Assignment #1

Scott M. Morgan

Introduction

The purpose of this assignment is to introduce how to perform Exploratory Data Analysis in preparation for building predictive models. To accomplish this objective, I use data from the Ames housing data set provided by the course instructor and SAS Studio to perform the analysis. Understanding the nuances of a data set is an integral part of the predictive modelling process. In this assignment, I use functionality within SAS Studio to examine variables, sort, produce Pearson correlation coefficients and generate various graphs. Please note all SAS code is included at the end of the assignment in the order in which it is discussed. I expect that this exercise will uncover interesting idiosyncrasies in what seems like an otherwise straightforward data set.

Results

I begin the analysis by examining the variables in the Ames housing data set using SAS content procedure. The data set contains 2930 observations and 82 variables. While this is too large to list individually here, the variables primarily describe physical characteristics of a home that would be of interest to potential buyers or sellers. Many of the variables in the data set appear to be categorical in nature (i.e. year built, paved drive way, kitchen quality, etc.). There are also a number of discrete variables which quantify the various features of the home (i.e. number of bathrooms, number of fireplaces, etc). Lastly, there are numerous continuous variables that describe the different physical dimensions (i.e. lot area in square feet, pool area in square feet, garage area in square feet, etc.).

While the naming convention of the variables is relatively effective, the data dictionary is a helpful resource. The data set itself has a wide variety of descriptive information that appears analytically interesting and appropriate to build a model with in order to predict SalePrice. In addition to more common discrete variables of a home such as number of bedrooms, for example, it would be interesting to include variables like SubClass to capture the impact of more qualitative dwelling characteristics on the SalePrice.

Outliers. Next, I use SAS to begin identifying potential outliers in 3 continuous variables. As mentioned in the instructions for this assignment, if this were being done for a professional project every continuous variable in the data set would need to be analyzed.

SalePrice. Using the SAS procedure PROC SORT, I first examined the SalePrice variable. As the data set is quite large and I do not wish to be inundated with data, I only returned the top 10 observations. In a corporate setting this would also save on IT resources. There did appear to be several potential outliers in the data set, especially at higher values. To put these potential outliers into context, I used to MEANS function to display a simple statistical summary. This output is displayed in Table 1 below.

Table 1: The MEANS Procedure

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
2930	180,796.06	79,886.69	12,789.00	755,000.00

As illustrated, there does appear to be outliers in the data set not only on the high end of SalePrice but also the low end. This is further corroborated by taking into consideration the standard deviation of SalePrice. If we operate under the assumption that an outlier is defined by an observation that is +/- 2 standard deviations from the mean, both the minimum and maximums of SalePrice, at the very least, are outliers. This same methodology is followed in the below discussion of outliers in the other continuous variables.

LotArea. LotArea, defined as lot size in square feet, is a continuous variable that also appears to contain outliers. PROC SORT reveals 3 large and 1 very large lot size. There are also a number of notably smaller lots, suggesting positive skewness. Given the standard deviation of the variables presented in Table 2 below, these smaller values do not appear to be outliers while the larger end of the range (+100,000 square feet) are likely outliers.

Table 2: The MEANS Procedure

Analysis Variable : LotArea				
N	Mean	Std Dev	Minimum	Maximum
2930	10,147.92	7,880.02	1,300.00	215,245.00

PoolArea. PoolArea, the pool area in square feet, is the last variable examined for outlier in this exercise. The statistical summary of this variable is provided below in Table 3. Many of the houses do not have pools while only a small number have large pools, distorting the output from the MEANS procedure. Utilizing the SUMMARY procedure, it is evident given the skewness and kurtosis that PoolArea is a non-normal distribution (Table 4). While we would assume those homes with larger pools would be useful predictors of sale price as they likely have larger lots, larger livable areas, etc., this variable might not be a suitable predictor to include in a predictive model without some sort of normalization / transformation.

Table 3: The MEANS Procedure

Analysis Variable : PoolArea				
N	Mean	Std Dev	Minimum	Maximum
2930	2.2433447	35.597180	0	800.0000000

Table 4: The SUMMARY Procedure

Analysis Variable : PoolArea				
Mean	**Std Dev**	**Median**	**Kurtosis**	**Skewness**
2.2433447	35.5971806	0	299.77494	16.939142

Exploring the functionality of SAS to understand the data set more, I also used PROC FREQ to decompose the continuous variables by "buckets" of categorical variables. For example, running the procedure on LotConfig by PoolArea tells us that the 800 square foot pool is a single corner lot and also confirms that 99.56% of the sample does not own a pool, suggesting it might not be appropriate to include in the model at all.

Correlation. Next, I used PROC CORR to produce the Pearson correlation coefficients and a scatterplot matrix of the potential continuous predictor variables with the response variable (SalePrice).

My expectation is that GrLivArea will be the single best predictor variable as the livable area of a home (excluding the basement) is the feature consumers value most when buying (or selling) a home. The worst predictor will be EnclosedPorch as this seems like an ancillary feature of a home. Diagnostics of the continuous variables indicate that multi-collinearity is generally not an issue; the highest coefficient being 0.80 between TotalBsmtSF and FirstFlrSF.

Overall, most of the correlations are positive. There are generally few negatively correlated combinations. LowQualFinSF, defined as low quality finished square feet (all floors), is the most negatively correlated of the continuous variables with SalePrice and it is only -0.037. The correlations of BsmtFinSF2 (Type 2 basement finished square feet), ThreeSsnPorch (three season porch area in square feet) and PoolArea are close to zero and thus have no linear relationship with SalePrice. The three continuous variables that have the strongest positive linear relationship with the response variable are GrLivArea (above ground living area square feet), GarageArea (size of garage in square feet) and TotalBsmtSF (total square feet of basement area). While the correlation coefficient is useful in choosing a predictor variable, it alone is not sufficient. Additional evaluation of the logic behind the variable's inclusion and the integrity of the variable itself is needed.

Exhibits 1 through 3 present scatterplots for the 3 continuous variables with the highest, lowest and closest to 0.5 correlations to the response variable. For reference, the 3 continuous variables with the highest correlation to the response variable are (from greater to least) GrLivArea (0.71), GarageArea (0.64), TotalBsmtSF (0.63). The 3 continuous variables with the lowest correlation to the response variable are (from least to greatest) EnclosedPorch (-0.13), LowQualFinSF (-0.04) and MiscVal (-0.02). Finally, the 3 continuous variables with correlations closest to 0.5 to the response variable are (from greater to least) FirstFlrSF (0.62), BsmtFinSF1 (0.43) and LotFrontage (0.36). While providing insight into the linear relationships of the variables, these charts also further assist in identifying possible outliers, which potentially exist in every variable displayed below. This represents an area were further analysis and action is needed.

Exhibit 1: Scatterplots for Continuous Variables with Highest Correlation to the Response Variable

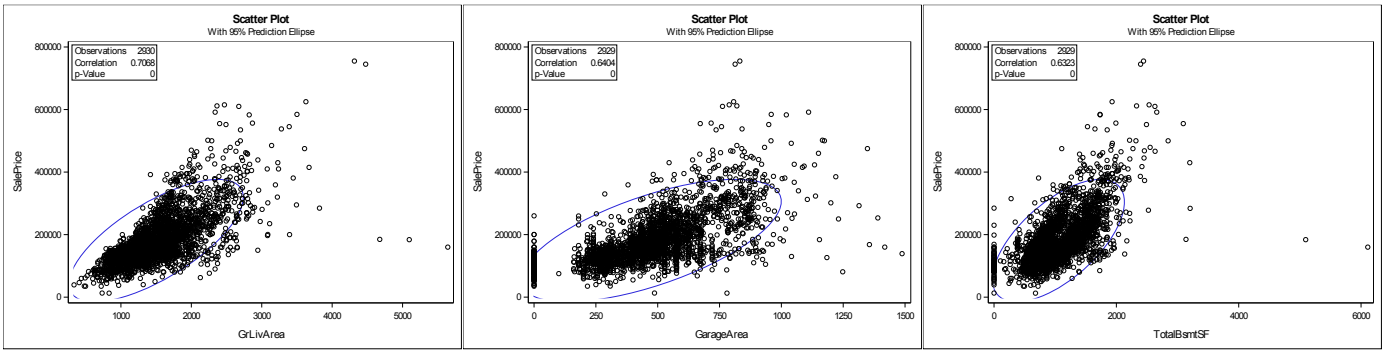


Exhibit 2: Scatterplots for Continuous Variables with Lowest Correlation to the Response Variable

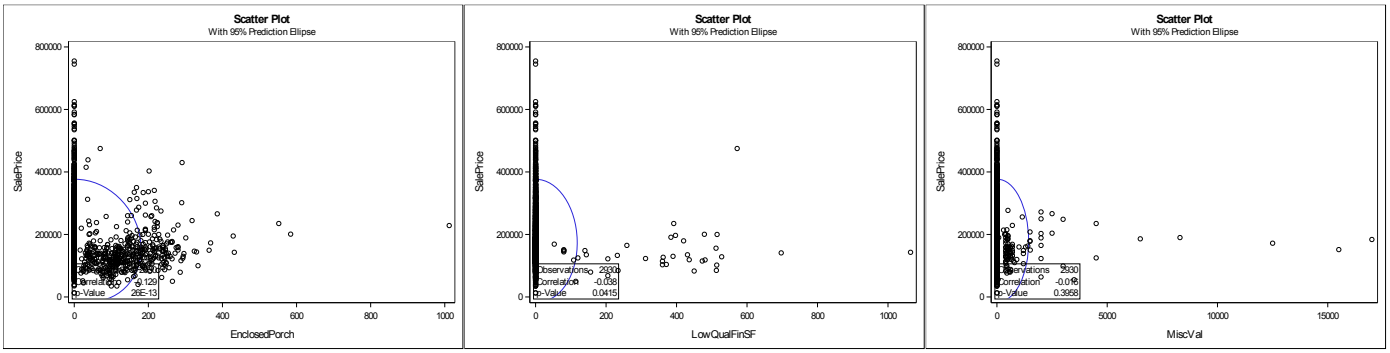
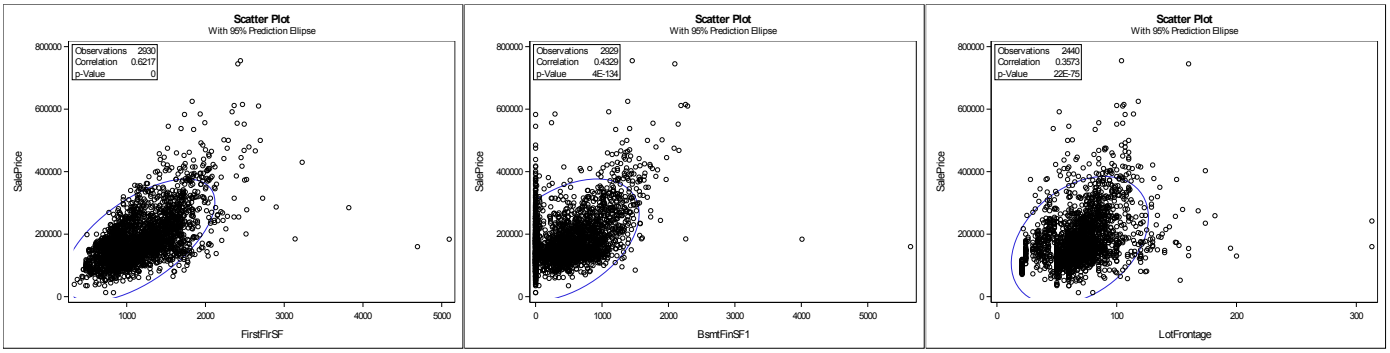
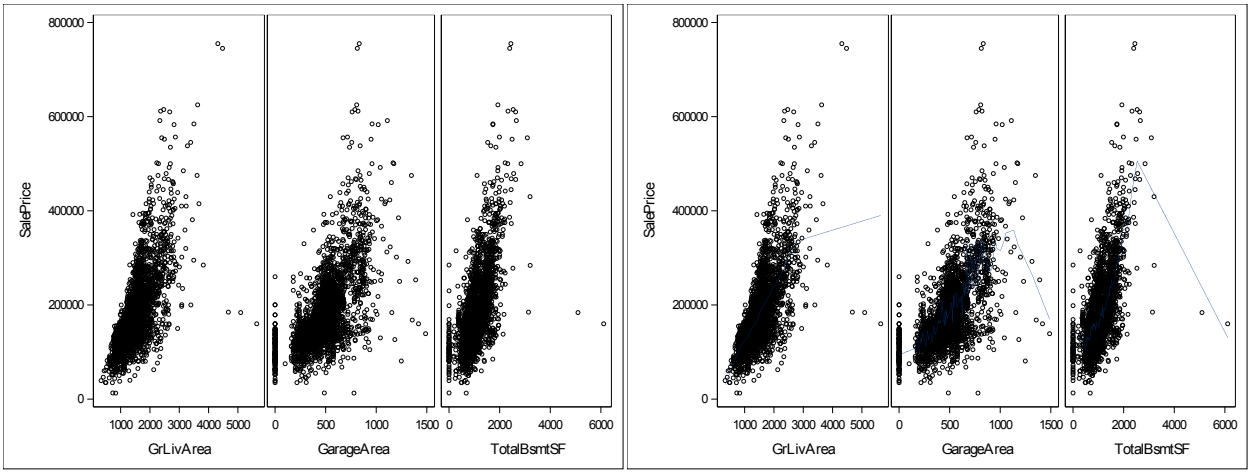


Exhibit 3: Scatterplots for Continuous Variables with ~0.5 Correlation to the Response Variable



Locally Estimated Scatterplot Smoother (LOESS). The above variables are further analyzed by graphing them as scatterplots and superimposed with the Locally Estimated Scatterplot Smoother (LOESS). Exhibit 4 illustrates the scatterplots of the 3 highly correlated variables with and without the LOESS curve. The LOESS scatterplot is of interest because it is a locally ("neighborhood") weighted, non-parametric fitting technique. The biggest advantage of LOESS is that it does not require a function to fit a model to all the data in a sample. As it pertains to the 3 variables in Exhibit 4, we can see how the distribution and presence of outliers is impacting the regression lines, especially in the case of GarageArea and TotalBsmtSF.

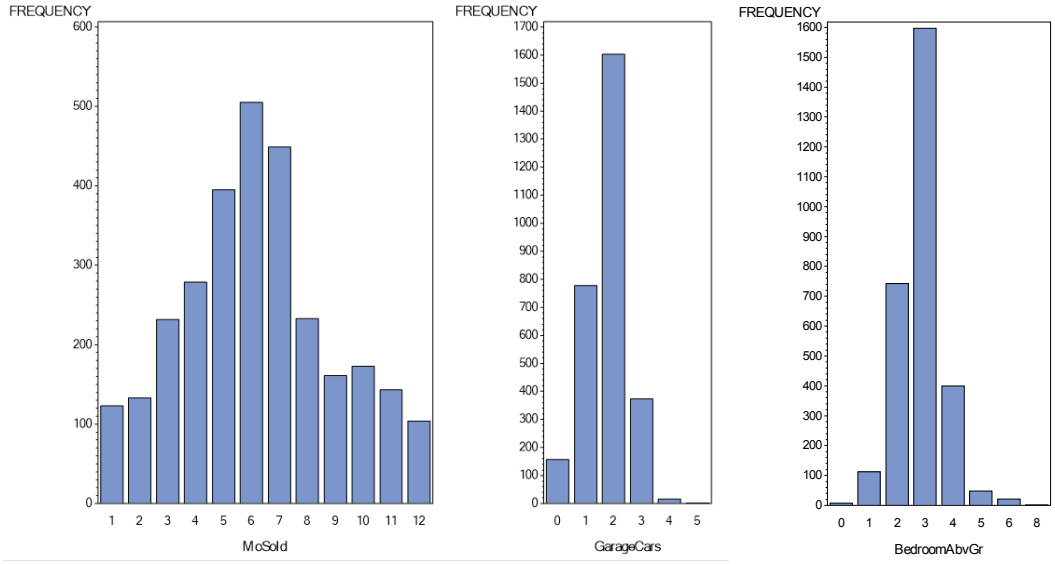
Exhibit 4: Scatterplots for Continuous Variables with Highest Correlation to the Response Variable with and without LOESS



Potential Categorical Predictor Variables. Next, I examine the categorical variables MoSold, GarageCars and BedroomAbvGr as potential predictors.

Distribution of Values. Exhibit 5 provides the distribution of values for the above mentioned categorical variables. We can see that there does seem to be seasonality in the months where homes are sold. This makes sense intuitively as families typically move during the summer months once children have finished the school year (i.e. a "school year effect"). The remaining 2 variables are more descriptive concerning the sample. Specifically, we can see that there are far more houses with 2 car garages as well as houses with 3 above ground bedrooms.

Exhibit 5: Distribution of Categorical Variables



Using PROC SORT and PROC MEANS on the 3 categorical variables, we observe several noteworthy items. The SAS code to run the statistical summaries is included at the end of the exercise and their findings discussed below.

- **MoSold.** SalePrice outliers (high values) occurred in January and July. This might be attributable to year end raises and bonuses being paid in January while July could be impacted by the aforementioned "school year effect." However, there does not appear to be a linear relationship between the month a home was sold and the sale price.
- **GarageCars.** SalePrice does appear to increase in a linear fashion with more garage capacity. More capacity implies a larger home and, therefore, a likely higher selling point.
- **BedroomAbvGr.** We would expect SalesPrice and BedroomAbvGr to exhibit a linear relationship in a similar fashion to GarageCars. However, after 4 above ground bedrooms the sale price unexpectedly drops. This is contrary to expectations.

Correlations. Lastly, we run the Pearson correlation coefficient on the categorical variables from the data set. The relationships are as expected: SalePrice has a higher correlation to GarageCars (0.64) than to BedroomAbvGr (0.14) and MoSold.

Conclusion

Exploratory data analysis provides a concise way to obtain a broad view of a data set and efficiently identify potential issues before subsequent analyses. For this exercise the Ames, Iowa data set was examined using a combination of statistical methods to identify and validate potential predictor variables.

While several prospective variables were identified, certain elements of the data set are problematic. First, there are certain data integrity issues when dealing with publicly available data sets. The original proprietors of the data could have altered and scaled the data differently than what was initially measured. Additionally, having not collected these data ourselves we are forced to rely on the original scientists in terms of collection techniques and accuracy. In terms of the data itself, several variables are potential candidates for transformation to achieve linearity. Specially, GarageArea, PoolArea and TotalBsmtSF could be transformed given the frequencies of zeros in the data.

Overall this exercise was a good introduction to data manipulation using SAS and interpreting statistical outputs. I found the nuances of the data set interesting, especially homes that didn't have certain features and considering how this could potentially skew distributions and, in turn, result in a subpar predictive model. The next steps in this process would be to further clean and transform the data where necessary, select the final predictor variables, run the initial model, fit/validate and then deploy the model to production.

Code

```
libname mydata "/scs/wtm926/"   access=readonly;


proc datasets library=mydata; run;


Data temp1;
        set mydata.ames_housing_data;


proc contents data=temp1;
run;



/* SECTION 2 */



/**Variable 1**/


/****Sort Variable 1 Descending***/
proc sort data=temp1;
        by descending saleprice;


proc print data=temp1 (obs=10);
        run;


/***Sort Variable 1 Ascending***/
```

```
proc sort data=temp1;
        by descending saleprice;

proc print data=temp1 (obs=10);
        run;

/*** Statistical Summary ***/

proc means data=temp1;
        var saleprice;
        run;



/**Variable 2**/

/*** Sort Variable 2 Descending***/

proc sort data=temp1;
        by descending lotarea;

proc print data=temp1 (obs=10);
        run;

/*** Sort Variable 2 Ascending***/
```

```
proc sort data=temp1;
        by lotarea;

proc print data=temp1 (obs=10);
        run;

/*** Statistical Summary ***/
proc means data=temp1;
        var lotarea;
        run;



/**Variable 3**/

/***Sort Variable 3 Descending***/

proc sort data=temp1;
        by descending poolarea;

proc print data=temp1 (obs=10);
        run;

/***Sort Variable 3 Ascending***/
proc sort data=temp1;
        by poolarea;
```

```
proc print data=temp1 (obs=10);
        run;

/*** Statistical Summary ***/
proc means data=temp1;
        var poolarea;
        run;

PROC SUMMARY PRINT MEAN STD MEDIAN KURTOSIS SKEWNESS MEDIAN;
        var poolarea;
        run;

proc freq data=temp1;
        tables lotconfig*poolarea;
        run;



/* SECTION 3 */



proc corr data=temp1 plot=matrix(histogram nvar=all);
        var _numeric_;
        with saleprice;
run;
```

```
proc corr data=temp1 plot=matrix(histogram nvar=all);
        var GrLivArea GarageArea TotalBsmtSF FirstFlrSF BsmtFinSF1 LotFrontage WoodDeckSF
OpenPorchSF SecondFlrSF LotArea BsmtUnfSF ScreenPorch PoolArea ThreeSsnPorch BsmtFinSF2 MiscVal
LowQualFinSF EnclosedPorch;
run; /** Multi-Collinearity check **/



/* SECTION 4 */



/** 3 Highest Correlations**/
proc corr data=temp1 nosimple rank plots=(scatter);
  var GrLivArea GarageArea TotalBsmtSF;
  with SalePrice;
run;

/** 3 Lowest Correlations**/
proc corr data=temp1 nosimple rank plots=(scatter);
  var EnclosedPorch LowQualFinSF MiscVal;
  with SalePrice;
run;

/** 0.5 Correlations**/
proc corr data=temp1 nosimple rank plots=(scatter);
  var BsmtFinSF1 FirstFlrSF LotFrontage;
```

```
  with SalePrice;
run;



/* SECTION 5 */



/**Without Loess**/
proc sgscatter data=temp1;
        compare x=(GrLivArea GarageArea TotalBsmtSF)
        y=saleprice;

/**Without Loess**/
proc sgscatter data=temp1;
        compare x=(GrLivArea GarageArea TotalBsmtSF)
        y=saleprice/loess;
run;

ods graphics off;

/* SECTION 6 */

proc freq data=temp1;
        tables MoSold GarageCars BedroomAbvGr ;
run;
```

```
PROC GCHART data=temp1;
        VBAR MoSold / discrete;

PROC GCHART data=temp1;
        VBAR GarageCars / DISCRETE;

PROC GCHART data=temp1;
        VBAR BedroomAbvGr / DISCRETE;



/* SECTION 7 */



/** MoSold **/
proc sort data=temp1;
        by MoSold ;

proc means data=temp1;
        by MoSold ;
        var saleprice;
run;

/** GarageCars **/
proc sort data=temp1;
```

```
        by GarageCars ;


proc means data=temp1;
        by GarageCars  ;
        var saleprice;
run;

/** BedroomAbvGr **/
proc sort data=temp1;
        by BedroomAbvGr;

proc means data=temp1;
        by BedroomAbvGr ;
        var saleprice;
run;



/* SECTION 8 */



proc corr data=temp1 plot(maxpoints=none)=matrix(histogram nvar=all);
        var saleprice MoSold GarageCars BedroomAbvGr ;
run;

proc corr data=temp1 plot=matrix(histogram nvar=all);
```

```
        var MoSold GarageCars BedroomAbvGr ;

        with saleprice;

run;
```