Assignment #1

Ji Jung

Introduction

This assignment will examine AMES_HOUSING_DATA and study on its data structure for future linear regression modeling and analysis.

Results

Examine the Variables in the Ames Housing Data Set

First, the data set needs to be loaded from the provided location. From the log, there are two important bounding characteristics of the overall data set. There are 2930 observations read from the data set MYDATA.AMES_HOUSING_DATA with 82 variables. From the descriptions of the assignment, this is a data dictionary. This dictionary will be examined soon as the exploration requires clarifications regarding a categorical variable or another ambiguity in the data collection.

Which Variables are Continuous and Which are Categorical?

82 variables have a handful of types, lengths and formats. From the data dictionary, a quick search produces the tally table:

Type	Tally
discrete	15
nominal	24
continuous	20
ordinal	23

The data dictionary makes it much more clear to understand what the variables intend to represent. Without this resource, a great deal of time would have been spent on examining the individual variables trying to infer what their individual meanings are.

Can We Develop a Model to Predict Sales Price from this Data Set?

From examining using the corr procedure to see if there are any variables that have a low *p*-value in relation to saleprice.

The list of all continuous variables is as follows:

LotFrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF FirstFlrSF SecondFlrSF LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch ThreeSsnPorch ScreenPorch PoolArea MiscVal

From here, the set of variables with a low *p*-value is selected.

LotFrontage LotArea MasVnrArea BsmtFinSF1 BsmtUnfSF TotalBsmtSF FirstFlrSF SecondFlrSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch PoolArea

All, except for PoolArea, have a *p*-value of < 0.0001. Pool Area has a *p*-value of 0.0002. Of these correlations, only a handful have strong Pearson correlation coefficients, where most are close to 0. Due to this, further selections will be made:

MasVnrArea BsmtFinSF1 BsmtUnfSF TotalBsmtSF FirstFlrSF GrLivArea GarageArea

From a simple Pearson correlation on the available continuous variables in this data set, there are seven variables of interest.

Examine Continuous Variables to Look for Questionable Observations

The continuous variables will be examined to look for questionable observations. The largest and smallest observations will be examined.

Examine Sales Price

Both the highest and lowest observations appear to be within a reasonable range. It is interesting, and potentially something to be examined further, that the lowest two values are almost a half lower than the third lowest value in the observations.

Examine LotFrontage

Although eliminated from modeling perspective, LotFrontage and LotArea will be examined for there are interesting indications.

From the data dictionary, LotFrontage is an observation of linear feet of street connected to property. Using the sort procedure for LotFrontage, the first 490 observations appear to be null for this variable.

Examine LotArea

LotArea is an observation of Lot size in square feet. Using the sort procedure for LotArea shows the highest four observations are significantly larger than the rest of the observations.

Questionable Observations Cleanup

Without any guidance as to the thresholds of observations that can be eliminated, cleanup can be difficult to assess. The act of Data Wrangling or Munging is an incredibly time consuming exercise that is necessary before constructing models. When the analyst is confronted with a tedious task, such as prepping and cleaning data, they will seek the most expressive and powerful tools to do this.

Investigate Potential Continuous Predictor Variables, with respect to Sales Price

The Pearson correlation on the selected variables during the exploratory data analysis phase results in:

Variable	Pearson Correlation Coefficients	Prob > $\|r\|$ under H_0: ρ=0	Number of Observations
GrLivArea	0.70678	<.0001	2930
GarageArea	0.64040	<.0001	2929
TotalBsmtSF	0.63228	<.0001	2929
FirstFlrSF	0.62168	<.0001	2930
MasVnrArea	0.50828	<.0001	2907
BsmtFinSF1	0.43291	<.0001	2929
BsmtUnfSF	0.18286	<.0001	2929

Further down-select to the five variables with a correlation coefficient greater than $|0.5|$.

MasVnrArea TotalBsmtSF FirstFlrSF GrLivArea GarageArea

Even though selections are based on *p*-value and Pearson correlation coefficient, these criteria are not alone enough to indicate in choosing a variable as the predictor variable. More examinations are required with more tools available.

The five variables from the corr process are graphed:

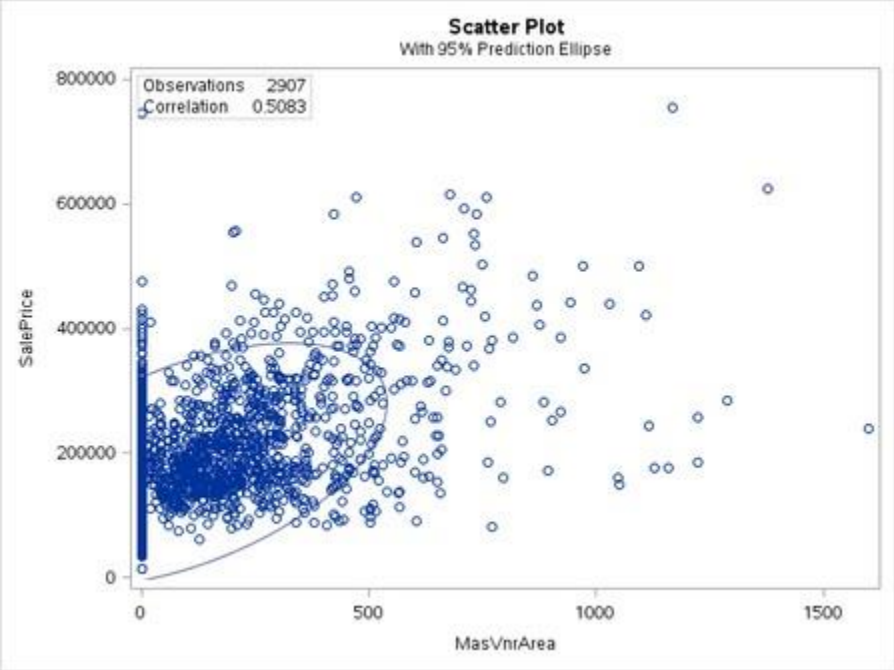

Figure 1: MasVnrArea vs SalePrice

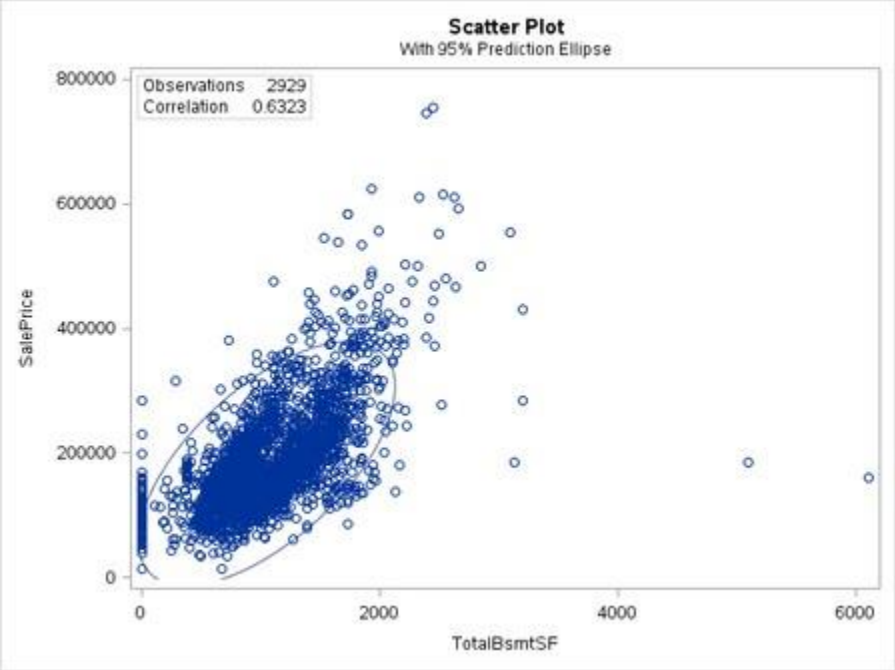

Figure 2: TotalBsmtSF vs SalePrice

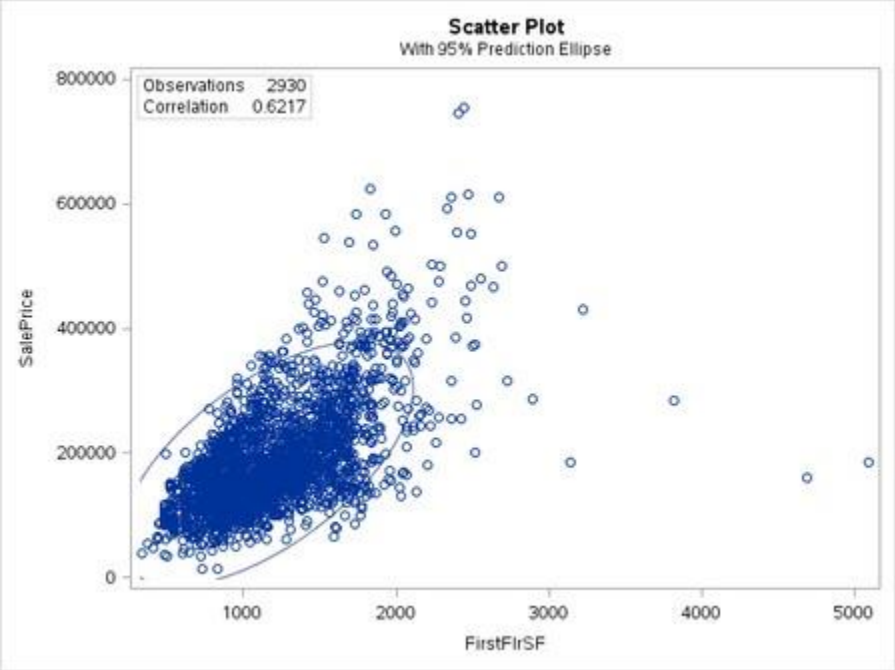

Figure 3: FirstFlrSF vs SalePrice

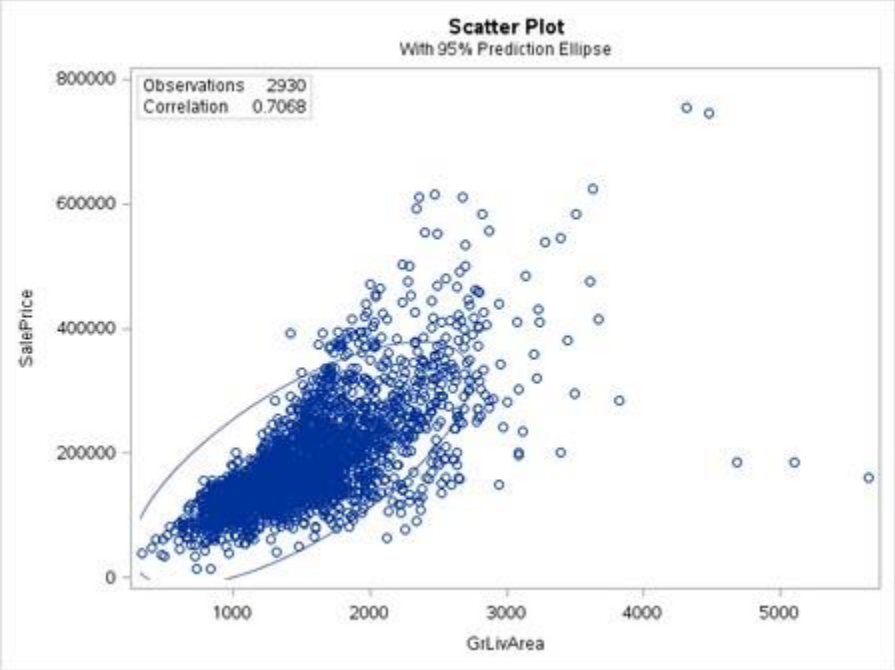

Figure 4: GrLivArea vs SalePrice

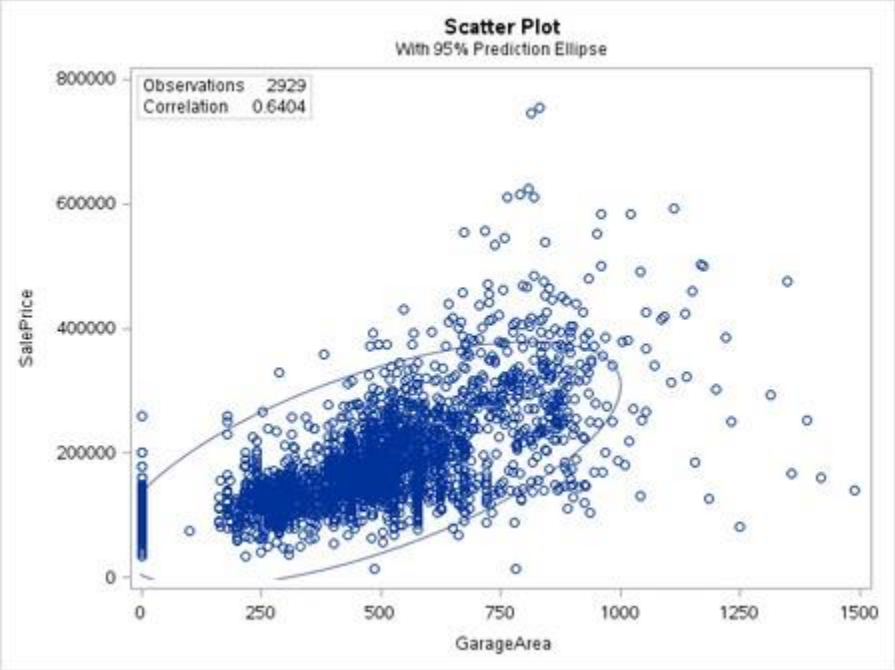

Figure 5: GarageArea vs SalePrice

Visualization of Selected Continuous Predictor Variables

The graphed variables above are scatterplots and overlaying the Locally Estimated Scatter plot Smoother (LOESS). Even though these variables have good Pearson correlation coefficients that the LOESS overlay for TotalBsmntSF, FirstFlrSF, and GrLivArea indicate chasing of some outlier. While the LOESS overlay for MsVnrArea and GarageArea look closer to expected from a regression line although they each seem to vector towards an outlier at the end.

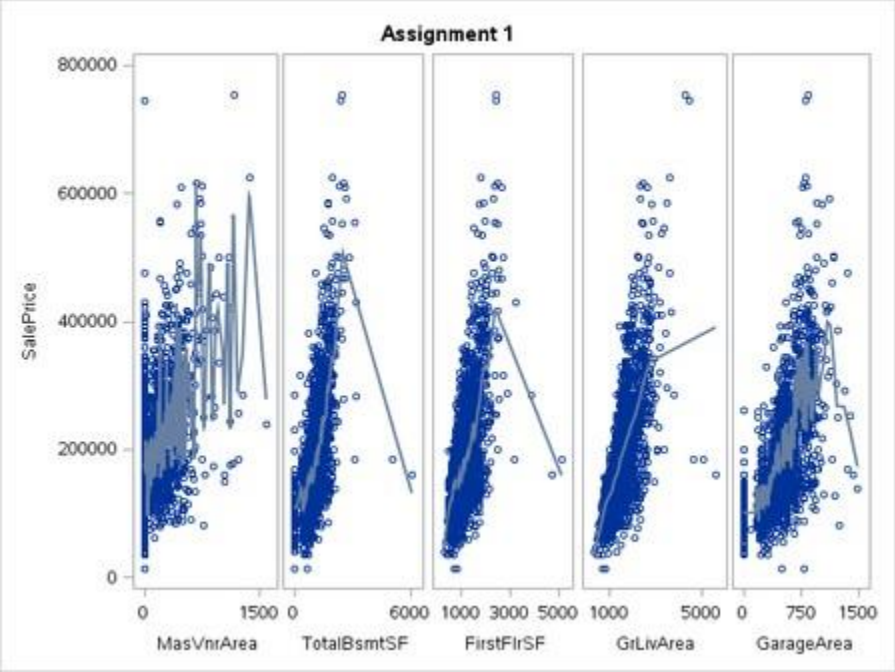

Figure 6: Five variables with LOESS Overlay

Conclusion / Reflection

The Exploratory Data Analysis indicates that there are some interesting variables to be examined for construction of a model. There are many parts of the data set that have missing values for observations. Much of the data set is categorical, and currently it can be difficult to assess whether these variables will be valuable predictors.

The relatively erratic LOESS overlay may be an indicator that can be considered in transformations of predictor variables. A smother LOESS overlay should indicate whether the relationship is approximately linear. Instead of using scatter plots, kernel density plots can be considered.