Noé Flores
Data Exploration:

Introduction:
The value of property, and houses, in particular, is determined by many different factors. These factors include variables such as the total number of full bathrooms, full bedrooms, half bathrooms, and even the month the house is sold has an impact on the sale price. The ability to effectively identify the factors that are most important in driving home values, and subsequently the final sales price of homes is critical. Our target goal is to get a glimpse of the data available on a sample population of housing data and use exploratory visualization methods to identify possible predictors for the final sales price of homes in our sample.

Results:
A quick examination of the contents of our dataset produces the results below. There are 2,930 total observations and 82 different variables.

Figure 1: Summary table contents.

Ames Housing Data	Contents
Observations	2930
Variables	82
Indexes	0
Observation Length	584
Deleted Observations	0
Compressed	NO
Sorted	NO

We need to examine the different variables types within the Ames Housing data set. A quick breakdown is provided in the table below.

Figure 2: Summary variable types.

Type	Tally
Numeric	39
Character	43
Total	82

Figure .3 below has a breakdown of some continuous and categorical values. There are 20 variables within the data set that are continuous, such as total basement square footage, lot area, open porch square footage, and first-floor square footage. We have 23 variables that are categorical in nature, like the neighborhood, paved drive, roof style, and garage type. There also some variables that fall somewhere in between, such as the total number of bedrooms and bathrooms.

Figure 3: Summary variable types.

Continuous Variables	Categorical
TotalBsmtSF	Neighborhood
LotArea	PavedDrive
OpenPorchSF	RoofStyle
FirstFlrSF	GarageType

Looking through the Ames housing data dictionary, it's relatively clear what most of the variables presented to us in this sample represent. Given previous experience with exploratory data analysis, I feel comfortable and confident in our ability to find variables within this data set that can help us indicate possible predictors of home prices. One variable that could possibly be included is location. A simple ordinal variable that gives us the ability to classify location as low, medium, or high demand could be beneficial in the modeling of home sales.

To move our analysis forward, we will examine the sales price records in an attempt to find any sales prices within our data that could be out of range or possibly erroneous, such as a negative value. Knowing that our data has 2,930 total observations, we limit the inquiry to the top and bottom 10 figures for sales in an effort not to overload the data consumer and bucket the data with similar observations. Figure .4 below provides a breakdown of those prices.

Figure 4: Summary Sales Prices.

Obs	SalePrice Descending	SalePrice Ascending
1	$ 755,000	$ 12,789
2	$ 745,000	$ 13,100
3	$ 625,000	$ 34,900
4	$ 615,000	$ 35,000
5	$ 611,657	$ 35,311
6	$ 610,000	$ 37,900
7	$ 591,587	$ 39,300
8	$ 584,500	$ 40,000
9	$ 582,933	$ 44,000
10	$ 556,581	$ 45,000

Looking at the breakdown in Figure.4, it is plain to see that there is a wide range between the maximum and minimum sale price. It is also worth mentioning that the top 10 and bottom 10 sales price observations appear to have values that are within range of one another, although the spread between the top and bottom is something that should potentially be explored further. As an added check, I performed the same analysis on lot area, and first-floor square footage. The breakdown is in Figure.5. Once again, we see that there is a wide range for the maximum and minimum values, but within themselves, everything appears to be in range.

Figure 5: Summary Sales Prices.

Obs	Lot Area Descending	Lot Area Ascending	First Floor sf Descending	First Floor sf Descending
1	215,245	1,300	215,245	1,300
2	164,660	1,470	164,660	1,470
3	159,000	1,476	159,000	1,476
4	115,149	1,477	115,149	1,477
5	70,761	1,477	70,761	1,477
6	63,887	1,484	63,887	1,484
7	57,200	1,488	57,200	1,488
8	56,600	1,491	56,600	1,491
9	53,504	1,495	53,504	1,495
10	53,227	1,504	53,227	1,504

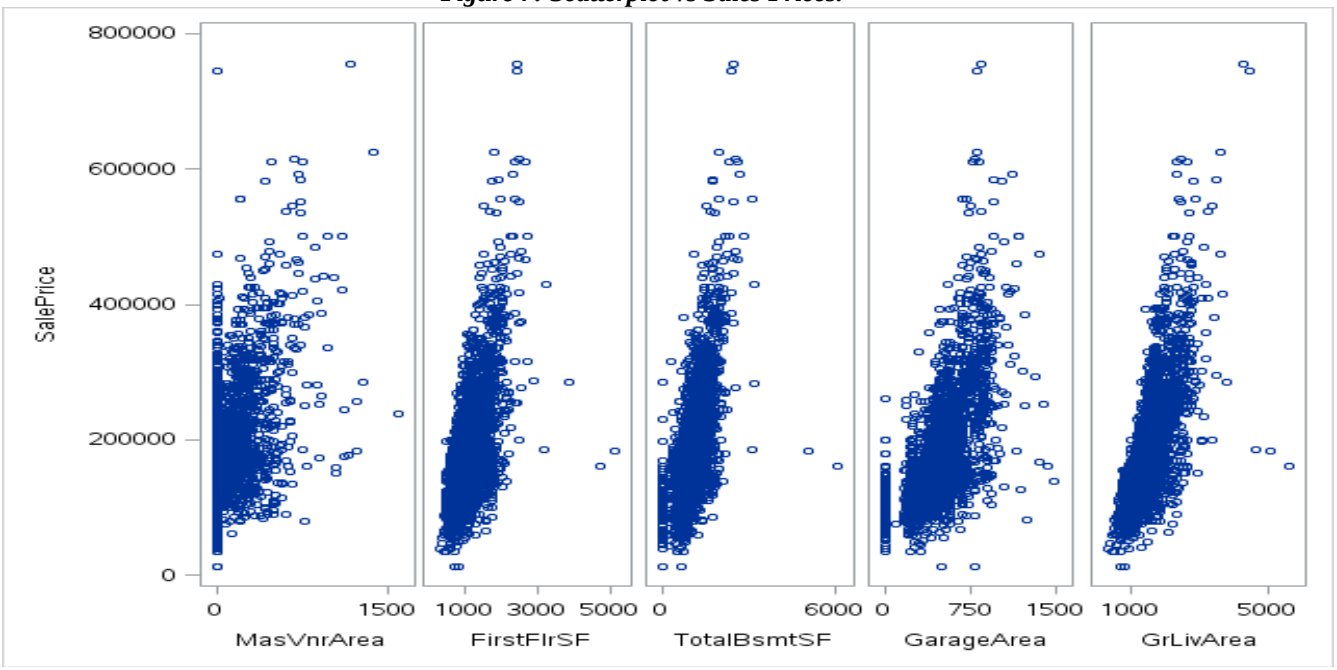
We move on to examine the correlation between various continuous predictor variables, and their relationship with the final sales price. We produce a scatterplot matrix in order to better observe and compare the different variables possible correlation to the sale price. In Figure.6 below you will see a breakdown of the top five correlated continuous variables. It would appear that the numeric correlation figures and the scatterplots appear to give us an indication of the strength of each variable. MasVnrArea has data points more widely dispersed, while GrLivArea has a more condensed formation, and we can see the better formation of a regression line. Given the previous observation, I would believe that GrLivArea will be the best predictor variable, while MasVnrArea will prove to be lacking and fall short as a predictor.

Figure 6: Summary Correlation matrix to Sales Prices.

Variable	GrLivArea	GarageArea	TotalBsmtSF	FirstFlrSF	MasVnrArea
Correlation to Sales Price	0.707	0.640	0.632	0.622	0.508
P- Value	<.0001	<.0001	<.0001	<.0001	<.0001
Observations	2930	2929	2929	2930	2907

The graphic in Figure.7 is a scatterplot of the top 5 correlated potential predictor variables with sales price.

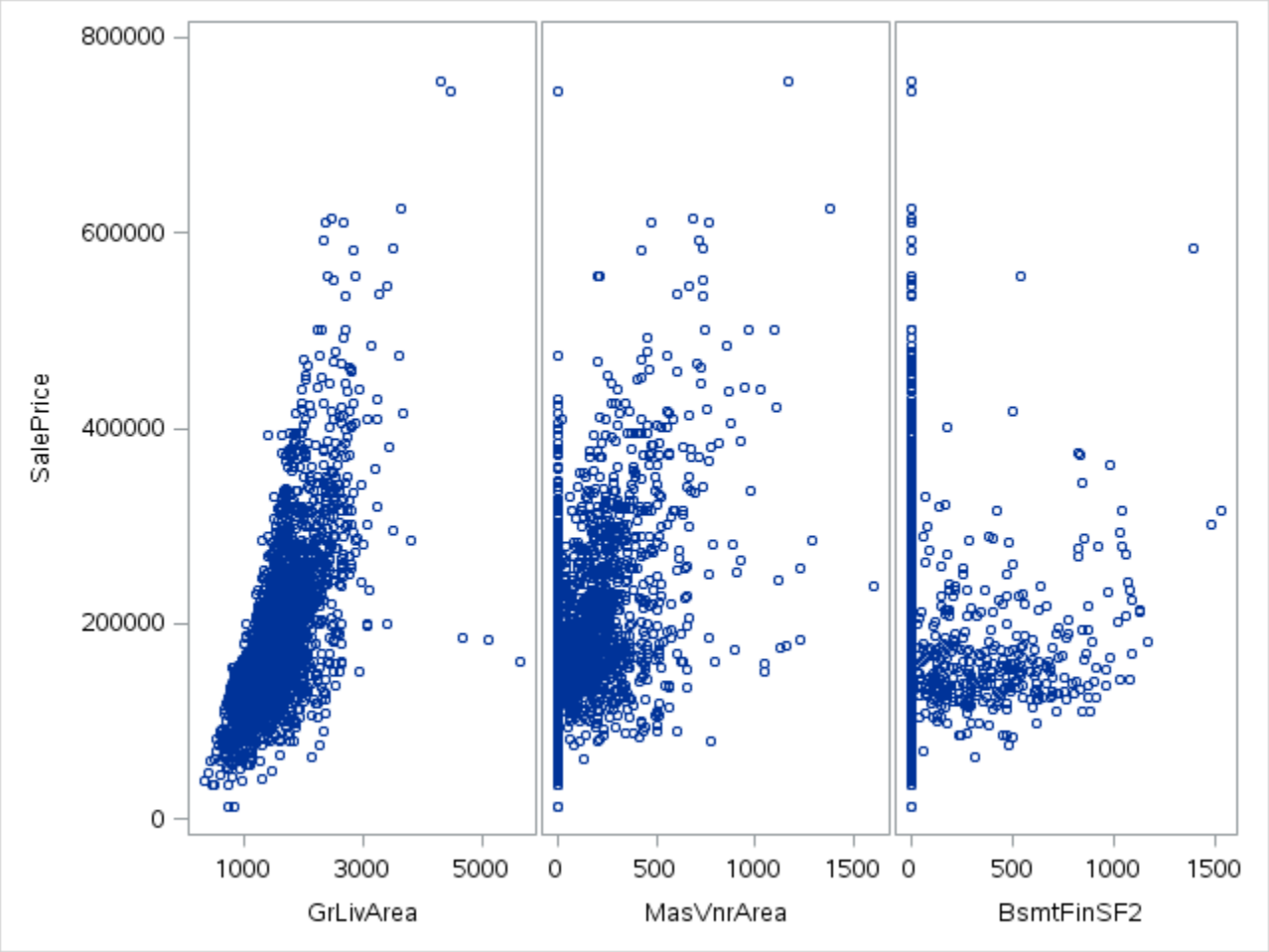
Figure 7: Scatterplot vs Sales Prices.



Regardless of the relative strength or weakness of the predictor variables we identified above, the correlation coefficient does not give us enough information to make a final decision on the strength of a predictor variable. The p-value and the number of observations, or sample size, also play a key role in determining the relative strength or weakness of a predictor value.

In order to get a better understanding of the relationship presented by the correlation coefficient of the predictor variables to sales price, we produced scatterplots with GrLivArea, the variable with the highest correlation to sale price, MasVnrArea, the variable with the most centric correlation to saleprice, and BsmtFinSF2, the variable with the lowest positive correlation in our dataset.

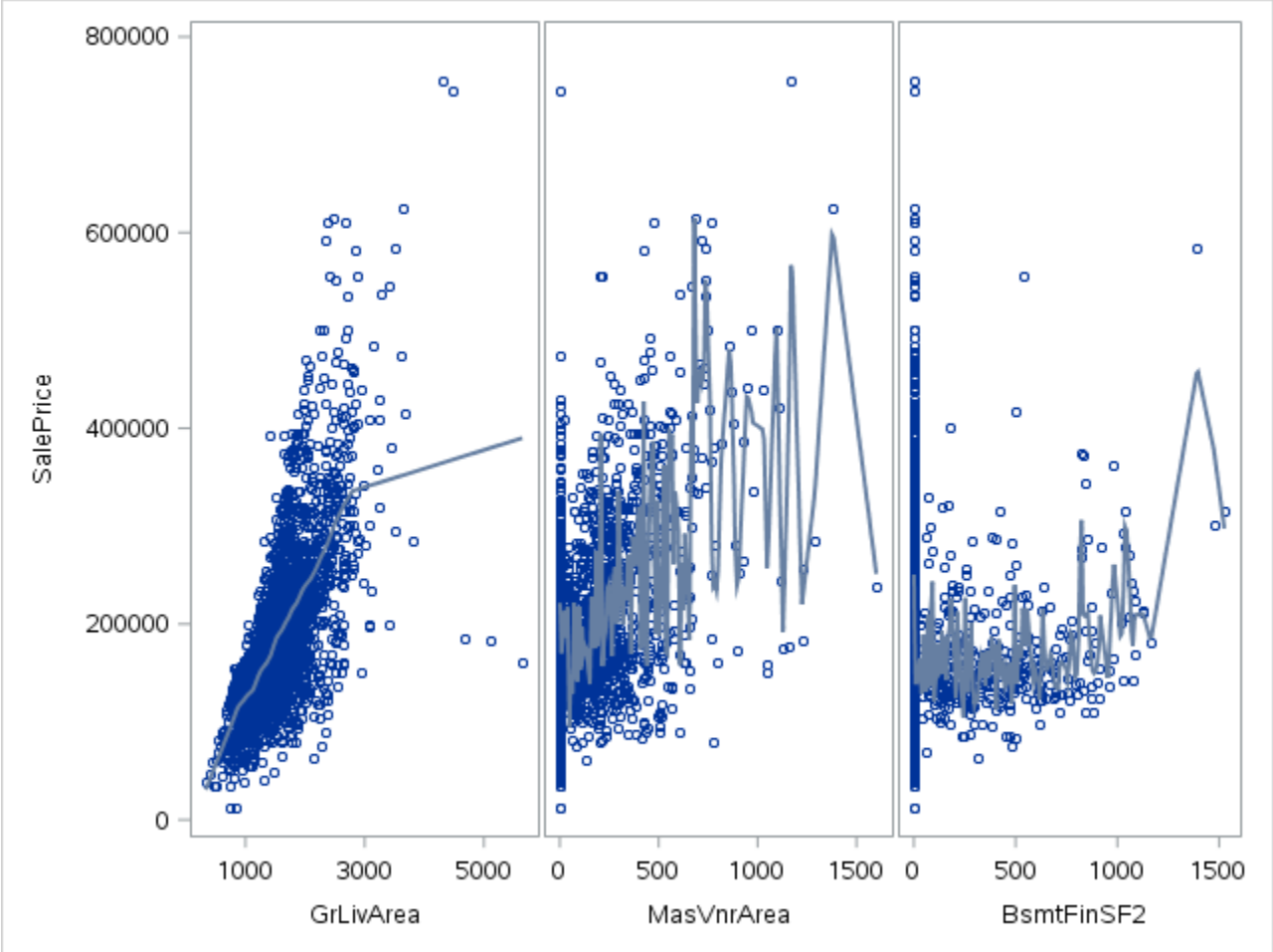
Figure 8: Sales Prices vs GrLivArea, MasVnrArea amd BsmtFinSF2.



Looking over the scatterplots displayed in Figure .8, we can see once again the tight concentration of data points between Sale Price and GrLivArea. The plot shows the beginnings of a linear relationship between the two variable. As we move on the Sale Price vs. MasVnrArea, we see that tight formation begins to break and perhaps the ability to fit a linear relationship falls apart. In the final display, the structure appears to have broken down completely.

We want to plot our variables once again, but this time we will incorporate the LOESS smoother for our predictor variable, sale price. The LOESS fits a local regression function to our data, within a chosen neighborhood (as cited in Bilenas, J.V. 2011. pg 5). The regression procedure performs a fit, weighted by the distance of points from the center of the neighborhood (as cited in Bilenas, J.V. 2011 pg 5). In Figure.9, we see the output from our new scatterplot, with the LOESS curve.

Figure 9: LOESS smoother.



There doesn't appear to be a clear indicator of a trend with respect to BsmtFinSF2, and that is what we expected, given the previous observations we made. Something interesting to note comes from the MasVnrArea. Thanks to fitting the LOESS smoother, we see that there appears to be a trend forming there in the direction of a more robust correlation. The GrLivArea displays a stronger pattern, but we should note that the LOESS smoother is going off towards some outlying data points.

We now move to select three different categorical variables from our date set in an effort to find possible relationships to the final sales price**.** We will use GargeCars, Fireplaces, and BedroomAbvGr. To begin the process, we start by collecting some necessary information regarding the distribution of these variables. The results are displayed below.

Figure 9: Distribution for Categorical Variables.

GarageCars	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	157	5.36	157	5.36
1	778	26.56	935	31.92
2	1603	54.73	2538	86.65
3	374	12.77	2912	99.42
4	16	0.55	2928	99.97
5	1	0.03	2929	100
BedroomAbvGr	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	8	0.27	8	0.27
1	112	3.82	120	4.1
2	743	25.36	863	29.45
3	1597	54.51	2460	83.96
4	400	13.65	2860	97.61
5	48	1.64	2908	99.25
6	21	0.72	2929	99.97
8	1	0.03	2930	100
Fireplaces	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	1422	48.53	1422	48.53
1	1274	43.48	2696	92.01
2	221	7.54	2917	99.56
3	12	0.41	2929	99.97
4	1	0.03	2930	100

Now that we have the information regarding the distribution of these variables, we can move on to produce some summary statistics on those categorical variables. We have the summary statistics for sales price with respect to the number of cars per garage below in Figure.10.

Figure 10: Summary Sales statistics by total Garages.

Garage Cars = NA Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1	$150,909.00		$150,909.00	$150,909.00
Garage Cars = 0 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
157	$104,949.25	$ 34,069.81	$ 34,900.00	$260,000.00
Garage Cars = 1 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
778	$127,267.42	$ 30,919.14	$ 35,000.00	$330,000.00
Garage Cars = 2 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1603	$183,562.10	$ 52,049.02	$ 12,789.00	$441,929.00
Garage Cars = 3 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
374	$310,304.62	$101,883.80	$ 81,000.00	$755,000.00
Garage Cars = 4 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
16	$228,748.69	$ 81,085.97	$123,000.00	$460,000.00
Garage Cars = 5 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1	$126,500.00		$126,500.00	$126,500.00

There appears to be a few outlying variables in this breakdown, but overall we can observe the presence of a trend. As the number of cars per garage increases, we see a steady increase in the mean Sales Price. The increase is also easily distinguished in the minimum and maximum values.

There is definitely an argument to made that a similar trend is present when we look at the total number of bedrooms in relation to the sales price, but there appears to be a bit more inconsistency here. While there is definitely a trend up as the number of bedrooms increase, we also seem to reach a ceiling of sorts with regard to the effect on the sale price. According to the information in Figure.11 below, 4 bedrooms looks like the sweet spot, but we admit there is much more that goes into the final price, and we are commenting solely on this specific variable.

Figure 11: Summary Sales statistics by total Bedrooms above ground.

BedroomAbvGr = 0 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
8	$ 218,494.88	$ 97,669.25	$ 108,959.00	$ 385,000.00
BedroomAbvGr = 1 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
112	$ 183,017.34	$105,616.39	$ 35,000.00	$ 615,000.00
BedroomAbvGr = 2 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
743	$ 162,167.68	$ 80,522.59	$ 12,789.00	$ 611,657.00
BedroomAbvGr = 3 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1597	$ 179,711.77	$ 65,577.75	$ 40,000.00	$ 500,000.00
BedroomAbvGr = 4 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
400	$ 216,357.05	$105,292.52	$ 62,500.00	$ 755,000.00
BedroomAbvGr = 5Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
48	$ 206,244.25	$ 94,056.23	$ 65,000.00	$ 545,224.00
BedroomAbvGr = 6Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
21	$ 159,701.71	$ 52,656.90	$ 84,900.00	$ 269,500.00
BedroomAbvGr = 8 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1	$ 200,000.00		$ 200,000.00	$ 200,000.00

The information in Figure.12 holds a similar breakdown of summary statistics, although this time with respect to sales price and the total number of fireplaces.

Figure 12: Summary Sales statistics by total Fireplaces.

Fireplaces = 0 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1422	$141,195.77	$ 44,546.56	$ 13,100.00	$360,000.00
Fireplaces = 1 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1274	$213,556.00	$ 81,459.23	$ 12,789.00	$625,000.00
Fireplaces = 2 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
221	$242,316.16	$113,124.75	$ 80,400.00	$755,000.00
Fireplaces = 3 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
12	$255,820.83	$ 96,637.06	$160,000.00	$462,000.00
Fireplaces = 4 Analysis Variable : Sales Price				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1	$260,000.00		$260,000.00	$260,000.00

The information we have presented so far from these 3 categorical variables would appear to indicate that there is some correlation between these variables and the sales price. In an effort to investigate this idea we can run a simple correlation. The information is presented below.

Figure 13: Correlation matrix of variables with sales price.

Variable	GarageCars	BedroomAbvGr	Fireplaces
Correlation to Sales Price	0.648	0.144	0.475
P- Value	<.0001	<.0001	<.0001
Observations	2929	2930	2930

Conclusion:

The possible future use of several continuous and categorical variables as predictors of the final sales price of homes within our Ames housing dataset have emerged from this exploratory data analysis. This was indicative in our preliminary correlation matrix of the continuous variables we found and the final correlation matrix against three specific categorical variables. While it's clear that some of the data and variables warrant further exploration, other variables we explored in this analysis left more doubt. Part of the difficulty of this data set is the large number of categorical variables present. The LOESS smoother gave us some indication of some possible linear trends. There is also the possibility that the methods and variables we used could be developed upon, perhaps through the use of a more expanse ANOVA analysis.

Appendix: SAS code

```
/*Read file into SAS and identify it as mydata*/

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

/*Problem 1; Examine variables, identify continuous and categorical variables
identify what they measure, is the data useful to predict property value*/

proc datasets library=mydata; run;

Data temp1;
  set mydata.ames_housing_data;

proc contents data=temp1 order=varnum;
run;

/*Problem 2; use proc sort to sort data by sale price*/

proc sort data=temp1; /*Problem 2 part 1*/
   by descending saleprice;

proc print data=temp1 (obs=10); /*Problem 2 part 1*/
   run;

 proc sort data=temp1; /*Problem 2 part 1*/
   by saleprice;

proc print data=temp1 (obs=10); /*Problem 2 part 1*/
   run;

/*Problem 2 part 2*/
proc sort data=temp1; /*Problem 2 part 1*/
   by descending Lotarea;

proc print data=temp1 (obs=10); /*Problem 2 part 1*/
   run;

proc sort data=temp1; /*Problem 2 part 1*/
   by Lotarea;

proc print data=temp1 (obs=10); /*Problem 2 part 1*/
   run;

/*Problem 2 part 3/*/
proc sort data=temp1; /*Problem 2 part 1*/
   by descending FirstFlrSf;
```

```
proc print data=temp1 (obs=10); /*Problem 2 part 1*/
   run;

proc sort data=temp1; /*Problem 2 part 1*/
   by FirstFlrSf;

proc print data=temp1 (obs=10); /*Problem 2 part 1*/
   run;

/*Problem 3 part 1*/
ods graphics on;

proc corr data=temp1 plot=matrix(histogram nvar=all);
   var _numeric_;
run;
ods graphics off;

proc corr data=temp1;
  var _numeric_;
    With saleprice;
run;

proc corr data=temp1 plot=matrix(histogram);
   var saleprice;
   with Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch
   PoolArea MiscVal;
run;

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

run;

proc sgscatter data=temp1;
  compare x=(MasVnrArea FirstFlrSF TotalBsmtSF GarageArea GrLivArea)
        y= saleprice;
run;

/*Problem .4 Scatterplot matrix of varibales vs saleprice for Problem 4*/
proc sgscatter data=temp1;
  compare x=(GrLivArea MasVnrArea BsmtFinSF2)
        y= saleprice;
run;
```

```
/*Individual scatterplot for overallqual vs saleprice*/
proc sgscatter data=temp1;
  compare x= GrLivArea
        y= saleprice;
run;

/*Individual scatterplot for PID vs saleprice*/
proc sgscatter data=temp1;
  compare x= MasVnrArea
        y= saleprice;
run;

/*Individual scatterplot for TotRmsAbvGrd vs saleprice*/
proc sgscatter data=temp1;
  compare x= BsmtFinSF2
        y= saleprice;
run;

/*Problem .5 Scatterplot matrix of varibales vs saleprice with LOESS smoother*/
proc sgscatter data=temp1;
  compare x=(GrLivArea MasVnrArea BsmtFinSF2)
        y= saleprice/ loess;
run;

/*Individual scatterplot for overallqual vs saleprice*/
proc sgscatter data=temp1;
  compare x= GrLivArea
        y= saleprice / loess;
run;

/*Individual scatterplot for PID vs saleprice*/
proc sgscatter data=temp1;
  compare x= MasVnrArea
        y= saleprice / loess;
run;

/*Individual scatterplot for TotRmsAbvGrd vs saleprice*/
proc sgscatter data=temp1;
  compare x= BsmtFinSF2
        y= saleprice / loess;
run;

/*Problem .6 */
proc freq data=temp1;
  tables GarageCars BedroomAbvGr Fireplaces ;
run;
```

```
/*Problem .7 Scatterplot matrix of varibales vs saleprice with LOESS smoother*/
/*Proc sort and proc mean by GarageCars*/
proc sort data=temp1;
  by GarageCars;

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

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

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

/*proc sort and proc mean by Fireplaces*/
proc sort data=temp1;
  by Fireplaces;

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

/*Problem .8 Scatterplot matrix of varibales vs saleprice with LOESS smoother*/
/*Proc sort and proc mean by GarageCars*/
ods graphics on;
  proc corr data=temp1 nosimple rank plot=matrix(histogram nvar=all)PLOTS(MAXPOINTS=none
);
    var GarageCars Fireplaces BedroomAbvGr;
    with saleprice;

run;
ods graphics off;
```

References

Black, K. (2008). *Business statistics: For contemporary decision making*. Hoboken, NJ: Wiley.

Wilcox, R. R. (2009). *Basic statistics: Understanding conventional methods and modern insights*. Oxford: Oxford University Press.

Cody, R. (2011). *SAS: Statistics by Example*. Carey, NC: SAS Institute Inc.

Bilenas, J,V. "*Scatterplot Smoothing Using PROC: LOESS and Restricted Cubic Splines*." http://jonasbilenas.com/Loess.pdf (accessed January 8, 2017)