is a reasonable section outline for the assignment report.

Section 1: Introduction

- Provide an introduction to your modeling problem and your analysis.

Section 2: A Basic Correlation Analysis

- Provide the EDA results with discussion for your six variables.

Section 3: Which Response Variable?

- Provide the EDA results with discussion for your three variables.

Section 4: Discrete or Continuous?

- Present your analysis and discussion of whether we should treat the example predictor variables as discrete or continuous predictor variables.