Assignment #2

Scott M. Morgan

Introduction:

The purpose of this assignment is to build and evaluate regression models to predict home sale price. To accomplish this objective, I use data from the Ames housing data set provided by the course instructor and SAS Studio to perform the analysis. Understanding the nuances of a data set and regression analysis is an integral part of the predictive modeling process. In this assignment, I use functionality within SAS Studio to examine variables and interpret output delivery system (ODS) reports to assess a series of predictive models. Please note all SAS code is included at the end of the assignment in the order in which it is discussed. Making decisions given imperfect information is a key skill in predictive analytics and I expect that this exercise will demonstrate that there is no singular metric that points to the "best" model, but rather a confluence of factors which includes statistical analysis and logical reasoning.

Results:

In the subsequent sections, we will generate and evaluate a series of predictive models. First, we will evaluate several variables and their predictive power using simple linear regression models. Building upon this foundation, we will then include these variables, and others, in several multiple regression models and observe their overall impact.

Simple Linear Regression. In this first section, we generate 3 simple linear regression models, assess model adequacy and determine which model is the best fit.

Model 1: MasVnrArea to Predict SalePrice. Using the SAS simple linear regression procedure, we generate the following parameter estimates (Table 1):

Table 1: Parameter Estimates for MasVnrArea to Predict SalePrice

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	157303	1466.89502	107.24	<.0001
MasVnrArea	1	226.47763	7.11940	31.81	<.0001

This results in the following model in equation form:

SalePrice = 157303 + 226.47763 x MasVnrArea

Within the context of this model, if MasVnrArea were equal to 0 then we expect the SalePrice to be $157,303. Additionally, for every increase of 1 unit in MasVnrArea we expect the SalePrice to

increase by $226.48. This is generally in line with expectations as more square footage, regardless of what it is specifically, generally costs more money.

Using the SAS procedure for correlation, we produce an analysis of variance table. The output is presented in Table 2 below. Of the 2930 observations in the sample, the number of observations with missing values is 23, which is acceptable. The F Value is large and statistically significant, so we can reject the null hypothesis that there is no linear relationship between SalePrice and MasVnrArea. R-Squared, which measures the percent of variance explained by the model, is only ~26%. This suggests that residuals or other potential variables are effecting the model. Phrased differently, 75% of the model variation is explained by other variables besides MasVnrArea. For reference, the Adjusted R-Squared will be particularly important in comparing the effectiveness of these models as it takes into account the number of predictors and increases only as new variables improve the overall model.

Table 2: Analysis of Variance

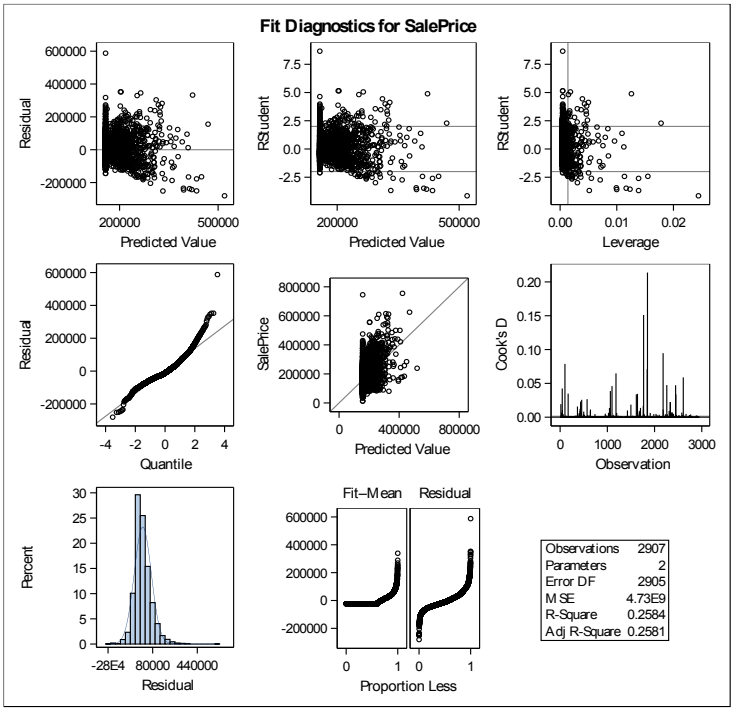
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.781879E12	4.781879E12	1011.96	<.0001
Error	2905	1.372718E13	4725361826		
Corrected Total	2906	1.850905E13			

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

We next shift our attention to the automatically generated ODS output (Exhibit 1) from SAS to assess the goodness-of-fit for this model. The scatterplot for residual values is slightly x-axis unbalanced with values pooling on the left side (i.e. smaller values). While this isn't necessarily bad it is something to take into consideration. This visual also indicates the existence of an unusual value. The Q-Q plot of the residuals shows there is a slight systematic pattern of progressive departure from normality, particularly in the upper right corner suggesting positive skewness. Again, the outlier mentioned in the scatterplot is present and obvious. The histogram is relatively normally distributed with slight positive skewness, corroborating our observations from the Q-Q plot. From these observations, we can posit this model is not homoscedastic. Cook's Distance (Cooks' D), which quantifies influential data points, illustrates 3 plausible outliers which is concerning. This begs the question if these extreme values are errors or legitimate values. The Fit-Mean Residual plot indicates the residuals have more of an impact on the

model than the independent variable. The corroborates our observation pertaining to the R-Squared value mentioned previously.

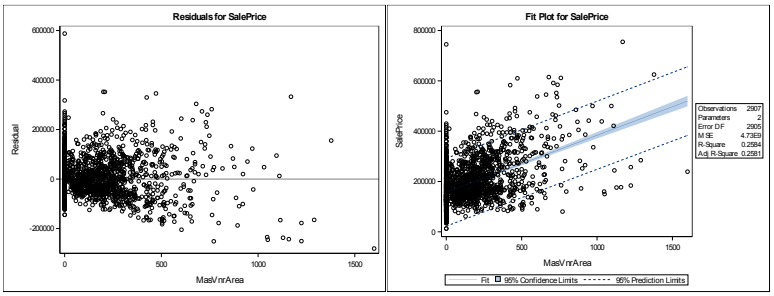
Exhibit 1: Fit Diagnostics for SalePrice Using MasVnrArea



To conclude our evaluation of the initial model, we analyze the residuals versus the independent variable and the fit plot for SalePrice. The visualization to the left in Exhibit 2 below shows the plots versus the independent variable; which looks relatively similar to the residual plot versus the predicted values from Exhibit 1 with pooling to the left side of the chart. Overall the shape of both scatterplots is not ideal as we are looking for no discernable pattern in the residuals. Lastly, the Fit Plot for SalePrice reveals a good portion of the observations fall outside of the 95% prediction limit, again indicating possible problems with the model.

Logically, using the masonry veneer area as a predictor for sale price does not make much sense as there are other features of a home that are likely more intrinsically valuable. Based on this analysis and intuition, we should conclude that MasVnrArea is not the most effective variable at predicting SalePrice and that this is not a very useful model in its current state. This is mainly due to the large number of MasVnrArea data being zero. Given its prevalence, however, we can assume they are legitimate data. In the subsequent sections, we seek to improve on this simple linear regression model by finding a better explanatory variable based on statistical analysis as well as rationale.

Exhibit 2: Residuals and Fit Plot for SalesPrice Using MasVnrArea



Model 2: "Better" Simple Linear Regression Model. In this portion of the analysis, our aim is to find a better simple linear regression model to predict Y = SalePrice. A "better" model in this case can primarily be thought of as exhibiting a superior Adjusted R-Squared as well as residuals with no pattern and normality in terms of distribution as well as data that falls within the 95% predict intervals. Using the SAS function *selection=rsquare* option PROC REG with *start =1* and *stop = 1*, we attempt to identify a better explanatory value X. First, we enter the top 5 most positively correlated continuous variables into the function and evaluate the R-Squared metric. These variables, in order of most correlated to least correlated with SalePrice are GrLivArea, GarageArea, TotalBsmtSF, FirstFlrSF and MasVnRArea. Table 3 below provides the model comparison. Based on the highest R-Squared metric, we will test GbLivArea, or the above ground living area in square feet, in our next model. It makes sense intuitively that more livable space, in general, equates to a higher sale price so I anticipate this to be a much more effective predictor.

Table 3: Model Comparison

Number in Model	R-Square	Variables in Model
1	0.5006	GrLivArea
1	0.4086	GarageArea
1	0.4002	TotalBsmtSF
1	0.3885	FirstFlrSF
1	0.2582	MasVnrArea

Table 4 below provides the parameter estimates for the new model based on GrLivArea:

Table 4: Parameter Estimates for GrLivArea to Predict SalePrice

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13290	3269.70277	4.06	<.0001
GrLivArea	1	111.69400	2.06607	54.06	<.0001

This results in the following model in equation form:

SalePrice = 13290+ 111.69400 x GrLivArea

Within the context of this model, if GrLivArea were equal to 0 then we expect the SalePrice to be $13,290. Additionally, for every increase of 1 unit in GrLivArea we expect the SalePrice to increase by $111.69. This is again in line with expectations that more square footage costs more money, however, it is surprising to see GrLivArea with a smaller coefficient relative to MasVnrArea.

Using the SAS procedure for correlation, we again produce an analysis of variance table. The output is presented in Table 5 below. Of the 2930 observations in the sample there were no missing values. The F Value is large and statistically significant, so we can reject the null hypothesis that there is no linear relationship between SalePrice and GrLivArea. R-Squared and Adjusted R-Squared are both ~50%, which is lower than we would like. We will continue to monitor the Adjusted R-squared as we build out the model with multiple independent variables.

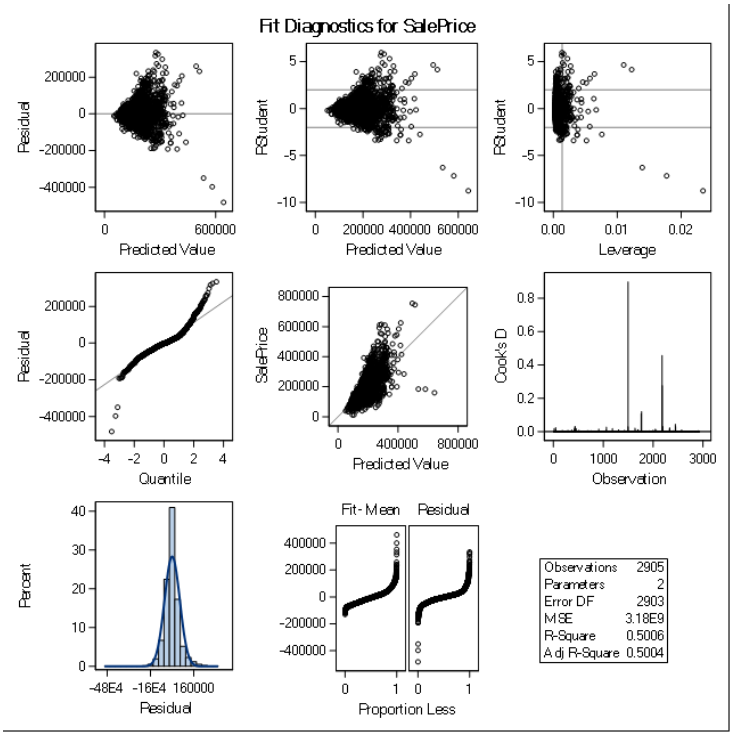
Table 5: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	9.33763E12	9.33763E12	2922.59	<.0001
Error	2928	9.354907E12	3194981962		
Corrected Total	2929	1.869254E13			

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

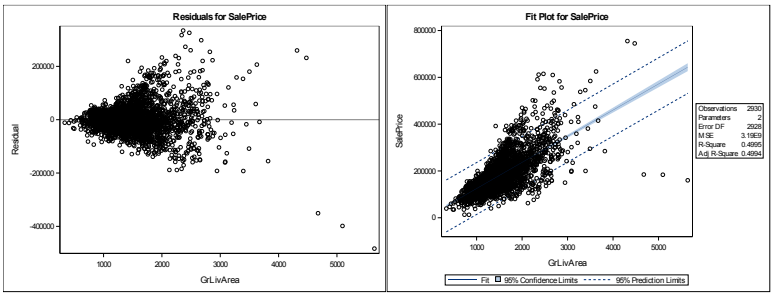
Studying the automatically generated ODS output (Exhibit 3) from SAS we assess the goodness-of-fit. The scatterplot for the residual values is clearly wedge shaped, suggesting that variance in the values increases as the SalePrice gets larger. There also appears to be 3 possible outliers. The Q-Q plot of the residuals suggests a non-normal distribution as the residuals trail off into the upper right corner of the chart. The presence of the possible outliers is also apparent. The histogram describes a largely normal distribution. It appears that the Q-Q plot and the histogram are giving conflicting information. Cook's D highlights the presence of 3 outliers that would require further investigation but that is currently outside of the scope of this project. The left pane of the Fit-Mean Residual plot is taller than the right, indicating that the independent variable accounts for a greater portion of the variation in the model, contrary to our initial model where the residuals had more impact. From these observations, we can conclude that this model is not homoscedastic.

Exhibit 3: Fit Diagnostics for SalePrice Using GrLivArea



Lastly, we analyze the residuals versus the independent variable and the fit plot for SalePrice. The visualization to the left in Exhibit 4 below shows the plots versus the independent variable; this again resembles a wedge-shape. The Fit Plot for SalePrice illustrates a positive linear relationship with a good portion of the data falling between the 95% prediction limits.

Exhibit 4: Residuals and Fit Plot for SalesPrice Using GrLivArea



While the intuition behind GrLivArea being a better predictor of SalePrice than MasVnrArea is sound and the model appears to be somewhat more effective, there is certainly room for improvement and there are still numerous inadequacies, including suspect residual accuracy and distribution shape. In the next sections, we explore whether categorical variables or multiple regression might result in a more robust predictive model.

Model 3: Simple Linear Regression Using Categorical Variables. In this section, we select a categorical variable to use as an explanatory variable to predict Y, SalePrice. The variables we can choose from are MoSold, GarageCars, FirePlaces and BedroomAbvGr. Table 6 below provides the model comparison. Instead of selecting a model solely based on R-Squared, my goal in this iteration is to select a variable based on rationale as well. We will test BedroomAbvGr in this third model. BedroomAbvGr is a discrete data type that represents the number of bedrooms above ground. The reasoning behind testing this variable is similar to Model 2; more bedrooms should be, at the very least, tangentially related to a higher sale price. I expect this to be a more effective predictor than the previous 2 models but I fully expect complications in using a categorical variable.

Table 6: Model comparison

Number in Model	R-Square	Variables in Model
1	0.4197	GarageCars
1	0.2252	Fireplaces
1	0.0207	BedroomAbvGr
1	0.0012	MoSold
1	0.4197	GarageCars

Table 7 below provides the parameter estimates for the new model based on BedroomAbvGr:

Table 7: Parameter Estimates for BedroomAbvGr to Predict SalePrice

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	141152	5245.39482	26.91	<.0001
BedroomAbvGr	1	13889	1765.04194	7.87	<.0001

This results in the following model in equation form:

SalePrice = 141152+ 13889 x BedroomAbvGr

Within the context of this model, if BedroomAbvGr were equal to 0 then we expect the SalePrice to be $141,152. Additionally, for every increase of 1 unit in BedroomAbvGr we expect the average SalePrice to increase by $13,889. This is in line with expectations that more bedrooms (above ground) equates to a higher sales price, however, we wouldn't expect a house with 0 bedrooms to cost $141,152. The exception could be a studio apartment.

Using the SAS procedure for correlation, we produce an analysis of variance table. The output is presented in Table 8 below. Of the 2930 observations in the sample there were no missing values. The F Value is smaller than we've observed in the previous models but still statistically significant, so we can reject the null hypothesis that there is no linear relationship between SalePrice and BedroomAbvGr. R-Squared and Adjusted R-squared are both 2%, which is obviously far lower than we would like.

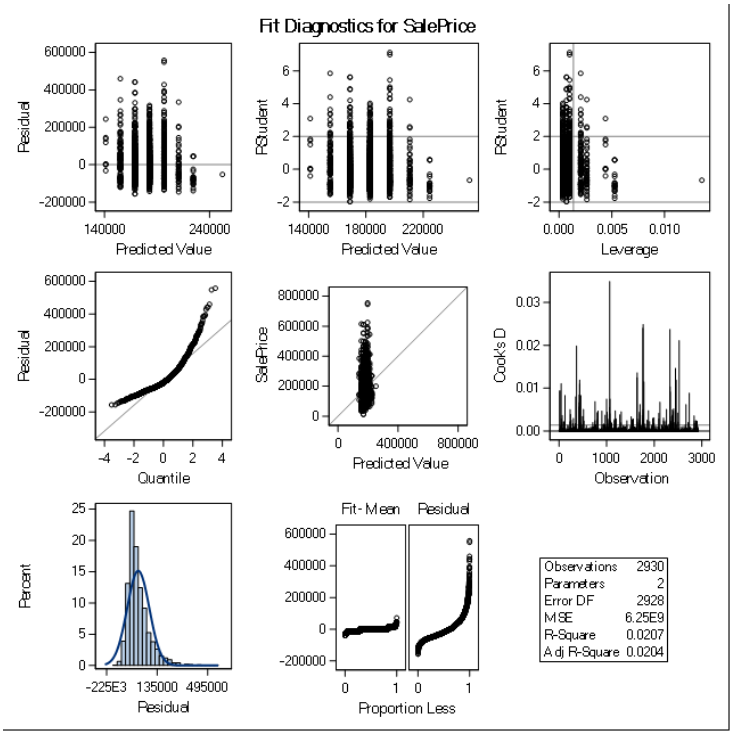
Table 8: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.871425E11	3.871425E11	61.92	<.0001
Error	2928	1.830539E13	6251842409		
Corrected Total	2929	1.869254E13			

Root MSE	79069	**R-Square**	0.0207
Dependent Mean	180796	**Adj R-Sq**	0.0204
Coeff Var	43.73358		

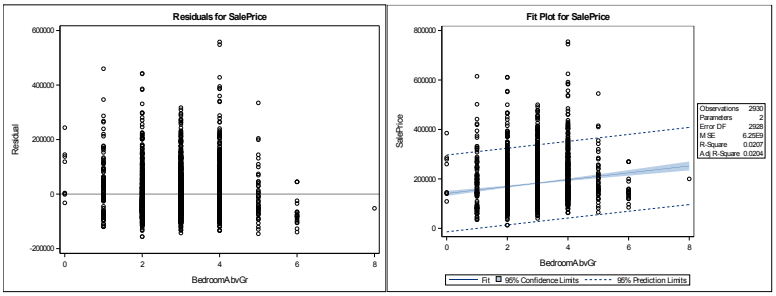
Studying the automatically generated ODS output (Exhibit 5) from SAS we assess the goodness-of-fit. The scatterplot for residual values appears to be somewhat concentrated; not entirely random. The Q-Q plot of the residuals suggests a non-normal distribution as the residuals trail off in both the lower left and upper right corner of the charts indicating positive skewness. The presence of 2 possible outliers is also apparent. The histogram of the residuals describes a normal distribution with positive skewness. Cook's D highlights the presence of at least one of the outliers as well numerous values that are above the threshold and potentially influential. The left pane of the Fit-Mean Residual plot is much lower than the right, indicating that there is a large amount of variation not explained by the model. The graph also confirms the residuals are not normally distributed. From these observations, we can conclude that this model is not homoscedastic.

Exhibit 5: Fit Diagnostics for SalePrice Using BedroomAbvGr



Lastly, we analyze the residuals versus the independent variable and the fit plot for SalePrice. The visualization to the left in Exhibit 6 below shows the plots versus the independent variable; this resembles the plot versus the predicted value. The Fit Plot for SalePrice illustrates a slight positive linear relationship with a surprising portion of the data falling between the 95% prediction limits. The largest amount of variability appears with 0, 5 and 6 above ground bedrooms. The presence of 1 of the outliers at 8 bedrooms is present at a below average cost. This suggests the need for further investigation of the legitimacy of this data point.

Exhibit 6: Residuals and Fit Plot for SalesPrice Using BedroomAbvGr



While the intuition behind BedroomAbvGr being a better predictor of SalePrice than MasVnrArea is sound, the statistical analysis indicates that this is not a good model. Using the SAS MEANS procedure, we check to see if the predicted model goes through the mean value for each of the categories. The model does pass through the mean values generally; this is important to monitor as we do not want the model to be over or under-fitted. In the next sections, we summarize our 3 simple linear regression models and then expand the analysis to include multiple variables.

Simple Linear Regression Model Comparison. Table 9 below provides a summary of the 3 models we constructed and their respective Adjusted R-Squares.

Table 9: Model Comparison

Model	Adjusted R-Square
1. SalePrice = 157303 + 226.47763 x MasVnrArea	0.2581
2. SalePrice = 13290+ 111.69400 x GrLivArea	0.4994
3. SalePrice = 141152+ 13889 x BedroomAbvGr	0.0204

While we highlighted various strengths and weaknesses of each model, if our conclusion is based solely off of Adjusted R-Squared values then we would conclude that the model which includes GrLivArea would be the best model to use for predicting SalePrice. It is important to note that relying on a single statistic is unwise but given our analysis of the residuals and the underlying rationale of GrLivArea being associated with higher sale prices, I think we can be comfortable in our conclusion that Model 2 is the 'best' choice out of the 3.

Multiple Linear Regression. In this second section, we generate 2 multiple linear regression models, assess model adequacy and determine which model is the best fit.

Model 4: 2 Continuous Variables. We will now fit a multiple regression model that uses 2 continuous explanatory variables to predict SalePrice. The two variables will be MasVnrArea and GrLivArea, which were examined independently in Model 1 and Model 2, respectively. Table 9 below provides the parameter estimates for the new model based on the two variables:

Table 9: Parameter Estimates for MasVnrArea and GrLivArea to predict SalePrice

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	26547	3141.93715	8.45	<.0001
MasVnrArea	1	118.54695	5.99550	19.77	<.0001
GrLivArea	1	94.60302	2.12104	44.60	<.0001

This results in the following model in equation form:

SalePrice = 26547+ 118.54695 x MasVnrArea + 94.60302 x GrLivArea

Within the context of this model, if MasVnrArea and GrLivArea were equal to 0 then we expect the SalePrice to be $26,547. Additionally, for every increase of 1 unit in MasVnrArea and GrLivArea we expect the average SalePrice to increase by $213.14. There are likely few circumstances where GrLivArea would equal 0 but it's possible for MasVnrArea to equal 0 if the house doesn't have a masonry veneer. Both coefficients have decreased from their original values when modeled independently.

Using the SAS procedure for correlation, we produce an analysis of variance table. The output is presented in Table 10 below. Of the 2930 observations in the sample 23 were missing values which is acceptable. The F Value is large and still statistically significant, so we can reject the null hypothesis that there is no linear relationship between SalePrice, MasVnrArea and GrLivArea. R-Squared and Adjusted R-Squared are both ~56%, which is the best figure we've seen so far but ideally it would be higher.

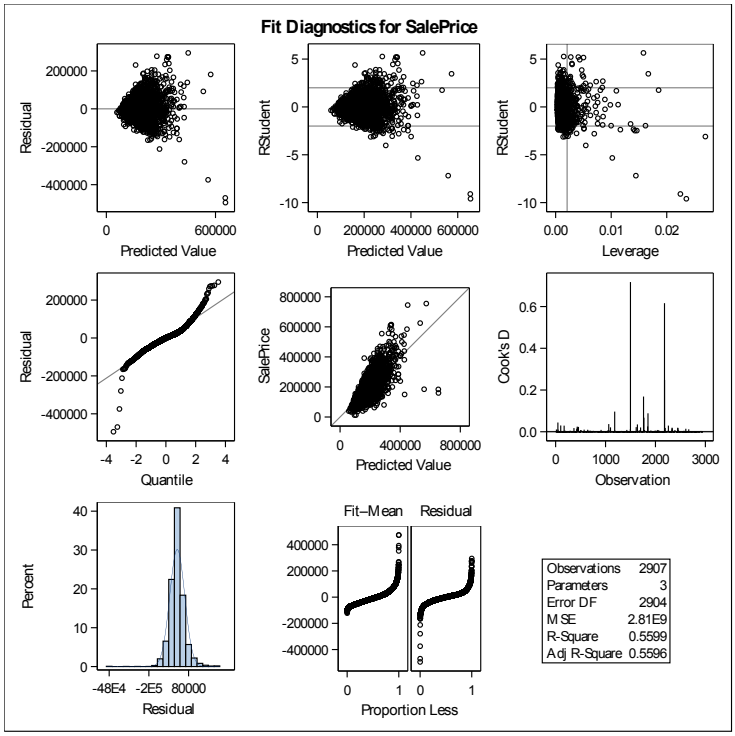
Table 10: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.036256E13	5.181279E12	1846.98	<.0001
Error	2904	8.146497E12	2805267561		
Corrected Total	2906	1.850905E13			

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

Studying the automatically generated ODS output (Exhibit 7) from SAS we assess the goodness-of-fit. The scatterplot for the residual values appears to be concentrated and wedge-shaped; not entirely random. The Q-Q plot of the residuals suggests a non-normal distribution with heavy tails. The presence of multiple possible outliers is also apparent. The histogram of the residuals suggests a normal distribution, contradicting the Q-Q plot. Cook's D denotes the presence of at least 2 possible outliers and numerous smaller influential points. The left pane of the Fit-Mean Residual plot is higher than the right, indicating that the predictor variables account for a greater portion of the variation in the model. From these observations, we can conclude that this model is not homoscedastic.

Exhibit 7: Fit Diagnostics for SalePrice Using MasVnrArea and GrLivArea



If we rely solely on the Adjusted R-Squared value, this multiple regression model fits better than the simple linear regression models constructed earlier.

Model 5: 3 Continuous Variables (2 'Good' + 1 'Worst'). We will now fit a multiple regression model that uses 3 continuous explanatory variables to predict SalePrice. The 3 variables will be MasVnrArea, GrLivArea and BsmtFinSF2. BsmtFinSF2 represents the Type 2 finished square footage of the basement and is the continuous variable with the smallest correlation of X and Y (0.01). Table 11 below provides the parameter estimates for the new model based on the 3 variables:

Table 11: Parameter Estimates for MasVnrArea, GrLivArea and BsmtFinSF2 to Predict SalePrice

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	25997	3159.53176	8.23	<.0001
MasVnrArea	1	118.65010	5.99412	19.79	<.0001
GrLivArea	1	94.62084	2.12098	44.61	<.0001
BsmtFinSF2	1	10.46253	5.78953	1.81	0.0708

This results in the following model in equation form:

SalePrice = 25997+ 118.65010 x MasVnrArea + 94.62084 x GrLivArea + BsmtFinSF2 x 10.46253

Within the context of this model, if MasVnrArea, GrLivArea and BsmtFinSF2 were all equal to 0 then we expect the SalePrice to be $25,997. Additionally, for every increase of 1 unit in MasVnrArea, GrLivArea and BsmtFinSF2 we expect the average SalePrice to increase by $223.73. There are likely few circumstances where GrLivArea would equal 0 but it's possible for MasVnrArea and BsmtFinSF2 to equal 0 if the house doesn't have a masonry veneer or a basement. The coefficients for MasVnrArea and GrLivArea changed marginally given the inclusion of BsmtFinSF2 to the model.

Using the SAS procedure for correlation, we produce an analysis of variance table. The output is presented in Table 12 below. Of the 2930 observations in the sample 24 were missing values which is acceptable. The F Value is large and still statistically significant, so we can reject the null hypothesis that there is no linear relationship between SalePrice, MasVnrArea, GrLivArea and BsmtFinSF2. R-Squared and Adjusted R-Squared are both ~56%, a very marginal improvement but essentially the same from Model 4.

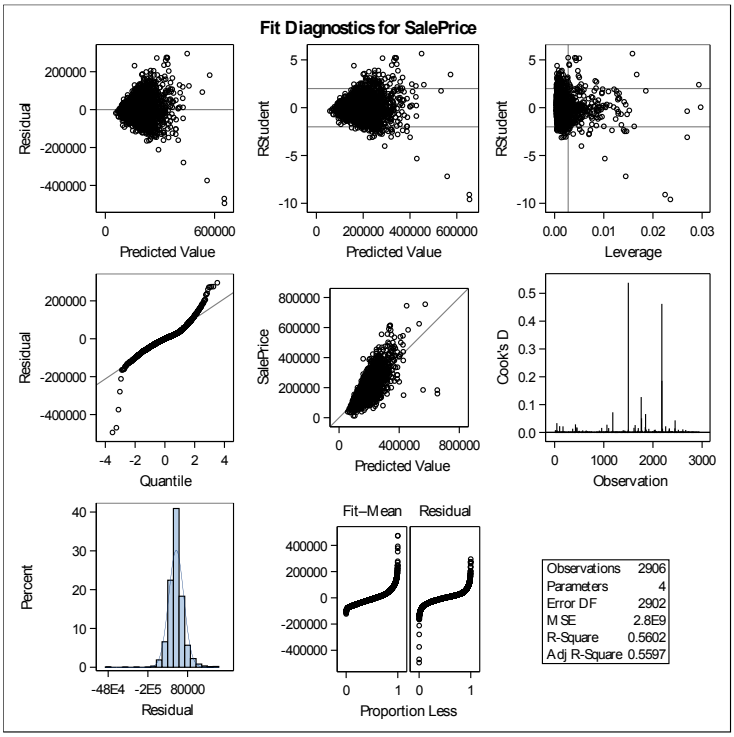
Table 12: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.036248E13	3.454159E12	1232.01	<.0001
Error	2902	8.136296E12	2803685698		
Corrected Total	2905	1.849877E13			

Root MSE	52950	R-Square	0.5602
Dependent Mean	180415	Adj R-Sq	0.5597
Coeff Var	29.34889		

Studying the automatically generated ODS output (Exhibit 8) from SAS we assess the goodness-of-fit. The scatterplot for residual values appears to be concentrated and wedge-shaped; not entirely random. The Q-Q plot of the residuals suggests a non-normal distribution with heavy tails. The presence of multiple possible outliers is apparent. The histogram of the residuals suggests a normal distribution, contradicting the Q-Q plot. Cook's D denotes the presence of at least 2 possible outliers and numerous smaller influential points. The left pane of the Fit-Mean Residual plot is higher than the right, indicating that the predictor variables account for a greater portion of the variation in the model. From these observations, we can conclude that this model is not homoscedastic.

Exhibit 8: Fit Diagnostics for SalePrice Using MasVnrArea, GrLivAream and BsmtFinSF2



If we rely solely on the Adjusted R-Squared value, this multiple regression model's fit is essentially the same as Model 5 (two variables). Interestingly, the inclusion of a poorly related variable (BsmtFinSF2) resulted in very marginal improvement in the R-Squared (0.5599 to 0.5602) and Adjusted R-Squared values (0.5596 to 0.5597). The F-Value also decreased notably, however it was still statistically significant. Overall, it appears that adding additional predictor variables, however unrelated, can improve the R-Squared value but does not necessarily translate into a better model fit.

Conclusions:

Building and validating regression models is a cornerstone of predictive analytics. For this exercise, the Ames, Iowa data set was used to generate and assess the effectiveness of simple linear and multiple regression models. We primarily used the Adjusted R-Squared statistic to compare our models but recognize that it is not the sole determinant in the selection process. Analysts need to take into account the strengths and weaknesses as well as the reasoning behind including each variable. A model with a series of variables that fits nicely but isn't supported by sensible rationale is a poorly specified model.

The first exercise was a good introduction to data manipulation using SAS and interpreting statistical outputs to identify potential predictor variables. This week's assignment expounded upon this foundation by using these potential predictor variables in initial models. This exercise also reinforced the importance of data visualization in validation activities. The next steps in the model building process would be to continue investigating additional predictors, transforming variables and addressing potential outliers.

Code:

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


proc datasets library=mydata; run;


Data temp1;
        set mydata.ames_housing_data;


*PART A*;


** Question 1**;


***Finding correlation for all numeric variables***;
proc corr data=temp1 rank;
        var _numeric_;
        with saleprice;
run;


***Correlation and simple linear regression for variable correlated ~0.5***;


***Model 1***;
proc corr data=temp1 rank;
        var _numeric_;
        with  MasVnrArea saleprice;
run;
```

```
proc reg data=temp1;
        model saleprice = MasVnrArea;

run;

** Question 2**;

***Evaluate New Continuous Variables***;
proc reg data=temp1;
        model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea /
        selection=rsquare start=1 stop=1;
run;

***Run Model 2***;

proc corr data=temp1 rank;
        var _numeric_;
        with  GrLivArea saleprice;
run;

proc reg data=temp1;
        model saleprice = GrLivArea ;
run;
```

```
** Question 3**;

***Evaluate New Categorical Variable***;
proc reg data=temp1;
        model saleprice = MoSold GarageCars Fireplaces BedroomAbvGr /
        selection=rsquare start=1 stop=1;
run;

***Run Model 3***;

proc corr data=temp1 rank;
        var _numeric_;
        with  BedroomAbvGr saleprice;
run;

proc reg data=temp1;
        model saleprice = BedroomAbvGr ;
run;

***Predited Mean Value go through average***;

proc sort data=temp1;
        by BedroomAbvGr ;

proc means data=temp1;
```

```
        by BedroomAbvGr ;
        var saleprice;
run;

proc corr data=temp1 nosimple rank plots=(scatter);
  var BedroomAbvGr ;
  with SalePrice;
run;

*PART B*;
** Question 5**;

proc reg data=temp1;
        model saleprice = MasVnrArea GrLivArea /
        selection=rsquare start=1 stop=2;
run;

***Run Model 4***;

proc corr data=temp1 rank;
        var _numeric_;
        with  MasVnrArea GrLivArea saleprice;
run;

proc reg data=temp1;
```

```
        model saleprice = MasVnrArea GrLivArea;
run;

** Question 6**;

proc reg data=temp1;
        model saleprice = MasVnrArea GrLivArea BsmtFinSF2  /
        selection=rsquare start=1 stop=3;
run;

***Run Model 5***;

proc corr data=temp1 rank;
        var _numeric_;
        with  MasVnrArea GrLivArea BsmtFinSF2 saleprice;
run;

proc reg data=temp1;
        model saleprice = MasVnrArea GrLivArea BsmtFinSF2;
run;
```