Assignment #1

Crystal M. Mosley

Introduction:

The purpose of this data analysis is to understand the Ames, Iowa housing data set by means of a data survey, data quality check and an exploratory data analysis. The data survey will allow for us to explore the Ames housing data to see what we have available, and what each piece of data is supposed to represent. This survey will help to ensure that we have the correct data to address the problem needing to be resolved, as well as help find out if other problems are occurring that can properly be addressed. During the data quality check, we will examine the data to ensure that it is *clean* and that any errors and outliers can be found and researched upon, or eliminated so our results are more accurate. Throughout the exploratory data analysis, we will be able to see what the housing data can tell us about the *typical* home values of Ames, Iowa.

Results:

Which variables are *continuous*, and which are *categorical*? Are some variables in-between?

The SAS proc (procedure) "contents" is used to show the list of variables within the dataset. Utilizing proc contents, we will be able to review the data *type* of each variable.

```
Proc contents data=temp1;
Run;
```

From this analysis, we see there are 82 different variables, and two data types: CHAR and NUM; however, we are not able to establish which variables are continuous, and which are categorical without knowledge of what each variable means and is measuring; therefore, we will utilize the data dictionary provided to decipher what each variable represents.

Type	Total
Nominal (categorical)	24
Ordinal (categorical)	23
Discrete (categorical)	15
Continuous	20
	82

Reviewing the data dictionary, we see that there are 62 total categorical variables, and 20 continuous variables; noting that SalePrice counts as the 20th continuous variable.

Do we have the right data to properly address our problem of developing a model to predict SALES PRICE?

To ensure that we have the proper data to create this model, we will examine the continuous variables by using the corr (correlation) proc to show which of these variables have the lowest *p-value* as it relates to the SALES PRICE.

The list of continuous variables to start:

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

```
proc corr data=temp1;
        var saleprice;
        with LotFrontage LotArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF FirstFlrSF
        SecondFlrSF LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF
        EnclosedPorch ThreeSsnPorch ScreenPorch PoolArea MiscVal;
run;
```

NOTE: MasVrnArea is listed as a continuous variable, but I received an error that this variable could not be found when I ran the code; therefore, I removed this variable.

Next, we select the variables with the lower p-value's:

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

After examining these correlations, we narrow the variables down to p-value's < 0.0001, and strong correlations coefficients (1.0 to 0.5). Seeing as how TotalBsmtSF contains the p-value required and a strong correlation coefficient, the BsmtFinSF1 and BsmtUnfSF are added to this list, as they are a part of the total square footage of the basement, and contain the required p-value.

BsmtFinSF1 BsmtUnfSF TotalBsmtSF FirstFlrSF GrLivArea GarageArea

We now have six variables of interest to create a model to predict the sales price.

Identify potential unusual values that may be outliers for SALES PRICE.

In further examining the continuous variables we will look for questionable observations. We will use the proc sort procedure to sort the data by sales price from smallest to largest to try and see if any data seems out of the typical range, or odd. Printing all of the data will not be necessary as we only need to evaluate the top and bottom 10 price ranges for our observations.

SalePrice:

Sort procedure for sales price:

```
proc sort data=temp1 out=sorted;
        by saleprice;
proc print data=sorted;
        var saleprice;
run;
```

Examining the lowest ten housing prices, the house range from observations 3 – 10 seem like a normal range, with the exceptions for observations 1 and 2; both of these pricings are less than half of the 3rd observations.
Examining the highest two housing prices, the house range from observations 2921 – 2930 also seem like a normal range, without any outliers in pricing.

PoolArea:

We're going to examine the PoolArea which is defined as the pool area in square feet. This variable was chosen because of the interest of having a pool is lot higher for buyers, than not; therefore, I deemed it an interest variable.

Sort procedure for pool area:

```
proc sort data=temp1 out=sorted;
        by poolarea;
proc print data=sorted;
        var poolarea;
run;
```

Interestingly enough, the first 2,917 observations show "0" in square footage for a pool.

WoodDeckSF:

The last variable chosen to be examined is the WoodDeskSF which is defined as the wood deck area in square footage. This variable was also chosen because of the interest of knowing how many homes have a wood deck, and the size of each deck.

Sort procedure for pool area:

```
proc sort data=temp1 out=sorted;
        by poolarea;
proc print data=sorted;
        var poolarea;
run;
```

The first 1,526 observations show "0" in square footage for a wood deck. Observations 1,527 – 2,929 all seem to have a normal range of growth, with the exception of observation 2,930; this

observation shows a jump in square footage by 554, whereas, the others show a steady growth in square footage.

Produce the Pearson correlation coefficients and a scatterplot matrix of the potential continuous predictor variables with the response variable SalePrice.

We will use the six variables that were narrowed down by lowest p-value and strong correlation coeffictient:

BsmtFinSF1 BsmtUnfSF TotalBsmtSF FirstFlrSF GrLivArea GarageArea

The results for each variable is as follows:

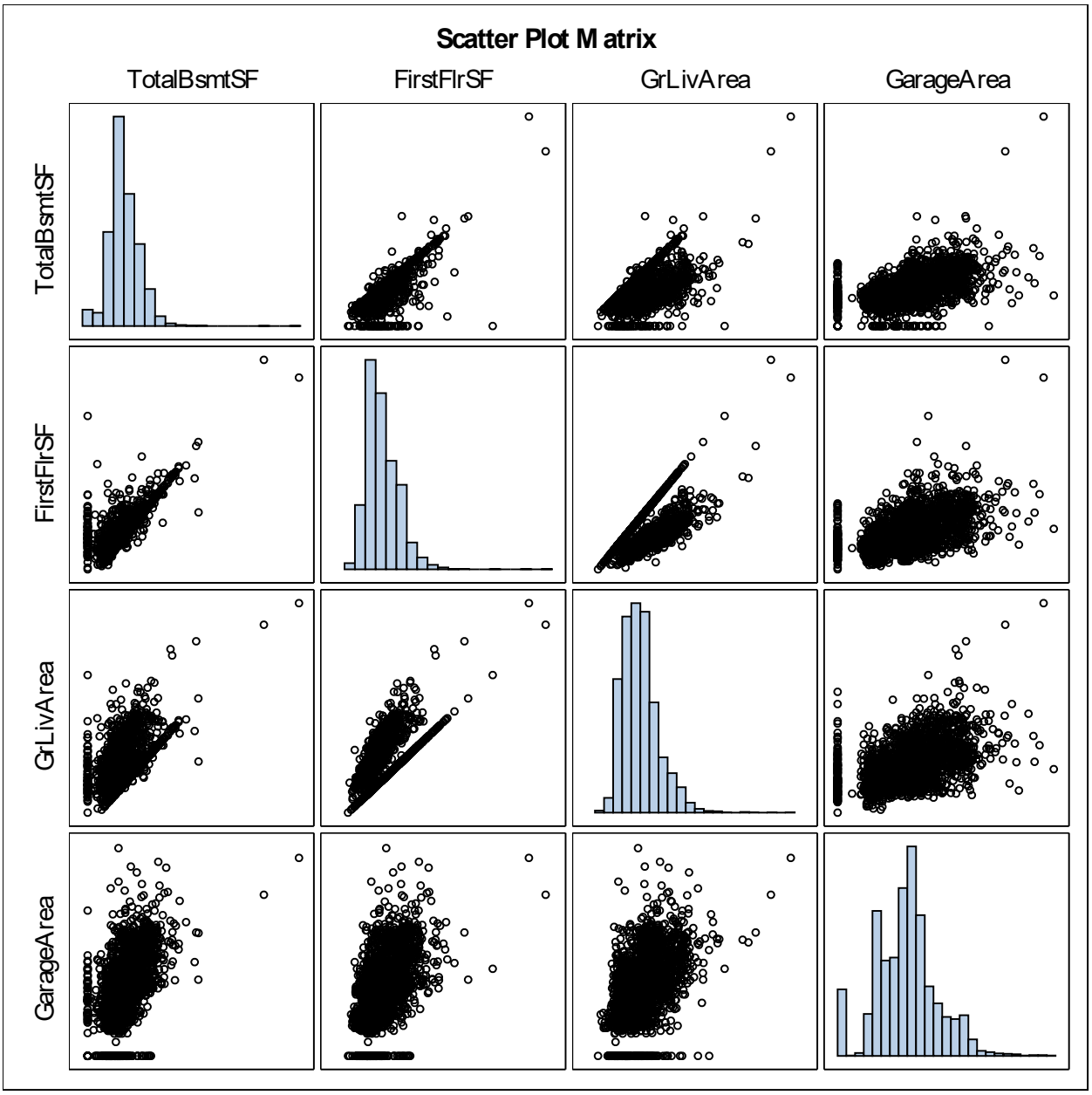
Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations						
	BsmtFinSF1	**BsmtUnfSF**	**TotalBsmtSF**	**FirstFlrSF**	**GrLivArea**	**GarageArea**
SalePrice	0.43291 <.0001 2929	0.18286 <.0001 2929	0.63228 <.0001 2929	0.62168 <.0001 2930	0.70678 <.0001 2930	0.64040 <.0001 2929

Now, we select the variables that have a correlation coefficient > 0.5:

TotalBsmtSF FirstFlrSF GrLivArea GarageArea

Scatter plot matrix of the four variables from the matrix corr proc:

```
proc corr data=temp1 plot(maxpoints=none)=matrix(histogram nvar=all);
        var TotalBsmtSF FirstFlrSF GrLivArea GarageArea;
run;
```

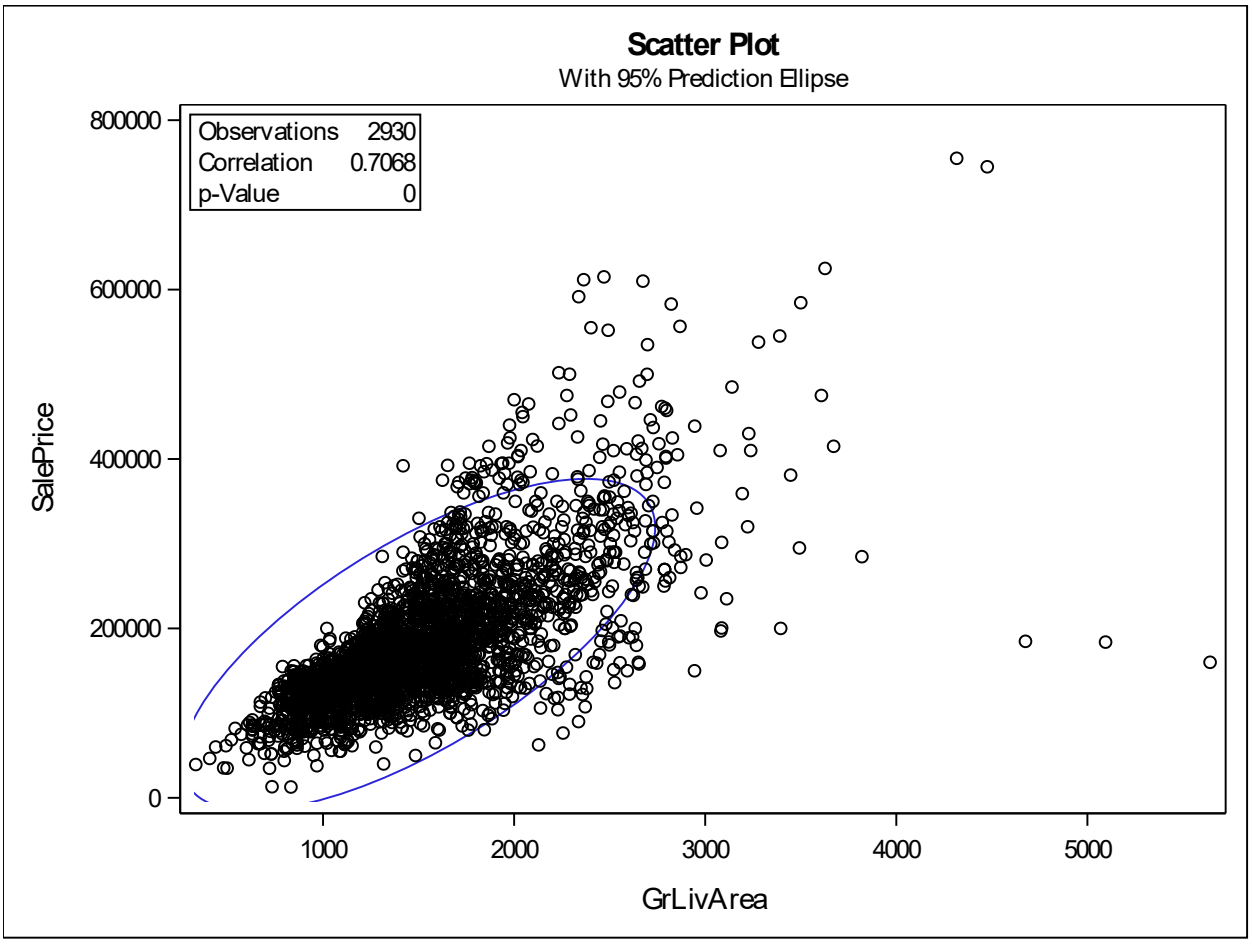


The predictor variable with strongest linear relationship with the response variable, SalePrice, is GrLivArea (above ground living area square feet). This variable has a positive correlation; whereas, GarageArea, FirstFlrSF, and TotalBsmtSF seem to have weak positive correlation.

These scatterplots may show that a relationship exists, but it does not, and cannot, prove that one variable is causing the other. There could be a third factor involved which is causing both, or some other systemic cause, or the apparent relationship could just be a fluke.

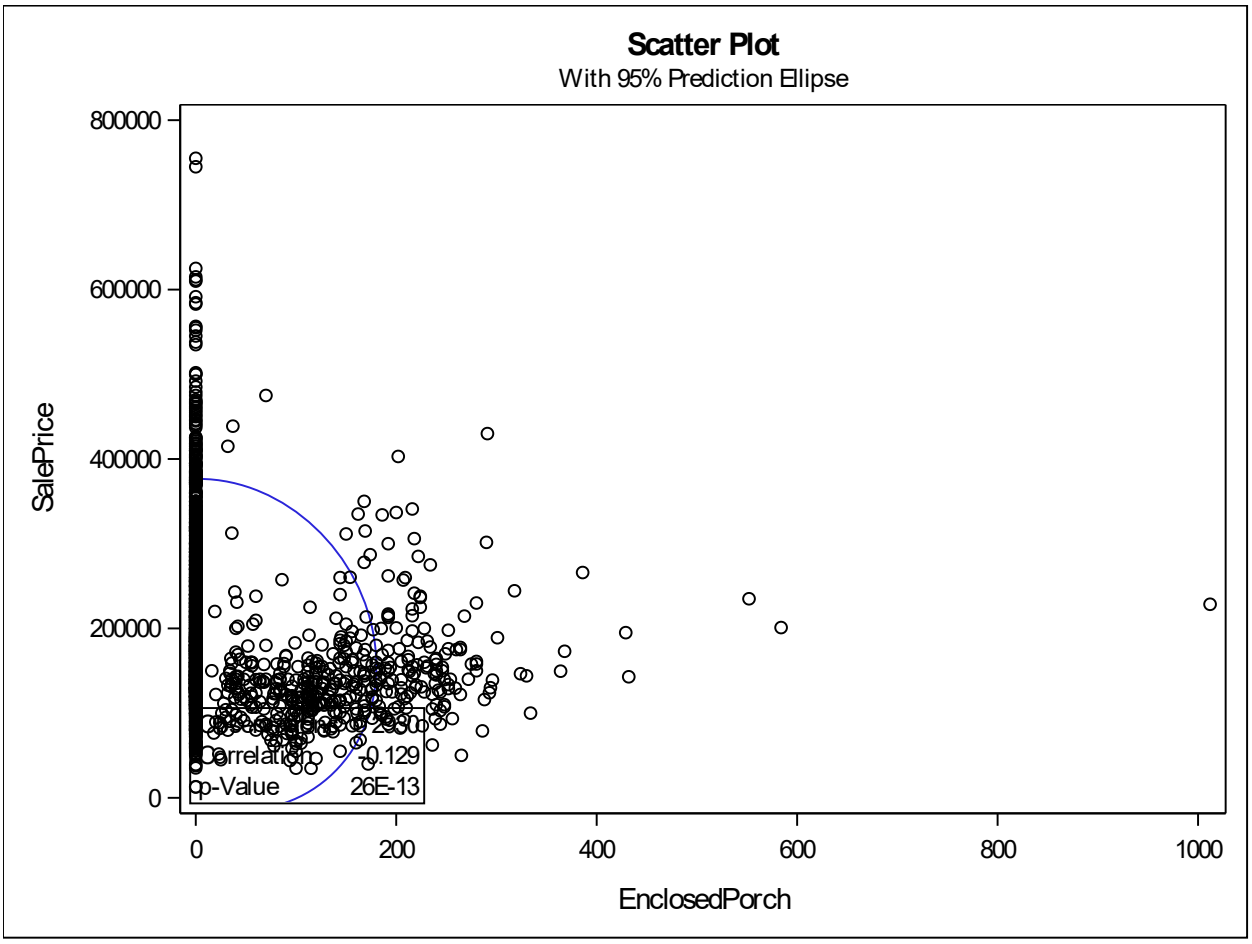
Scatterplot for the continuous variable with the highest correlation with SalePrice:

```
proc corr data=temp1 plots=(scatter);
        var GrLivArea;
        with SalePrice;
run;
```



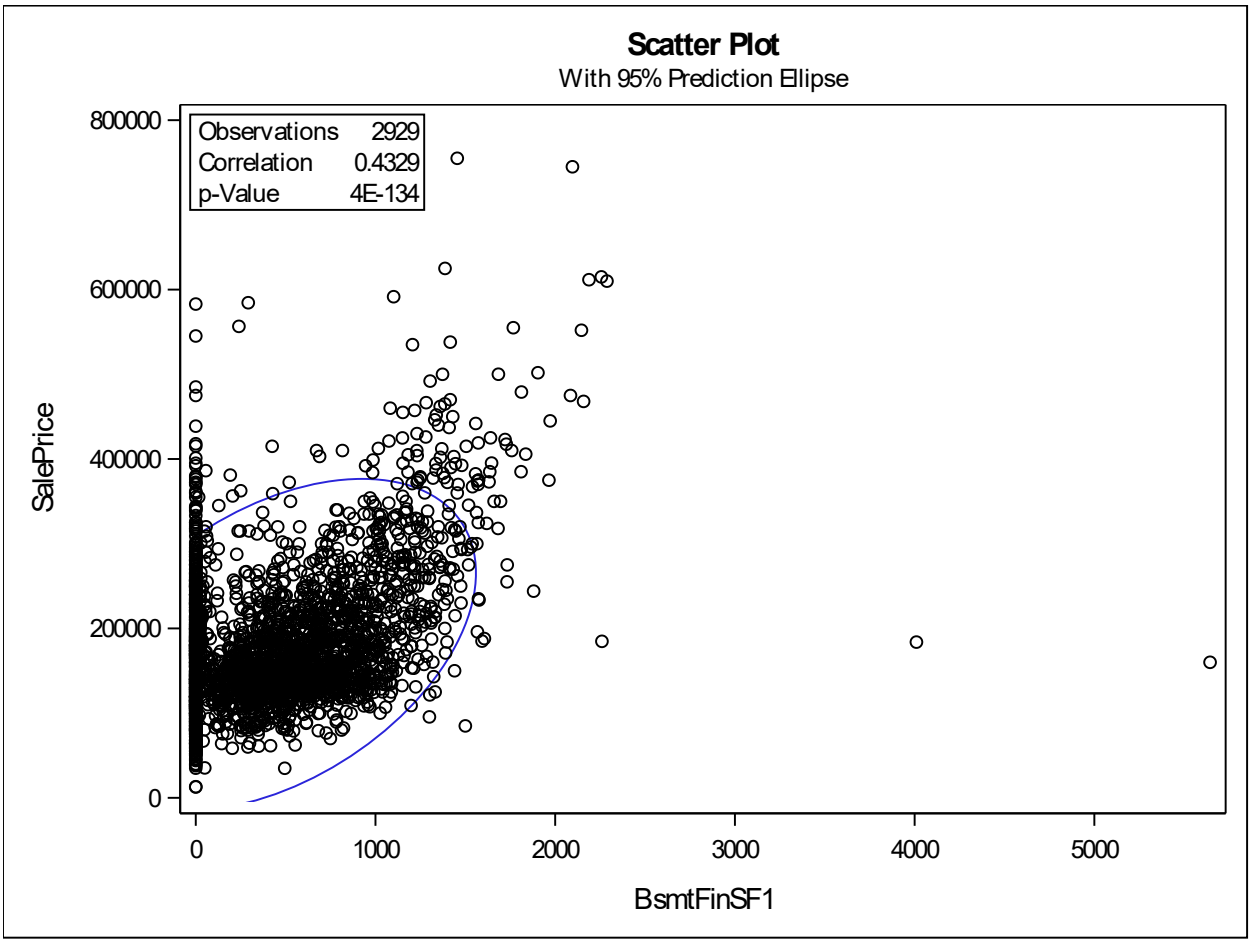
Scatterplot for the continuous variable with the lowest correlation with SalePrice:

```
proc corr data=temp1 plots=(scatter);
        var EnclosedPorch;
        with SalePrice;
run;
```



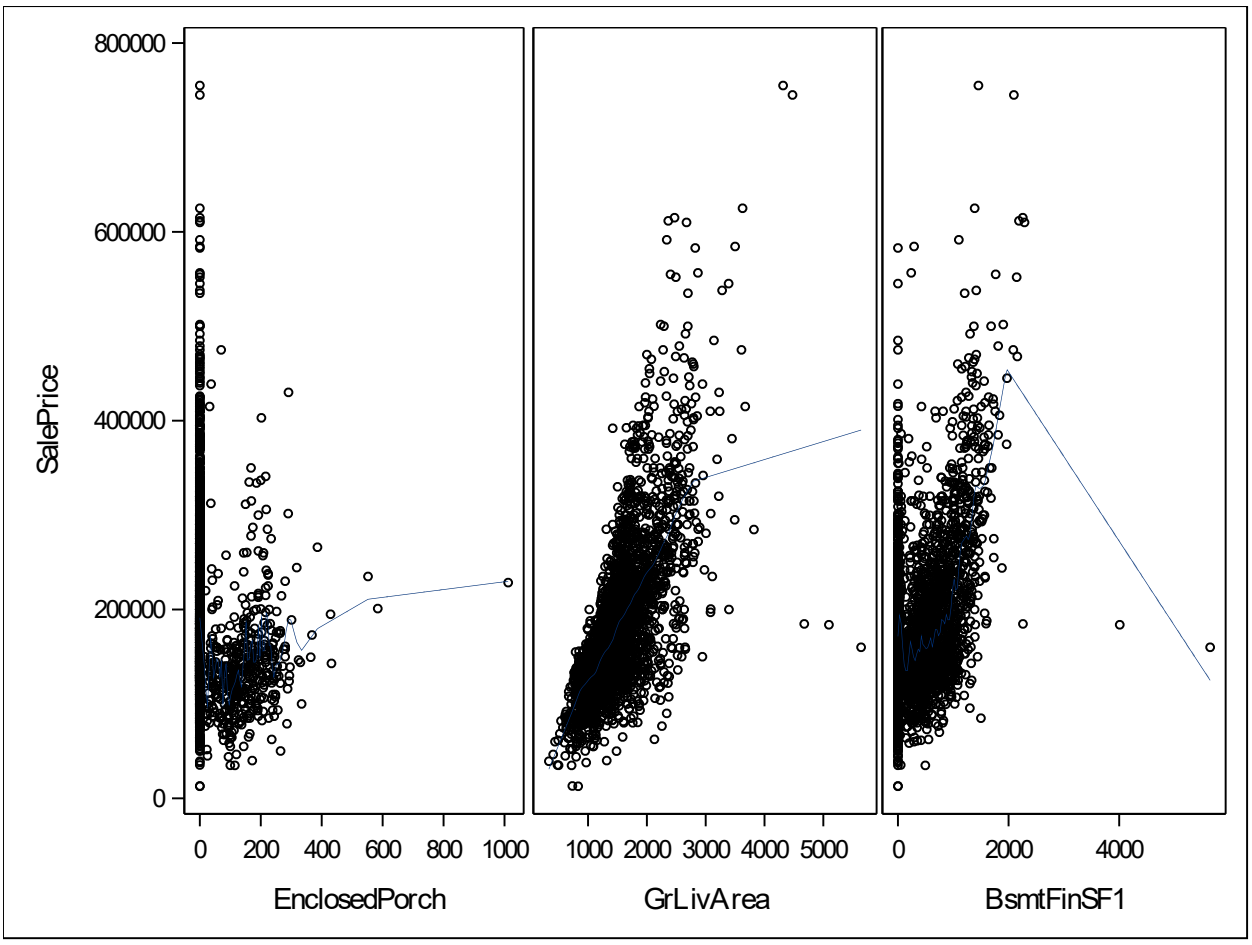
Scatterplot for the continuous variable with the correlation closest to 0.5 with SalePrice:

```
proc corr data=temp1 plots=(scatter);
        var BsmtFinSF1;
        with SalePrice;
run;
```



Scatterplot utilizing a LOESS smoother for SalePrice:

```
ods graphics on;
proc sgscatter data=temp1;
        compare x=(EnclosedPorch GrLivArea BsmtFinSF1)
        y=saleprice / loess;
run;
quit;
ods graphics off;
```



The reasons for using the LOESS is that it does not require the specification of a function to fit a model to all of the data in the sample. Instead the analyst only has to provide a smoothing parameter value and the degree of the local polynomial. In addition, LOESS is very flexible, making it ideal for modeling complex processes for which no theoretical models exist. These two advantages, combined with the simplicity of the method, make LOESS one of the most attractive of the modern regression methods for applications that fit the general framework of least squares regression but which have a complex deterministic structure.

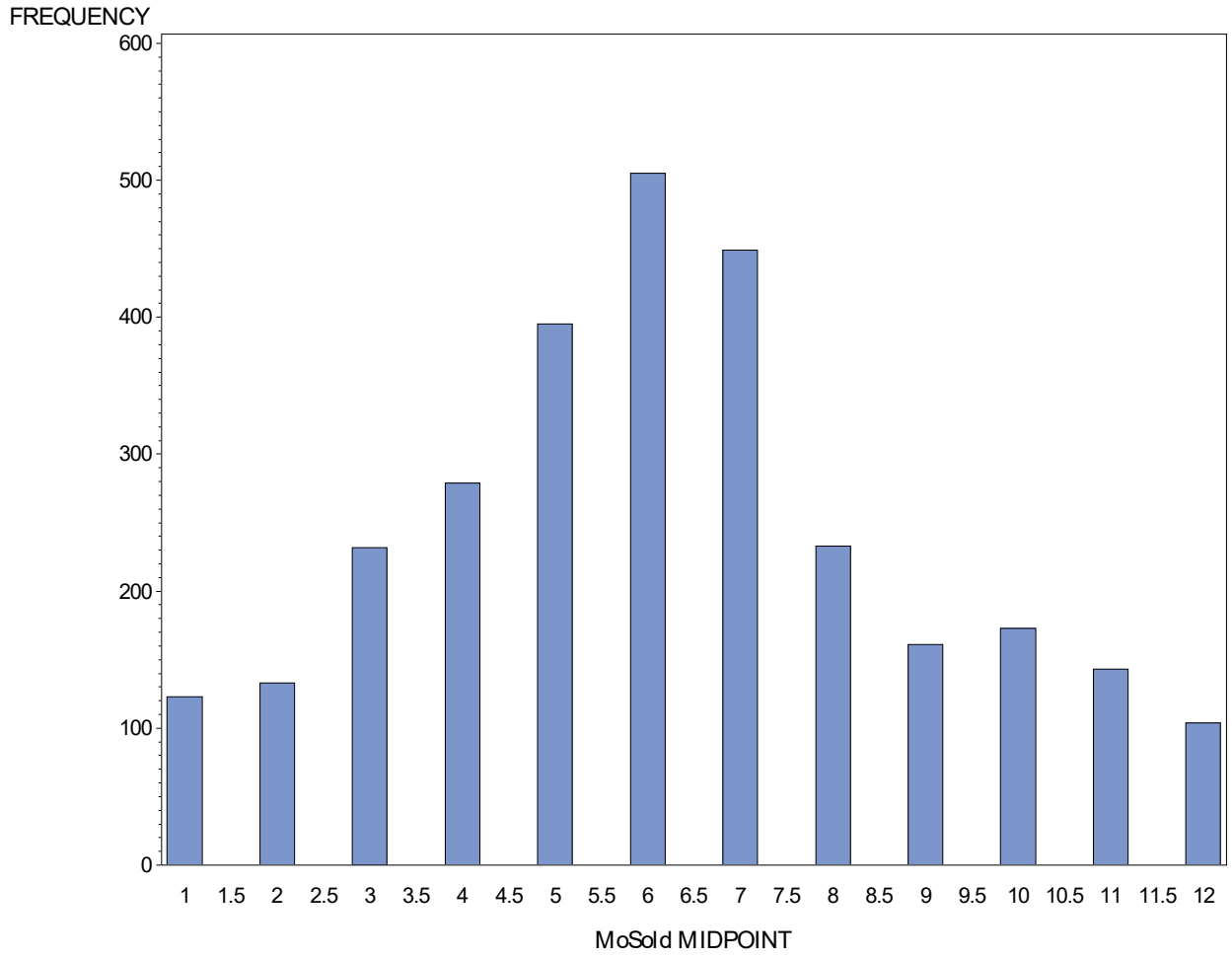
Examine the distribution of values for categorical variables utilizing proc freq (frequency) procedure.

FREQ procedure for MoSold, GarageCars, and Fireplaces:

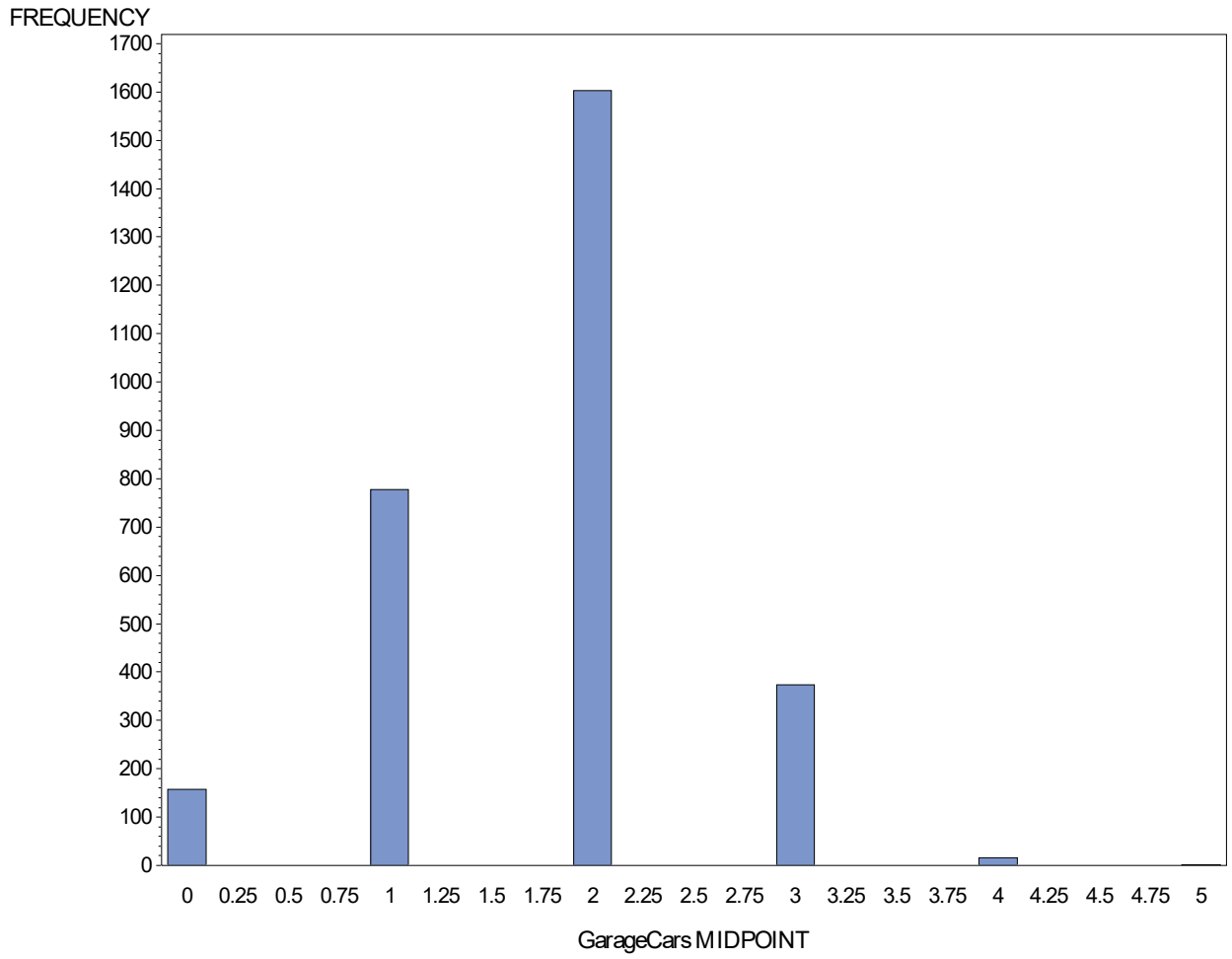
```
proc freq data=temp1;
        tables MoSold GarageCars Fireplaces;

proc gchart data=temp1;
        vbar MoSold GarageCars Fireplaces;
run;
```

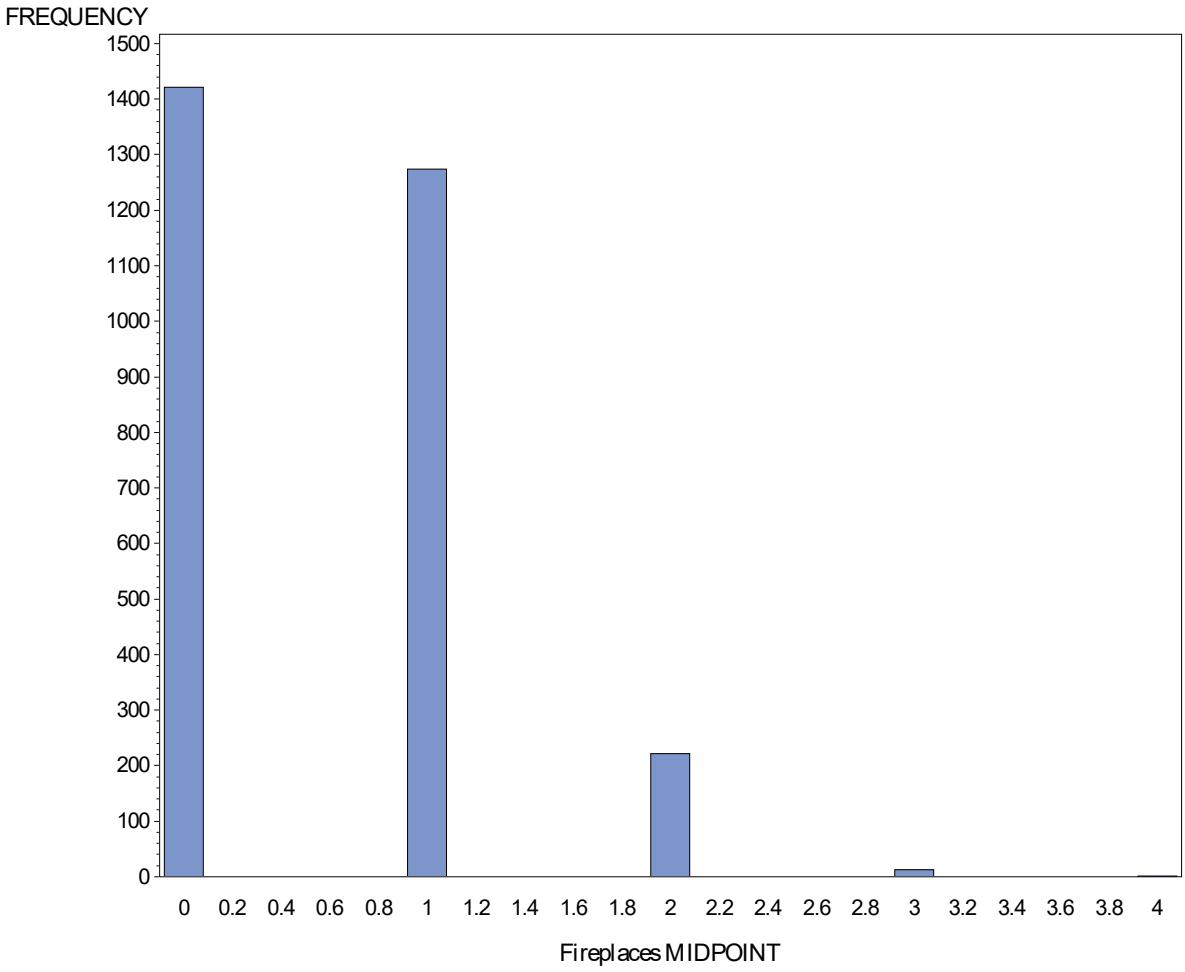
MoSold:



GarageCars:


FREQUENCY
1700
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
0
0.25
0.5
0.75
1
1.25
1.5
1.75
2
2.25
2.5
2.75
3
3.25
3.5
3.75
4
4.25
4.5
4.75
5
GarageCars MIDPOINT

Fireplaces:



Examine the linear relationship between the values of the categorical variable and the SalePrice.

SORT and MEANS procedures for MoSold, GarageCards, and Fireplaces:

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

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

proc sort data=temp1;
        by Fireplaces;
proc means data=temp1;
        by Fireplaces;
        var saleprice;
run;
```

MoSold=1

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
123	194210.02	105421.30	39300.00	755000.00

MoSold=2

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
133	178364.35	78949.16	13100.00	615000.00

MoSold=3

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
232	176130.46	83261.85	46500.00	611657.00

MoSold=4

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
279	167711.99	66342.49	64500.00	555000.00

MoSold=5

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
395	173700.22	69406.22	37900.00	584500.00

MoSold=6

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
505	181542.56	78468.36	12789.00	591587.00

MoSold=7

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
449	184366.86	84335.04	40000.00	745000.00

MoSold=8

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
233	186222.46	78548.37	64000.00	500067.00

MoSold=9

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
161	191552.14	86930.67	44000.00	545224.00

MoSold=10

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
173	180057.06	82780.11	35311.00	479069.00

MoSold=11

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
143	187651.27	79914.34	34900.00	446261.00

MoSold=12

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
104	184454.05	78826.11	62383.00	492000.00

GarageCars=.

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1	150909.00	.	150909.00	150909.00

GarageCars=0

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
157	104949.25	34069.81	34900.00	260000.00

GarageCars=1

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
778	127267.42	30919.14	35000.00	330000.00

GarageCars=2

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1603	183562.10	52049.02	12789.00	441929.00

GarageCars=3

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
374	310304.62	101883.80	81000.00	755000.00

GarageCars=4

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
16	228748.69	81085.97	123000.00	460000.00

GarageCars=5

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1	126500.00	.	126500.00	126500.00

Fireplaces=0

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1422	141195.77	44546.56	13100.00	360000.00

Fireplaces=1

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1274	213556.00	81459.23	12789.00	625000.00

Fireplaces=2

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
221	242316.16	113124.75	80400.00	755000.00

Fireplaces=3

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
12	255820.83	96637.06	160000.00	462000.00

Fireplaces=4

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1	260000.00	.	260000.00	260000.00

Examine the correlation between MoSold, GarageCars, and Fireplaces.

CORR procedure:

```
proc corr data=temp1;
        var saleprice;
        with MoSold GarageCars Fireplaces;
run;
```

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	
	SalePrice
MoSold	0.03526 0.0563 2930
GarageCars	0.64788 <.0001 2929
Fireplaces	0.47456 <.0001 2930

Reviewing the above results, the GarageCars and Fireplaces variables produce the strongest, positive correlation in the dataset. MoSold seems to be the weaker correlation of the three, due to the high p-value, and low correlation coefficient.

Conclusions:

The exploratory data analysis conducted on the Ames housing data shows that there will be many different variables that need to be examined before creating a model. The continuous variables utilized for this analysis shows that there is quite a bit of missing data from many observations, which could lead to outliers, errors, and a non-so-reliable model. As for the categorical data, I would not rely on those variables as being valuable predictors; mostly because I'm still unsure how they would work effectively without being a continuous type.

Being that the LOESS plots show inconsistency, we may consider transformation of our predictor variables during the model building process.

Code:

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

proc datasets library=mydata;
run;

Data temp1;
        set mydata.ames_housing_data;

proc contents data=temp1;

proc corr data=temp1;
        var saleprice;
        with LotFrontage LotArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF FirstFlrSF
        SecondFlrSF LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF
        EnclosedPorch ThreeSsnPorch ScreenPorch PoolArea MiscVal;

        proc sort data=temp1 out=sorted;
        by wooddecksf;

proc print data=sorted;
        var wooddecksf;
run;

proc means data=temp1;
        var poolarea;
run;

proc corr data=temp1 plot=matrix(histogram nvar=all);
        var BsmtFinSF1 BsmtUnfSF TotalBsmtSF FirstFlrSF GrLivArea GarageArea;
        with saleprice;
```

```
run;

proc corr data=temp1 plots=(scatter);
	var TotalBsmtSF FirstFlrSF GrLivArea GarageArea;
	with SalePrice;

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

proc corr data=temp1 plot(maxpoints=none)=matrix(histogram nvar=all);
	var TotalBsmtSF FirstFlrSF GrLivArea GarageArea;
run;

proc corr data=temp1 plots=(scatter);
	var EnclosedPorch;
	with SalePrice;

proc corr data=temp1 plots=(scatter);
	var GrLivArea;
	with SalePrice;

proc corr data=temp1 plots=(scatter);
	var BsmtFinSF1;
	with SalePrice;


ods graphics on;

proc sgscatter data=temp1;
	compare x=(EnclosedPorch GrLivArea BsmtFinSF1)
	y=saleprice / loess;
run;
quit;
ods graphics off;

proc freq data=temp1;
	tables MoSold GarageCars Fireplaces;

proc gchart data=temp1;
	vbar MoSold GarageCars Fireplaces;
run;

proc sort data=temp1;
	by MoSold;

proc means data=temp1;
```

```
        by MoSold;
        var saleprice;
run;

proc sort data=temp1;
        by GarageCars;

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

proc sort data=temp1;
        by Fireplaces;

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

proc corr data=temp1;
        var saleprice;
        with MoSold GarageCars Fireplaces;
run;
```