**FE 517: SAS FOR FINANCE**

# FINAL PROJECT REPORT ON AMES HOUSING DATASET


**BY:**

**HETANSH MADHANI | 10413329 | hmadhani@stevens.edu**

**RUDRESH BHATT | 10419384 | rbhatt2@stevens.edu**

**PRIYANK GUPTA | 10422451 | pgupta8@stevens.edu**


## INTRODUCTION:

We are going to use Ames Housing Dataset available on Kaggle to perform some insightful data analysis. We plan to plot the data trying to identify the important factors on which the price of the house depends on. The dataset has over 70 explanatory variables describing every aspect of residential homes in Ames, Iowa. We want to find out which of these 70 explanatory variables have the most impact on the final house price. We hope to find if there are variables which have a linear relation to the final house price and we plan to find correlation between the variables and eventually try to build a regression model and see how well the model performs.

## DATASET:

The data set contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010. The data has 82 columns which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers).

## APPROACH:

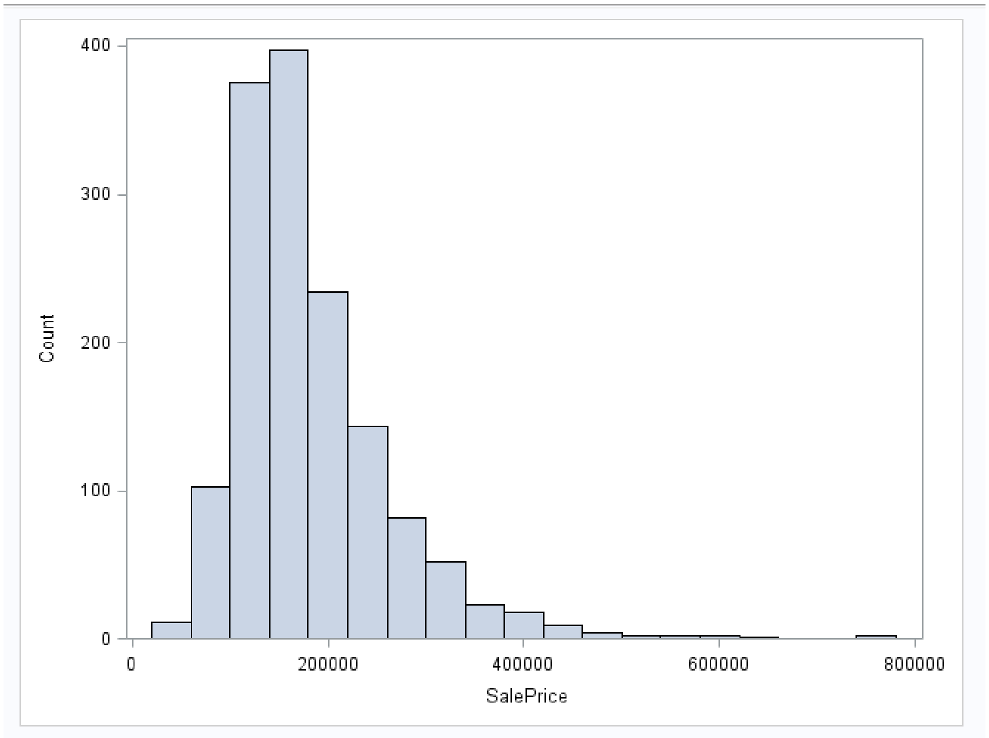
We will initially load the data set and examine the variables in the Ames Housing Data set. We have a data dictionary and we will use this dictionary as soon as our exploration brings us to the point where we need clarification about a categorical variable or another ambiguity in the data collection.

We can use the SAS procedure 'contents' to examine a list of the variables and their types, lengths, and formats respectively.

```
proc contents data = train order = varnum;
run;
```

We then analyze the data by visualizing important variables and target variable. We start by plotting the histogram of our target variable SalePrice.

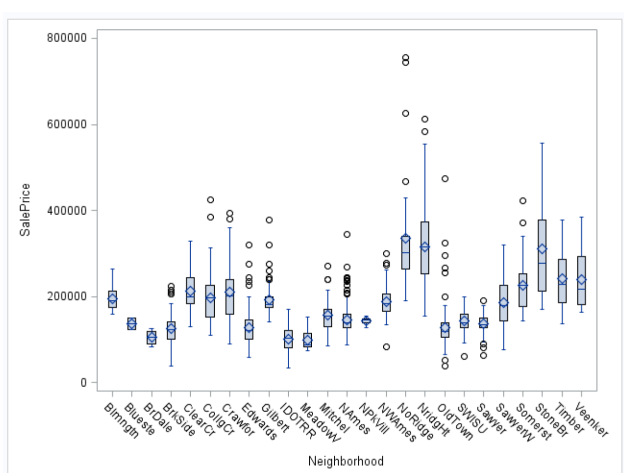
```
PROC SGPLOT;
      HISTOGRAM SalePrice / SCALE = COUNT;
      run;
```

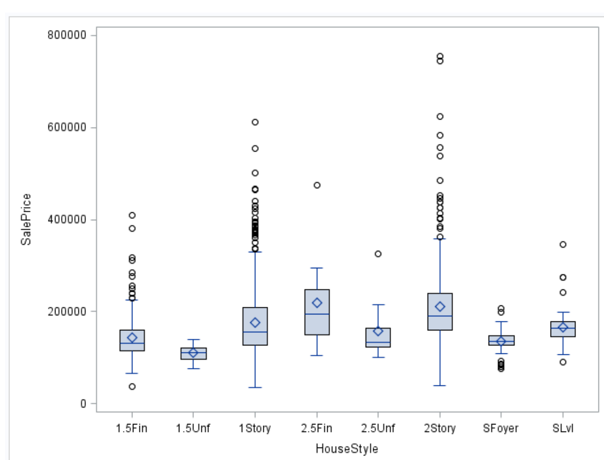



From the histogram, we can observe that most of the house prices are ranged around the $200000 mark, with very few exceeding $400000.

Next, we try to analyze the trend of the Sale Price with respect to the Neighborhood and the House Style.

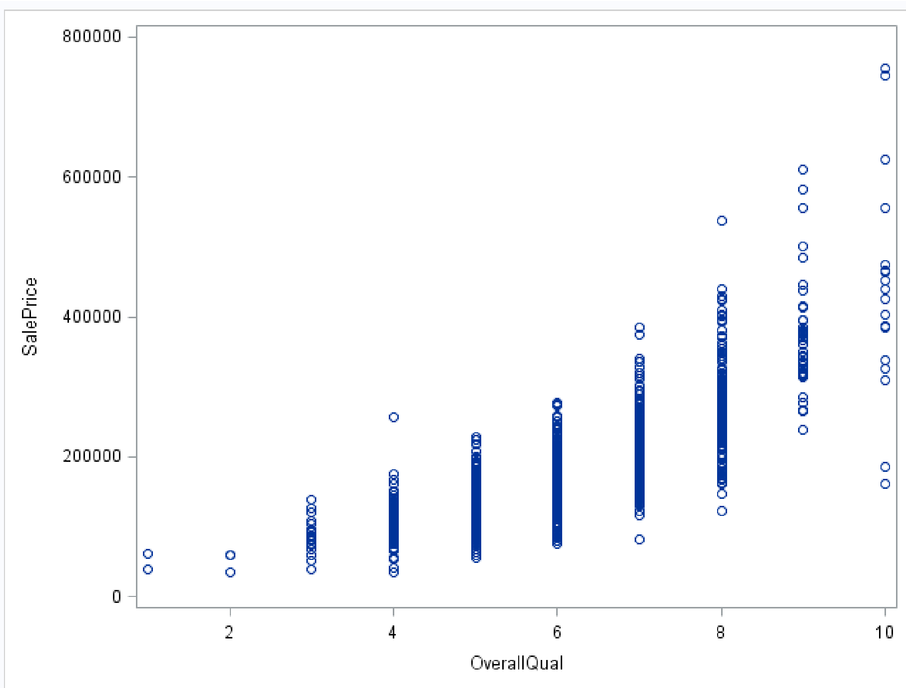
```
PROC SGPLOT;
	VBOX SalePrice / CATEGORY = Neighborhood;
	run;
PROC SGPLOT;
	VBOX SalePrice / CATEGORY = HouseStyle;
	run;
```



From the two box plots we observe that Northridge and Northridge Heights account for the highest price along with Stone Brook. As per the plot of HouseStyle and SalePrice we observe the range for the 1 Story houses, 2.5Fin: Two and one-half story: 2nd level finished and 2 story houses are almost the same, indication it is not much of an impact as compared to the Neighborhood.

Lastly, we plot the Sale Price vs Overall Quality of the house, 1 being the lowest and 10 being the best.




As expected the Houses with Overall Quality as 8, 9 or 10 have a high Sale Price. But key points to observe here is that, there are some points where having a Overall Quality of 10 results in low price than those with Overall Quality of 9. This indicates there are some other variables that resulted in a lower price for that house.

We now start to pick out variables to see if we can build a model to predict Sale Price. For that we start with all the continuous variables for now. We examine those variables using the 'corr' procedure and pick out the variables with low p-value.

```
proc corr data=train;
      var saleprice;
      with MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd
      MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
      _1stFlrSF _2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath
      FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd
      Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF _3SsnPorch
      ScreenPorch PoolArea MiscVal MoSold YrSold
      EnclosedPorch;
      run;
```

Of these correlations, only a handful have strong pearson correlation coefficients, where most are close to 0. Due to this we down select even further and end up selecting the following 5 Variables.

## The SAS System

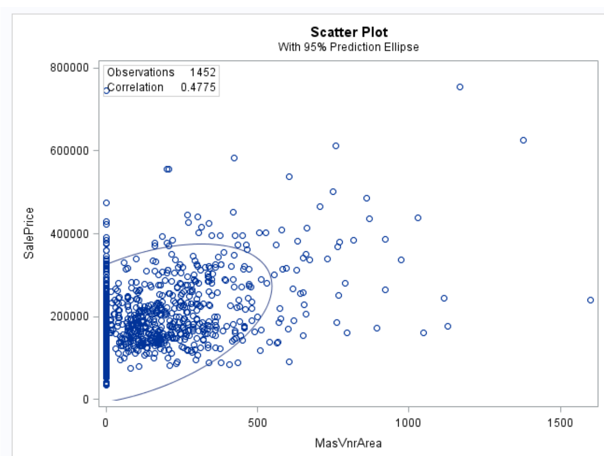
### The CORR Procedure

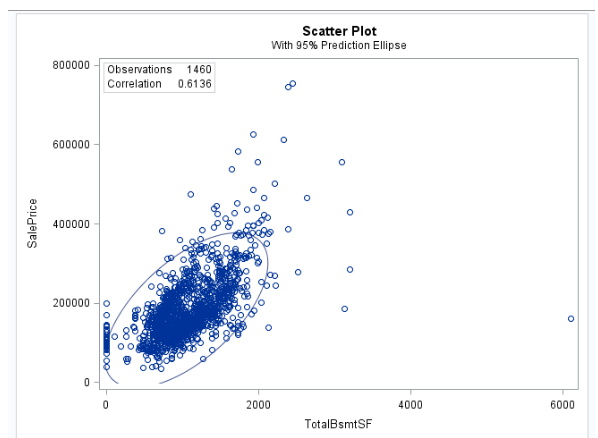
| | |
|---|---|
| **1 With Variables:** | SalePrice |
| **5 Variables:** | MasVnrArea TotalBsmtSF _1stFlrSF GrLivArea GarageArea |

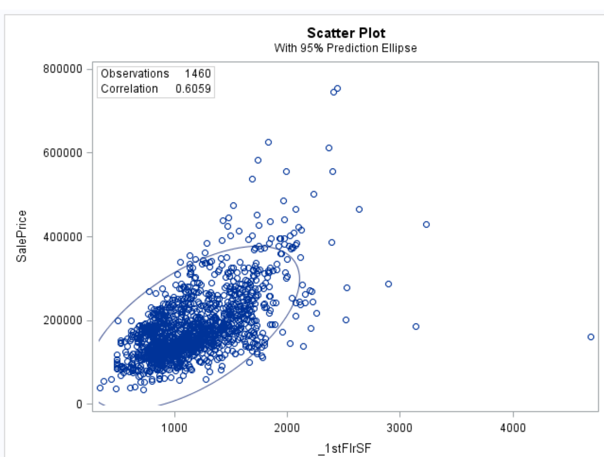
**Pearson Correlation Coefficients**
**Prob > |r| under H0: Rho=0**
**Number of Observations**

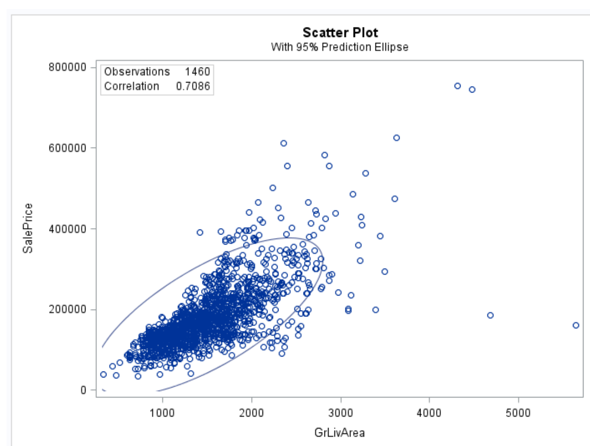
| | GrLivArea | GarageArea | TotalBsmtSF | _1stFlrSF | MasVnrArea |
|---|---|---|---|---|---|
| **SalePrice** | 0.70862<br><.0001<br>1460 | 0.62343<br><.0001<br>1460 | 0.61358<br><.0001<br>1460 | 0.60585<br><.0001<br>1460 | 0.47749<br><.0001<br>1452 |

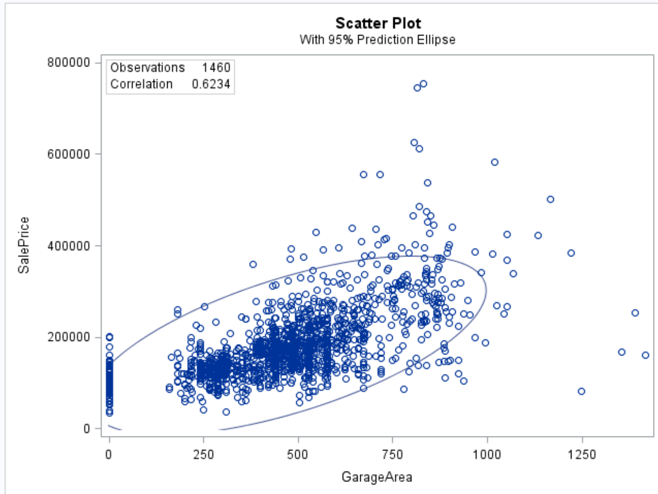
We provide graphs of the five variables from the corr process.






**BUILDING THE ACTUAL REGRESSION MODEL:**

We use the continuous variables we picked in the previous section and first choose to model MasVnrArea, which correlated approximately 0.47 with SalePrice. We will use this variable to build a simple linear regression model and comment on the model adequacy.

```
proc reg;
  model SalePrice = MasVnrArea;
run;
```

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 2.079646E12 | 2.079646E12 | 428.24 | <.0001 |
| Error | 1450 | 7.041626E12 | 4856293464 | | |
| Corrected Total | 1451 | 9.121272E12 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 69687 | R-Square | 0.2280 |
| Dependent Mean | 180615 | Adj R-Sq | 0.2275 |
| Coeff Var | 38.58322 | | |

**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 158936 | 2107.61360 | 75.41 | <.0001 |
| MasVnrArea | 1 | 209.08537 | 10.10372 | 20.69 | <.0001 |

Within the context of these variables, the model coefficients indicate that if MasVnrArea was 0 the SalePrice of the house would be $158936. We look into our data dictionary to find that MasVnrArea is described ambiguously as 'Masonry veneer area in square feet'. We find many observations in the data set where MasVnrArea is 0. Therefore, there are likely many observations in this data set that have a SalePrice and do not have a masonry veneer. This does make us feel quite poorly about the model as not all the houses would have the same SalePrice with MasVnrArea as 0.

Next, we pick the variable with a signification correlation coefficient to use in the model. We select the best variable depending on the R-square Value and examine the regression model.

```
proc reg;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF _1stFlrSF MasVnrArea
  BsmtFinSF1 BsmtUnfSF/
    selection=rsquare start=1 stop=1;
run;
```

| Number in Model | R-Square | Variables in Model |
|---|---|---|
| 1 | 0.5042 | GrLivArea |
| 1 | 0.3875 | GarageArea |
| 1 | 0.3757 | TotalBsmtSF |
| 1 | 0.3683 | _1stFlrSF |
| 1 | 0.2280 | MasVnrArea |
| 1 | 0.1474 | BsmtFinSF1 |
| 1 | 0.0465 | BsmtUnfSF |

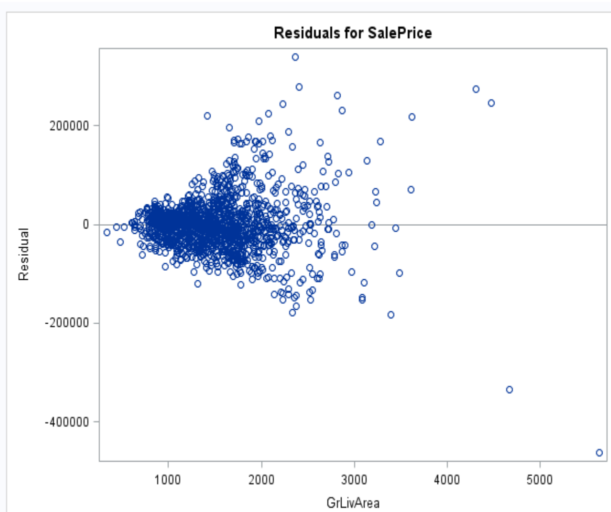
We select GrLivArea variable to examine its regression model as it has the highest R-square value among the various variables.

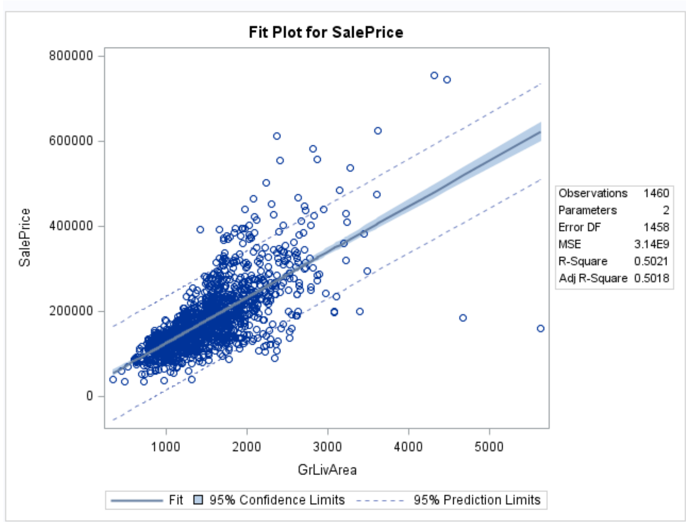
```
proc reg;
  model SalePrice = GrLivArea;
run;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 4.62374E12 | 4.62374E12 | 1470.59 | <.0001 |
| Error | 1458 | 4.584171E12 | 3144150265 | | |
| Corrected Total | 1459 | 9.207911E12 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 56073 | R-Square | 0.5021 |
| Dependent Mean | 180921 | Adj R-Sq | 0.5018 |
| Coeff Var | 30.99290 | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | 18569 | 4480.75455 | 4.14 | <.0001 |
| GrLivArea | 1 | 107.13036 | 2.79362 | 38.35 | <.0001 |






From the residual plot and fit plot we find some linear association with the observations. The residual plot has random points and no specific pattern. It can be said that there is a linear relationship between GrLivArea and SalePrice.

We now look at the categorical values and find the variable with the best correlation coefficient to analyze the regression model with that variable.

```
proc corr data=train nosimple rank plots=(scatter);
   var OverallQual GarageCars YearBuilt FullBath GarageYrBlt Fireplaces;
   with SalePrice;
   RUN;
```

**The SAS System**

**The CORR Procedure**

| | |
|---|---|
| **1 With Variables:** | SalePrice |
| **6 Variables:** | OverallQual GarageCars YearBuilt FullBath GarageYrBlt Fireplaces |

**Pearson Correlation Coefficients**
**Prob > |r| under H0: Rho=0**
**Number of Observations**

| | | | | | | |
|---|---|---|---|---|---|---|
| **SalePrice** | OverallQual | GarageCars | FullBath | YearBuilt | GarageYrBlt | Fireplaces |
| | 0.79098 | 0.64041 | 0.56066 | 0.52290 | 0.48636 | 0.46693 |
| | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 |
| | 1460 | 1460 | 1460 | 1460 | 1379 | 1460 |

Look at the results we select OverallQual as our variable to build the regression model with.

```
proc reg;
  model SalePrice = OverallQual;
run;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 5.760947E12 | 5.760947E12 | 2436.77 | <.0001 |
| Error | 1458 | 3.446964E12 | 2364172965 | | |
| Corrected Total | 1459 | 9.207911E12 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 48623 | R-Square | 0.6257 |
| Dependent Mean | 180921 | Adj R-Sq | 0.6254 |
| Coeff Var | 26.87511 | | |

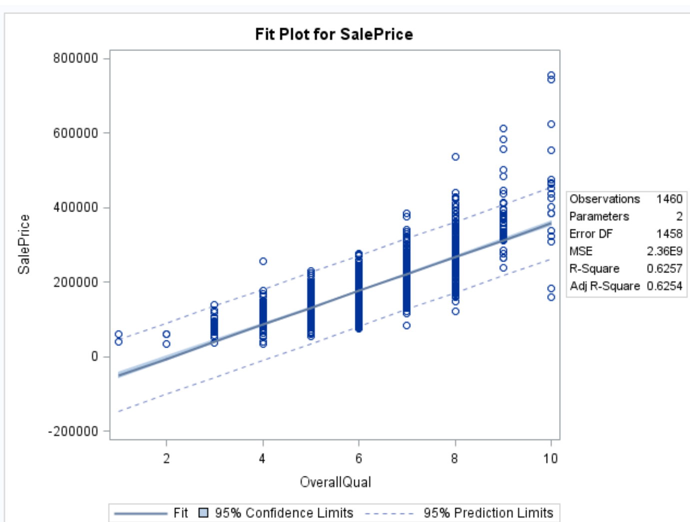
| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | -96206 | 5756.40739 | -16.71 | <.0001 |
| OverallQual | 1 | 45436 | 920.43025 | 49.36 | <.0001 |

As such, our fitted model is SalePrice = 45436 x OverallQual - 96206

Within the context of these variables, the model coefficients indicate that if OverallQal was 0 the SalePrice of the house would be $-96,206. As OverallQual is categorical, and as its a 10 way scale that begins with 1, it is not reasonable to think of OverallQual being 0.

OverallQual is a categorical variable, a one unit change will result in a much larger jump than the previous continuous variables. If there is a one unit change, our model tells us that the average change in the mean of SalePrice is about $45,436. Quite a large slope on this due to the categorical variable being between 1-10.

The model also has some goodness-of-fit information. We look at our R-Square to see that this regression model only explains ~62% of the variability in SalePrice using OverallQual. We will pay attention to the Adjusted R-Square as we continue to build models so that we can compare model performance with consideration to the size of the sample and number of variables are included in the model.



There appears to be a positive linear trend. This model looks quite linear with the largest amount of variability in observation of sale price coming from when a house is rated a 10. Intuitively we'd expect for there to be some high-priced outliers that, due to the survey characterization method, would have to be assigned a value of 10. There are some interesting outliers at 6, 9, and 10, with some even being low outliers at 10.

If we are to simply compare models based on R-Square values, this model explains the most amount of variability in SalePrice at 62%. Even though this is a categorical variable, we find this to be a highly associative variable.

**MULTI LINEAR REGRESSION MODELS:**

Model: GrLivArea, MasVnrArea predicts SalePrice

```
proc reg; /*MODEL 1*/
  model SalePrice = GrLivArea MasVnrArea;
run;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 5.029505E12 | 2.514753E12 | 890.54 | <.0001 |
| Error | 1449 | 4.091767E12 | 2823855693 | | |
| Corrected Total | 1451 | 9.121272E12 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 53140 | R-Square | 0.5514 |
| Dependent Mean | 180615 | Adj R-Sq | 0.5508 |
| Coeff Var | 29.42167 | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | 28796 | 4335.43879 | 6.64 | <.0001 |
| GrLivArea | 1 | 93.19394 | 2.88342 | 32.32 | <.0001 |
| MasVnrArea | 1 | 103.34373 | 8.37046 | 12.35 | <.0001 |

Our fitted model is {SalePrice} = 28796 + 93.19394 x {GrLivArea} + 103.34373 x {MasVnrArea}

Within the context of these variables, the model coefficients indicate that if GrLivArea was 0, and MasVnrArea was 0 the SalePrice of the house would be $28,796. The GrLivArea would never be 0, however it is possible that MasVnrArea could be 0. We look at our R-Square to see that this regression model only explains ~55% of the variability in SalePrice using GrLivArea and MasVnrArea.

Model: GrLivArea + MasVnrArea + OverallQual predicts SalePrice

```
proc reg;
  model SalePrice = GrLivArea MasVnrArea OverallQual;
run;
```

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 3 | 6.617752E12 | 2.205917E12 | 1275.87 | <.0001 |
| Error | 1448 | 2.50352E12 | 1728950283 | | |
| Corrected Total | 1451 | 9.121272E12 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 41581 | R-Square | 0.7255 |
| Dependent Mean | 180615 | Adj R-Sq | 0.7250 |
| Coeff Var | 23.02169 | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | -90630 | 5199.43722 | -17.43 | <.0001 |
| GrLivArea | 1 | 51.91836 | 2.63535 | 19.70 | <.0001 |
| MasVnrArea | 1 | 53.70922 | 6.75129 | 7.96 | <.0001 |
| OverallQual | 1 | 30702 | 1012.97762 | 30.31 | <.0001 |

Our fitted model is {SalePrice} = 51.91836 x {GrLivArea} + 53.70922 x {MasVnrArea} + 30702 x {OverallQual} – 90630.

The model has a R-Square value of 0.7255 meaning it can explain ~72% variation in sale price using GrLivArea, MasVnrArea and OverallQual.

It appears that from some perspectives adding more predictor variables, results in a better R-Square value.

We now use our first multi regression model to predict some values and compare it with the actual sale Price.

| SalePrice | predSales |
|---|---|
| 208500 | 208413.01 |
| 181500 | 146406.75 |
| 223500 | 211982.06 |
| 140000 | 188809.99 |
| 250000 | 269806.59 |
| 143000 | 155726.15 |
| 307000 | 205888.47 |

We see that from the first observation the prediction is very close. But there are some obversations where our prediction values are way off the real Sale Price, for example the last one, our predicted price is off by a $100000, which is bad.

**CONCLUSION:**

The next steps would be so see if there any other important categorical variables that could be used by us to begin a better attempt at modelling. Our initial assessment is that a categorical variable performs the best for explaining the variability of sale price. Still the maximum we could achieve was a ~72% of variability explanation.

Our future work is to explore more variables and try to design better models for predicting the Sale Price more precisely.