Noé Flores
Regression Model Building

Introduction:
In this analysis, our target goal is to build upon our exploratory visualization methods to identify possible predictor variables for the final sales price of homes in our Ames Housing data sample. The dataset contains 2,930 total observations and 82 different variables. Those variables are broken down into 23 categorical variables and 20 continuous variables. To build upon the exploratory analysis and identify possible predictor variables, we will build and compare linear regression models based on the variables available in our dataset in an effort to identify a model with a good fit for our dependent variable, final sale price.

Model#1 Results and Analysis:
In our Exploratory Data Analysis, we observe and compare different continuous variables with potential correlation to SalePrice.

Figure 1: Summary Correlation matrix to Sale Prices.

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

We begin by fitting a simple linear regression model using MasVnrArea as our predictor variable. The equation for the model we are seeking to build is the following:

$$\textbf{SalePrice} = \beta_0 + \beta_1 \textbf{(MasVnrArea)} + \varepsilon$$

The results of running the model produce the following parameter results:

Figure 2: Parameter estimates of Regression model #1 SalePrice with MasVnrArea.

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	157,303	14666.89502	107.24	<.0001
MasVnrArea	1	226.47763	7.1194	31.81	<.0001

Taking the results from our analysis, we can now state the complete fitted model as follows:

SalePrice = 157,303 + 226.47763 * MasVnrArea

Keeping in mind that the predicted value is SalePrice, the model coefficients in the previous equation indicate to us that if MasVnrArea was 0, the final SalePrice of the house would be $157,303.00, with MasVnrArea functioning as the predictor variable. A one unit change in MasVnrArea would result in a SalePrice mean difference of $226.48. Also of note, the variables in our model each have t-values that are greater than zero, and the P-values are also significant, coming in below the desired significance level of .05.

We move towards an examination of the goodness of fit information from our regression analysis on SalePrice with respect to MaVnrArea. The information in Figure.3 below provides the results of our Analysis of Variance (ANOVA).

Figure 3: Summary of Analysis of Variance in model #1 SalePrice with MasVnrArea

Source	Degrees of Freedom	Sum of Squares	Main Square	F Value	Pr > F
Model	1	4.781879E+12	4.781879E+12	1011.96	<.0001
Error	2905	1.372718E+13	4725361826		
Corrected Total	2906	1.850905E+13			

We can interpret the information above as statistically significant with respect to the model we ran. The F-test, or F-Value is a test statistic used to decide whether the model as a whole has statistically significant predictive capability. Generally speaking, we want this number to be high. The Pr > F is the P-value associated with the F statistic in our analysis. It is used in testing the null hypothesis of the model. In order for us to accept that our predictor variable (MasVnrArea) has some value in the prediction of SalePrice, we need a P-value less than .05.

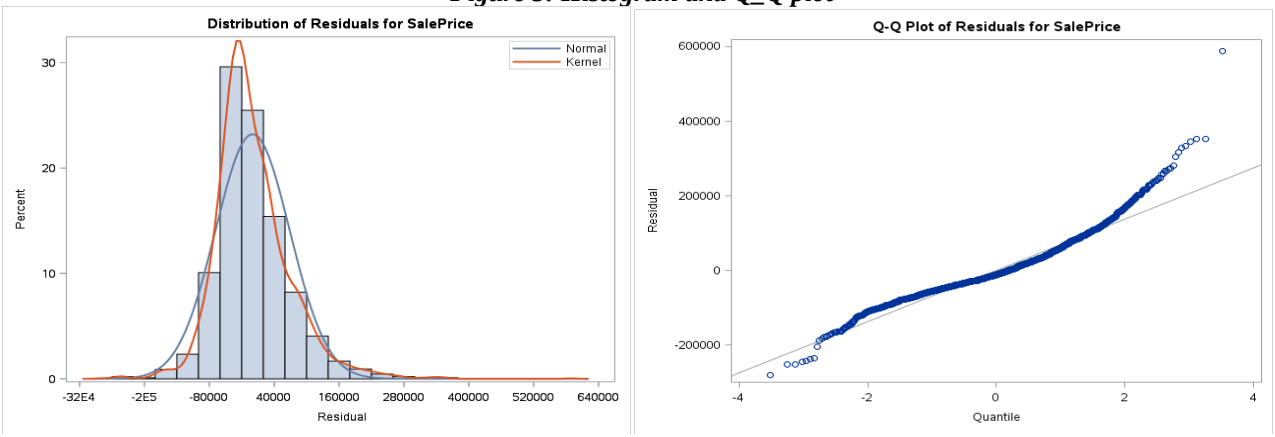
The R-Square in this regression model only explains 26% of the variability in SalePrice when using MasVnrArea as a predictor variable. This is displayed in Figure.4 below. We usually look for our models to have an R-Square in the range of 50%. Our adjusted R-Square also indicates how well our model fits, adjusting for the number of variables. We can see from the results that these numbers aren't quite what we want, but it definitely requires us to investigate the model further.

Figure 4: R-Square and Adjusted R-Square model #1 SalePrice with MasVnrArea

Root MSE	68741	**R-Square**	0.2584
Dependent Mean	180380	**Adj R-Sq**	0.2581
Coeff Var	38.10908		

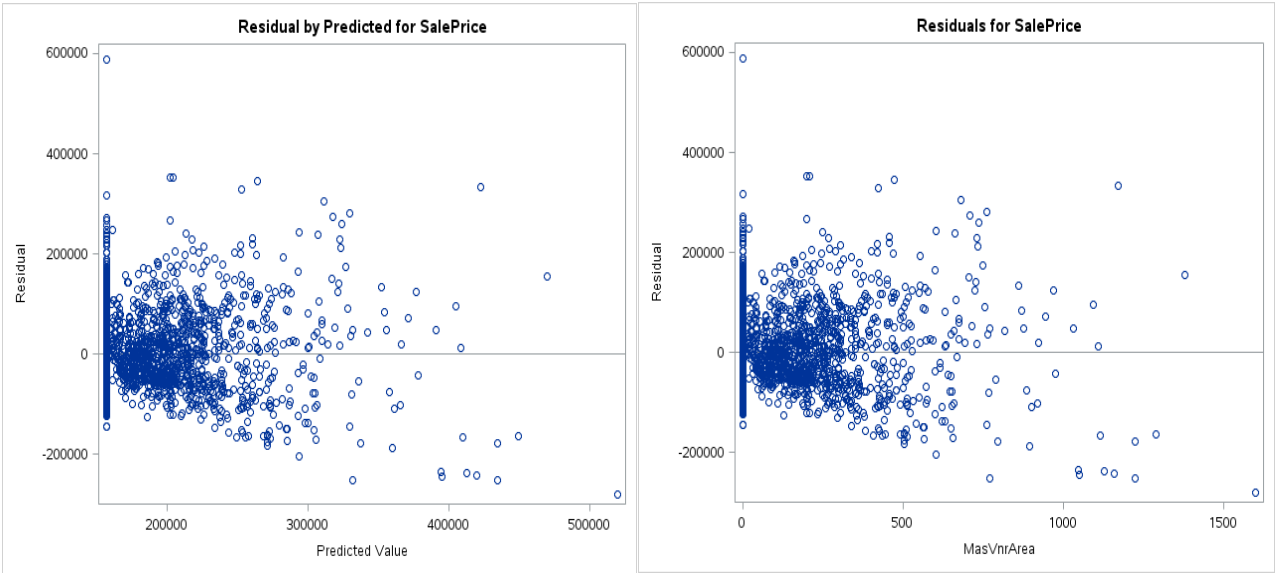
A primary assessment tool for a model's goodness of fit and overall quality can come from graphical diagnostics outputs of the model. In order to make an assessment of normality, we can use the histogram and Q_Q plots from our model. The histogram enables us to examine the distribution of the data. In this particular case, it looks like our data is mostly normal in its distribution. The Q_Q plots, however, paint a slightly different picture. If we had a normal distribution, we would expect our data to lie flat and on a straight line. We can see that there are visible tails in our data that call into question the normality of our distribution and there is also a noteworthy outlier at the top.

Figure 5: Histogram and Q_Q plot



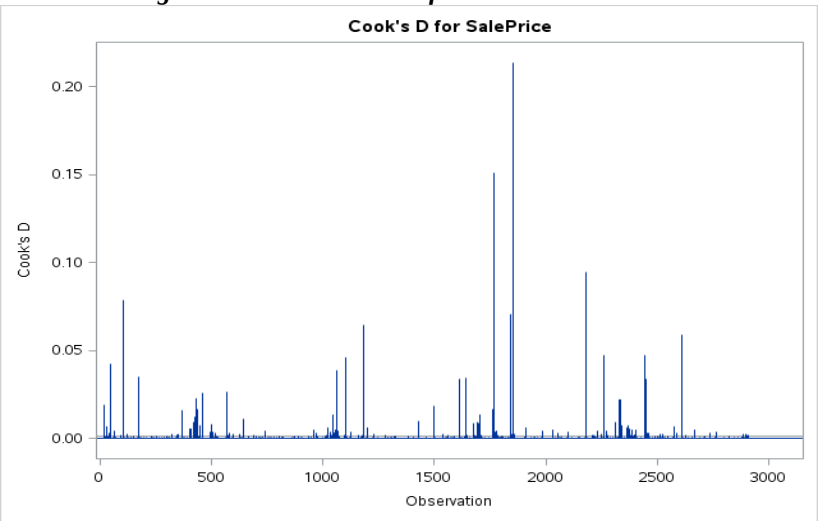
The plot of residuals versus the predicted, and residual versus the independent variable are looking for a random dispersion of data points. If the points in a residual plot are random, then we can assume that a linear regression model is a good fit. What we see here in Figure.6 is that there is a heavy concentration of observations in both the residual versus predicted value and the residual versus the independent variable plots at the 0 level, meaning MasVnrArea is equal to 0. There is also some bulking of data.

Figure 6: Residual vs Predicted Value and Residual vs Independent Variable



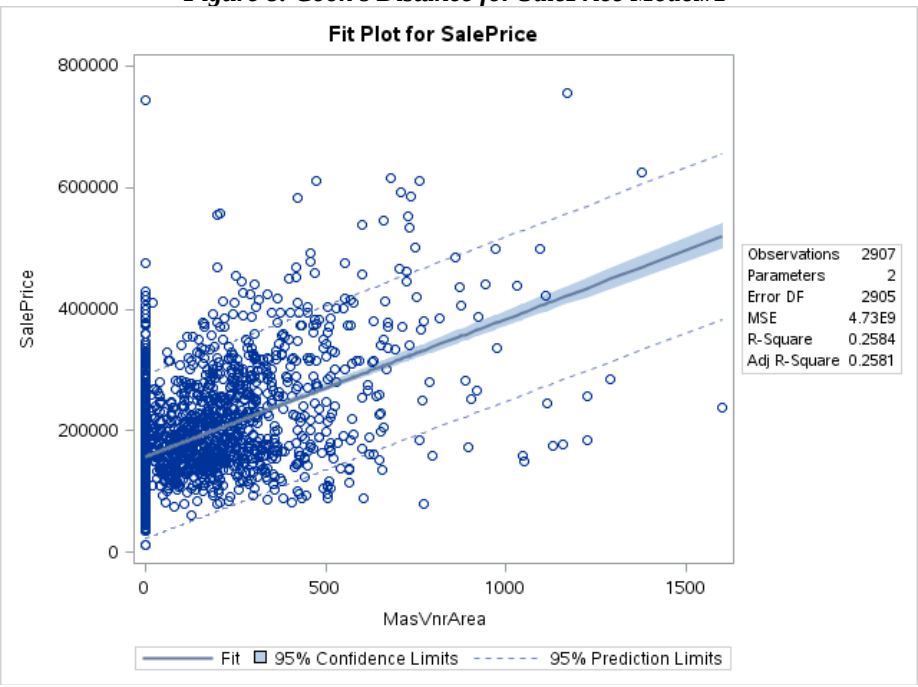
The Cook's Distance (Cook's D) is a measure of the influence of an observation on regression analysis. We typically want to see minimal spikes in this display. What we find in our review of this model is that there is spiking across the range, but only two points in particular in Figure.7 are troublesome, which merit some added investigation.

Figure 7: Cook's Distance for SalePrice Model#1



The Fit Plot below assists us in identifying a linear trend for our model. As you can see from the graphic in Figure.8, there is some evidence of a trend present here, but once again we see a significant amount of our observations fall at 0 MasVnr Area. We also recognize that most of our data points fall within the 95% confidence interval level depicted in the plot by the dotted lines. It is still evident that there are data points outside of those limits.

Figure 8: Cook's Distance for SalePrice Model#1



Summary for Model#1:
There are some positive observations to make with respect to this particular model. The P-Values and F-statistic indicate that there are some statistically significant indications that our predictor variable (MasVnrArea) is useful towards predicting SalePrice, but our adjusted R-Square is relatively low at 25%. The Q_Q plot and histogram indicate normality, but there is some notable diversion at the tails, and there are some noticeable spikes in our Cook's D plot. The residuals and fit plan appear decent with respect to what they are presenting, but there are many observations of MasVnrArea at zero. Overall all there are some concerns with this model.

Model#2 Analysis and Results:
In our second model, we use the results from a SAS regression analysis to select the variable with the highest R-Square. The top five results from that analysis are displayed below.

Figure 9: Top 5 variables according to R-Square.

Variable in Model	R-Square
OverallQual	0.6463
GrLivArea	0.5100
GarageCars	0.4373
GarageArea	0.4197
TotalBsmtSF	0.4169

Our intent with this model was to use the predictor variable in our dataset with the highest value of R-Square, which is OverallQual or Overall Quality. We will not use the OverallQual variable for our second model. The reason for this is that OverallQual is ordinal in nature, meaning that it draws its value from ranking. While the variable is nominal in nature, there is a minimum and maximum ceiling. Instead, we will use the second-ranked variable, GrLivArea in this model, which is a real continuous variable.

Once again, our analysis begins with fitting a simple linear regression model using GrLivArea as our predictor variable.

$$\textbf{SalePrice} = \beta_0 + \beta_1 (\textbf{GrLivArea}) + \varepsilon$$

The results of running the model produce the following parameter results:

Figure 10: Parameter estimates of Regression model #2 SalePrice with GrLIvArea.

Variable	Degrees of Freedom	Paramter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13,290	3269.70277	4.06	<.0001
GrLivArea	1	111.694	2.06607	54.06	<.0001

The complete fitted equation of the model is the following:

SalePrice = 13,290 + 111.694 * GrLivArea

Once again, keeping in mind that the desired predicted value is SalePrice, the model coefficients in the equation indicate to us that if GrLivArea was 0, the final SalePrice of the house would be $13,290, and a one unit change in GrLivArea would result in a SalePrice mean increase or decrease of $111.69. Our t-values are greater than zero, although less so than in our first model, but P-values remain significantly below the benchmark .05 level.

Our goodness of fit information from our regression analysis on SalePrice with respect to GrLivArea is below in Figure.11 and stems from the results of our Analysis of Variance (ANOVA).

Figure 11: Summary of Analysis of Variance in Model #2 SalePrice with GrLivArea

Source	Degrees of Freedom	Sum of Squares	Main Square	F Value	Pr > F
Model	1	9.33763E+12	9.33763E+12	2922.59	<.0001
Error	2905	9.35491E+13	3.19498E+09		
Corrected Total	2906	1.86925E+13			

The information is statistically significant. The F-Value is significantly higher than our first model and the P-value remains low. According to these results, it is normal to accept that our predictor variable (GrLivArea) adds value in the prediction of SalePrice.

The R-Square in this second model displayed in Figure.12, explains nearly 50% of the variability in SalePrice when using GrLivArea as the predictor variable. We are typically looking for our models to have an R-Square in the 50% range. The adjusted R-Square is also in line with our desired results.

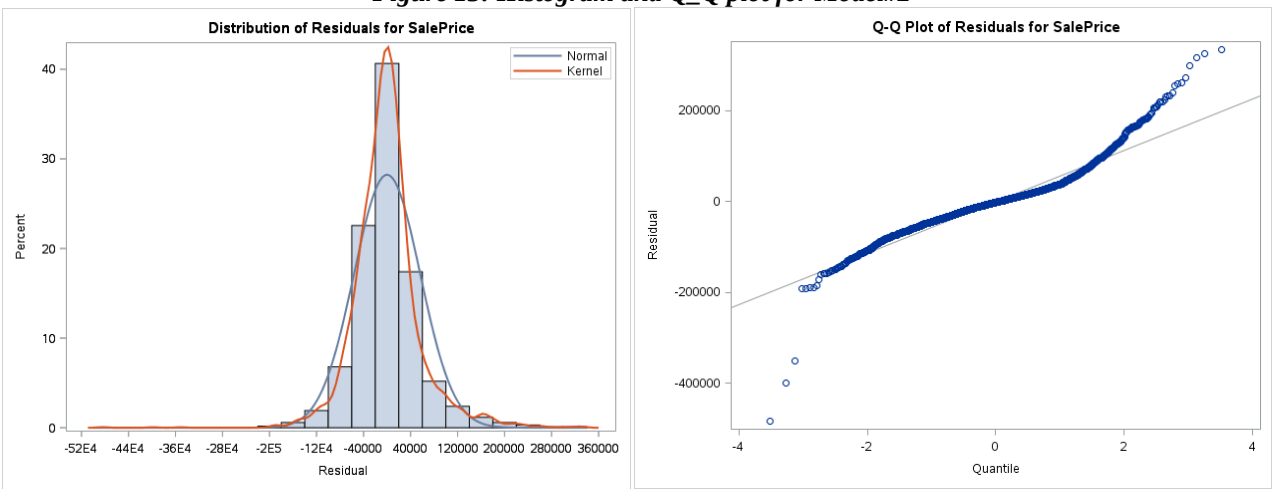
Figure 12: R-Square and Adjusted R-Square Model #2 SalePrice with GrLivrArea

Root MSE	56524	**R-Square**	0.4995
Dependent Mean	180796	**Adj R-Sq**	0.4994
Coeff Var	31.26405		

At this point in our analysis of the second model, it would appear that our numerical indicators are more favorable than in the first model. We can move into our assessment of the statistical graphics hopeful that they can corroborate the initial findings.

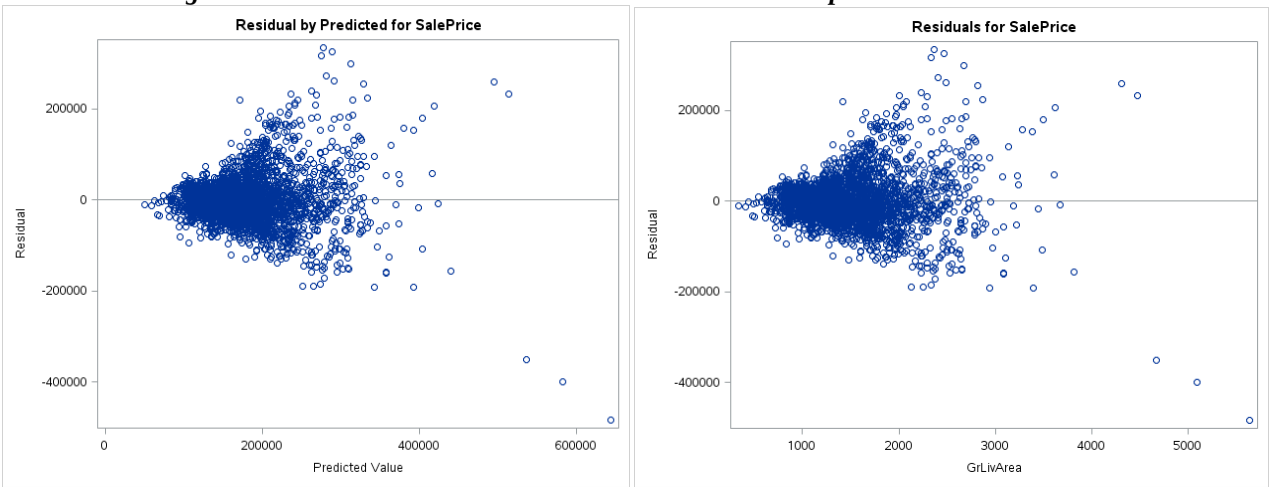
Our normality assessment for this model will once again come from our observations of the histogram and Q_Q plots. In this particular model, the histogram is painting a distribution that would appear to be slightly less normal than our first model, although not by much. The evidence is in the slight increase in tailing. The Q_Q plots appear to agree with our previous statement. Instead of having our data points fall on a straight line, we see clearly visible tailing, calling into question the normality of our distribution. There are also more outliers present in this Q_Q plot.

Figure 13: Histogram and Q_Q plot for Model#2



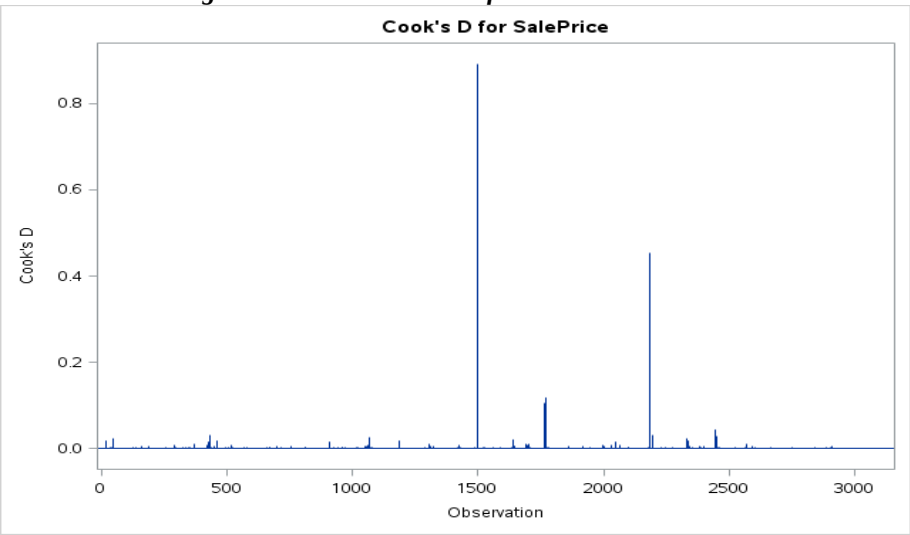
Our plots of residual versus the predicted, and residual versus the independent variable, are looking for a random spread of data points. If the spread of data points is random, we assume that a linear regression model is a good fit. What we see here in Figure.14 is that there is a heavy concentration of observations in both plots. There doesn't appear to be any one significant data point on the plots to draw our attention as we saw in the first model, but there is cause for concern given the cluster of data we see.

Figure 14: Residual vs Predicted Value and Residual vs Independent Variable Model#2



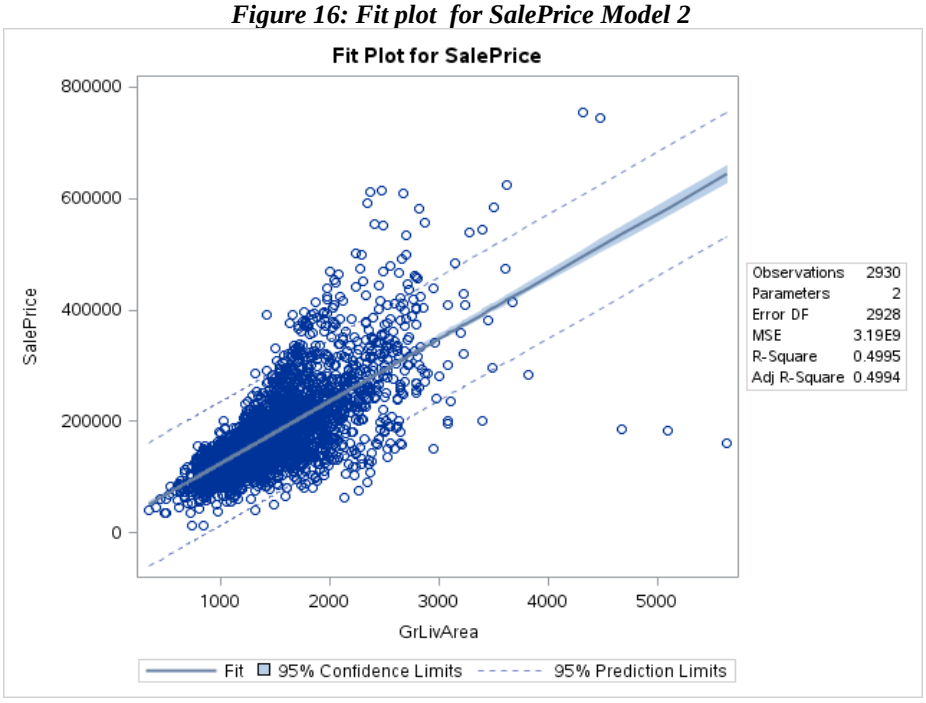
Our Cook's Distance (Cook's D) plot in Figure.15 for our second model holds some promise. There are two points on the plot that appear to bear significant influence on our regression analysis, but overall it looks much steadier than our first model. We want to see as few spikes in this display as possible, and besides the two significant departures I mentioned, it looks relatively smooth.

Figure 15: Cook's Distance for SalePrice Model 2



The Fit Plot below is indicating stronger evidence of a linear trend present in this model. Most of our data points in this model fall within the 95% confidence interval lines depicted by the dotted lines. While there are still data points outside of those limits, this plot represents a better fitting model when compared to our first.

Figure 16: Fit plot for SalePrice Model 2



Summary for Model#2:
Model 2 appears to have a better fit as a predictor for SalePrice. The P-Values and F-statistic are once again indicative that there is some statistical significance that our predictor variable (GrLivArea) is useful towards predicting SalePrice. Our adjusted R-Square is much better, coming in at nearly 50% as opposed to 25% in model 1. The histogram indicates normality, but there is some notable diversion at the tails in our Q_Q plot in this model, more so than our first model. By contrast, our Cook's D output is much smoother. The residuals show more bulking of data points than we would like, but the fit plot is much better in model 2. Overall, there appears to be a more positive linear trend in model 2.

Model#3 Results and Analysis:
For our third simple linear regression model, we are going to shift focus from continuous variables and moved towards fitting a model for SalePrice using one of the categorical variables in our data set. We will be choosing from the following variables: MoSold, GarageCars, FirePlaces, and BedroomAbvGr. In order to make a decision, I will run a simple correlation matrix against SalePrice, and choose the variable with the highest correlation. The results are below.

Figure 17: Top 5 variables according to R-Square.

Variable	GarageCars	BedroomAbvGr	Fireplaces	MoSold
Correlation to Sales Price	0.648	0.144	0.475	0.035
P- Value	<.0001	<.0001	<.0001	0.0563
Observations	2929	2930	2930	2930

We move forward and build our simple linear regression for SalePrice using GarageCars fitting the simple linear regression model using GarageCars as our predictor variable.

$$\textbf{SalePrice} = \beta_0 + \beta_1 (\textbf{GarageCars}) + \varepsilon$$

The results of running the model produce the following parameter results:

Figure 18: Parameter estimates of Regression model #3 SalePrice with GarageCars.

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	60,556	2845.07928	21.28	<.0001
GarageCars	1	68,060	1479.10676	46.01	<.0001

The entire fitted equation of the model is the following:

SalePrice = 60,556 + 68,060 * GarageCars

Our desired predicted variable is once again SalePrice. The model coefficients in the equation indicate to us that if GarageCars was 0, the final SalePrice of the house would be $60,556, and a one unit change in GarageCars would result in a SalePrice mean increase or decrease of $68,060. Our t-values that are greater than zero and our P-values remain below the benchmark of .05.

Our goodness of fit information from our regression analysis on SalePrice with respect to GarageCars is below in Figure.19 and stems the Analysis of Variance (ANOVA).

Figure 19: Summary of Analysis of Variance in Model #3 SalePrice with GarageCars

Source	Degrees of Freedom	Sum of Squares	Main Square	F Value	Pr > F
Model	1	7.845707E+12	7.845707E+12	2117.33	<.0001
Error	2927	1.08459E+13	3705478914		
Corrected Total	2928	1.86916E+13			

The information remains statistically significant. The F-Value is higher than Model.1, but falls short of the results in Model.2. The P-value, however, remains low and significant. According to these results, we can once again accept that our predictor variable (GarageCars) has possible value in the prediction of SalePrice.

The R-Square in Model.3 displayed in Figure.20 below, explains nearly 42% of the variability in SalePrice when using GrLivArea as the predictor variable. As we previously stated, we are looking for our models to have an R-Square in the 50% range. The R-Square in this model falls short of the metric but hits higher than our first model. The adjusted R-Square shares similar characteristics.

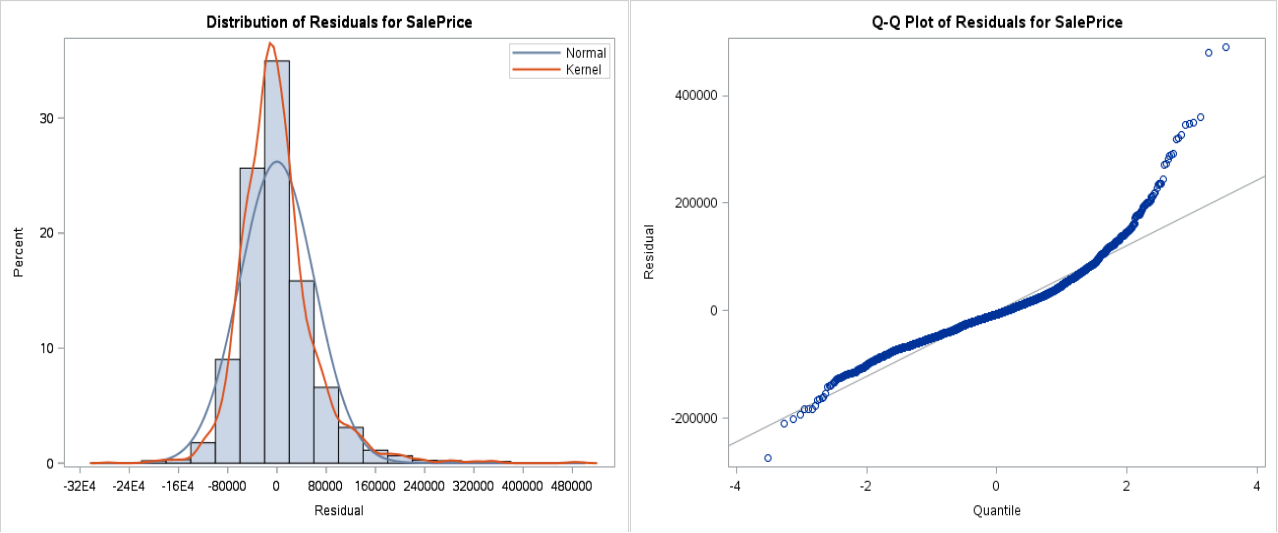
Figure 20: R-Square and Adjusted R-Square Model #3 SalePrice with GarageCars

Root MSE	60873	**R-Square**	0.4197
Dependent Mean	180806	**Adj R-Sq**	0.4195
Coeff Var	33.66733		

At this point in our analysis of the third simple linear regression model, our numerical indicators are falling into the middle ground with respect to our enthusiasm of the models predictive potential. The assessment of the statistical graphics will be instrumental as a deciding factor for the model's fit and usefulness.

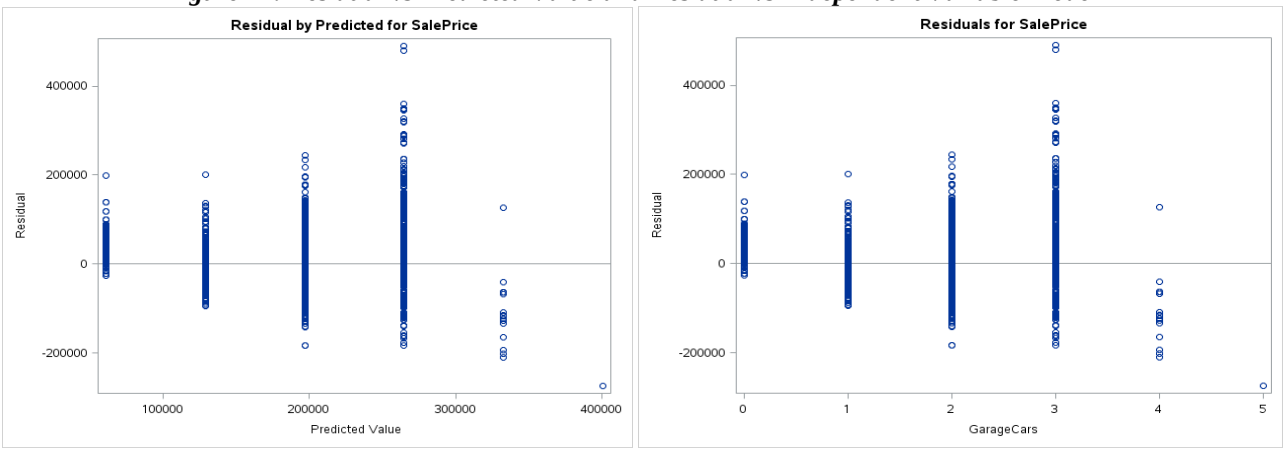
Normality assessments from our histogram and Q_Q plots in model appear to fall in line with the previous three. The histogram is painting a distribution that would seem to be mostly normal, although there is some evident skewness. The Q_Q plots however really highlight the skewing in the histogram. There is some heavy tailing in the Q_Q plot for this particular model, especially toward the top of the graph. We also see the presence of two outliers at the top right of the graph.

Figure 21: Histogram and Q_Q plot for Model#3



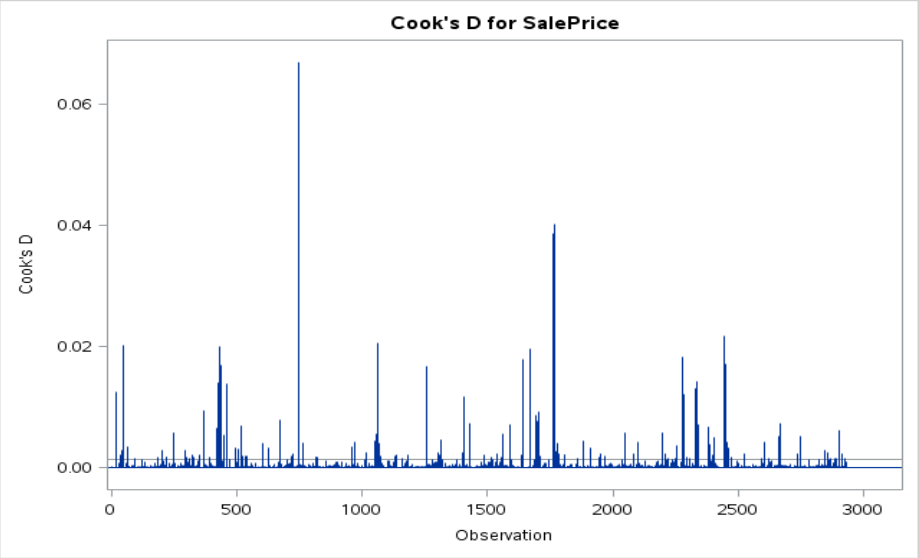
Our plots of residuals versus the predicted and residual versus the independent variable in Figure.22 present a bit of change from what we have dealt with in the first two models. We are now dealing with a categorical variable, and as is plainly evident from our graphs, it would appear that there is a pattern. This is not unusual. When dealing with categorical variables, we want to examine the scatter present within the vertical lines. What we see in the figure below is that there is a fair amount of scattering in those vertical lines corresponding to GarageCars, but there is some bulking as well.

Figure 22: Residual vs Predicted Value and Residual vs Independent Variable Model 2



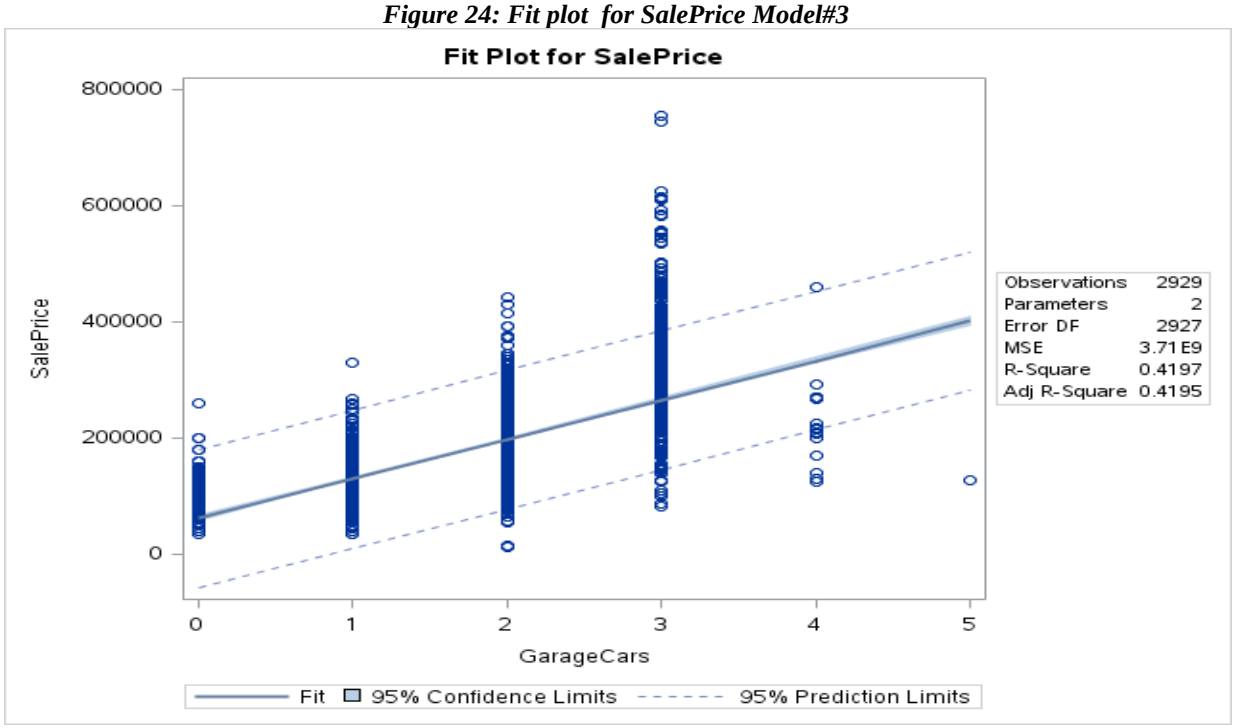
The Cook's Distance (Cook's D) plot in Figure.23 for this third model is actually fairly promising, despite the visual it presents. There are two points on the plot that appear to bear some influence on our regression analysis, but overall the spiking present in this model is relatively stable.

Figure 23: Cook's Distance for SalePrice Model#3



The Fit Plot in Figure.24 is again a bit different from what we have seen in the first two models. We have the 6 different possibilities for the number of GarageCars in our data set and setting aside the pattern those 6 different levels appear to represent; we do see a linear trend. As the number of GargeCars increase, so does the independent variable SalePrice. A good portion of the data falls within the 95% prediction limits, but there is a peculiarity in this display stemming from the last point corresponding to 5 car garages.

Figure 24: Fit plot for SalePrice Model#3



Since we are working with categorical data, we will use the predictive model to make one last fit assessment by finding out if our predicted model actually goes through the mean value of SalePrice for each category group. The data would suggest that in some cases the predicted model comes close to the means of the GarageCars but misses the mark in several instances. This isn't necessarily a problem, given that our model appears to fall close to the mark with the means when we have substantial observations.

Figure 25: Predicted Model vs Means for GarageCars

Garage Cars Mean		**Predicted Model Price**
Garage Cars = 0		**SalePrice = 60,556 + 68,060 * 0**
Obs	**Mean Price**	
157	$104,949.25	$ 60,556.00
Garage Cars = 1		**SalePrice = 60,556 + 68,060 * 1**
Obs	**Mean**	
778	$127,267.42	$ 128,616.00
Garage Cars = 2		**SalePrice = 60,556 + 68,060 * 2**
Obs	**Mean**	
1603	$183,562.10	$ 196,676.00
Garage Cars = 3		**SalePrice = 60,556 + 68,060 * 3**
Obs	**Mean**	
374	$310,304.62	$ 264,736.00
Garage Cars = 4		**SalePrice = 60,556 + 68,060 * 4**
Obs	**Mean**	
16	$228,748.69	$ 332,796.00
Garage Cars = 5		**SalePrice = 60,556 + 68,060 * 5**
Obs	**Mean**	
1	$126,500.00	$ 400,856.00

Summary for Model#3:

Model 3 appears to be a decent predictor for SalePrice. The P-Values and F-statistic indicate that there is statistical significance that our predictor variable (GarageCars) is useful towards predicting SalePrice. Our adjusted R-Square is decent, coming in at 42%, but short of where we want it. The histogram indicates some normality, but there is heavy tailing in our Q_Q plot. The Cook's D output is relatively smooth, with the exception of two points and the residuals show variance across the vertical markers. The fit plot is perhaps the best indicator, presenting a clear linear trend.

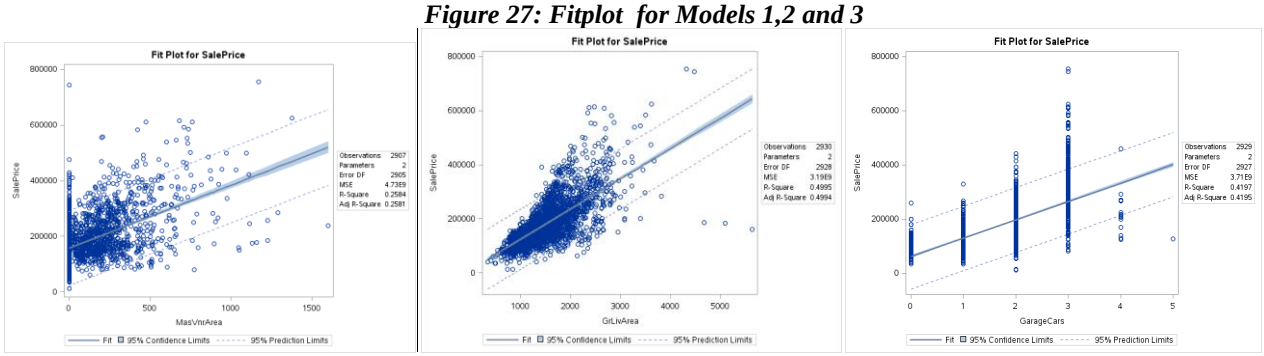
Summary analysis of fit for Simple Linear Regression Models:
Having gone through each model in detail, we will now sum up which one we believe provides a better fit and predictive ability for our independent variable, SalePrice. Figure.26 has a breakdown of some of the essential summary statistics for our models.

Figure 26: Summary for Models 1,2 and 3

Models	R-Square	F-Statistic	Pr > F
(1) SalePrice = 157,303 + 226.47763 * MasVnrArea	0.2584	1011.96	<.0001
(2) SalePrice = 13,290 + 111.694 * GrLivArea	0.4995	2922.59	<.0001
(3) SalePrice = 60,556 + 68,060 * GarageCars	0.4197	2117.3	<.0001

In the display below you will see the graphical output of the fit plot for each model in respective order from model 1, through model 3.

Figure 27: Fitplot for Models 1,2 and 3



Given the information presented in the summary of statistics for each model, and taking into account not only the Fit Plots we see above in Figure.27, but all the graphical information tools we analyzed, we can determine that Model.2 is likely to be the best fit for determining the future value of SalePrice. Each model has noticeable strengths and weaknesses. Model.2 which incorporates GrLivArea as a predictor for SalePrice appears to present the best combination of attributes for our desire to find the best fitting model.

Multiple Linear Regression Models:
We previously constructed and examined the benefits of simple linear regression models to predict SalePrice in our data set. We will now move forward and incorporate a multiple linear regression model into our analysis.

MLR Model#1 Results and Analysis:
In our first model, we use the continuous variables from our simple linear regression.

Figure 28: Continuous Variable R-Square and Correlation to SalePrice.

Variable	Correlation	R-Square
GrLivArea	.707	0.4995
MasVnrArea	0.508	0.2584

We now begin our analysis by fitting a multiple linear regression model using MasVnrArea and GrLivArea as our predictor variables for SalePrice.

$$\textbf{SalePrice} = \beta_0 + \beta_1\,(\textbf{MasVnrArea}) + \beta_2\,(\textbf{GrLivArea}) + \varepsilon$$

The results of running the model produce the following parameter results:

Figure 29: Parameter estimates of MLR Model#1.

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	26547	3141.93715	8.45	<.0001
GrLivArea	**1**	94.60302	2.12104	44.6	<.0001
MasVnrArea	**1**	118.54695	5.9955	19.77	<.0001

The complete fitted equation of the model is the following:

SalePrice = 26,547 + 94.60 * GrLivArea + 118.55 * MasVnrArea

Our desired predicted variable is once again SalePrice. The model coefficients in the equation indicate to us that if GrLivArea was 0, and MasVnrArea was 0, the final SalePrice of the house would be $26,547. A one unit change in GrLivArea and MasVnrArea would result in a SalePrice mean increase or decrease of $213.15 ($94.60+118.55). Our t-values are greater than zero and our P-values remain below the benchmark of .05. We notice that the coefficients have more impact relative to the total price of the home, meaning that in the instances where our predictor variables could be 0, our final SalePrice is significantly depleted.

Our goodness of fit information from our Multiple Linear Regression analysis (MLR), on SalePrice with respect to GrLivArea and MasVnrArea is below and stems from our analysis of variance.

Figure 30: Summary of Analysis of Variance on MLR Model #1.

Source	Degrees of Freedom	Sum of Squares	Main Square	F Value	Pr > F
Model	2	1.036256E+13	5.181279E+12	1846.98	<.0001
Error	2904	8.146497E+12	2805267561		
Corrected Total	2906	1.850905E+13			

The information is statistically significant. The F-Value is high and the P-value remains low and significant. According to these results, we can accept that our predictor variables of GrLivArea and MasVnrArea have value in the prediction of SalePrice.

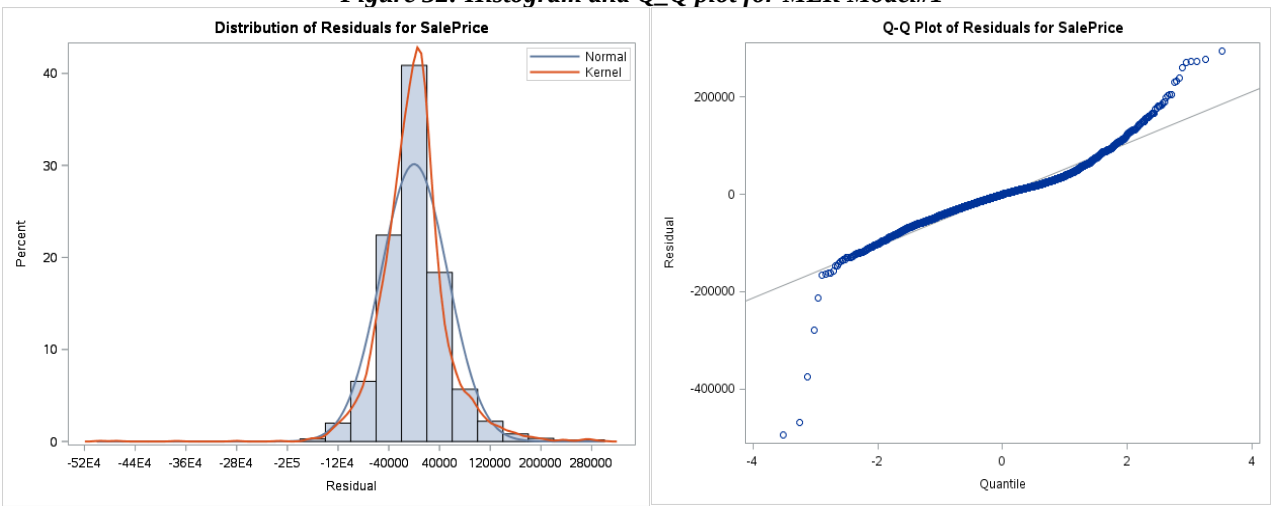
The R-Square in our MLR displayed in Figure.31 explains nearly 56% of the variability in SalePrice comes from using GrLivArea and MasVnrArea as our predictor variables. This is the highest R-Square number we have seen yet. As a reminder, we are looking for an R-Square that is at or above the 50% percent mark. The adjusted R-Square shares similar characteristics.

Figure 31: R-Square and Adjusted R-Square MLR Model #1

Root MSE	52965	**R-Square**	0.5599
Dependent Mean	180380	**Adj R-Sq**	0.5596
Coeff Var	29.36284		

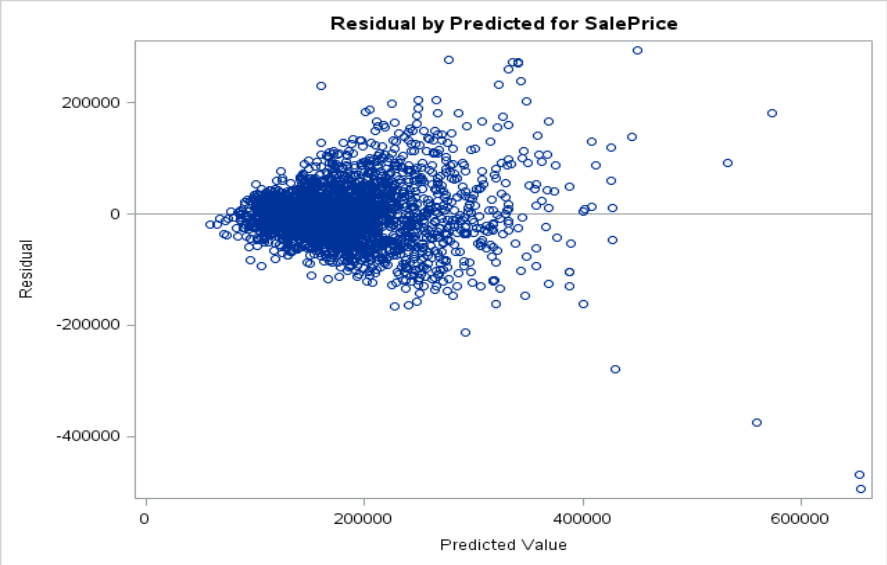
Normality assessments from our histogram in the MLR model appear to fall in line with a normal distribution. We can see some skewness, but overall the fit is good. The Q_Q plots, however, paint a different picture. There is heavy tailing on the right side, and we have a lot of outliers present at the bottom of the graph. While it appears that the bulk of the data is in line, we cannot ignore the tails in this case. Two conflicting displays of normality for our MLR model.

Figure 32: Histogram and Q_Q plot for MLR Model#1



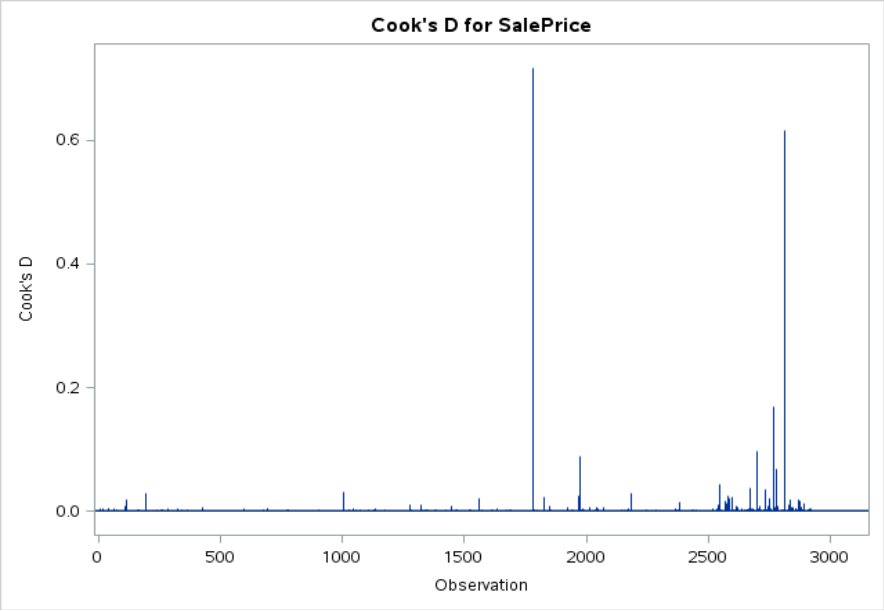
Our plot of the residuals versus the predicted, look for a random spread of data points. If the spread of data points is random, we assume that a linear regression model is a good fit. This is true in MLR as well. What we see here in Figure.33 is that there is a heavy concentration of observations in the plot. There are some significant data points here that draw our attention. If we examine the lower right-hand side of the graph, we see two distinct outliers. We also see a trail of outlying data points seemingly leading to the outliers in the corner of our graph.

Figure 33: Residual vs Predicted Value for MLR Model#1



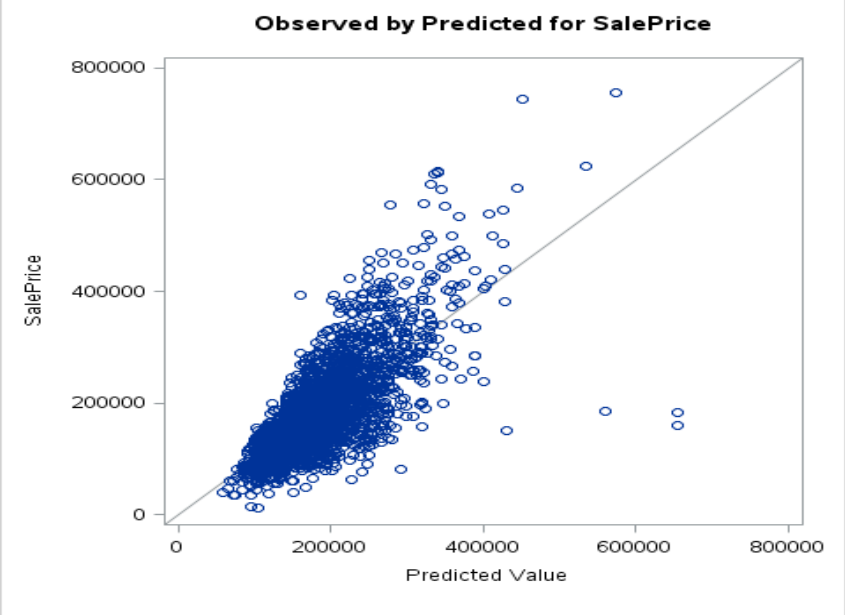
The Cook's Distance (Cook's D) plot in Figure.34 for this MLR model is actually quite promising. There is some spiking present, but really only two major spikes that are troublesome. These points likely have some influence on our regression analysis and likely coincide with the outliers we have seen in the previous displays.

Figure 34: Cook's Distance for SalePrice MLR Model #1



The actual versus predicted value plot in Figure.35 provides one more input for our analysis of the model's fit. We are looking to see if our data fall close to the diagonal line, meaning that we should see most of our points close to the regression line in the middle of the plot. We can also visualize the plot as wanting to see similar data points for Predicted values and the Actual SalePrice. As you can see in Figure.35, most of the points appear to be collecting around the lower end of the graph, but close to the regression line we mentioned.

Figure 35: Actual versus Predicted for SalePrice MLR Model#1



Summary for MLR Model#1:

Our first multiple linear regression model (MLR), has some difficulties. The P-Values and F-statistic indicate that there is statistical significance that our predictor variables are in fact useful towards predicting SalePrice. Our R-Square is good, coming in at 56%, but our Q_Q plot has some troubling evidence. The heavy tails in the Q_Qplot are difficult to overlook, and we see the presence of more outliers than what we previously observed in our simple linear regression models. The Cook's D output is decent, with the exception of two points. If we compare this MLR to some of our previous simple linear regression models, it is clear some of these models are better suited for fit, particularly GrLivArea as a predictor for SalePrice. The combination of R-Square and lack of outliers in this particular simple linear regression model allows us to make this distinction.

Multiple Linear Regression Model#2 Results and Analysis:
Our final Multiple Linear Regression Model will incorporate the previous two variables, GrLivArea and MasVnrArea, as well as one final variable. The final variable we will include is BsmtUnfSf(Basement-Unfinished Square Footage). Our reasons for incorporating this variable stems from the fact that it has the smallest positive correlation to SalePrice, its P-Value is significant, and it is continuous.

Figure 36: BsmtUnfSf correlation to SalePrice.

Variable	BsmtUnfSF
Correlation to Sales Price	0.183
P- Value	<.0001
Observations	2929

We now begin our analysis by fitting a multiple linear regression model using MasVnrArea, GrLivArea, and BsmtUnfSf as our predictor variables for SalePrice.

$$\textbf{SalePrice} = \beta_0 + \beta_1\,(\textbf{MasVnrArea}) + \beta_2\,(\textbf{GrLivArea}) + \beta_3\,(\textbf{BsmtUnfSF}) + \varepsilon$$

The results of running the model produce the following parameter results:

Figure 37: Parameter estimates of MLR model #.

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	25788	3195.16634	8.07	<.0001
GrLivArea	1	93.91175	2.17379	43.2	<.0001
MasVnrArea	1	118.5739	5.9952	19.78	<.0001
BsmtUnfSF	1	3.23138	2.30311	1.4	0.1607

The complete fitted equation of the model is the following:

SalePrice = 25,788 + 93.91 * GrLivArea + 118.57 * MasVnrArea+ 3.23 * BsmtUnSF

Our desired predicted variable is once again SalePrice. The model coefficients in the equation indicate to us that if GrLivArea was 0, MasVnrArea was 0, and BsmtUnSF was 0, the final SalePrice of the house would be $25,788. A one unit change in GrLivArea, MasVnrArea, and BsmtUnSF would result in a SalePrice mean change of $215.71 ($93.91+118.57+3.23). Our t-values that are greater than zero for all three variables. The P-values remain below the benchmark of .05 for GrLivArea and MasVnrArea, but BsmtUnSF has a noticeably high P-Value of .16. This means we have to reject the hypothesis that BsmtUnSF is predictive in SalePrice.

Our goodness of fit information from our second Multiple Linear Regression analysis (MLR), on SalePrice with respect to GrLivArea, MasVnrArea, and BsmtUnfSF is below and stems from our analysis of variance.

Figure 38: Summary of Analysis of Variance on MLR Model#2.

Source	Degrees of Freedom	Sum of Squares	Main Square	F Value	Pr > F
Model	3	1.035884E+13	3.452948E+12	1231.02	<.0001
Error	2902	8.139930E+12	2.804938E+09		
Corrected Total	2905	1.849877E+13			

The information is statistically significant. The F-Value is high and the P-value is low and significant. According to these results, we can accept that the combination of our predictor variables of GrLivArea, MasVnrArea, and BsmtUnSF have value in the prediction of SalePrice.

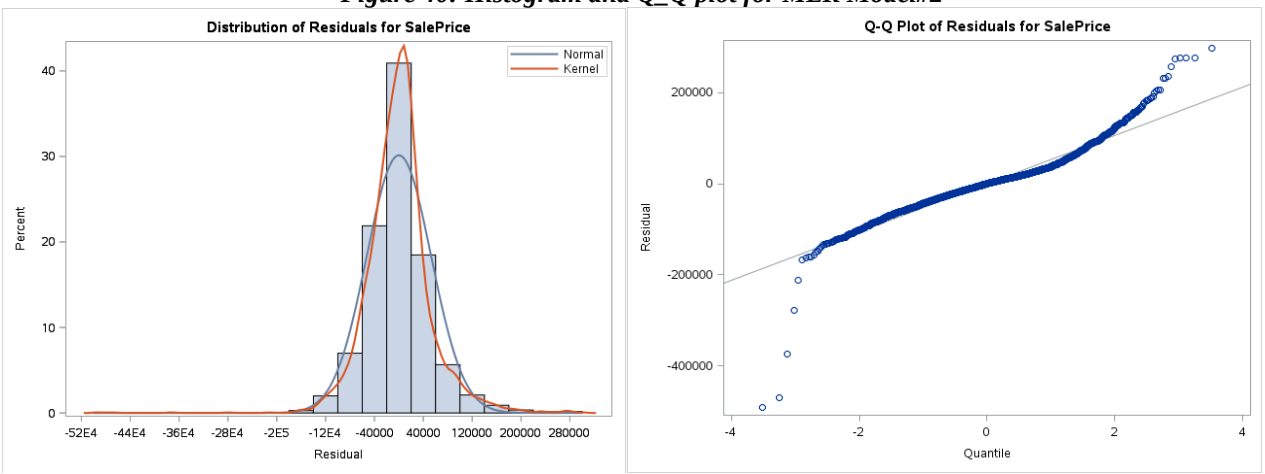
The R-Square in our second MLR is displayed in Figure.31 explains nearly 56% of the variability in SalePrice when using GrLivArea and MasVnrArea is explained by our predictor variable. This is the highest R-Square number we have seen yet. As a reminder, we are looking for an R-Square that is at or above the 50% percent mark. The adjusted R-Square shares similar characteristics.

Figure 39: R-Square and Adjusted R-Square for MLR model #2

Root MSE	52962	**R-Square**	0.56
Dependent Mean	180415	**Adj R-Sq**	0.5595
Coeff Var	29.35544		

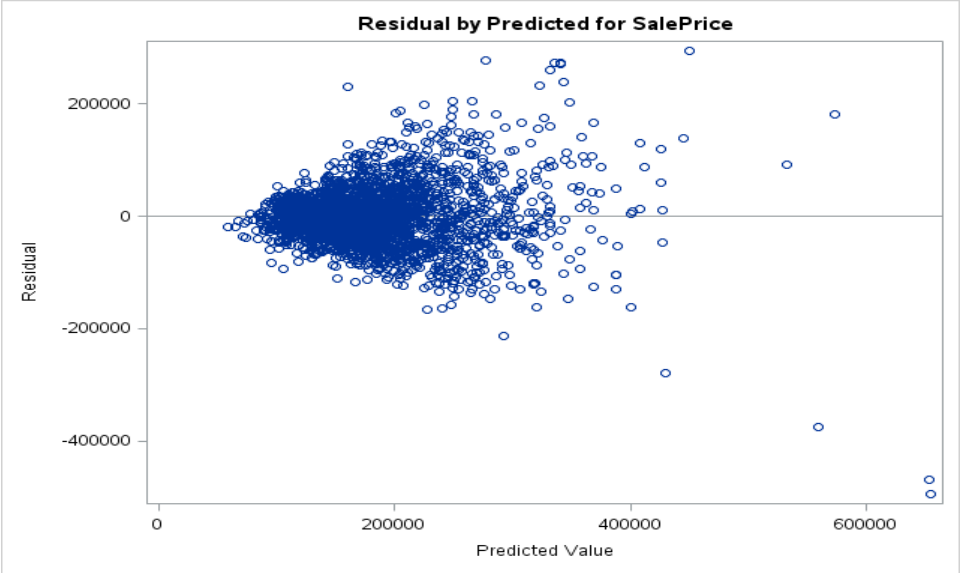
Normality assessments from our histogram in the second MLR model appear to fall in line with a normal distribution. This is similar to what we saw in the first MLR model. The Q_Q plot once again proves to be problematic. There is heavy tailing on both ends, and we have outliers present at the left-hand bottom of the graph. While it appears that the bulk of the data is in line, we cannot ignore the outliers and the tail on the right. Conflicting displays once again.

Figure 40: Histogram and Q_Q plot for MLR Model#2



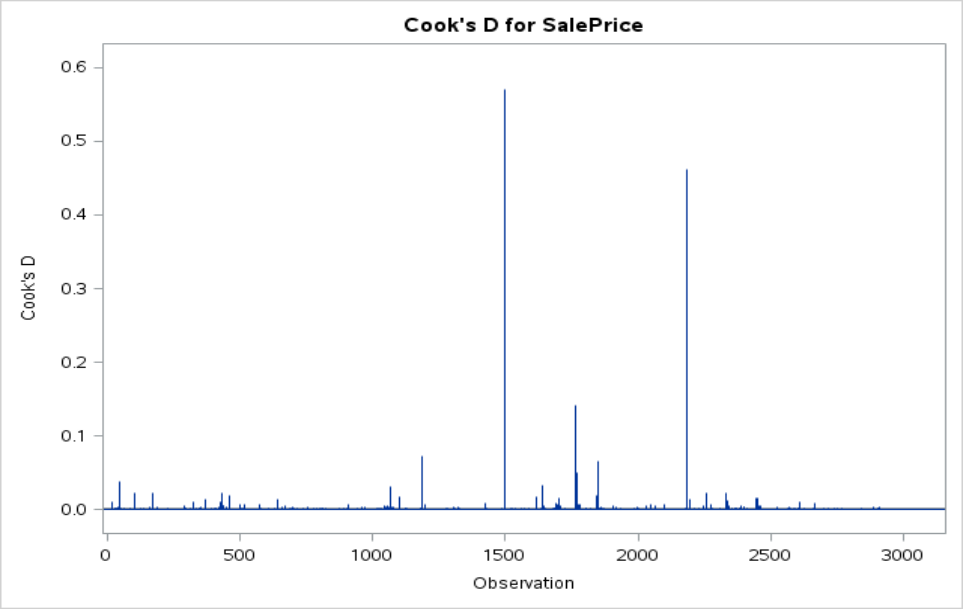
The residual versus the predicted looks for a random spread of data points. If the spread of data points is random, we assume that a linear regression model is a good fit. In Figure.41 there is a heavy concentration of observations in the plot and not a great deal of dispersion. Once again, there are some significant data points here that draw our attention. If we examine the lower right-hand side of the graph, we see two distinct outliers, and what appears to be a trail of outlying data points appearing to lead to the outliers in the corner of our graph.

Figure 41: Residual vs Predicted Value for MLR Model#2



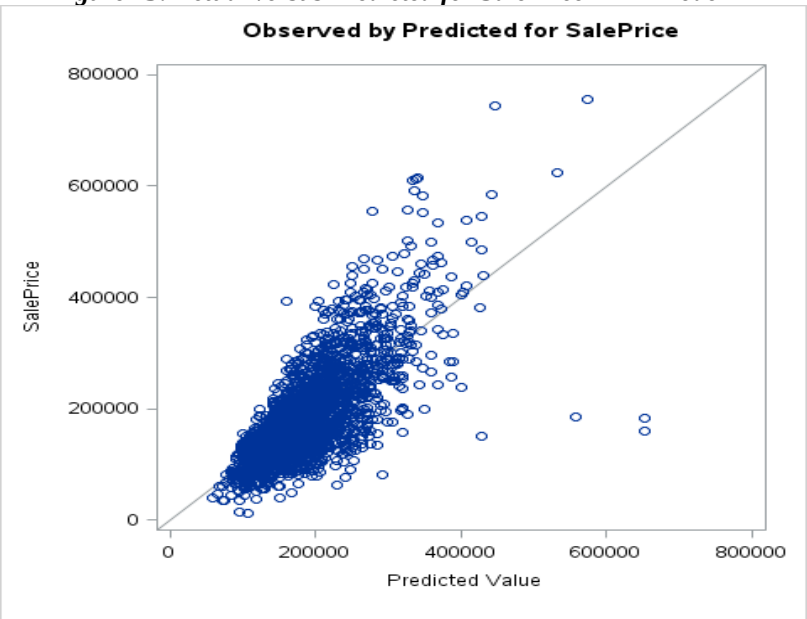
The Cook's Distance (Cook's D) plot in Figure.42 for our second MLR model is promising. There is some spiking present, but only two significant spikes that are troublesome. These points likely have some influence on our model and probably coincide with the outliers we noticed in the previous displays.

Figure 42: Cook's Distance for SalePrice MLR Model #2



The actual versus predicted value plot in Figure.43 provides our final input into our analysis of the model's fit. Again, we are looking to see if our data fall close to the diagonal line. We should see most of our points close to the regression line in the middle of the plot. The plot can also be understood and visualized by checking to see if similar data points for predicted values and the actual SalePrice coincide.

Figure 43: Actual versus Predicted for SalePrice MLR Model#2



Summary for MLR Model#2:
Our second multiple linear regression model (MLR) provides some interesting observations. The P-Values and F-statistic indicated that there is statistical significance that our predictor variables are in fact useful towards predicting SalePrice, and our R-Square is good, coming in at 56%, The Q_Q plot once again proved to be troubling with respect to validating our model. The heavy tails in the Q_Qplot are difficult to overlook, and we continue to see the presence of more outliers. The Cook's D output is decent, with the exception of two points, but the residuals weren't useful.

Summary analysis of fit for MLR Models#1 and#2:
Having gone through each model we come to some interesting realizations. Incorporating the last variable did not change our model by much. You can see in the breakdown below that with the exception of F-Statistic actually going down, not much changed. Some of this could be attributed to our deliberate use of a variable with low correlation. It would appear that each of these Multiple Linear Regression models provides similar effectiveness when it comes to modeling SalePrice.

Figure 44: Summary for Models 1

Models	R-Square	F-Statistic	Pr > F
(1)SalePrice = 26,547 + 94.60 * GrLivArea + 118.55 * MasVnrArea	0.5599	1846.98	<.0001
(2)SalePrice = 25,788 + 93.91 * GrLivArea + 118.57 * MasVnrArea+ 3.23 * BsmtUnSF	0.56	1231.02	<.0001

Conclusion:
We examined several linear models in this analysis. In all the cases, we found the models to be statistically significant when it came to the regression analysis numeric output. In all 5 models, we found that looking into the ODS output and graphical displays led us to question and dig deeper into the results. There wasn't a single model in this analysis that did not have a flaw. If a model lacked statistically significant P-Values and R-Square values, this would certainly prove difficult to overlook. but the numbers themselves should never be the sole indicators of a model with a good fit. A good next step would be to refine the predictive variables in an attempt to create a better fitting model for our data.

Appendix: SAS code

```
/*Read file into SAS and identify it as mydata*/
libname mydata '/scs/wtm926/' access = readonly;

Data temp1;
  set mydata.ames_housing_data;

/***Problem 1 Y=saleprice=dependent variable X= MasVnrArea predictor variable correl=(.50)
EDA #1***/
/***Fit simple linear regression model using X to predict Y***/
  /***Check correlation from EDA 1 ***/
proc corr data=temp1;
   var saleprice MasVnrArea FirstFlrSF TotalBsmtSF GarageArea GrLivArea;
run;
/***Problem 1-a plot fitted regression model over scatterplot saleprice***/
proc reg data=temp1 plots(unpack)=(diagnostics fitplot residualplot);
   model saleprice = MasVnrArea;
run;
proc reg data=temp1 plots=(diagnostics fitplot residualplot);
   model saleprice = MasVnrArea;
run;

/***Problem 2-a find a better simple linear regression model to predict Y=Saleprice***/
/***using the selection = rsquare option in PROC REG***/
proc reg data=temp1;
   model saleprice = SID PID SubClass LotFrontage LotArea OverallQual OverallCond YearBuilt
   YearRemodel MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF FirstFlrSF
SecondFlrSF LowQualFinSF
   GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr
TotRmsAbvGrd Fireplaces
   GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch
   PoolArea MiscVal MoSold YrSold/
   selection=rsquare start=1 stop=1;
run;
/***Problem 2-b Fit simple linear regression model using this better X to predict Y***/
proc reg data=temp1 plots(unpack)=(diagnostics fitplot residualplot);
   model saleprice = GrLivArea;
run;


/***Problem 3-a Selct one of the categorical variables to use as explanatory ***/
/***variable (X) to Predict (Y=SalePrice)Cat=MoSold GarageCars Fireplaces BedroomAbvGr***/
/***Check correlation from EDA***/
proc corr data=temp1 rank;
```

```
    var MoSold GarageCars Fireplaces BedroomAbvGr;
    with saleprice;
run;
proc reg data=temp1 plots(unpack)=(diagnostics fitplot residualplot);
    model saleprice = GarageCars;
run;
/*proc sort and proc mean by GarageCars*/
proc sort data=temp1;
   by GarageCars;
proc means data=temp1;
   by GarageCars;
   var saleprice;
run;

/***Problem .5 Part B-Fit a multiple regression model that uses two continuous explanatory
Variables***/
/***to predict (Y=Saleprice)Use Variables from Step1=MasVnrArea and Step2=GrLivArea***/
proc reg data=temp1 plots(unpack)=(diagnostics fitplot residualplot);
  model SalePrice = GrLivArea MasVnrArea;
run;

/***Problem .6 Part B-Add a third continuous predictor (X variable) to your multiple
regression***/
/***from step=5to predict (Y=Saleprice)Use Variables from Step1=MasVnrArea and
Step2=GrLivArea***/
/***Check correlation in order to select variable with lowest correlation to Y***/
proc corr data=temp1 rank;
    var SID PID SubClass LotFrontage LotArea OverallQual OverallCond YearBuilt
    YearRemodel MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF FirstFlrSF
SecondFlrSF LowQualFinSF
    GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr
TotRmsAbvGrd Fireplaces
    GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
    MiscVal MoSold YrSold;
    with saleprice;
run;

/***Run Multiple regression Analysis with GrlivArea MasVnrArea BsmtUnfSF***/
proc reg data=temp1 plots(unpack)=(diagnostics fitplot residualplot);
  model SalePrice = GrLivArea MasVnrArea BsmtUnfSf;
run;
```

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.

Cooks distance. https://en.wikipedia.org/wiki/Cook's_distance (accessed January 13, 2017)