**Assignment #1**
**Nitin Gaonkar**

**Introduction:**

The purpose of this assignment is to understand the data for future statistical modelling, there are three important parts in understanding the data, 1) a data survey, (2) a data quality check, and (3) an initial exploratory data analysis.

**Results:**

1. **Variables description:**

**Below table has the variables classified:**

| Nominal | Ordinal | Continuous | Discrete |
|---|---|---|---|
| PID | Lot Shape | Lot Frontage | Year Built |
| MS SubClass | Utilities | Lot Area | Year Remod/Add |
| MS Zoning | Land Slope | Mas Vnr Area | Bsmt Full Bath |
| Street | Overall Qual | BsmtFin SF 1 | Bsmt Half Bath |
| Alley | Overall Cond | BsmtFinType 2 | Full Bath |
| Land Contour | Exter Qual | BsmtFin SF 2 | Half Bath |
| Lot Config | Exter Cond | Bsmt Unf SF | Bedroom |
| Neighborhood | Bsmt Qual | Total Bsmt SF | Kitchen |
| Condition 1 | Bsmt Cond | 1st Flr SF | TotRmsAbvGrd |
| Condition 2 | Bsmt Exposure | 2nd Flr SF | Functional |
| Bldg Type | BsmtFin Type 1 | Low Qual Fin SF | Fireplaces |
| House Style | HeatingQC | Gr Liv Area | Garage Yr Blt |
| Roof Matl | Electrical | Garage Area | Garage Cars |
| Exterior 1 | KitchenQual | Wood Deck SF | Mo Sold |
| Exterior 2 | FireplaceQu | Open Porch SF | Yr Sold |
| Mas Vnr Type | Garage Finish | Enclosed Porch | |
| Foundation | Garage Qual | 3-Ssn Porch | |
| Heating | Garage Cond | Screen Porch | |
| Central Air | Paved Drive | Pool Area | |
| Garage Type | Pool QC | Misc Val | |
| Misc Feature | Fence | SalePrice | |
| Sale Type | | | |
| Sale Condition | | | |

Yes, I think the variables such as yearbuild, year remod, lot area would definetly help in building a statistical model for predicting the sale price, although the data contains the other variables which also play a major role in predicting the sale price, I feel that school district(ordinal) variable can be a included in the analysis, since it influences sale price.

**2 .**

we should not print the entire observations, since we have too many observations, because of this the system may crash.

After sorting the data with sale price ascending and descending and by looking the max and min value of sale price I could say that the range of the sale price was pretty high. The difference between the maximum sale price and sale price of most of the other observations was large, which implied that the values around the maximum sale price were kind of fishy.

In order to analyze further I ran the mean proc and got the below values of mean and std deviation, by analyzing the mean and the std deviation it implied that the values around the maximum values of sale price were outliers.

| Analysis Variable : SalePrice | | | | |
|---|---|---|---|---|
| N | Mean | Std Dev | Minimum | Maximum |
| 2930 | 180796.06 | 79886.69 | 12789.00 | 755000.00 |

The second continuous variable which I used is TotalBsmtSF.

After sorting the data I could see that the max value is 6110 and the min value is Zero, after running the proc mean I could see that the mean and the std deviation is 1051.61 and 440.61, the values around the maximum like 6110, 5095 looked fishy.

| Analysis Variable : TotalBsmtSF | | | | |
|---|---|---|---|---|
| N | Mean | Std Dev | Minimum | Maximum |
| 2929 | 1051.61 | 440.6150670 | 0 | 6110.00 |

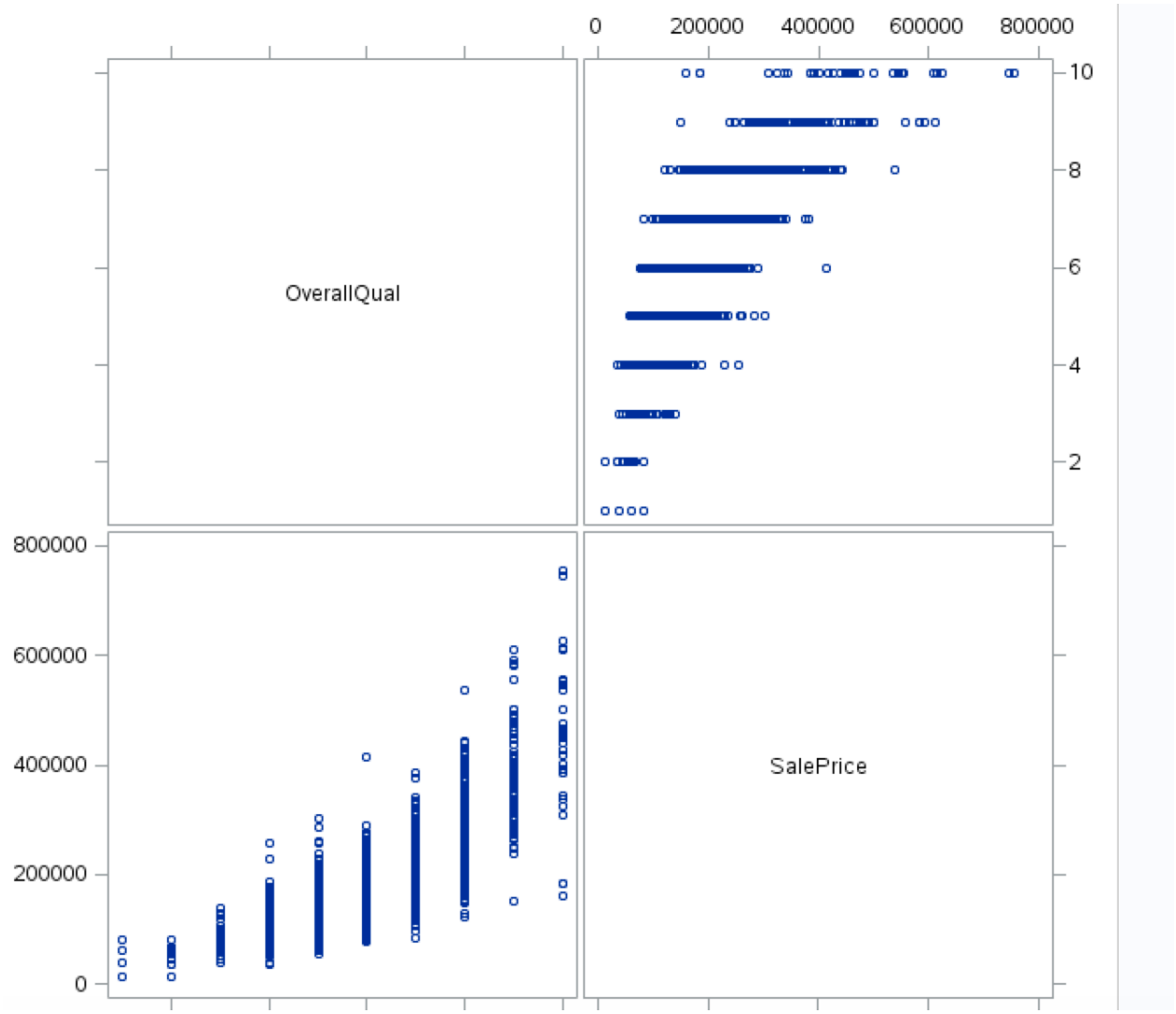
After sorting the data with firstflrsf ascending and descending and by looking the max and min value of sale price I could say that the range of the firstflrsf was pretty high. The difference between the maximum sale price and sale price of most of the other observations was large, which implied that the values around the maximum firstflrsf were kind of fishy.

In order to analyze further I ran the mean proc and got the below values of mean and std deviation, by analyzing the mean and the std deviation it implied that the values around the maximum values of firstflrsf were outliers.

| Analysis Variable : FirstFlrSF | | | | |
|---|---|---|---|---|
| N | Mean | Std Dev | Minimum | Maximum |
| 2930 | 1159.56 | 391.8908853 | 334.0000000 | 5095.00 |

**3.**

Overall Qual predictor variable has a strong linear relationship with the response variable. By looking at the numeric correlation measure we can see that there is a strong relation, as per the scatter plot we can see that the sale price increases with the higher value of overall Qual.

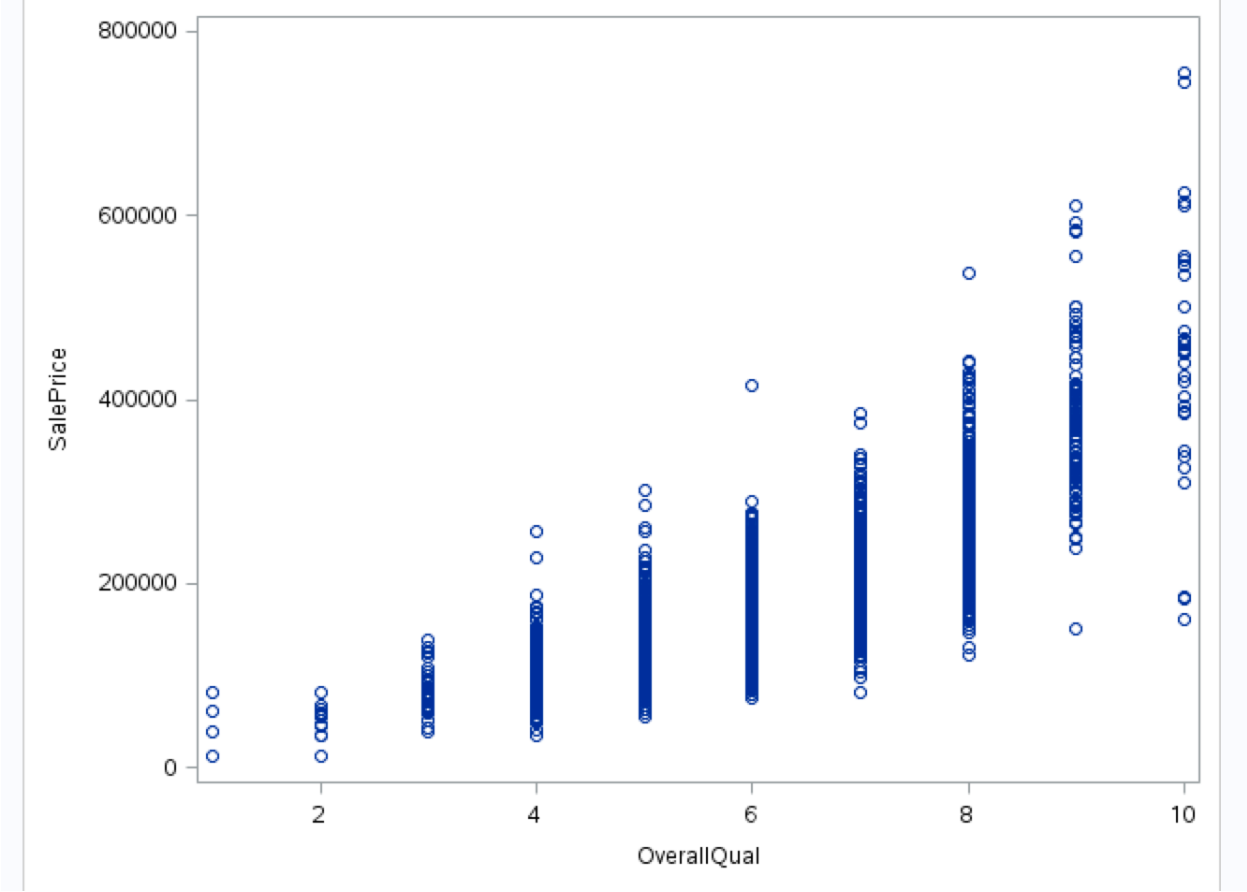


I think overall qual will be the best predictor variable since it has a good linear correlation ship with the response variable and the worst would be mosold since the correlation with the response variable is pretty low.

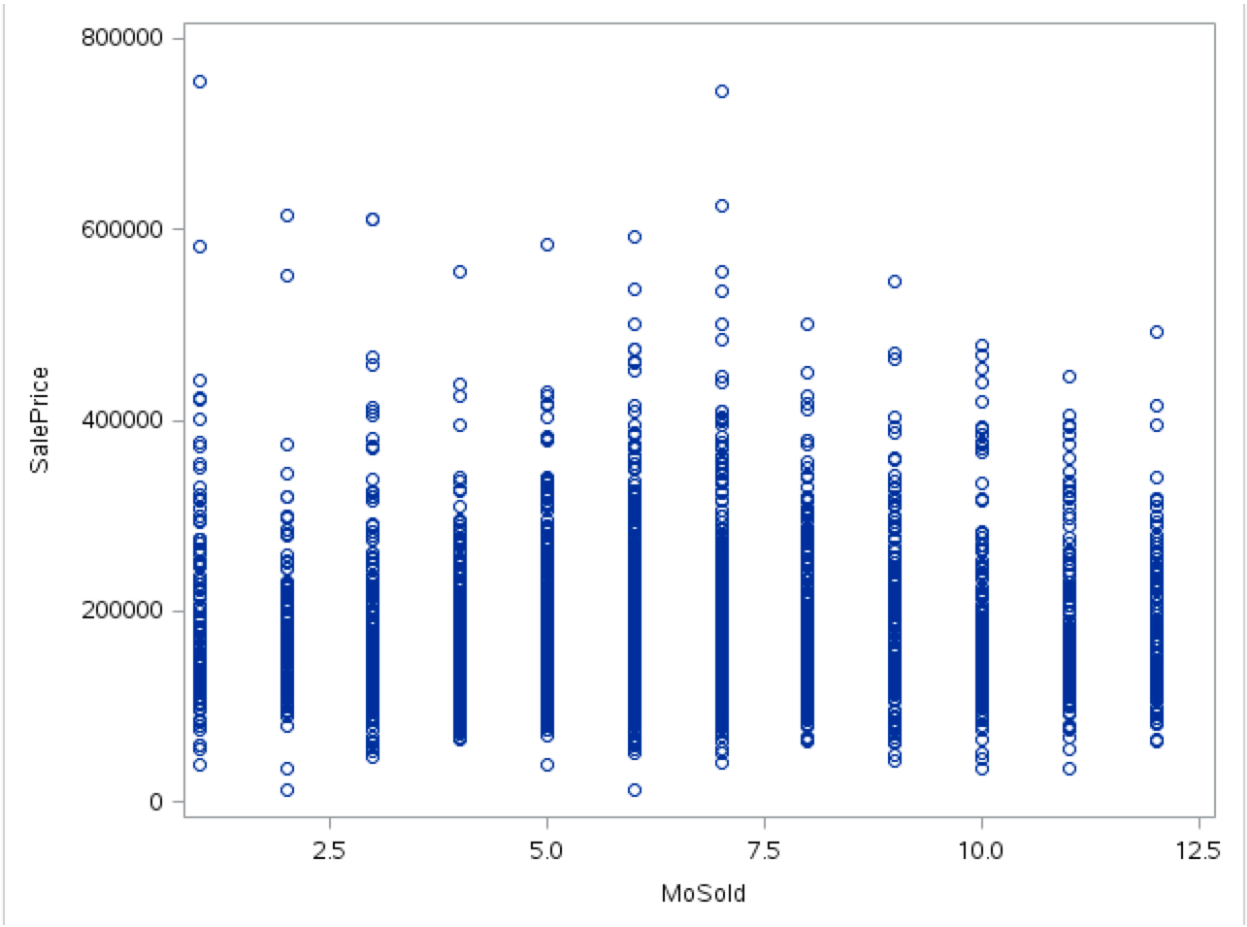
I did not find any high correlations within the set of potential variables, coefficient is not sufficient to make a decision regarding the predictor variable and its usefulness because the reliability of the linear model also depends on how many observed data points are in the sample.

4.

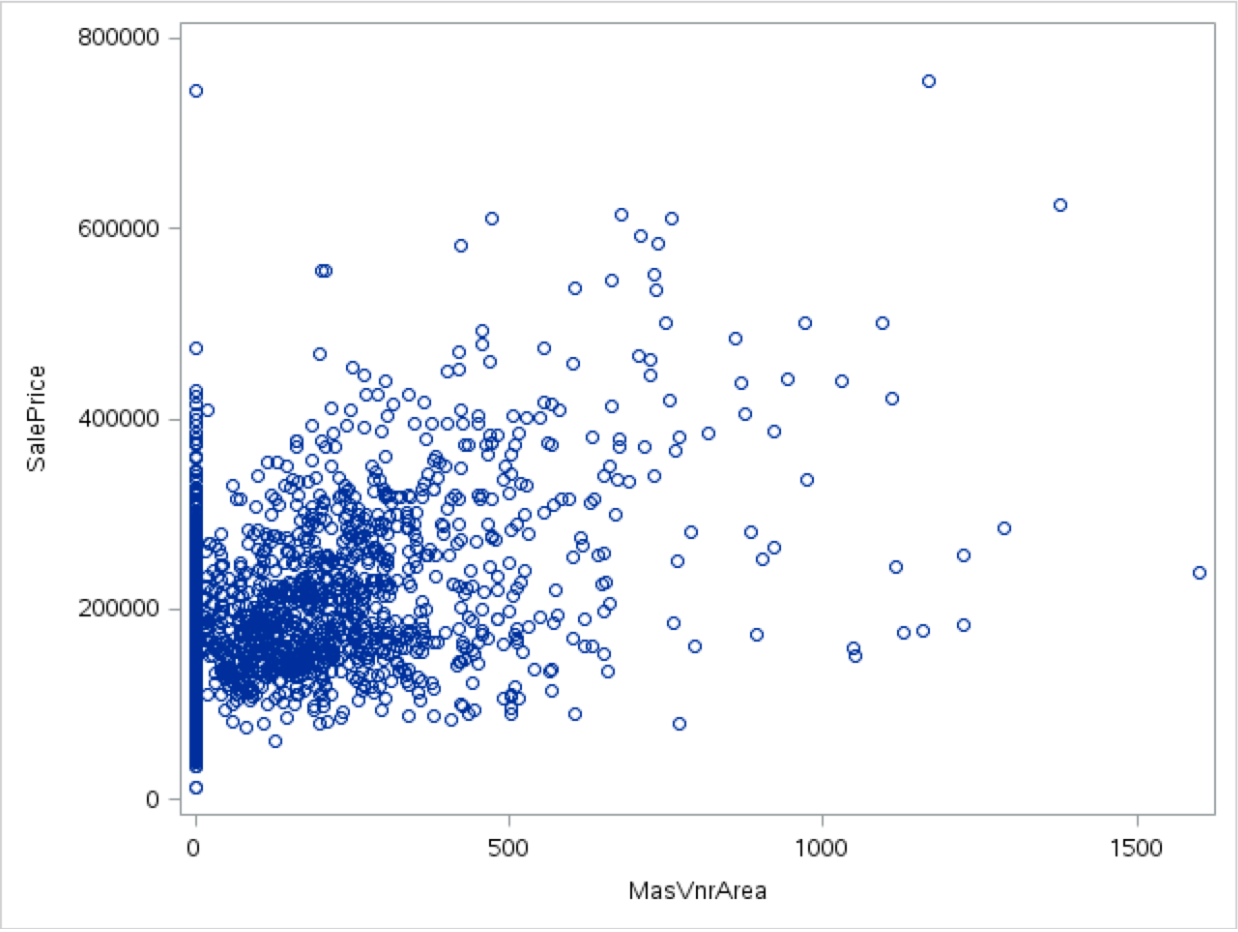
Below is the scatter plot for the overall qual and sale price as the overall qual has highest correlation with Y.



**As we can see from the scatter plot that the correlation between the X and Y is at the lowest.**



**MasVnrArea** variable has a corr co-efficient as 0.5, so let's plot a scatter plot for this as X with the sale price.



**5.**

Below is the scatter plot with loess.

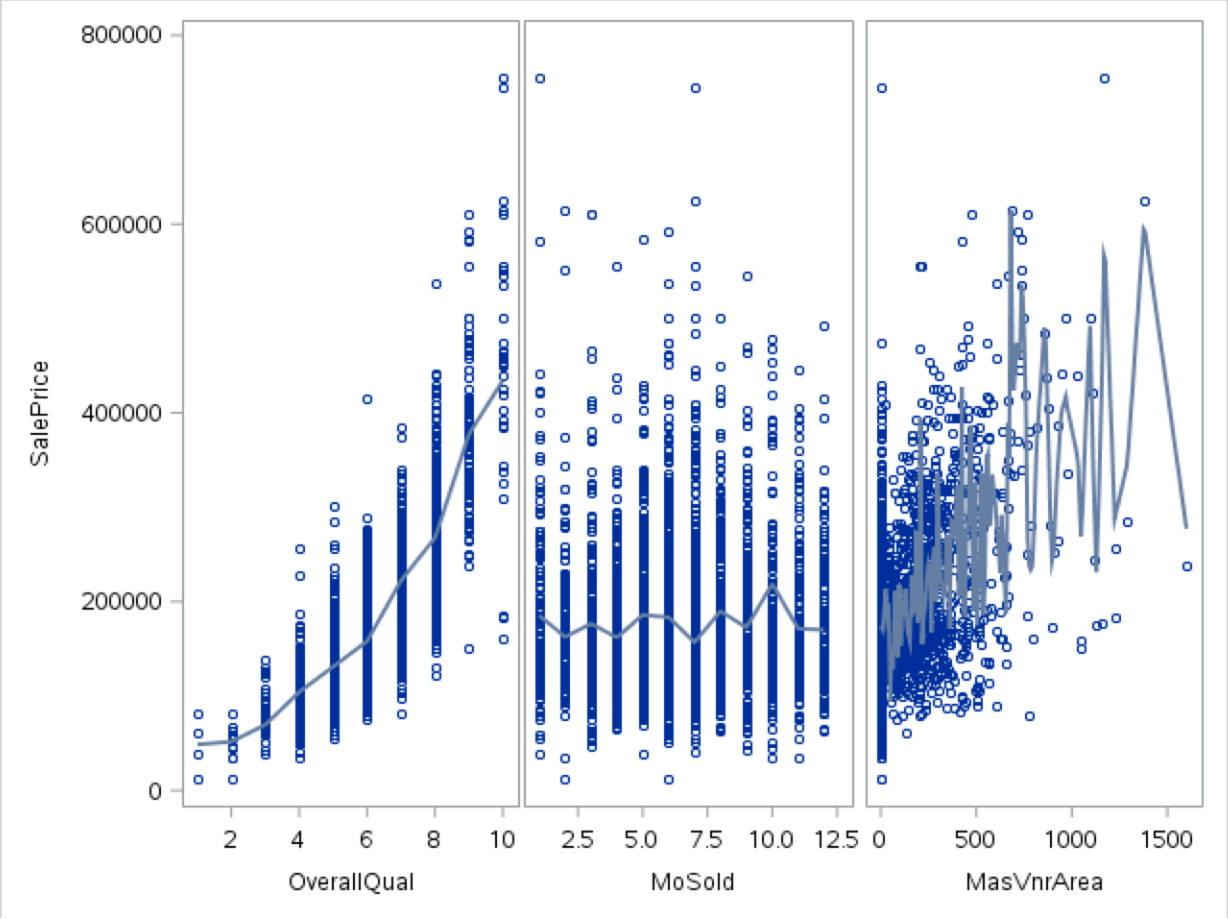
Here I have used three predictor variables overllqual, mosold and masvbrArea.

We can see that the overallqual is strongly correlated and we have a linear relationship with the sale price, where as the mosold is not correlated and no linear relationship with the sale price, where as the masvnr area is considered we can see that there is some correlation with the sale price.

Loess curve is weighted regression and it helps us to have a look at the chart and understand the big picture, from the graph one it shows that the saleprice increase as the overallqual increases.

Also we don't see much relationship between the mosold and the sale price.

For masvnr area we can see that there is some correlation between the sale price.



Below is the table produced by proc freq using the variable Mosold, here we can see that the frequency is high between mosold 5 to 7 implying that frequency of selling are high during these months, because of this the percent are also pretty high among these months., also the data looks like normally distributed.

6.

Below are the frequency tables for the 3 categorical variables.

1. Mosold
2. Fireplaces
3. Bedroombvgr

Below is the table produced by proc freq using the variable Mosold, here we can see that the frequency is high between mosold 5 to 7 implying that frequency of selling are high during these months, because of this the percent are also pretty high among these months., also the data looks like normally distributed.

**The FREQ Procedure**

| MoSold | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 | 123 | 4.20 | 123 | 4.20 |
| 2 | 133 | 4.54 | 256 | 8.74 |
| 3 | 232 | 7.92 | 488 | 16.66 |
| 4 | 279 | 9.52 | 767 | 26.18 |
| 5 | 395 | 13.48 | 1162 | 39.66 |
| 6 | 505 | 17.24 | 1667 | 56.89 |
| 7 | 449 | 15.32 | 2116 | 72.22 |
| 8 | 233 | 7.95 | 2349 | 80.17 |
| 9 | 161 | 5.49 | 2510 | 85.67 |
| 10 | 173 | 5.90 | 2683 | 91.57 |
| 11 | 143 | 4.88 | 2826 | 96.45 |
| 12 | 104 | 3.55 | 2930 | 100.00 |

**Below is the table produced by proc freq using the variable Fireplaces, here we can see that the frequency is high between 0 to 1 implying that the there are lot of houses without fireplaces and also with only one fireplace, there are few observations where the fireplaces are more than 2. So almost 91 % of the houses in the observations have one or no fireplaces.**

**The FREQ Procedure**

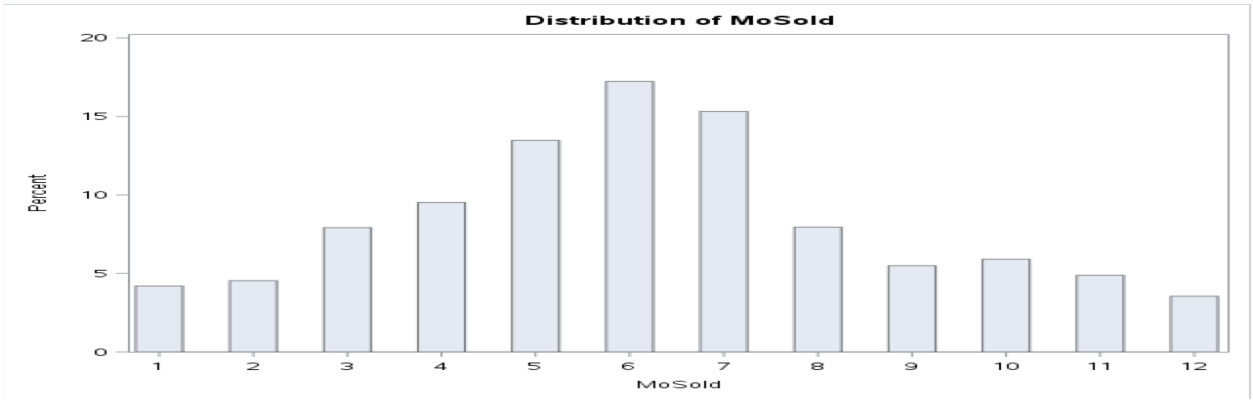
| Fireplaces | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 1422 | 48.53 | 1422 | 48.53 |
| 1 | 1274 | 43.48 | 2696 | 92.01 |
| 2 | 221 | 7.54 | 2917 | 99.56 |
| 3 | 12 | 0.41 | 2929 | 99.97 |
| 4 | 1 | 0.03 | 2930 | 100.00 |

**Below is the table produced by proc freq using the variable bedroomabvgr, here we can see that the frequency is high between 2 to 3 implying that the there are lot of houses with 2 to 3 bedrooms above ground about 50 % of the houses have 3 bedrooms above the ground.**

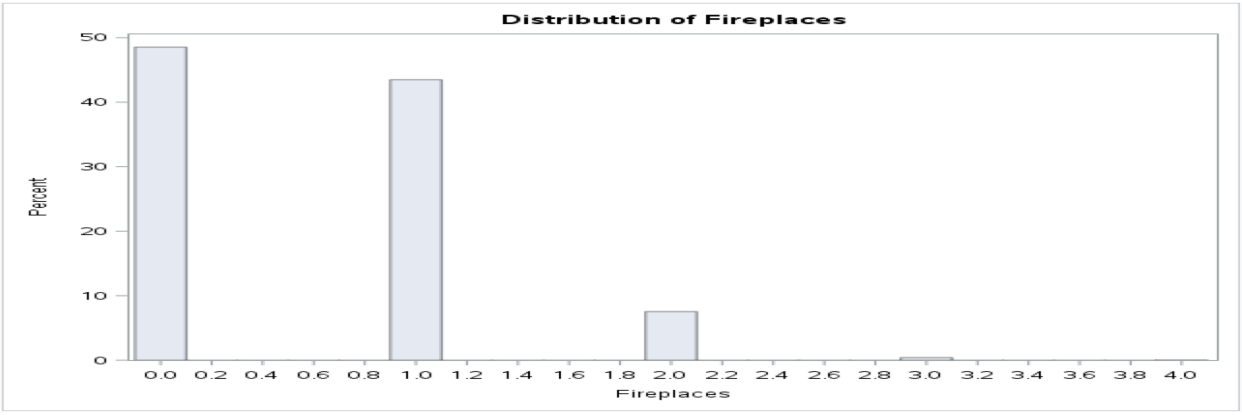
**The FREQ Procedure**

| BedroomAbvGr | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 8 | 0.27 | 8 | 0.27 |
| 1 | 112 | 3.82 | 120 | 4.10 |
| 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 | 100.00 |

**Below are the histogram for the data distribution of these 3 categorical variables:**








**7.**

The first categorical variable which I used for this analysis is the Mosold.

After going through the data I could see that the mean sale price is high for the first month, also I could see that the std deviation is also pretty high for the month 1 so we have wide range of values.

Also noticed that the sale price is high around year end..

By having a look at the correlation values of mosold and sale price and also having a look at the scatter plot I can say that I don't see a linear relationship with the mosold and sale price.

The second categorical variable which I used for this analysis is the BedroomAbvGr

After going through the data I could see that the mean sale price is high even when the Bedroomabgr is zero that implies that the no of bedroomabvgr do not influence that much in the sale price. As we can also see that the mean sale price is around 200 k for 8 Bedroomabgr which is not on the higher side of the sale price.

By having a look at the correlation values of BedroomAbvGr and sale price and also having a look at the scatter plot I can say that I don't see a strong linear relationship with the Bedroomabvgr and sale price.

The third categorical variable which I used for this analysis is the Fireplaces

After going through the data I could see that the mean sale price is high whenever there are more fireplaces, for ex: average sale price is around 140 k without fireplaces where as the average sale price is around 260 k with 4 fireplaces, as we can see the sale price increase with the increase in fireplaces

By having a look at the correlation values of Fireplaces and sale price and also having a look at the scatter plot I can say that I see a linear relationship with the fireplaces and sale price.

**8.**

The CORR Procedure

**1 With Variables:** SalePrice

**3 Variables:** MoSold BedroomAbvGr Fireplaces

| Simple Statistics | | | | | | |
|---|---|---|---|---|---|---|
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** |
| **SalePrice** | 2930 | 180796 | 79887 | 529732456 | 12789 | 755000 |
| **MoSold** | 2930 | 6.21604 | 2.71449 | 18213 | 1.00000 | 12.00000 |
| **BedroomAbvGr** | 2930 | 2.85427 | 0.82773 | 8363 | 0 | 8.00000 |
| **Fireplaces** | 2930 | 0.59932 | 0.64792 | 1756 | 0 | 4.00000 |

| Pearson Correlation Coefficients, N = 2930<br>Prob > \|r\| under H0: Rho=0 | | | |
|---|---|---|---|
| | **MoSold** | **BedroomAbvGr** | **Fireplaces** |
| **SalePrice** | 0.03526<br>0.0563 | 0.14391<br><.0001 | 0.47456<br><.0001 |

By looking at the results we can say that the only variable that has a strong correlation is the fireplaces which can have an influence on the sale price other two variables doest look like they have strong correlation with the response variable (saleprice)

**Conclusion:**

i don't see any major potential difficulties or concerns in the model building, the variables provided would be definitely sufficient and effective in building the process, only variable i think could have been added would have been the neighborhood with good school district as this factor definitely affects the sale price. I think overallqual variable has a major influence on the sale price and it has a strong correlation with the sale price, also we may require few transformation for few of the predictor such as the Bedroomabvgr while modeling to check if they have any influence on the sale price. i think total bsmt sf and gr liv area these two variables can be a major factor in the variation of the sale price. Overall i think we have enough variables to build a model.

**Code:**

Paste your code in at the end.

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

proc datasets library=mydata;

run;

 quit;


data my_assign;


set mydata.ames_housing_data;

proc contents data=my_assign;

run;
```

Question 2:

```
proc sort data=my_assign;

by descending saleprice;

proc print data=my_assign(obs=10);

run;

proc means data=my_assign;
var TotalBsmtSF;

run;


proc sort data=my_assign;
```

```
by descending TotalBsmtSF;
proc print data=my_assign(obs=100);
run;
```

**Question 3:**

```
ods graphics on;
proc corr data=my_assign plot=matrix(histogram nvar=all);
with saleprice;
run;
ods graphics off;

proc sgscatter data=my_assign;
 title "Scatterplot Matrix for Ames housing Data";
 matrix OverallQual saleprice;
run;
```

**4.**

```
proc sgplot data=my_assign;
scatter x=OverallQual y=saleprice;
run;
proc sgplot data=my_assign;
scatter x=OverallQual y=saleprice;
```

```
run;
```

5

6.

```
 proc freq data=my_assign;
tables Bedroomabvgr ;
run;
proc freq data=my_assign;
tables Fireplaces ;
run;
proc freq data=my_assign;
tables mosold ;
run;
```

7.

```
proc sort data=my_assign;
by mosold;

proc means data=my_aasign;
by mosold;
var saleprice;
run;
proc sort data=my_assign;
```

```
by Bedroomabvgr;


proc means data=my_assign;

by Bedroomabvgr;

var saleprice;

run;


proc sort data=my_assign;

by fireplaces;


proc means data=my_assign;

by fireplaces;

var saleprice;

run;
```

**8.**

```
ods graphics on;

proc corr data=my_assign plot=matrix(histogram);

var mosold Bedroomabvgr Fireplaces;

with saleprice;

run;
```