

|  |
| --- |
| Assignment #1: Getting to Know Your Data  *PREDICT 410* |

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

**Introduction:**

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.

1. A Data Survey

* What data do we have and what is it supposed to represent?
* Do we have the right data to properly address our problem?

- What kinds of problems can we properly address given the data that we have?

1. 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 sales prices should not 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 or under some value such that you would question the true price of the home, 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 or as 14,000. 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. Trouble is what do you do? Ideally you would go back to the original source of the data points and verify. But, in the absence of doing this, should you impute different values or remove the data points?

1. An Initial Exploratory Data Analysis

- 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.

- To do this, one typically examines the data using graphical or visual displays of the data, summative statistics, or various analytical techniques like data sorting.

This first assignment is meant to be an introduction into how to begin to analyze a dataset in preparation for building predictive models. In the process, it will give you an introduction to SAS and SAS programming.

**Assignment Tasks**

1. Examine the variables in the Ames housing dataset. Which are continuous and which are categorical? Are there any that are somewhere in between? Is it clear what these variables measure? If not, please see the data dictionary for the Ames Housing Dataset. From your prior knowledge and experience, do we have the right data to properly address our problem of developing a model to predict SALES PRICE? Identify at least one other variable that you think should be included and discuss why.
2. Use the PROC SORT procedure to sort the data by sales price. Then print the data to determine if there are records with SALES PRICE that are out of range or fishy. You may NOT want to print the entire dataset, but limit your printed observations. Why? Identify potential unusual values that may be outliers for SALES PRICE. Identify and discuss in your write up. Do the same thing for at least 2 other continuous variables of your choice from the Ames housing dataset. *NOTE: If you were doing this as a professional project, you would do this same thing for EVERY continuous variable in your dataset. In addition, if this were a professional project, you would have to develop a set of logical tests or criteria that can be coded into a SAS DATA step to “clean” the sales price variable and all the other variables in the dataset. You would need to decide if you want to DELETE or DROP a record with inappropriate data or impute different values. You are NOT asked to do this, yet! However, you should know that this kind of activity typically takes up about 90% of your time as a predictive modeler.*

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. 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 the remainder of this assignment we will be developing an exploratory data analysis for a regression problem with a continuous response variable.

1. Use PROC CORR to produce the Pearson correlation coefficients and a scatterplot matrix of the potential continuous predictor variables with the response variable Y (***sale price).***  All continuous variables should be included in this analysis. For help with SAS see Chapter 8 pp. 111-117 in *SAS Statistics By Example*. The following snippet of code may help.

ods graphics on**;**

**proc** **corr** **data**=temp plot=matrix**(**histogram nvar=all**)** plots(maxpoints=NONE;

**run;**

ods graphics off**;**

Comment on which predictor variables have the strongest linear relationships with the response variable,Y? What do you notice about the relationship between the numeric correlation measure and the graphical relationship in the scatterplot? Which predictor variable do you think will be the best single predictor variable. Why? Which will be the worst and why? Are there high correlations within the set of potential predictor variables? This is a primary way to see/identify multicollinearity. Is the correlation coefficient sufficient information to make a decision regarding a predictor variable and it’s usefulness in developing a predictive model? The answer is, NO – why is that?

1. At a minimum, make a scatter plot for the X continuous variable with the highest correlation with Y. Do the same for the X variable that has the lowest correlation with Y. Finally, make a scatter plot between X and Y with the correlation closest to 0.5. You should have a better perspective on the patterns in the data using the individual scatter plots, rather than the scatterplot matrix. *Note: Technically, if you were doing this as a professional analysis of this data, you would generate scatterplots for EVERY potential X variable paired with Y.*
2. For the 3 predictor (X) variables you used in Step 4, produce a scatterplot with a LOESS (Locally Estimated Scatterplot Smoother) smoother for Y. Comment on why we are interested in the LOESS scatterplots and what they are showing us. One item to notice is how the LOESS curve deviates from the regression curve. Here is a snippet of SAS code to help – but you will have to edit this code for your specific program.

ods graphics on**;**

**proc** **sgscatter** **data**=temp**;**

compare x=**(**x1 - x2**)**

y=Y / loess**;**

**run;** **quit;**

ods graphics off**;**

1. Select 3 categorical variables of your choice and use the PROC FREQ procedure to see the distribution of values for these variables. Display the distribution using an appropriate graph for each of these 3 categorical variables.
2. For the 3 categorical variables selected in step 6), do the following analysis for each variable - one categorical variable at a time. Use PROC SORT to sort the data by the categorical variable. Then use PROC MEANS with the BY command to obtain mean values for Y (sales price) by the categorical variable. Do you see a linear relationship between the values of the categorical variable and the response Y? Discuss.
3. For your own learning follow some of the examples in SAS Statistics By Example and learn how to make a histogram, a bar chart, and a scatterplot matrix using PROC SGPLOT and PROC SGSCATTER. Typically, a preliminary EDA entails an exhaustive set of graphics and summary statistics to illustrate the distributions and patterns in ALL of the variables of the dataset that one is using in the modeling process. You are NOT asked to do that here (unless you want to), but you are asked to practice all of the different types of basic graphs and summary statistics computations. Is there anything interesting you see in this data?
4. Please summarize your findings in at least 1 paragraph in a conclusion / reflection section. You should, at a minimum, address the following questions:

* 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?

**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, graphics, tables or other non-essential information. The document should be submitted in pdf format. Name your file: Assignment1\_LastName.pdf.