Assignment #3: Data Analysis and Regression

Harini K. Anand

1. Introduction

Models for accurate prediction of housing prices are in heavy demand by homeowners, home sellers, mortgage lenders, building developers, tax assessment offices, insurance companies, etc. It is because home buying is a costly investment.

In this report, we present the results of the data analysis and the regression models built using Ames, Iowa Housing dataset from DeCock (2011), for the prediction of home sale prices. The report is structured to present the analysis that went into the selection of the sample data first, followed by the exploratory data analysis (EDA) to select the predictor variables, and then the analysis of the simple and multiple linear regression models with different types of predictor variables with and without transformations to the response variable. As part of the model analysis, the report presents the regression output analysis, Goodness-of-Fit (GOF) analysis, predictive accuracy analysis, and comparison of models. In conclusion, the regression model that fits the data better in contrast to all the models that are built, for the prediction of home sale prices is presented.

2. Data

The dataset used contains the assessed values for individual residential properties sold in Ames, IA from 2006 to 2010. The data dictionary for the dataset is available at this location: http://www.amstat.org/publications/jse/v19n3/decock/DataDocumentation.txt. Per the data dictionary, the dataset has 2,930 observations. It contains 82 variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables, and 2 observation identifiers).

3. Sample Definition

As stated above, the goal was to build and evaluate linear regression models for the prediction of price of a 'typical' home. Towards that goal, we have chosen to exclude any observations in the population of apartments, condominiums, commercial properties, farm properties, and industrial properties (those that are not in the residential zones) as these observations present huge variations in the sale prices. For similar reasons, observations that are not of "normal" sales (such as sales between family members, abnormal sales or partial sales) and building types that are not single-family homes (such as townhomes, duplex, 2 family conversions) are excluded. Additionally, we also excluded any observations of 'atypical' homes such as those with no paved street access, basements (vast majority of houses in mid-west have basements to protect against freeze damage) or no water or just ELO (Electricity only) (in these cases, homes would be un-inhabitable). Due to the housing boom in the 1950s, huge differences are noted in the sale prices of homes built before 1950 compared to those built after. So, for this assignment, the regression models were developed to predict prices of houses built only after 1950. Consequently, the observations of houses built before 1950 are excluded from the sample. The excluded observations do not reflect a 'typical' home and the variations in the data they present could achieve poor results for the model.

Additionally, we also examined the special notes in the data dictionary which states that houses with more than 4000 square feet should be removed from the observations due to the sale price not

representing the actual market values and also due to some observations being from unusual sales. These observations were excluded as they are outliers and would affect the model.

Drop conditions

Based on the above discussion of the appropriate sample population for the problem to predict the value of a house in Ames, Iowa, the below exclusion rule set (drop conditions) is defined:

a. ***01: Not Residential:*** Excluded observations that are not in residential zones (observations with MS Zoning values A, C, I, or FV)
b. ***02: Non-Normal Sale:*** Excluded observations with any sale condition value that is other than "Normal" (observations with Sale Condition that is not 'Normal')
c. ***03: Not Single-Family Home:*** Excluded observations with building types that are other than "Single-family" (observations with Bldg Type values '2FmCon', 'Duplx', 'TwnhsE', or 'TwnhsI')
d. ***04: Street Not Paved:*** Excluded observations that do not have "paved" road access to property (observations with Street value 'Gravel')
e. ***05: No Water:*** Excluded observations with utilities that do not include water (observations with Utilities values 'NoSeWa' or 'ELO')
f. ***06: Built Pre-1950:*** Excluded observations of houses built before the year 1950 (observations with Year Built value < 1950)
g. ***07: No Basement:*** Excluded observations with no basements (observations with BsmtFin Typ1 1 value 'No Basement')
h. ***08: LT 800 OR GT 4000 SqFt***: Excluded observations with ground living area of less than 800 sq.ft or more than 4000 sq.ft (observations with Gr Liv Area with value < 800 sq. ft or > 4000 sq.ft)

Waterfall Drop Condition	Count
01: Not Residential	168
02: Non-Normal Sale	457
03: Not Single Family Home	362
04: Street Not Paved	2
06: Built Pre-1950	480
07: No Basement	27
08: LT 800 OR GT 4000 SqFt	10
99: Eligible Sample	1424

Table 01: Distribution table of the records excluded from the eligible sample based on drop conditions. Also included is the size of eligible Sample.

From Table 01, we are to determine that a total of 1506 observations were excluded from the original dataset. There is one observation in the dataset with Utilities value of 'NoSeWa' (No Sewage or Water). But this observation also has the Non-Normal sale. As a result, it is excluded as part of the condition 02. After the exclusions using the waterfall drop conditions, the eligible sample 1424 observations.

Code Snippet 01 in section 9 contains the R code for the creation of the eligible sample population.

4. Simple Linear Regression Models

In this section, we present the results of fitting two simple linear regression models for the prediction of home sale price. SalePrice is the response variable.

As a first step, to determine the continuous predictor variables for the two models, we utilized exploratory data analysis techniques to find the two most potential variables among the continuous variables in the dataset.

4.1. Selection of Predictor variables for the two simple linear regression models

The predictor variables for the two models were selected after evaluating 6 continuous variables in the dataset. Table 02 below shows the variables that were evaluated. These 6 variables were picked out based on the intuition that they would typically affect the sale prices of homes. Both graphical and non-graphical exploratory data analysis techniques were employed in the evaluation of these variables.

Prior to performing the analysis, observations with any missing values were removed from the sample. This resulted in a reduction of 334 observations in the sample.

The criteria employed to select a predictor variable are:

a) the variable shows a correlation (linearity trend) with the response variable (SalePrice),

b) has a strong correlation coefficient (positive or negative) in comparison to the rest of the variables, and

c) does not have 'adversely' skewed data or large number of outliers (presence of outliers or skewed data impacts the model)

A skewed predictor variable could be made symmetric using transformation. However, it is not done in the simple linear regression models described in this section. Note: Transformations to predictor variables are evaluated and applied in section 7.3.

Variable Name	Data Type	Description
GrLivArea	Continuous	Above grade (ground) living area square feet
LotArea	Continuous	Lot size in square feet
TotalBsmtSF	Continuous	Total square feet of basement area
MasVnrArea	Continuous	Masonry veneer area in square feet
LotFrontage	Continuous	Linear feet of street connected to property
GarageArea	Continuous	Size of garage in square feet

Table 02 shows the variables evaluated for the selection of predictor variables in the two simple linear regression models

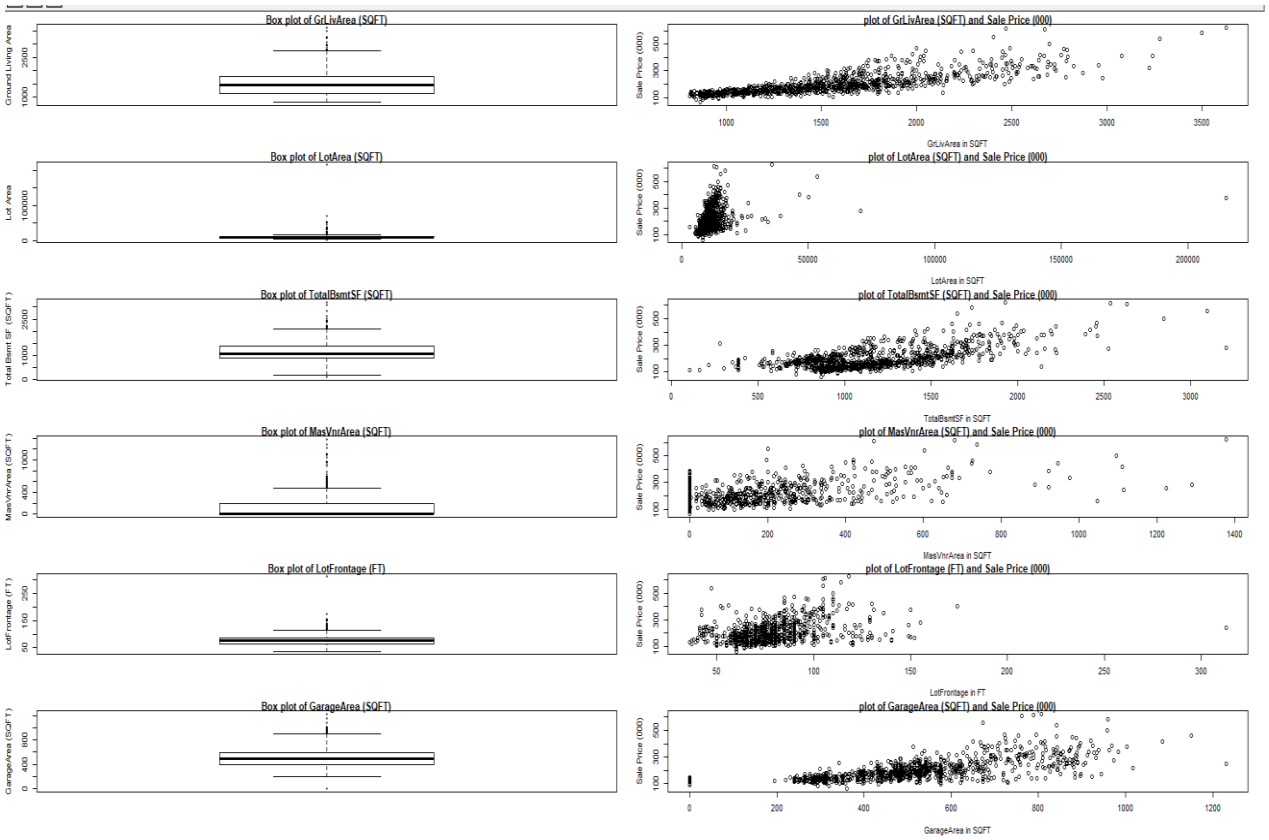

FIG 01. Single variable box plot and scatter plot against Sale Price for the continuous variables GrLivArea, LotArea, TotalBsmtSF, MasVnrArea, LotFrontage, and GarageArea illustrated from top to bottom for comparison

In FIG 01, the variables GrLivArea, TotalBsmtSF, and GarageArea are noted to capture a linearity trend with the response variable, SalePrice. An examination of the single variable box plots of these variables to the left shows the presence of outliers in the data.

In Table 03, among the above visually identified potential predictor variables, we noted that GrLivArea has the largest Correlation Coefficient (0.8039091). Its closeness to 1 indicates a strong positive correlation between GrLivArea and SalePrice. GarageArea follows second with the second largest Correlation Coefficient value (0.6800787). It also indicates a strong positive correlation between GarageArea and SalePrice.

Also can be derived from Table 03, is that both these variables (GrLivArea and GarageArea) have mean > median indicating right skewness in the data.

Variable Name	Minimum	Q1	Median	Mean	Q3	Maximum	Correlation Coefficient (measured against SalePrice)
GrLivArea	808	1,127	1,478	1,519	1,780	3,627	0.8039091
LotArea	3,182	8,724	10,012	11,038	11,926	215,245	0.2954364
TotalBsmtSF	105	880	1,066	1,146	1,382	3,206	0.6465373
MasVnrArea	0.0	0.0	0.0	123.1	196.0	1378.0	0.5775225
LotFrontage	36.0	65.0	75.0	76.96	85.0	313.0	0.3629618
GarageArea	0.0	393.2	490.0	509.7	596.0	1231.0	0.6800787

Table 03 shows the five part summary, mean and the correlation coefficient computed against SalePrice

Based on the criteria laid out at the beginning of the section, we selected GrLivArea and GarageArea as the predictor variables in the two simple linear regression models. Both GrLivArea and GarageArea have right skewness and some outliers in the data. The variables could be made symmetric using transformations. However, in the next two sections, we proceeded with the fitting of the simple linear regression models using the raw values without any transformation to the variables with the full understanding that these characteristics will have an effect on the models.

Code Snippet 02 in section 9 contains the R code for the selection of two continuous predictor variables.

4.2. Simple Linear Regression Model #1 (GrLivArea)

Fitting a linear model to the observations of GrLivArea (X) to predict SalePrice(Y) using the ordinary least squares estimation method provided us the following fitted equation:

SalePrice = 13365.222 + 118.335 * (GrLivArea)

Term	Coefficient	Standard Error	t-value	p-value (Pr(>|t|))
(Intercept)	13365.222	4614.559	2.896	0.00386 **
GrLivArea	118.335	2.874	41.174	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 42310 on 979 degrees of freedom
Multiple R-squared: 0.6339, Adjusted R-squared: 0.6335
F-statistic: 1695 on 1 and 979 DF, p-value: < 2.2e-16
Residuals:

Min	1Q	Median	3Q	Max
-106987	-24333	-4104	15678	280207

Table 04 Regression Output for the Simple Regression Model SalePrice~GrLivArea

Analysis of the Regression Output:

Based on the regression output in Table 04, we are able to conclude that the overall regression is strongly statistically significant since the probability (p-value) of the overall F-Test is less than 2.2e-16 (very low value). Based on this, at a level of significance 0.001, we can reject the null hypothesis that **all** the predictor variables have a zero coefficient and have no effect on the regression.

The coefficient of determination (Adjusted R-squared) value of 0.6335 indicates approximately 63.35% of the variation in the SalePrice can be explained by variation in the GrLivArea values.

Among the coefficient T-test results, the predictor variable, GrLivArea is statistically significant with a p-value of less than 2.2e-16 (very low value at a level of significance of 0.001). The coefficient for GrLivArea indicates an increase of $118.335 in SalePrice for a unit change in GrLivArea. The standard error for the coefficient of GrLivArea is also small indicating higher precision in the GrLivArea coefficient estimation.

Contrary to the coefficient of GrLivArea, the intercept (constant term) is statistically significant with a p-value (Pr(>|t|) of 0.00386 at a slightly higher level of significance (0.01). Also, the standard error

which indicates the precision of the intercept estimate is quite large (4614.559) indicating the error can be considerable. The coefficient for the intercept informs that $13365.222 is the estimated value of SalePrice when GrLivArea=0.

Assessment of Goodness-of-Fit of the model:

Next, we performed the residual analysis for Goodness-of-Fit (GoF) using the diagnostic plots to validate the normality assumption and the homoscedasticity assumption.

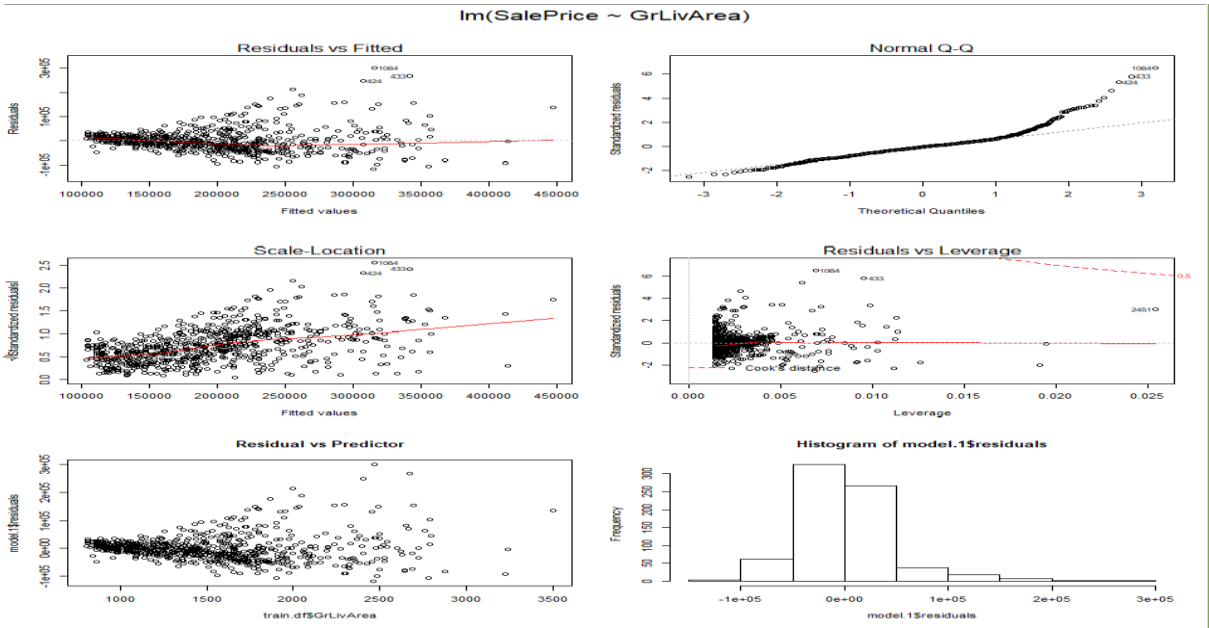

FIG 02. Residual diagnostic plots (Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage, Residual vs Predictor, Histogram of Residuals cascaded left to right from top to bottom) for the simple linear regression model (SalePrice ~ GrLivArea)

In FIG 02, from the Residuals vs Fitted scatter plot, we noticed the residuals do not show a horizontal band formation around the residual = 0 indicating the error variances are not constant. The plot shows a somewhat funnel distribution with outliers.

From the Residuals vs Predictor scatter plot, we see the residuals showing some structure starting out as a horizontal band but fanning out in a funnel distribution. This questions the predictor effect on the regression.

From the Normal Q-Q plot in FIG 02, we see the residuals deviate from normal (Gaussian) distribution. The plot also indicates the residuals are right-skewed with tails.

The histogram plot for the residuals in FIG 02 does not show a symmetric bell-shaped histogram indicating the normality assumption for the errors is not likely to be true.

Based on the analysis of the above diagnostic plots, there are concerns with using this model for statistical inference since all the GOF conditions are not met. As a result, we are not confident to use the t-test scores, p-values, and overall F-test score provided in Table 04.

Analysis for predictive accuracy:

Metric	Value using training data	Value using test data
MSE (measure of average error between the predicted values and observed values)	1,786,550,952	2,059,442,213
MAE (the average magnitude of the errors in the predictions, without considering their direction)	29,333.89	29,701.36

Table 05 MSE and MAE values to assess predictive accuracy of the SalePrice ~ GrLivArea model

For predictive accuracy, the mean square error and the mean absolute error were computed using the residuals. Table 05 lists the values of MSE and MAE obtained using training and test data. The values are higher for test data versus training data. The large values of both metrics indicate a large model prediction error.

4.3. Simple Linear Regression Model #2 (GarageArea)

Fitting a linear model to the observations of GarageArea (X) to predict SalePrice(Y) using the ordinary least squares estimation method provided us the following fitted equation:

SalePrice = 62366.4 + 259.8 * (GarageArea)

Term	Coefficient	Standard Error	t-value	p-value (Pr(>|t|))
(Intercept)	62366.4	5235.6	11.91	<2e-16 ***
GarageArea	259.8	9.7	26.79	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 53120 on 979 degrees of freedom
Multiple R-squared: 0.4229, Adjusted R-squared: 0.4223
F-statistic: 717.5 on 1 and 979 DF, p-value: < 2.2e-16
Residuals:

Min	1Q	Median	3Q	Max
-155316	-29577	-3722	22043	352964

Table 06 Regression Output for the Simple Regression Model SalePrice~GarageArea

Analysis of the Regression Output:

Based on the regression output in Table 06, we are able to conclude that the overall regression is strongly statistically significant since the probability (p-value) of the overall F-Test is less than 2.2e-16 (a very low value). Based on this, at a level of significance of 0.001, we can reject the null hypothesis that **all** the predictor variables have a zero coefficient and have no effect on the regression.

The coefficient of determination (Adjusted R-squared) value of 0.4223 indicates approximately 42.23% of the variation in the SalePrice can be explained by variation in the GarageArea values.

From Table 06, for the coefficient T-test results, we gather the predictor variable, GarageArea, is statistically significant with a p-value of less than 2e-16 (for the computed T test statistic). At a significance level of 0.001, we can reject that the coefficient of GarageArea is 0. The coefficient for GarageArea indicates an increase of $259.8 in SalePrice for a unit change in GarageArea. The standard error for the coefficient of GarageArea is 9.7 (quite low).

Similarly, the intercept (constant term) is also statistically significant with a p-value less than 2e-16. At a significance level of 0.001, we can reject that the constant term is 0. The standard error which indicates the precision of the intercept estimate is quite large (5235.6) indicating the error in the constant term can be considerable. The coefficient for the intercept informs that $62366.40 is the predicted value of SalePrice when GarageArea=0.

Assessment of Goodness-of-Fit of the model:

Next, we performed the residual analysis for Goodness-of-Fit (GoF) using the diagnostic plots to validate the normality assumption and the homoscedasticity assumption.

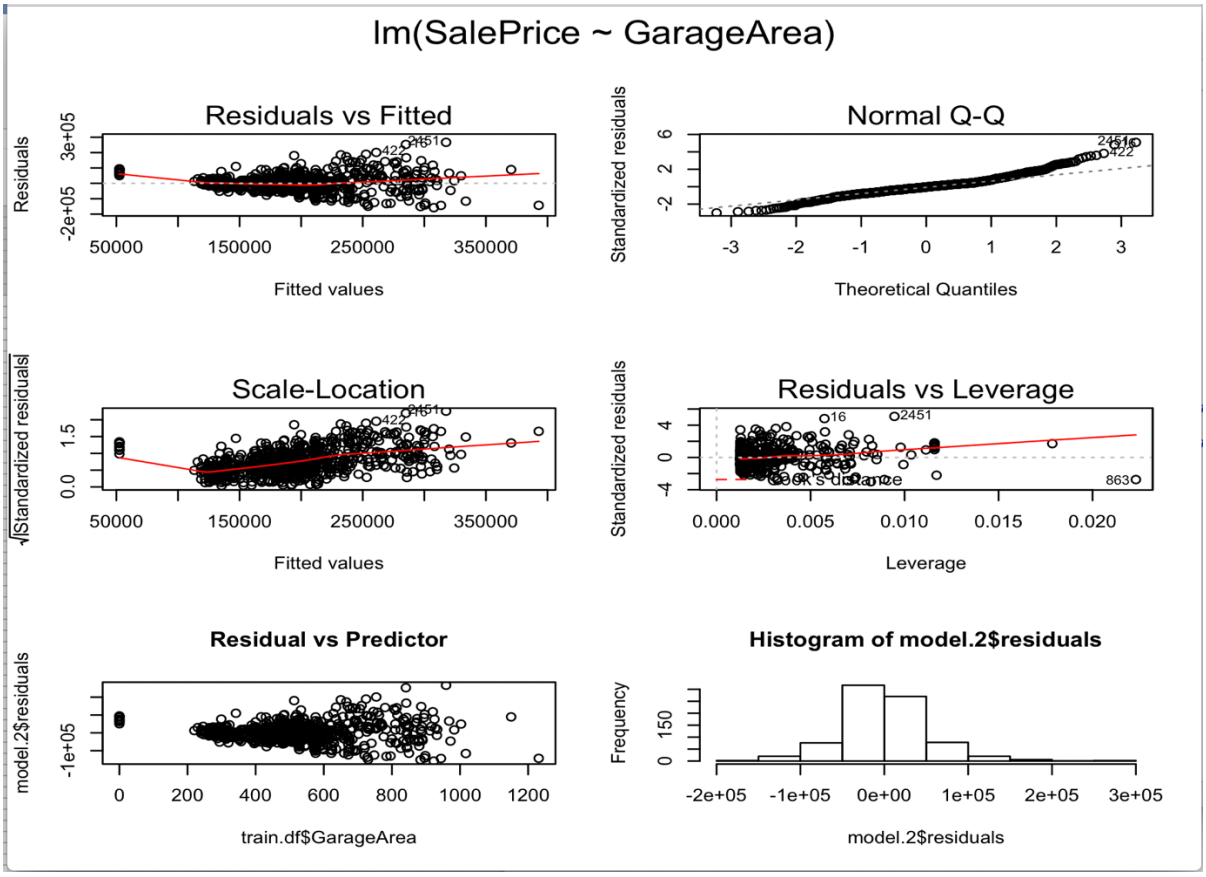

FIG 03. Residual diagnostic plots (Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage, Residual vs Predictor, Histogram of Residuals cascaded left to right from top to bottom) for the simple linear regression model (SalePrice ~ GarageArea)

In FIG 03, the Residuals vs Fitted scatterplot does not show a consistent random pattern of the residual errors. The plot also does not show horizontal band formation of the residuals around residual=0. It also shows a spread out in a funnel distribution as the fitted values increase. The smoother

line shows a slight curve rather than a line consistent with residual =0. The constant variance assumption is not validated.

The Residuals vs Predictor scatter plot shows a horizontal band but fanning out in a funnel distribution. This plot indicates a problem with the model.

From the Normal Q-Q plot in FIG 03 indicates the residuals deviate somewhat slightly from normal (Gaussian) distribution. The residuals also have slight skewness to right with tail.

The histogram plot of the residuals in FIG 03 does not show a truly symmetric bell-shaped histogram indicating the normality assumption for the errors is not likely to be true.

Therefore, there are concerns with using this model for statistical inference since all the GOF conditions are not met. As a result, we are not confident in the t-test scores, p-values, and overall F-test score provided in Table 06.

Analysis for predictive accuracy:

Metric	Value using training data	Value using test data
MSE (measure of average error between the predicted values and observed values)	2,816,332,003	3,212,738,186
MAE (the average magnitude of the errors in the predictions, without considering their direction)	37,166.56	38,496.85

Table 07 MSE and MAE values to assess predictive accuracy of the SalePrice ~ GarageArea model

For predictive accuracy, the mean square error and the mean absolute error were computed using the residuals. Table 07 lists the values of MSE and MAE obtained using training and test data. The values of MSE and MAE are higher for test data compared to the values of training data. The large values of both errors indicate a large model prediction error.

Code Snippet 04 in section 9 contains the R code for fitting the two simple linear regression models with GrLivArea and GarageArea to predict the SalePrice.

5. Multiple Linear Regression Model – Model #3 (GrLivArea+GarageArea)

Fitting a multiple linear model to the observations of GrLivArea (X1), GarageArea (X2) to predict the SalePrice(Y) using the ordinary least squares estimation method provided us the following fitted equation:

SalePrice = 93.355 * (GrLivArea) + 134.919 * (GarageArea) - 17180.474

Term	Coefficient	Standard Error	t-value	p-value (Pr(>|t|)
(Intercept)	-17180.474	4408.725	-3.897	0.000104 ***
GrLivArea	93.355	2.901	32.182	< 2e-16 ***
GarageArea	134.919	7.798	17.302	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 37040 on 978 degrees of freedom
Multiple R-squared: 0.7197, Adjusted R-squared: 0.7191
F-statistic: 1256 on 2 and 978 DF, p-value: < 2.2e-16
Residuals:

Min	1Q	Median	3Q	Max
-132731	-21637	-2507	15637	274742

Table 08 Regression Output for the Multiple Regression Model SalePrice~GrLivArea+GarageArea

Analysis of the Regression Output:

Based on the regression output in Table 08, we are able to conclude that the overall regression is strongly statistically significant since the probability (p-value) of the F-Test is less than 2.2e-16. Based on this, at a level of significance of 0.001, we can reject the null hypothesis that **all** the predictor variables have a zero coefficient and have no effect on the regression.

The coefficient of determination (Adjusted R-squared) value of 0.7191 indicates approximately 71.91% of the variation in the SalePrice can be explained by variation in the predictor variables GrLivArea and GarageArea.

From Table 08 using the T-test results, we determine that the predictor variable GarageArea, is statistically significant with a p-value of less than 2e-16 (for the computed T test statistic). At a significance level of 0.001, we can reject that the coefficient of GarageArea is 0. The coefficient implies that given all else in the model is held fixed, one additional Sqft in the GarageArea is associated with an estimated SalePrice that is $134.919 higher. The standard error for the coefficient of GarageArea is 7.798 which is small.

Similarly, we also determine that the predictor variable GrLivArea is statistically significant with a p-value corresponding to the t test statistic 32.182 with a value of less than 2e-16. So, at a significance level of 0.001, we reject the null hypothesis that the coefficient of GrLivArea is 0. Similar to GarageArea, that given all else in the model is held fixed, one additional Sqft in the GrLivArea is associated with an estimated SalePrice that is $93.355 higher. The standard error for the coefficient of GrLivArea is 2.901 which is considerably small.

The intercept (constant term) is also statistically significant with p-value of 0.000104. At a significance level of 0.001, we can reject that the constant term is 0. The standard error which indicates the precision of the intercept estimate is quite large (4408.725) indicating the error in the constant term can be considerable. The coefficient for the intercept informs an estimate of -$17180.474 in the predicted value of SalePrice when GarageArea=0 and GrLivArea=0.

Assessment of Goodness-of-Fit of the model:

The goodness of fit of the multiple regression model is determined by checking whether the assumption of normality and homoscedasticity are met using the diagnostic residual plots in FIG 04.

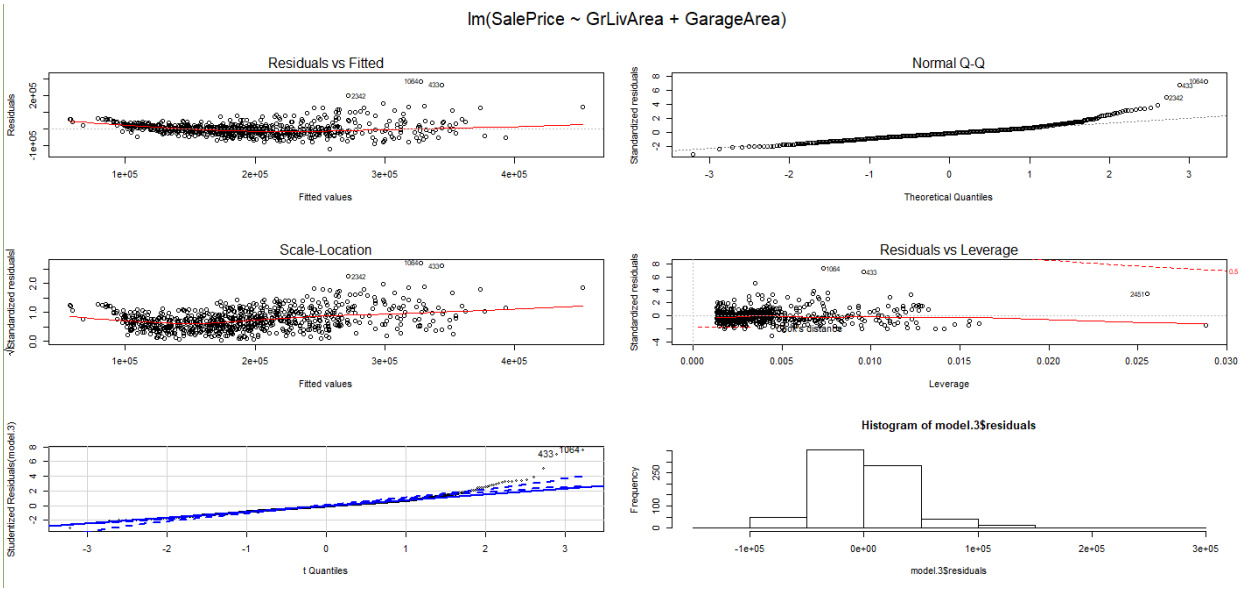

FIG 04. Residual diagnostic plots (Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage, studentized QQ plot, Histogram of Residuals) for the linear regression model (SalePrice ~ GrLivArea+GarageArea)

In FIG 04, from the Residuals vs Fitted scatter plot, the residuals do not fall in a symmetrical pattern and do not have a constant spread throughout the range. There are also outliers in the plot. This disputes the constant variance assumption.

From the Normal Q-Q plot and studentized Q-Q plots in FIG 04, the residuals deviate from normal (Gaussian) distribution. This indicates the residuals have a slight right tail with skewness.

The histogram plot of the residuals in FIG 04 does not show a symmetric bell-shaped histogram. It also confirms the presence of right tail in the residuals indicating the normality assumption does not hold.

Therefore, there are concerns with using this model for statistical inference since all the GOF conditions are not met. As a result, we are not confident in the t-test scores, p-values, and overall F-test score provided in Table 08.

Analysis for predictive accuracy:

Metric	Value using training data	Value using test data
MSE (measure of average error between the predicted values and observed values)	1,367,844,829	1,610,439,286
MAE (the average magnitude of the errors in the predictions, without considering their direction)	27,475.48	26,838.85

Table 09 MSE and MAE values to assess predictive accuracy of the SalePrice ~ GrLivArea+GarageArea model

For predictive accuracy, the mean square error and the mean absolute error were computed using the residuals. Table 09 lists the values of MSE and MAE obtained using the training and test data. The large values of both error indicate a large model prediction error and the values for test data are higher than those for training data.

Determination of the better fit model to predict SalePrice – the multiple linear regression model or the simple linear regression models

The criteria employed to evaluate the better fit model to predict SalePrice are:

1. the adjusted R-squared values
2. coefficient P-values for the predictor variables
3. partial F-test
4. the MSE and MAE errors

	Model #3 SalePrice ~ GrLivArea+GarageArea	**Model #1 SalePrice ~ GrLivArea**	**Model #2 SalePrice ~ GarageArea**
Adjusted R-squared	0.7191	0.6335	0.4223
Coefficient p-values	Intercept 0.000104*** GrLivArea < 2e-16 *** GarageArea < 2e-16 ***	Intercept 0.00386 ** GrLivArea <2e-16 ***	Intercept <2e-16 *** GarageArea <2e-16 ***
MSE & MAE Errors (using test data)	MSE - 1,610,439,286 MAE – 26,838.85	MSE – 2,059,442,213 MAE - 29,701.36	MSE – 3,212,738,186 MAE – 38,496.85

Table 10 Model comparison table with adjusted R-squared, p-values, MSE&MAE errors

Based the information in Table 10, we gathered that Model #3 (SalePrice ~ GrLivArea+GarageArea) has the best adjusted R-squared value among the 3 models. Only Model #3 and Model#2 have the P-values for the single coefficients are statistically significant at a significance level of 0.001. For Model #1, the intercept is not statistically significant for at the same level of significance. The intercept is significant at a higher level of significance of 0.01 Also, the MSE error and MAE errors (which measure the predictive accuracy) for Model #3 are the lowest compared to the simple linear regression models (Model #1 and Model #2)

Additionally, the results of the partial F-test show that at a significance level of 0.001, that the additional coefficients in Model #3 are statistically significant.

Partial F-test result between Model #1 and Model #3 showed that at significance level of 0.001, it allows us to reject the reduced model (null hypothesis which is Model#1). Table 11 shows the Anova output.

Model 1: SalePrice ~ GrLivArea
Model 3: SalePrice ~ GrLivArea + GarageArea

Model	**Res.Df**	**RSS**	**Df**	**Sum of Sq**	**F**	**Pr(>F)**
1	979	1.7526e+12				
3	978	1.3419e+12	1	4.1075e+11	299.37	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 11 Anova output of Model #1 and Model #3

Partial F-test result between Model #2 and Model #3 is that at significance level of 0.001, it allows us to reject the reduced model (null hypothesis which is Model#2). Table 12 shows the Anova output.

Model 2: SalePrice ~ GarageArea
Model 3: SalePrice ~ GrLivArea + GarageArea

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2	979	2.7628e+12				
3	978	1.3419e+12	1	1.421e+12	1035.7	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 12 Anova output of Model #2 and Model #3

Based on this evaluation, we conclude Model #3 (SalePrice ~ GrLivArea + GarageArea), the multiple linear regression model with more predictors is the better fit model to predict the SalePrice.

Code Snippet 05 in section 9 contains the R code for fitting the multiple linear regression model to predict the SalePrice with two continuous predictor variables (GrLivArea and GarageArea).

6. Neighborhood accuracy

For the neighborhood accuracy, we began with fitting linear regression model of Neighborhood (X) values (a nominal variable) to predict the SalePrice (Y). The fitted equation is

SalePrice = 159895 -37395* BrkSide + 73248* ClearCr + 35648* CollgCr + 59109* Crawfor -16267* Edwards + 30482* Gilbert -6465* IDOTRR + 6863* Mitchel -12460* NAmes + 171974* NoRidge + 180272* NridgHt + 33752* NWAmes -25033* OldTown -22670* Sawyer + 27197* SawyerW + 106519* Somerst + 174983* StoneBr -22395 * SWISU + 87683*Timber + 98415* Veenker

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	159895	46007	3.475	0.000540 ***
NeighborhoodBrkSide	-37395	56347	-0.664	0.507123
NeighborhoodClearCr	73248	49184	1.489	0.136852
NeighborhoodCollgCr	35648	46202	0.772	0.440618
NeighborhoodCrawfor	59109	48496	1.219	0.2233
NeighborhoodEdwards	-16267	46539	-0.35	0.726789
NeighborhoodGilbert	30482	46402	0.657	0.511448
NeighborhoodIDOTRR	6465	50399	0.128	0.897965
NeighborhoodMitchel	6863	46642	0.147	0.883055
NeighborhoodNAmes	-12460	46134	-0.27	0.787168
NeighborhoodNoRidge	171974	46794	3.675	0.000255 ***
NeighborhoodNridgHt	180272	46484	3.878	0.000115 ***
NeighborhoodNWAmes	33752	46505	0.726	0.468213
NeighborhoodOldTown	-25033	47622	-0.526	0.599293
NeighborhoodSawyer	-22670	46539	-0.487	0.626326
NeighborhoodSawyerW	27197	46465	0.585	0.558514
NeighborhoodSomerst	106519	47622	2.237	0.025605 *
NeighborhoodStoneBr	174983	49694	3.521	0.000456 ***
NeighborhoodSWISU	-22395	56347	-0.397	0.691155
NeighborhoodTimber	87683	46918	1.869	0.062048 .
NeighborhoodVeenker	98415	48253	2.04	0.041755 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 46010 on 726 degrees of freedom
Multiple R-squared: 0.632, Adjusted R-squared: 0.6219
F-statistic: 62.35 on 20 and 726 DF, p-value: < 2.2e-16

Table 13 Regression output for the linear regression model (SalePrice ~ Neighborhood)

Table 13 shows the overall F-test for the regression model with Neighborhood variable is statistically significant with p-value < 2.2e-16. So, at a significance level of 0.001, we can reject the null hypotheses that **all** the predictor variables have no effect on the regression.

However, Table 13 also shows for the individual coefficient T-test, many predictor variables are NOT statistically significant even at very high significance levels indicating no effect of these variables on the model.

Only NoRidge, NridgHt, StoneBr, and the intercept are significant at a significance level of 0.001. Somerst and Veenker are significant at higher level of significance of 0.05. Also, the adjusted R-squared is higher than that of Model #3.

Analysis of the fit of the neighborhoods by the model:

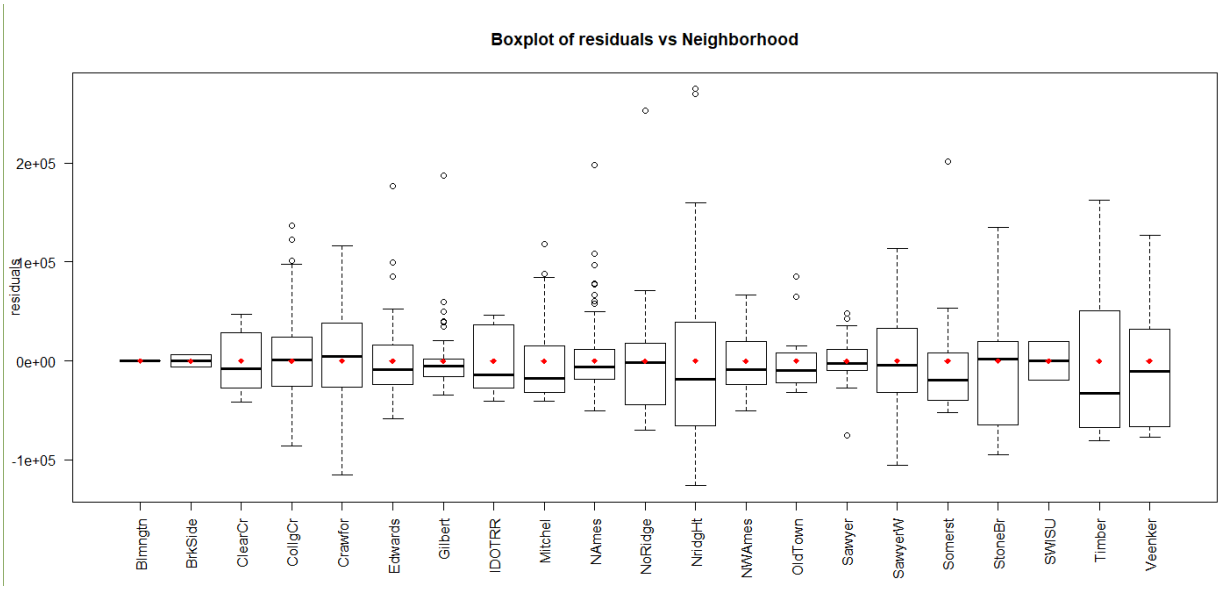

FIG 05. Residual Vs Neighborhood box plot

In general, residual values that are close to the horizontal line are predicted well. The residual values above the line residual=0 (positive) indicate underprediction and the residual values below the line residual=0 indicate overprediction.

Based on the above criteria, in FIG 05, we use the boxplots to determine whether the residuals have pronounced positive and negative values. Those boxplots with pronounced positive residual values indicate underpredicted neighborhoods. The boxplots with pronounced residual negative values indicate overpredicted neighborhoods. The boxplots are drawn with the mean values shown in red.

So, using the above criteria and the skewness of the boxplots, we determined that the neighborhoods ClearCr, Edwards, Gilbert, IDOTRR, Mitchel, NAmes, Nridgeit, NWAmes, OldTown, SwayerW, SomerSt, Timber, and VeenKer are consistently overpredicted and the neighborhood Crawfor is consistently overpredicted. The rest of the neighborhoods Blmngtn, BrkSide, CollgCr, NoRidge, Sawyer, StoneBr and SWISU are fitted well.

Plot of mean price/sqft vs MAE for all the neighborhoods:

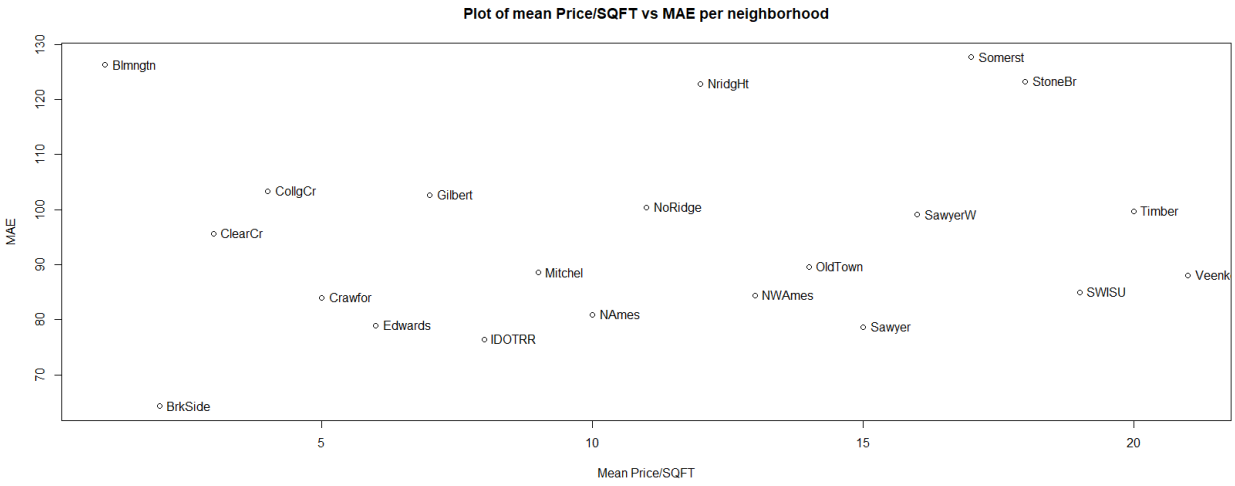

FIG 06 plot with mean price/sqft vs MAE computed for each of the 21 neighborhoods

The plot in FIG 06 is created by determining the price/sqft for each of the 21 neighborhoods along with the MAE values. However, no distinct pattern or relationship is noticed in the plot for price/sqft against MAE.

Fitting of model with indicator variables:

Next, we grouped the neighborhoods by price per square foot. In the sample data set, the price per square foot varies from 36.10 to 207.37 (units $/sqft). We divided this range of price/sqft into four categories:

Category 1: price/sqft >= 30.0 & price/sqft <=80.0

Category 2: price/sqft > 80.0 & price/sqft <=130.0

Category 3: price/sqft >130.0 & price/sqft <=180.0

Category 4: price/sqft >180.0 & price/sqft <=230.0

We then created a 4 indicator variables (group1, group2, group3, group4) corresponding to 4 categories above.

Table 14 shows how the neighborhoods spread across the groups. Table 14 shows group2 has most observations (464) and 20 out of the 21 neighborhoods have group 2 houses. Group 1 follows next with 225 observations and 16 out of 21 neighborhoods have group 1 houses. Group 3 has 56 observations and 11 neighborhoods have group 3 houses. Group 4 has 2 observations with 2 neighborhoods associated with it.

Neighborhood	Group 1	Group 2	Group 3	Group 4
Blmngtn	0	1	0	0
BrkSide	2	0	0	0
ClearCr	3	4	0	0
CollgCr	11	91	16	0
Crawfor	5	4	0	0
Edwards	24	18	1	0
Gilbert	7	48	3	0
IDOTRR	4	1	0	0
Mitchel	10	25	1	0
NAmes	97	80	4	0
NoRidge	1	27	1	0
NridgHt	0	33	14	1
NWAmes	18	28	0	0
OldTown	4	10	0	0
Sawyer	22	21	0	0
SawyerW	8	38	4	0
Somerst	0	7	7	0
StoneBr	0	3	3	0
SWISU	0	2	0	0
Timber	6	16	2	1
Veenker	3	7	0	0

Table 14 Table showing distribution among the 4 groups based on price/Sqft per neighborhood

We then re-fitted the multiple regression model in Section 5 with the group indicator variables. Among the group indicator variables, we use group 4 (the category of houses with price/sqft > 180.0 but less than <=230.0) as the ***base reference category***.

The resulted fitted multiple regression equation is

SalePrice = 9.01E+04 + 9.91E+01 * GrLivArea + 1.18E+02 * GarageArea -1.25E+05 * group1 - 9.96E+04 * group2 - 8.74E+04 * group3

Table 15 below shows the regression output for the model.

The coefficients of the regression equation depend on the category left out (the base reference category which is group4 in our case) because each indicator variable coefficient is interpreted relative to the base category.

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.01E+04	2.71E+04	3.33	0.000912 ***
GrLivArea	9.91E+01	3.43E+00	28.907	< 2e-16 ***
GarageArea	1.18E+02	9.34E+00	12.621	< 2e-16 ***
group1	-1.25E+05	2.65E+04	-4.73	2.69e-06 ***
group2	-9.96E+04	2.64E+04	-3.781	0.000169 ***
group3	-8.74E+04	2.67E+04	-3.275	0.001106 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 37060 on 741 degrees of freedom
Multiple R-squared: 0.7564, Adjusted R-squared: 0.7547
F-statistic: 460.1 on 5 and 741 DF, p-value: < 2.2e-16
Residuals:

Min	1Q	Median	3Q	Max
-100955	-21976	-3789	16032	274487

Table 15 Regression output for the linear regression model (SalePrice ~ GrLivArea + GarageArea + group1 + group2 + group3)

Analysis of the regression output:

Based on the regression output in Table 15, we are able to conclude that the overall regression is strongly statistically significant since the probability (p-value) of the overall F-Test is 2.2e-16 indicating at a level of significance of 0.001 we can reject the null hypothesis that **all** the predictor variables have no effect on the regression.

The coefficient of determination (Adjusted R-squared) value of 0.7547 indicates approximately 75.47% of the variation in the SalePrice can be explained by the variation in the group indicator variables and the GrLivArea and GarageArea variables.

In Table 15, we determine GrLivArea and GarageArea are statistically significant with p-value < 2e-16. So, at a significance level of 0.001, we can reject the null hypothesis that individual predictor variable GrLivArea or GarageArea has no effect on the regression.

The same thing is also true from the T-test results of group1, group2, and the intercept. They are also statistically significant at significance level 0.001. Only group 3 is statistically significant at a slightly higher significance level of 0.01.

Except for the coefficients of GrLivArea and GarageArea, the standard error for the intercept, group1, group2 and group3 variables is high (> 26500) which indicates considerable error in them.

Analysis of Goodness-of-Fit:

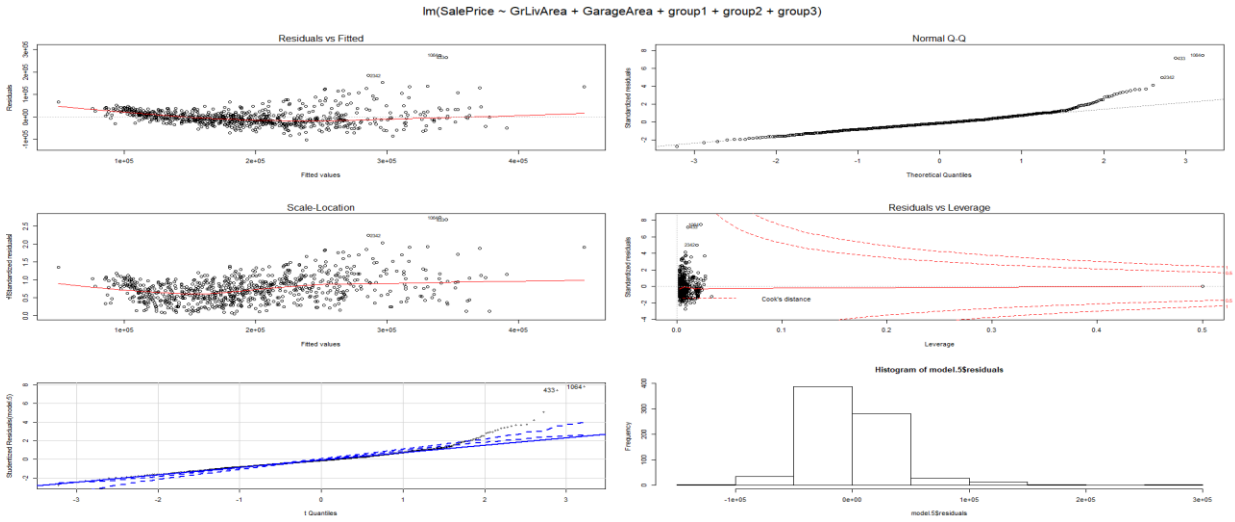

FIG 07. Residual diagnostic plots (Residual vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage, Residual vs, studentized qqplot, Histogram of Residuals) for the linear regression model (SalePrice ~ GrLivArea + GarageArea + group1 + group2 + group3 + group4)

In FIG 07, from the Residual vs Fitted scatter plot, we noticed the residuals do not show a horizontal band formation around the residual = 0 indicating the error variances are not constant. The plot shows a somewhat funnel distribution with outliers.

Both the Normal Q-Q plot and the studentized Q-Q plot in FIG 07 indicate the residuals deviate from normal (Gaussian) distribution. The residuals have long tails with right skewness.

The histogram plot of the residuals in FIG 07 does not show a symmetric bell-shaped histogram. It also confirms the presence a tail to the right of the histogram.

Therefore, there are concerns with using this model for statistical inference since all the GOF conditions are not met. As a result, we are not confident in the t-test scores, p-values, and overall F-test score provided in Table 15.

Analysis of predictive accuracy:

Error	Value using training data	Value using test data
MSE (measure of average error between the predicted values and observed values)	1,362,084,594	1,455,623,515
MAE (the average magnitude of the errors in the predictions, without considering their direction)	25,572.85	26,651.3

Table 16 MSE and MAE values to assess predictive accuracy of SalePrice ~ GrLivArea + GarageArea + group1 + group2 + group3 model

For predictive accuracy, the mean square error and the mean absolute error were computed using the residuals. Table 16 lists the values of MSE and MAE using training data. The test data shows higher MSE and MAE values. The large values of both errors indicate a large model prediction error.

Determination of the better fit model to predict SalePrice – multiple regression model (SalePrice ~ GrLivArea+GarageArea) or multiple regression model (SalePrice ~ GrLivArea + GarageArea + group1 + group2 + group3)

The criteria employed to evaluate the better fit model to predict SalePrice are:

1. the adjusted R-squared values
2. coefficient P-values for the predictor variables
3. the MSE and MAE errors
4. Partial F-Test results

	Model #3 **SalePrice ~ GrLivArea + GarageArea**	**Model #5** **SalePrice ~ GrLivArea + GarageArea + group1 + group2 + group3**
Adjusted R-squared	0.7191	0.7547
Coefficient p-values	Intercept 0.000104*** GrLivArea < 2e-16 *** GarageArea < 2e-16 ***	Intercept 0.000912 *** GrLivArea < 2e-16 *** GarageArea < 2e-16 *** Group1 2.69e-06 *** Group 2 0.000169 *** Group 3 0.001106 ***
MSE & MAE Errors (using test data)	MSE - 1,610,439,286 MAE - 26,838.85	MSE - 1,455,623,515 MAE - 26,651.3

Table 17 Model comparison table with adjusted R-squared, p-values, MSE&MAE errors

Based the information in Table 17, we determined that Model #5 has a better adjusted R-squared value than Model #3. Both models have the T-tests for the individual coefficients statistically significant at a level of significance 0.001. Additionally, Model #5 has better MSE and MAE values using the test data.

Additionally, the results of the partial F-test for multiple variable exclusion show that at a significance level of 0.001, that the additional coefficients in Model #5 are statistically significant.

Partial F-test result between Model #3 and Model #5 showed that at significance level of 0.001, it allows us to reject the reduced model (null hypothesis which is Model#3). Table 18 shows the Anova output.

Model 3: SalePrice ~ GrLivArea + GarageArea
Model 5: SalePrice ~ GrLivArea + GarageArea + group1 + group2 + group3

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
3	744	1.1474e+12				
5	741	1.0175e+12	3	1.299e+11	31.534	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 18 Anova output of Model #3 and Model #5

Based on the above analysis, we conclude Model #5 (SalePrice ~ GrLivArea+GarageArea + group1 + group2 + group3) is the better fit model to predict SalePrice compared to Model #3 (SalePrice ~ GrLivArea+GarageArea)

Code Snippet 06 in section 9 contains the R code for neighborhood accuracy analysis.

7. SalePrice versus Log SalePrice as the Response

In this section, we present 3 models:

a. A multiple regression model with SalePrice as response variable with 4 continuous predictor variables and 2 discrete variables.
b. A multiple regression model with log(SalePrice) as response variable with same 4 continuous predictor variables and 2 discrete variables.
c. A multiple regression model with log(SalePrice) as response variable with a transformation applied to one of the predictor variables used in a. and b.

7.1. SalePrice Model

For this model, we chose the continuous and discrete variables listed in Table 19:

Variable Name	DataType	Description
TotalSqftCalc	Continuous	=BsmtFinSF1 + BsmtFinSF2 +GrLivArea. Total Area in Sqft in the house
MasVnrArea	Continuous	Masonry veneer area in square feet
LotFrontage	Continuous	Linear feet of street connected to property
GarageArea	Continuous	Size of garage in square feet
YearBuilt	Discrete	Original construction date
TotRmsAbvGrd	Discrete	Total rooms above grade

Table 19 List of variables to use in the model

The fitting of the linear regression model with the observations of the predictor variables to predict SalePrice provided the below fitted equation:

SalePrice = -2.340e+06 + 5.01e+01 * TotalSqftCalc + 5.76e+01 * MasVnrArea + 3.29e+02 * LotFrontage + 6.93e+01 * GarageArea + 1.17e+03 * YearBuilt + 5.32e+03 * TotRmsAbvGrd

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.34E+06	1.38E+05	-17.02	< 2e-16 ***
TotalSqftCalc	5.01E+01	2.26E+00	22.142	< 2e-16 ***
MasVnrArea	5.76E+01	8.00E+00	7.206	1.42e-12 ***
LotFrontage	3.29E+02	7.42E+01	4.43	1.08e-05 ***
GarageArea	6.93E+01	8.43E+00	8.219	9.14e-16 ***
YearBuilt	1.17E+03	7.07E+01	16.603	< 2e-16 ***
TotRmsAbvGrd	5.32E+03	1.17E+03	4.56	5.98e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 31410 on 740 degrees of freedom
Multiple R-squared: 0.8252, Adjusted R-squared: 0.8238
F-statistic: 582.3 on 6 and 740 DF, p-value: < 2.2e-16

Residuals:

Min	1Q	Median	3Q	Max
-112047	-18951	-2206	15909	200304

Table 20 Regression output for the linear regression model (SalePrice ~ TotalSqftCalc + MasVnrArea + LotFrontage + GarageArea + YearBuilt + TotRmsAbvGrd)

We next analyzed the results of the Overall F-test and the T-test results (local tests) for a regression effect of the predictor variables using results in Table 20.

Analysis of the regression output:

The overall F test is to understand the regression effect of all the predictor variables. From Table 18, we noted since the Overall F test p-value is very low. So, the overall regression is statistically significant at a level of significance indicating at least one of the predictor variables have an effect on the regression.

The adjusted R-squared value is 0.8238 which indicates 82.38% of the variation in the response variable can be explained by variations in the predictor variables.

The coefficients listed in Table 20 indicate that given all else in the model is held fixed, one additional sqft for TotalSqftCalc is associated with an estimated SalePrice increase of $50.1. Similarly, under similar conditions where everything else in the model is held fixed, an additional sqft for MasVnrArea gives an estimated SalePrice increase of $57.6. Likewise, under similar conditions, when other variables are held constant, LotFrontage, GarageArea, YearBuilt, and TotRmsAbvGrd contribute to an increase of $329, $69.3, $1170, and $5320 for a unit increase in the variables respectively.

The value of the intercept is estimated price of the SalePrice when all the variables are set to 0.

All the coefficients except for TotRmsAbvGrd and the intercept have really low standard error indicating high precision in the coefficients. TotRmsAbvGrd has a standard error of 1170 and the intercept has a standard error of 138000 indicating quite a high error in both the terms.

The t values and p values indicate that all the coefficients (including the intercept) are statistically significant at a level of significance of 0.001 rejecting the null hypothesis that the individual coefficients of the predictor variables do not have an effect.

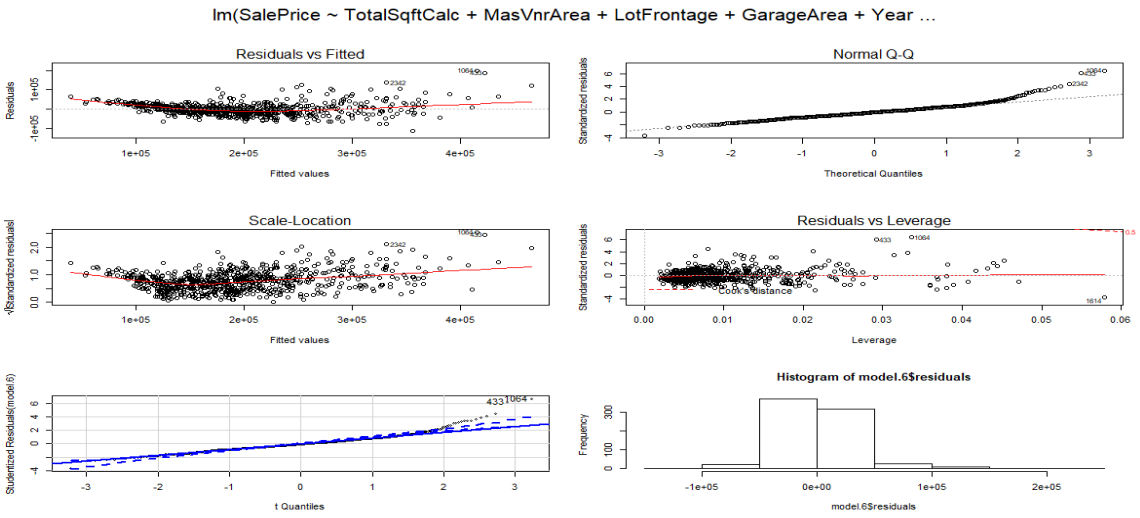

FIG 08. Residual diagnostic plots (Residual vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage, Residual vs, studentized Q-Q plot, Histogram of Residuals) for the simple linear regression model (SalePrice ~ TotalSqftCalc+MasVnrArea+LotFrontage+GarageArea+YearBuilt+TotRmsAbvGrd)

Analysis of the Goodness-of-fit of the model:

In the FIG 08, we studied the residual vs fitted plot and see a residuals form a slight curve instead of horizontal bar of random bouncing residuals around residual=0. There is also a spread out of the residuals as the fitted values increase. This indicates there is non-constant variance.

In the Normal Q-Q plot, a deviation from normal(Gaussian) distribution is noted with a right tail. The same is noted in the histogram of the residuals and the studentized residual Q-Q plot.

Since the analysis of the Goodness-of-fit for the regression reveals some of the underlying probabilistic assumptions for OLS are not met, we are not confident in the statistical inference provided in Table 20.

Analysis of predictive accuracy:

Error	Value using training data	Value using test data
MSE (measure of average error between the predicted values and observed values)	977,087,849	990,272,260
MAE (the average magnitude of the errors in the predictions, without considering their direction)	22,601.73	22,379.01

Table 21 MSE and MAE values to assess predictive accuracy of the SalePrice ~ TotalSqftCalc + MasVnrArea + LotFrontage + GarageArea + YearBuilt + TotRmsAbvGrd model

For predictive accuracy, the mean square error and the mean absolute error were computed using the residuals. Table 21 lists the values of MSE and MAE obtained using training and test data. The values of MSE and MAE are higher for test data compared to training data. The large values of both errors indicate a large model prediction error.

However, we noticed, adding more predictor variables is reducing the MSE and MAE as this model has the lowest values compared to the single linear regression models and the multiple linear regressions models described thus far. Next, we apply a transformation to the response variable to determine if it aids in the improvement of the predictive accuracy.

7.2. Log(SalePrice) Model

We built another model with Log(SalePrice) as the predictor variable but with the same list of predictor variables listed in Table 19 (with 4 continuous variables and 2 discrete variables).

The fitting of the linear regression model with the observations of the predictor variables to predict Log(SalePrice) gives the below fitted equation:

Log(SalePrice) = -1.57 + 2.13E-04 *TotalSqftCalc + 1.65E-04 * MasVnrArea + 1.41E-03 * LotFrontage + 3.22E-04 * GarageArea + 6.42E-03 * YearBuilt + 3.40E-02 * TotRmsAbvGrd

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.57E+00	5.50E-01	-2.846	0.00455 **
TotalSqftCalc	2.13E-04	9.06E-06	23.458	< 2e-16 ***
MasVnrArea	1.65E-04	3.20E-05	5.144	3.44e-07 ***
LotFrontage	1.41E-03	2.97E-04	4.756	2.38e-06 ***
GarageArea	3.22E-04	3.38E-05	9.534	< 2e-16 ***
YearBuilt	6.42E-03	2.83E-04	22.682	< 2e-16 ***
TotRmsAbvGrd	3.40E-02	4.67E-03	7.279	8.60e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.1257 on 740 degrees of freedom
Multiple R-squared: 0.8625, Adjusted R-squared: 0.8614
F-statistic: 773.6 on 6 and 740 DF, p-value: < 2.2e-16
Residuals:

Min	1Q	Median	3Q	Max
-0.74538	-0.07492	-0.00167	0.06943	0.52185

Table 22 Regression Output for the multiple regression model (LogSalePrice ~ TotalSqftCalc + MasVnrArea + LotFrontage + GarageArea + YearBuilt + TotRmsAbvGrd)

Analysis of the regression output:

The overall F test is to understand the regression effect of all the predictor variables. From Table 22, we concluded that the overall regression to predict LogSalePrice is statistically significant indicating at least one of the predictor variables have an effect on the regression since the overall F test p-value is very low compared to the level of significance of 0.001.

The adjusted R-squared value is 0.8614 which indicates 86.14% of the variation in the response variable can be explained by variations in the predictor variables.

The coefficients listed in Table 22 indicate that given all else in the model is held fixed, one additional sqft for TotalSqftCalc is associated with an estimated LogSalePrice increase of 0.000213. Similarly, under similar conditions when all else is held constant with the model, an additional sqft for MasVnrArea gives an estimated LogSalePrice increase of 0.000165. Likewise, when other variables are held constant, LotFrontage, GarageArea, YearBuilt, and TotRmsAbvGrd contribute to an increase of 0.00141, 0.000322, 0.00642, and 0.034 to the LogSalePrice for a unit increase in the variables respectively. The value of the intercept (-1.57) is estimated value of the LogSalePrice when all the variables are set to 0. The standard error for the coefficients is low except for the intercept which is several magnitudes higher in comparison.

The t values and p values indicate that all the coefficients (except for the intercept) are statistically significant at the level of significance of 0.001. Therefore, we reject the null hypothesis that the individual coefficient for these predictor variables has no effect on the regression. The p-value for the intercept is 0.00455. We can only reject the null hypothesis that the intercept is 0 only at a higher level of significance of 0.01.

Next we looked at the diagnostic plots to determine if the regression output can be used with confidence.

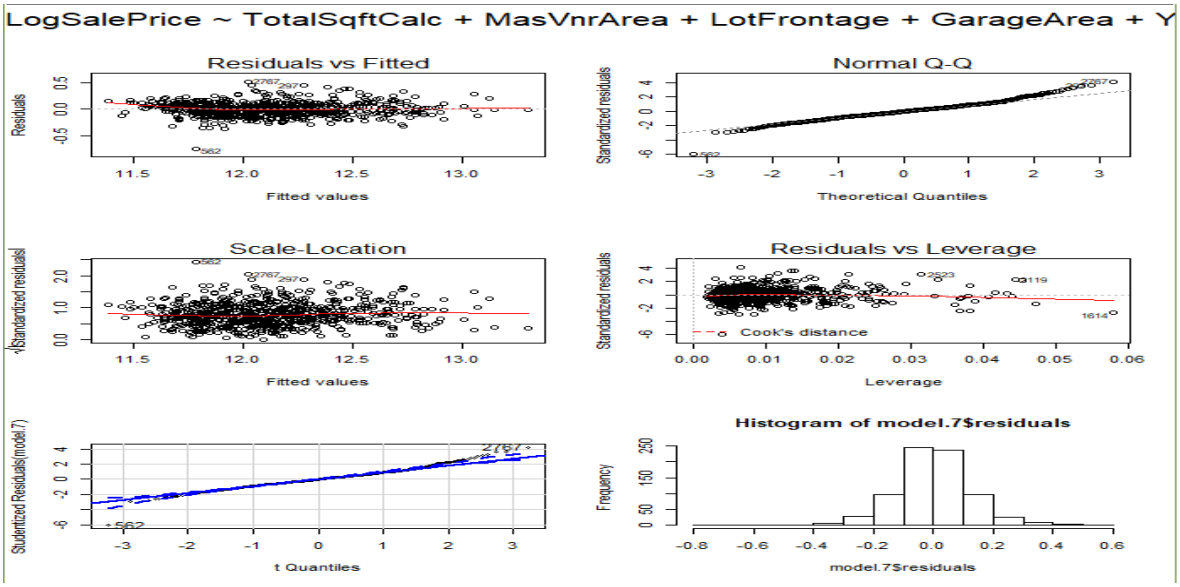

FIG 09. Residual diagnostic plots (Residual vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage, Residual vs, studentized Q-Qplot, Histogram of Residuals) for the simple linear regression model (LogSalePrice ~ TotalSqftCalc+MasVnrArea+LotFrontage+GarageArea+YearBuilt+TotRmsAbvGrd)

Analysis of the Goodness-of-fit of the model:

Using the FIG 09, we studied the residual vs fitted plot and see the residuals form a slight curve initially but subsequently there is a horizontal bar of random bouncing residuals around residual=0. There are outliers in the plot. This indicates there are some non-constant variations in the variance. But as the fitted values increase (closer to 12.0 and higher), a somewhat linear pattern emerges indicating between the response and the predictor variables. There is no distinct funnel type distribution seen as in the case of previous model diagnostic residual plots which indicates heteroskedasticity.

In the Normal Q-Q plot, only a slight deviation from normal (Gaussian) distribution is noted with small tails. The same is noted in the histogram of the residuals (for most extent, there is symmetry) and also in the studentized residual Q-Q plot.

This model comes closest to the meeting the underlying probabilistic assumptions for ordinary least squares regression. Though slight deviations exist, we conclude that we can use this model to infer the log value of the SalePrice.

Next, we evaluate the predictive accuracy metrics of the model.

Analysis of predictive accuracy:

Error	Value using training data (in the scale of SalePrice)	Value using test data (in the scale of SalePrice)
MSE (measure of average error between the predicted values and observed values)	743,768,509	728,227,238
MAE (the average magnitude of the errors in the predictions)	18,677.9	18,560.56

Table 23 MSE and MAE values to assess predictive accuracy of the LogSalePrice ~ TotalSqftCalc + MasVnrArea + LotFrontage + GarageArea + YearBuilt + TotRmsAbvGrd model

For predictive accuracy, the mean square error and the mean absolute error were computed using the residuals in the same scale of SalePrice. Table 23 lists the values of MSE and MAE obtained using training and test data. The MSE and MAE values for model are better for the test data than for the training data. **As mentioned in the earlier section, adding more predictor variables thus far had the effect of improving the predictive accuracy. This also turned out to be true when a transformation is applied to the response variable**.

7.3. Log(SalePrice) Model with applied transformation to TotalSqftCalc

Though there is no heteroscedasticity evident in the residual plots shown in section 7.2, as discussed in section 4.1, we determined among the continuous predictor variables used in the multiple linear regression model, there are variables with some right skewness. One of the techniques that help to make the variables symmetric is transformation. One of the commonly used transformation for right-skewed data is a logarithmic transformation. We applied that to the variable TotalSqftCalc and refit the multiple linear regression model.

The resultant fitted linear regression equation is

LogSalePrice = -4.37 + 4.29E-01 * LogTotalSqftCalc + 2.15E-04 * MasVnrArea + 1.58E-03 * LotFrontage + 3.31E-04 * GarageArea + 6.40E-03 * YearBuilt + 3.59E-02 * TotRmsAbvGrd

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.37E+00	6.08E-01	-7.186	1.63e-12 ***
LogTotalSqftCalc	4.29E-01	2.05E-02	20.917	< 2e-16 ***
MasVnrArea	2.15E-04	3.31E-05	6.491	1.57e-10 ***
LotFrontage	1.58E-03	3.10E-04	5.099	4.34e-07 ***
GarageArea	3.31E-04	3.54E-05	9.355	< 2e-16 ***
YearBuilt	6.40E-03	2.97E-04	21.582	< 2e-16 ***
TotRmsAbvGrd	3.59E-02	4.91E-03	7.322	6.38e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.1316 on 740 degrees of freedom
Multiple R-squared: 0.8493, Adjusted R-squared: 0.8481
F-statistic: 695.2 on 6 and 740 DF, p-value: < 2.2e-16
Residuals:

Min	1Q	Median	3Q	Max
-0.75022	-0.07717	-0.00406	0.07669	0.51148

Table 24 Regression Output for the Simple Regression Model LogSalePrice ~ LogTotalSqftCalc + MasVnrArea + LotFrontage + GarageArea + YearBuilt + TotRmsAbvGrd model

Analysis of the regression output:

Based on the regression output in Table 24, we are able to conclude that the overall regression is strongly statistically significant since the probability (p-value) of the F-Test is less than 2.2e-16. Based on this, at a level of significance of 0.001, we can reject the null hypothesis that **all** the predictor variables have a zero coefficient and have no effect on the regression.

The coefficient of determination (Adjusted R-squared) value of 0.8481 indicates approximately 84.81% of the variation in the LogSalePrice can be explained by variation in the predictor variables which include the LogTotalSqftCalc variable.

The coefficients listed in Table 24 indicate that given all else in the model is held fixed, one unit increase in LogTotalSqftCalc is associated with an estimated LogSalePrice increase of 0.429. Likewise, when other variables are held constant, MasVnrArea, LotFrontage, GarageArea, YearBuilt, and TotRmsAbvGrd contribute to an increase of 0.000215, 0.00158, 0.000331, 0.0064, and 0.0359 to the LogSalePrice for a unit increase in the variables respectively. The value of the intercept (-4.37) is estimated value of the LogSalePrice when all the variables are set to 0. The standard errors are low for all the coefficients except for the intercept in comparison.

The t values and p values indicate that all the coefficients (including the intercept) are statistically significant at the level of significance of 0.001. Therefore, we reject the null hypothesis that the individual coefficient for these predictor variables and the intercept has no effect on the regression.

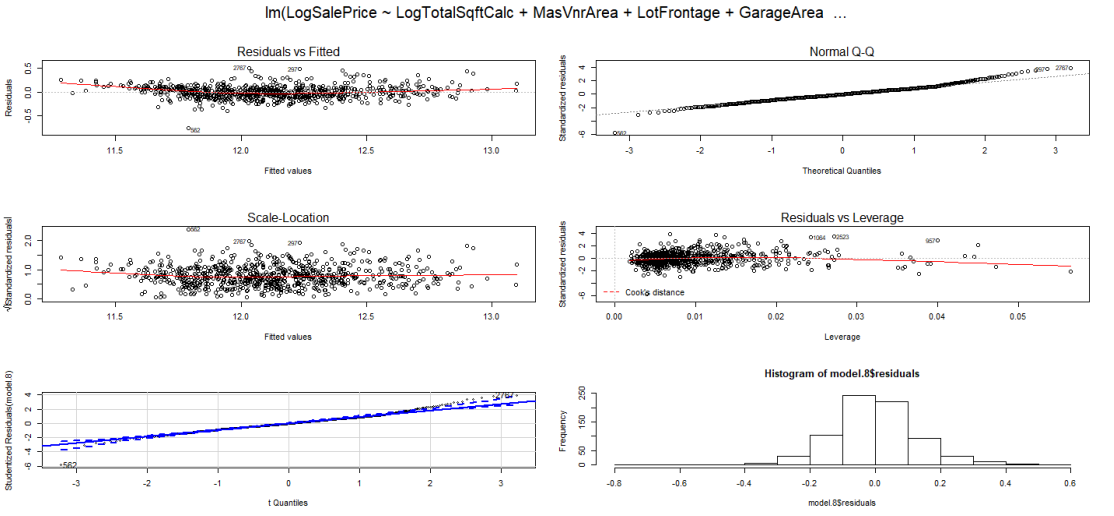

FIG 10. Residual diagnostic plots (Residual vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage, Residual vs, studentized Q-Qplot, Histogram of Residuals) for the simple linear regression model (LogSalePrice ~ LogTotalSqftCalc+MasVnrArea+LotFrontage+GarageArea+YearBuilt+TotRmsAbvGrd)

Analysis of the Goodness-of-fit of the model:

Using the FIG 10, in the residuals vs Fitted plot, the residuals form a slight curve initially. There are also outliers in the plot. The middle part of the plot shows random horizontal bar pattern around residual=0 line. However, at the ends, there is departure from that and the smoother line shows a slight curvature. This indicates there are some non-constant variations in the variance.

In the Normal Q-Q plot, only a slight deviation from normal (Gaussian) distribution is noted with small tails. The same is noted in the histogram of the residuals (there is only a slight deviation from symmetry) and also in the studentized residual Q-Q plot.

Next, we evaluate the predictive accuracy metrics of the model.

Analysis of predictive accuracy:

Error	Value using training data (in the scale of SalePrice)	Value using test data (in the scale of SalePrice)
MSE (measure of average error between the predicted values and observed values)	862,793,223	860,555,955
MAE (the average magnitude of the errors in the predictions)	19,570.12	19,602.15

Table 25 MSE and MAE values to assess predictive accuracy of the LogSalePrice ~ LogTotalSqftCalc + MasVnrArea + LotFrontage + GarageArea + YearBuilt + TotRmsAbvGrd model

For predictive accuracy, the mean square error and the mean absolute error were computed using the residuals in the same scale of SalePrice. Table 25 lists the values of MSE and MAE obtained using training and test data. The MSE value for the model is better for the test data than for the training data whereas MAE value for the model is worse with test data.

7.4. Comparison and Discussion of Model Fits

In this section, we present the results of comparison between the three models described in sections 7.1, 7.2 and 7.3.

The criteria used for the evaluation of the models are:

Adjusted R-squared scores
Coefficient P-values
MSE/MAE error values

	Model # 6 (SalePrice ~ TotalSqftCalc + MasVnrArea + LotFrontage + GarageArea + YearBuilt + TotRmsAbvGrd)	Model #7 (LogSalePrice ~ TotalSqftCalc + MasVnrArea + LotFrontage + GarageArea + YearBuilt + TotRmsAbvGrd)	Model #8 (LogSalePrice ~ LogTotalSqftCalc + MasVnrArea + LotFrontage + GarageArea + YearBuilt + TotRmsAbvGrd)
Adjusted R-squared	0.8238	0.8614	0.8481
Coefficient p-values	(Intercept) < 2e-16 *** TotalSqftCalc < 2e-16 *** MasVnrArea 1.42e-12 *** LotFrontage 1.08e-05 *** GarageArea 9.14e-16 *** YearBuilt < 2e-16 *** TotRmsAbvGrd 5.98e-06 ***	(Intercept) 0.00455 ** TotalSqftCalc < 2e-16 *** MasVnrArea 3.44e-07 *** LotFrontage 2.38e-06 *** GarageArea < 2e-16 *** YearBuilt < 2e-16 *** TotRmsAbvGrd 8.60e-13 ***	(Intercept) 1.63e-12 *** TotalSqftCalc < 2e-16 *** MasVnrArea 1.57e-10 *** LotFrontage 4.34e-07 *** GarageArea < 2e-16 *** YearBuilt < 2e-16 *** TotRmsAbvGrd 6.38e-13 ***
MSE & MAE (using test data)	MSE - 990,272,260 MAE - 22,379.01	MSE - 728,227,238 MAE - 18,560.56	MSE - 860,555,955 MAE - 19,602.15

Table 26 Comparison table for models with adjusted R-squared, coefficient P-values, predictive accuracy errors

From Table 26, we can determine that Model #7 with the LogSalePrice as the response variable and no transformations on the predictor variables has the best adjusted R-square. The Coefficient

P-values for the predictor variables for all three models as shown in the table are statistically significant at the level of significance of 0.001. The intercept for Model #7 is however statistically significant at a higher level of significance of 0.01. Also, for predictive accuracy (computed using test data), Model #7 has the best MSE and MAE values indicating it has the least predictive error of the 3 models. The transformation applied to the response variable has certainly shown improvement in the prediction of the home price values. Contrary to that, transformation of the predictor variable did not help. It affected the model adversely which is evident from the reduced adjusted R-squared value and also from the MSE and MAE values.

Based on this, we concluded that Model #7 is the better model among all the models described in this report to predict home price values.

Code Snippet 07 in section 9 contains the R code for SalePrice vs LogSalePrice models

8. Summary

We began the analysis with the definition of the sample for which we used a waterfall based drop conditions. The drop conditions were created based on the definition of a 'typical' house and also by reading the data dictionary.

After defining the sample, we applied exploratory data techniques to identify the two potential continuous predictor variables for use in the single linear regression models. We used both graphical and non-graphical techniques to identify the correlation between response and predictor variables for the selection of the variables.

Subsequently, we built two simple linear regression models using the identified predictor variables. After that, we built a multiple linear regression model with the same predictor variables. When a comparison was done, it was evident the multiple linear regression model fared better among the 3 models in terms of the amount of variation that is explained in the response variable and the predictive accuracy.

After that, we performed neighborhood accuracy tests and added indicator variables to the multiple linear regression model which improved the predictive accuracy.

Next, we built three models with four continuous predictor variables and two discrete variables - one with no transformations to either predictor or response variables, one with logarithmic transformation to the response variable, and one with a logarithmic transformation to the response and a predictor variable.

The comparison study of the 3 models provided an interesting result that the model with the logarithmic transformation of the response variable is the better model of the 3 models to predict the price of the home sales. The model with the logarithm transformation to the response variable has the highest adjusted r-squared value and the lowest predictive accuracy error values.

9. Code

9.1. Sample Definition

```
# Read in csv file for Ames housing data;
path.name <- 'C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week3\\Assignment-03\\';
#path.name <- '//Users//ttg043//Desktop//Assignment_03//externalsender//';
file.name <- paste(path.name,'ames_housing_data.csv',sep='');

# Read in the csv file into an R data frame;
ames.df <- read.csv(file.name,header=TRUE,stringsAsFactors=FALSE);

# Create a waterfall of drop conditions to define a sample population
ames.df$dropCondition <- ifelse((ames.df$Zoning!='RH' & ames.df$Zoning!='RL' & ames.df$Zoning!='RP' &
ames.df$Zoning!='RM'),'01: Not Residential',
  ifelse(ames.df$SaleCondition!='Normal','02: Non-Normal Sale',
  ifelse(ames.df$BldgType!='1Fam','03: Not Single Family Home',
  ifelse(ames.df$Street!='Pave','04: Street Not Paved',
  ifelse((ames.df$Utilities!='AllPub' & ames.df$Utilities!='NoSewr'),'05: No Water',
  ifelse(ames.df$YearBuilt <1950,'06: Built Pre-1950',
  ifelse(ames.df$TotalBsmtSF <1,'07: No Basement',
  ifelse((ames.df$GrLivArea <800 | ames.df$GrLivArea >4000),'08: LT 800 OR GT 4000 SqFt',
  '99: Eligible Sample'
  ))))))));

# Save the counts of the drop condition rules
waterfall <- table(ames.df$dropCondition);

# convert waterfall table as a matrix
as.matrix(waterfall,9,1)

# Eliminate all observations that are not part of the eligible sample population;
eligible.population <- subset(ames.df,dropCondition=='99: Eligible Sample');

# Check that all remaining observations are eligible and obtain the count of the eligible population
table(eligible.population$dropCondition)

#save the eligible.population as RDS file
saveRDS(eligible.population,file='C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week3\\Assignment-03\\population.rds');
```

Code Snippet 01: R code for creation of the eligible population sample

9.2. Exploratory data analysis for the selection of the continuous predictor variables

```
#read the eligible.population from RData file
eligible.population <- readRDS('C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week3\\Assignment-03\\population.rds')

# Make a vector of the names;
keep.vars <- c('GrLivArea','LotArea','TotalBsmtSF','MasVnrArea','LotFrontage','GarageArea','SalePrice')

# Restrict the sample to just the columns of interest
skinny.df <- eligible.population[,keep.vars];

# Use the structure command to view the contents of the data frame;
str(skinny.df)

##############################################################################################
# Delete observations with missing values
##############################################################################################
sample.df <- na.omit(skinny.df);

# parameters for the display of the plots for continuous variables
par(mfrow=c(6,2), mar=c(4.1,4.1,1.1,2.1))

# EDA for the GrLivArea variable
# five number summary for GrLivArea
summary(sample.df$GrLivArea)
# boxplot of GrLivArea
boxplot(sample.df$GrLivArea, ylab='Ground Living Area', main='Box plot of GrLivArea (SQFT)')
# scatterplot of GrLivArea vs SalePrice
plot(sample.df$GrLivArea,sample.df$SalePrice/1000,xlab='GrLivArea in SQFT',ylab='Sale Price (000)',main='plot of GrLivArea (SQFT) and Sale Price (000)')
# correlation coefficient between GrLivArea and SalePrice
cor(sample.df$GrLivArea, sample.df$SalePrice)

# EDA for the LotArea variable
# five number summary for LotArea
summary(sample.df$LotArea)
# boxplot of LotArea
boxplot(sample.df$LotArea, ylab='Lot Area', main='Box plot of LotArea (SQFT)')
# scatterplot of LotArea vs SalePrice
plot(sample.df$LotArea,sample.df$SalePrice/1000,xlab='LotArea in SQFT',ylab='Sale Price (000)',main='plot of LotArea (SQFT) and Sale Price (000)')
# correlation coefficient between LotArea and SalePrice
cor(sample.df$LotArea, sample.df$SalePrice)

# EDA for the TotalBsmtSF variable
# five number summary for TotalBsmtSF
summary(sample.df$TotalBsmtSF)
# boxplot of TotalBsmtSF
boxplot(sample.df$TotalBsmtSF, ylab='Total Bsmt SF (SQFT)', main='Box plot of TotalBsmtSF (SQFT)')
# scatterplot of TotalBsmtSF vs SalePrice
plot(sample.df$TotalBsmtSF,sample.df$SalePrice/1000,xlab='TotalBsmtSF in SQFT',ylab='Sale Price (000)',main='plot of TotalBsmtSF (SQFT) and Sale Price (000)')
# correlation coefficient between TotalBsmtSF and SalePrice
cor(sample.df$TotalBsmtSF, sample.df$SalePrice)
```

```
# EDA for the MasVnrArea variable
# five number summary for MasVnrArea
summary(sample.df$MasVnrArea)
# boxplot of MasVnrArea
boxplot(sample.df$MasVnrArea, ylab='MasVnrArea (SQFT)', main='Box plot of MasVnrArea (SQFT)')
# scatterplot of MasVnrArea vs SalePrice
plot(sample.df$MasVnrArea,sample.df$SalePrice/1000,xlab='MasVnrArea in SQFT',ylab='Sale Price (000)',main='plot of
MasVnrArea (SQFT) and Sale Price (000)')
# correlation coefficient between MasVnrArea and SalePrice
cor(sample.df$MasVnrArea, sample.df$SalePrice)


# EDA for the LotFrontage variable
# five number summary for LotFrontage
summary(sample.df$LotFrontage)
# boxplot of LotFrontage
boxplot(sample.df$LotFrontage, ylab='LotFrontage (FT)', main='Box plot of LotFrontage (FT)')
# scatterplot of LotFrontage vs SalePrice
plot(sample.df$LotFrontage,sample.df$SalePrice/1000,xlab='LotFrontage in FT',ylab='Sale Price (000)',main='plot of
LotFrontage (FT) and Sale Price (000)')
# correlation coefficient between LotFrontage and SalePrice
cor(sample.df$LotFrontage, sample.df$SalePrice)


# EDA for the GarageArea variable
# five number summary for GarageArea
summary(sample.df$GarageArea)
# boxplot of GarageArea
boxplot(sample.df$GarageArea, ylab='GarageArea (SQFT)', main='Box plot of GarageArea (SQFT)')
# scatterplot of GarageArea vs SalePrice
plot(sample.df$GarageArea,sample.df$SalePrice/1000,xlab='GarageArea in SQFT',ylab='Sale Price (000)',main='plot of
GarageArea (SQFT) and Sale Price (000)')
# correlation coefficient between GarageArea and SalePrice
cor(sample.df$GarageArea, sample.df$SalePrice)
```

Code Snippet 02: R code for the EDA analysis for the selection of two continuous predictor variables

9.3. Creation of training and test datasets

```
#read the eligible.population from RData file
eligible.population <- readRDS('C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week3\\Assignment-03\\population.rds')

# Make a vector of the names;
keep.vars <-
c('BsmtFinSF1','BsmtFinSF2','GrLivArea','LotArea','TotalBsmtSF','MasVnrArea','LotFrontage','GarageArea','Neighborhood','YearBuilt','TotRmsAbvGrd','SalePrice')

# Restrict the sample to just the columns of interest
skinny.df <- eligible.population[,keep.vars];

# Use the structure command to view the contents of the data frame;
str(skinny.df)

#################################################################################################
# Delete observations with missing values
#################################################################################################
sample.df <- na.omit(skinny.df);

#################################################################################################
# Add additional composite variables
#################################################################################################
# Define total square footage of each house above ground
sample.df$TotalSqftCalc <- sample.df$BsmtFinSF1 + sample.df$BsmtFinSF2 + sample.df$GrLivArea
sample.df$SalePrice.SQFT <- sample.df$SalePrice/sample.df$TotalSqftCalc
sample.df$LogSalePrice <- log(sample.df$SalePrice)
sample.df$LogTotalSqftCalc <- log(sample.df$TotalSqftCalc)
sample.df$LogGarageArea <- log(sample.df$GarageArea)

#################################################################################################
# Add a train/test flag to split the sample
#################################################################################################
sample.df$u <- runif(n=dim(sample.df)[1],min=0,max=1)
sample.df$train <- ifelse(sample.df$u<0.70,1,0)

# Check the counts on the train/test split
table(sample.df$train)

# Check the train/test split as a percentage of whole
table(sample.df$train)/dim(sample.df)[1]

# create a data frame to store the training data
train.df <- subset(sample.df,train==1)

#save the train.df as RDS file
saveRDS(train.df,file='C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week3\\Assignment-03\\train.rds');

# create a data frame to store the training data
test.df <- subset(sample.df,train==0)

#save the test.df as RDS file
saveRDS(test.df,file='C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week3\\Assignment-03\\test.rds');
```

Code Snippet 03: R code for the creation of the training and test datasets

9.4. Simple Linear Regression Models

```
#Read the train.df
train.df <- readRDS('C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week3\\Assignment-03\\train.rds')

#Read the test.df
test.df <- readRDS('C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week3\\Assignment-03\\test.rds')

# include 'car' library
library('car')

##############################################################################################
#       model.1 - SIMPLE REGRESSION MODEL - SalePrice~GrLivArea                              #
##############################################################################################

# Fit a linear regression model with R
model.1 <- lm(SalePrice ~ GrLivArea, data=train.df)

# Display model summary
summary(model.1)

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) for training data
mse.1 <- mean(model.1$residuals^2)
mae.1 <- mean(abs(model.1$residuals))

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) for test data
mse.test.1 <- mean((test.df$SalePrice - predict.lm(model.1, test.df)) ^ 2)
mae.test.1 <- mean(abs(test.df$SalePrice - predict.lm(model.1, test.df)))

# Diagnostic plots for SalePrice~GrLivArea simple regression plot
par(mfrow = c(3, 2), oma = c(0, 0, 2, 0))
plot(model.1)

# Make a scatterplot
plot(train.df$GrLivArea,model.1$residuals)
title('Residual vs Predictor')

# Histogram of the residuals
hist(model.1$residuals)


##############################################################################################
#       model.2 - SIMPLE REGRESSION MODEL - SalePrice~GarageArea                             #
##############################################################################################

# Fit a linear regression model with R
model.2 <- lm(SalePrice ~ GarageArea, data=train.df)

# Display model summary
summary(model.2)

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) for training data
mse.2 <- mean(model.2$residuals^2)
mae.2 <- mean(abs(model.2$residuals))
```

```
# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) for test data
mse.test.2 <- mean((test.df$SalePrice - predict.lm(model.2, test.df)) ^ 2)
mae.test.2 <- mean(abs(test.df$SalePrice - predict.lm(model.2, test.df)))


# Diagnostic plots for SalePrice~GarageArea simple regression plot
par(mfrow = c(3, 2), oma = c(0, 0, 2, 0))
plot(model.2)

# Make a scatterplot
plot(train.df$GarageArea,model.2$residuals)
title('Residual vs Predictor')

# Histogram of the residuals
hist(model.2$residuals)
```

Code Snippet 04: R code for fitting the simple linear regression models to predict the SalePrice

9.5. Multiple Linear Regression Model with two variables

```
#Read the train.df
train.df <- readRDS('C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week3\\Assignment-03\\train.rds')

#Read the test.df
test.df <- readRDS('C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week3\\Assignment-03\\test.rds')

# include 'car' library
library('car')

##############################################################################################
#       model.3 - MULTIPLE REGRESSION MODEL - SalePrice~GrLivArea+GarageArea                 #
##############################################################################################

# Fit a linear regression model with R
model.3 <- lm(SalePrice ~ GrLivArea+GarageArea, data=train.df)

# Display model summary
summary(model.3)

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) using train.df
mse.3 <- mean(model.3$residuals^2)
mae.3 <- mean(abs(model.3$residuals))

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) using test.df
mse.3.test <- mean((test.df$SalePrice - predict.lm(model.3, test.df)) ^ 2)
mae.3.test <- mean(abs(test.df$SalePrice-predict.lm(model.3,test.df)))

# Diagnostic plots for SalePrice~GarageArea simple regression plot
par(mfrow = c(3, 2), oma = c(0, 0, 2, 0))
plot(model.3)

# Use the function qqPlot() in the car package to assess the Studentized residuals
qqPlot(model.3)

# Histogram of the residuals
hist(model.3$residuals)

# partial f-test for model #1 and model.3
anova(model.1, model.3)
# partial f-test for model #2 and model.3
anova(model.2, model.3)
```

Code Snippet 05: R code for fitting the multiple linear regression model to predict the SalePrice with two continuous predictor variable

9.6. Neighborhood Accuracy

```
#Read the train.df
train.df <- readRDS('C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-
DL\\Week3\\Assignment-03\\train.rds')

#Read the test.df
test.df <- readRDS('C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week3\\Assignment-
03\\test.rds')

#Use the library car
library('car')

##############################################################################################
#       model.4 - SIMPLE REGRESSION MODEL - SalePrice~Neighborhood                           #
##############################################################################################

# Fit a linear regression model with R
model.4 <- lm(SalePrice ~ Neighborhood, data=train.df)

# Display model summary
summary(model.4)

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) using train.df
mse.4 <- mean(model.4$residuals^2)
mae.4 <- mean(abs(model.4$residuals))

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) using test.df
mse.4.test <- mean((test.df$SalePrice - predict.lm(model.4, test.df)) ^ 2)
mae.4.test <- mean(abs(test.df$SalePrice-predict.lm(model.4,test.df)))

# create the boxplot residuals vs Neighborhood a
boxplot(model.4$residuals~Neighborhood, data=train.df, las=2, ylab='residuals', main='Boxplot of residuals vs
Neighborhood')
means <- tapply(model.4$residuals,train.df$Neighborhood,mean)
points(means,col="red",pch=18)

#compute mae using aggregate function
neighborhood.mae <- aggregate(abs(model.4$residuals), by=list(Neighborhood=train.df$Neighborhood), FUN=mean)
colnames(neighborhood.mae) <- c('Neighborhood','MAE')

#compute pricesqft using aggregate function
neighborhood.pricesqft <- aggregate((train.df$SalePrice/train.df$TotalSqftCalc),
by=list(Neighborhood=train.df$Neighborhood), FUN=mean)
colnames(neighborhood.pricesqft) <- c('Neighborhood','meanPriceSQFT')

#combine the above two data frames into one data frame
neighborhood.mae.pricesqft <- cbind(neighborhood.mae,neighborhood.pricesqft)

#create a plot of the mean pricseSQFT vs mae
plot(neighborhood.mae.pricesqft$meanPriceSQFT,neighborhood.mae.pricesqft$mae,xlab='Mean Price/SQFT',
ylab='MAE', main='Plot of mean Price/SQFT vs MAE per neighborhood')
text(neighborhood.mae.pricesqft$meanPriceSQFT, neighborhood.mae.pricesqft$mae, labels =
neighborhood.mae.pricesqft$Neighborhood, pos = 4)

#creates the four categories based on price/sqft
group.ind <- function(val) {
```

```
 y <-
ifelse((val>=30.0&val<=80.0),1,ifelse((val>80.0&val<=130.0),2,ifelse((val>130.0&val<=180.0),3,ifelse((val>180.0&val<=23
0.0),4,5))))
 return(y)
}
# generate the group for each observations in the training data
train.df$group <- group.ind(train.df$SalePrice.SQFT)

# distribution table groups vs neighborhood in the training data
table(train.df$group,train.df$Neighborhood)

# generate the indicator variables in the training data
train.df$group1 <- ifelse(train.df$group==1,1,0);
train.df$group2 <- ifelse(train.df$group==2,1,0);
train.df$group3 <- ifelse(train.df$group==3,1,0);
train.df$group4 <- ifelse(train.df$group>=4,1,0);


# generate the group for each observations in the test data
test.df$group <- group.ind(test.df$SalePrice.SQFT)

# distribution table groups vs neighborhood in the test data
table(test.df$group,test.df$Neighborhood)

# generate the indicator variables in the test data
test.df$group1 <- ifelse(test.df$group==1,1,0);
test.df$group2 <- ifelse(test.df$group==2,1,0);
test.df$group3 <- ifelse(test.df$group==3,1,0);
test.df$group4 <- ifelse(test.df$group>=4,1,0);

#########################################################################################################
# model.5 - MULTIPLE REGRESSION MODEL - GrLivArea+GarageArea+group1+group2+group3+group4                 #
#########################################################################################################
# Fit a linear regression model with R
model.5 <- lm(SalePrice ~ GrLivArea+GarageArea+group1+group2+group3, data=train.df)

# Display model summary
summary(model.5)

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) using train.df
mse.5 <- mean(model.5$residuals^2)
mae.5 <- mean(abs(model.5$residuals))

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) using test.df
mse.5.test <- mean((test.df$SalePrice - predict.lm(model.5,test.df))^2)
mae.5.test <- mean(abs(test.df$SalePrice-predict.lm(model.5,test.df)))

# Diagnostic plots for SalePrice~GrLivArea simple regression plot
par(mfrow = c(3, 2), oma = c(0, 0, 2, 0))
plot(model.5)
# Use the function qqPlot() in the car package to assess the Studentized residuals
qqPlot(model.5)
# Histogram of the residuals
hist(model.5$residuals)
```

Code Snippet 06: R code for neighborhood accuracy analysis

9.7. SalePrice versus Log SalePrice as the Response

```
#Read the train.df
train.df <- readRDS('C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-
DL\\Week3\\Assignment-03\\train.rds')

#Read the test.df
test.df <- readRDS('C:\\Users\\harinikanand\\Desktop\\NorthwesternUniversity\\MSDS-410-DL\\Week3\\Assignment-
03\\test.rds')


# include the 'car' library
library('car')

#########################################################################################################
# model.6 - MULTIPLE REGRESSION MODEL -                                                                 #
# SalePrice~TotalSqftCalc+MasVnrArea+LotFrontage+GarageArea+YearBuilt+TotRmsAbvGrd                      #
#########################################################################################################

# Fit a linear regression model with R
model.6 <- lm(SalePrice ~ TotalSqftCalc+MasVnrArea+LotFrontage+GarageArea+YearBuilt+TotRmsAbvGrd,
data=train.df)

# Display model summary
summary(model.6)

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) using train.df
mse.6 <- mean(model.6$residuals^2)
mae.6 <- mean(abs(model.6$residuals))

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) using test.df
mse.6.test <- mean((test.df$SalePrice - predict.lm(model.6, test.df))^2)
mae.6.test <- mean(abs(test.df$SalePrice-predict.lm(model.6,test.df)))

# Diagnostic plots for SalePrice~GarageArea simple regression plot
par(mfrow = c(3, 2), oma = c(0, 0, 2, 0))
plot(model.6)
# Use the function qqPlot() in the car package to assess the Studentized residuals
qqPlot(model.6)
# Histogram of the residuals
hist(model.6$residuals)

#########################################################################################################
# model.7 - MULTIPLE REGRESSION MODEL –                                                                 #
# LogSalePrice~TotalSqftCalc+MasVnrArea+LotFrontage+GarageArea+YearBuilt+TotRmsAbvGrd                   #
#########################################################################################################

# Fit a linear regression model with R
model.7 <- lm(LogSalePrice ~ TotalSqftCalc+MasVnrArea+LotFrontage+GarageArea+YearBuilt+TotRmsAbvGrd,
data=train.df)

# Display model summary
summary(model.7)

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) using train.df
mse.7 <- mean((train.df$SalePrice-exp(model.7$fitted.values))^2);
mae.7 <- mean(abs(train.df$SalePrice-exp(model.7$fitted.values)));
```

```
# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) using train.df
mse.7.test <- mean((test.df$SalePrice-exp(predict.lm(model.7, test.df)))^2);
mae.7.test <- mean(abs(test.df$SalePrice-exp(predict.lm(model.7, test.df))));


# Diagnostic plots for SalePrice~GarageArea simple regression plot
par(mfrow = c(3, 2), oma = c(0, 0, 2, 0))
plot(model.7)
# Use the function qqPlot() in the car package to assess the Studentized residuals
qqPlot(model.7)
# Histogram of the residuals
hist(model.7$residuals)

#########################################################################################################
# model.8 - MULTIPLE REGRESSION MODEL –                                                                 #
# LogSalePrice~LogTotalSqftCalc+MasVnrArea+LotFrontage+GarageArea+YearBuilt+TotRmsAbvGrd                #
#########################################################################################################

# Fit a linear regression model with R
model.8 <- lm(LogSalePrice ~ LogTotalSqftCalc+MasVnrArea+LotFrontage+GarageArea+YearBuilt+TotRmsAbvGrd,
data=train.df)

# Display model summary
summary(model.8)

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) using train.df
mse.8 <- mean((train.df$SalePrice-exp(model.8$fitted.values))^2);
mae.8 <- mean(abs(train.df$SalePrice-exp(model.8$fitted.values)));

# Access residuals to compute Mean Square Error (MSE) and Mean Absolute Error (MAE) using train.df
mse.8.test <- mean((test.df$SalePrice-exp(predict.lm(model.8, test.df)))^2);
mae.8.test <- mean(abs(test.df$SalePrice-exp(predict.lm(model.8, test.df))));


# Diagnostic plots for SalePrice~GarageArea simple regression plot
par(mfrow = c(3, 2), oma = c(0, 0, 2, 0))
plot(model.8)
# Use the function qqPlot() in the car package to assess the Studentized residuals
qqPlot(model.8)
# Histogram of the residuals
hist(model.8$residuals)
```

Code Snippet 07: R code for SalePrice vs LogSalePrice models