Assignment #1: Getting to Know Your Data (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:

Before we can begin to build statistical models, we will always need to get to know our data.

Knowing our data typically consists of three components: (1) a data survey, (2) a data quality check, and (3) an initial exploratory data analysis. Here is a breakdown of how we should view each of these components. Perform each of these steps using the guidance provided for each step. As you read this assignment and perform these tasks, note that not everything that we do in this assignment will translate directly to a section in our assignment report. However, much of what we learn about our data will show up indirectly in our report, such as comments in the Introduction. In this assignment we have to take a more mature view of statistical analysis. We must perform the necessary data obligations that are part of every data analysis project, and then we must pick and choose what is important for discussion, which is also part of every data analysis project.

(1) A Data Survey

- Take some time to take a broad overview of the Ames housing data set. Read over the data documentation. What data do we have, and what is it supposed to represent?

- In the linear regression component of this course we want to build linear regression models to predict the value of a property (or home). Do we have the right data to properly address our problem? Are there observations in the data that should be excluded?

- What kinds of problems can we properly address given the data that we have? In particular if we were to build a regression model with the variable SalePrice as the response variable, what types of properties would we be valuing? Do we need to be careful about what we are doing here?

(2) Define the Sample Population

- When building statistical models we have to define the population of interest, and then sample from THAT population. Frequently we will not actively perform the sampling function. Instead, the data will be made available and we will have to sample from it retrospectively, i.e. we will need to carve out the population of interest. In our case the objective of our application is to be able to provide estimates of home values for 'typical' homes in Ames, Iowa. We may not be able to define what 'typical' is, but we can use the data to find out what is atypical. Any values which are not atypical are then considered to be typical.

- Define the appropriate sample population for your statistical problem. Hint: 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.

The definition of your sample data should be its own section in your assignment report.

(3) A Data Quality Check

- In practice your data will not be 'clean'. You will need to examine your data for errors and outliers. Errors will not always show as outliers, and outliers are not necessarily errors.

- If you have a data dictionary that states the set of proper values for each field, then you will want to check your data against the data dictionary.

- If you do not have a data dictionary, then you will need to reason and explore your way to a proper data set.

Example 1: In this project you will be modeling the sales price of housing transactions. It should be obvious that none of these sales prices should be zero or negative. Observations with a zero or negative sales price should logically be considered to be errors.

Example 2: Suppose we had a 'small' number of housing transactions with a sale price over one million dollars, should we consider these sales prices to be valid? In this case these values could be valid data points, which would make them outliers, or they could be errors, such as 140,000.00 entered as 1,400,000. In either case they are not relevant data points if the objective is to model the 'typical' home price for the area. If million dollar homes were normal data points, then we would have many conforming data points.

- Consider the use of Python functions across a Python DataFrame when forming your data quality check. **Pick twenty variables that you want to consider and run a data quality check on these twenty variables.**

Put your data quality check in an appendix in your assignment report.

(4) An Initial Exploratory Data Analysis

- **Pick ten variables from the twenty variables from your data quality check to explore in your initial exploratory data analysis.** Perform an initial exploratory data analysis. How do we perform an exploratory data analysis for continuous versus discrete (or categorical) data? Consider the use of scatterplots, scatterplot smoothers such as LOESS, and boxplots to produce relevant graphics when appropriate.

Now that we have a basic understanding, we can begin the modeling building process. Note that in the model building EDA we are particularly interested in the relationships between the response variable and the predictor variables.

Your initial exploratory data analysis should be at least one section in your assignment report. It might be wise to split your EDA into two sections in your report – one section for continuous variables and one section for discrete variables.

After we have performed the necessary prerequisite data work, we can then begin the modeling process. Every modeling process begins with an initial exploratory data analysis that is oriented for the problem at hand. Different statistical models require different types of exploratory analysis. In this assignment we will be developing an exploratory data analysis for a regression problem with a continuous response variable.

(5) An Initial Exploratory Data Analysis for Modeling

- What is the response variable in this problem? In addition to the raw response variable should we consider a transformation of the response variable? Consider SalePrice and log(SalePrice).

- Pick three variables from the ten variables from your initial exploratory data analysis and explore their relationship with SalePrice and log(SalePrice).

- Note that the correct eda technique depends on the type of the predictor variable – discrete or continuous.

Items to discuss in this section of your report.

- Does your EDA suggest any potential difficulties or concerns for the model building process?

- Does your EDA suggest that there may be a need to consider transformations in the predictor variables at some point during the model building process?

These results should be in a separate section from your initial exploratory data analysis results.

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: Sample Definition

- Provide the waterfall of your drop conditions with counts.

Section 2: Data Quality Check

- Provide a table listing out your twenty variables.
- Provide the data quality results with discussion for your twenty selected variables.

Section 3: Initial Exploratory Data Analysis

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

Section 4: Exploratory Data Analysis for Modeling

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

Documentation

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_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 **Assignment1_YourLastName.zip**