Assignment #3
Jennifer M. Wanat

Introduction:

In this report our objective is to be able to provide estimates of home values for the typical home in Ames, Iowa. The data set was obtained from the Ames, Iowa Assessor's Office and assembled by Dr. Dean De Cock at Truman State University. In order to build a model to provide estimates of home values, an exploratory data analysis (EDA) of the data set was completed, which allowed for the selection of two of the most promising predictor variables.

The data documentation was reviewed to understand the types of variables collected. As the goal is to predict the typical home value, some conditions were dropped so that the data set represented the typical single-family home. Then an EDA was conducted to demonstrate why the two variables were selected for modeling purposes. Descriptive statistics was performed on the variables when possible. The variables were tabulated and examined for null entries or errors.

Next, the two predictor variables were each fitted in a simple linear regression model, and then combined into a multiple linear regression model. The models were assessed by neighborhood to determine certain neighborhoods were consistently over- or under-predicted, and the neighborhoods were grouped accordingly. The multiple regression model was refitted with the neighborhood groups and compared to the original multiple linear regression model. The model summaries and parameters were provided, and the goodness-of-fit was assessed.

Finally, two models were fit using the same set of predictor variables, but the response variables would be sale price and the log transformation of sale price. The predictor variables for these final two models did not need to be based upon prior selection. The EDA and linear regression modeling were conducted with the R programming language.

Data:

The data set contains 82 variables measured from 2930 individual residential properties sold in Ames, IA from 2006 to 2010. Refer to the data documentation for description of the variables.

Sample Definition:

From the data set, the conditions listed in Table 1 were dropped and not used for the sample population. This process of elimination is referenced as waterfall conditions, as the first condition dropped will result in a smaller sample size in which the subsequent condition to be dropped would be applied. This process continues until all waterfall conditions have been processed.

Table 1: Waterfall conditions.

Variable	Drop Condition	Number of Properties Dropped	Sample Population
			2930
Building Type	Not equal to single-family detached	505	2425
Sale Condition	Not equal to normal	423	2002
Street	Not paved	6	1996
Above Grade Living Area	Greater than 4,000 square feet	1	1995
Lot Area	Greater than 100,000 square feet	3	1992
Bedroom	No bedrooms	4	1988
Full Bath	No full baths	1	1987
Pool	Pool area greater than 0	9	1978

The resulting sample population data set (a data frame called eligible.population) contained 82 variables from 1978 individual residential properties. All subsequent data quality checks and exploratory data analysis were conducted on this data set.

The waterfall conditions were selected to create a data set that represented typical single-family, detached homes in Ames, IA. The data documentation indicated that there were 5 observations from the original data set that were either outliers or unusual sales. The drop condition for observations greater than 4,000 square feet dropped these observations from the data set, if the prior waterfall conditions had not already done so.

Exploratory Data Analysis:

Two variables were selected from the original 82 variables of the data set for an exploratory data analysis. These variables were selected as a result of an initial EDA, and are thought to be the two most promising predictor variables for predicting sale price. The two variables selected were Gr Liv Area (the above grade (ground) living area in square feet) and Tot Rms Abv Grd (the total rooms above grade; does not include bathrooms). The EDA utilized boxplots, histograms and scatterplots, as appropriate.

As Gr Liv Area is a continuous variable, a basic descriptive statistical and quantile summary of the variable was conducted. The descriptive statistical summary in Table 2 includes the minimum value (Min.), first quantile value (1st Qu.), median, mean, third quantile value (3rd Qu.), and maximum value (Max.) of the variable. The quantile summary in Table 3 includes the variable value at the 0, 25, 50, 75 and 100% quantiles.

Table 2: Descriptive statistics of GrLivArea variable.

Variable	Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
GrLivArea	334	1111	1444	1491	1755	3820

The statistics listed in Table 2 are measures of central tendency. The mean or average is the sum of the measurements divided by the total number of measurements. The median is the middle value from the ordered set of measurements. The mean is affected more by outliers than the median. Quartiles (Q) divide the group of data into four equal parts. Q1 and Q3 are the first and third quartiles, and divide the data into the 25th and 75th percentiles, respectively. The median is also known as Q2 and is located at the 50th percentile. Together, the data between Q1 and Q3 represent the middle 50% of the data. Minimum is the smallest value in the data set and maximum is the largest value in the data set.

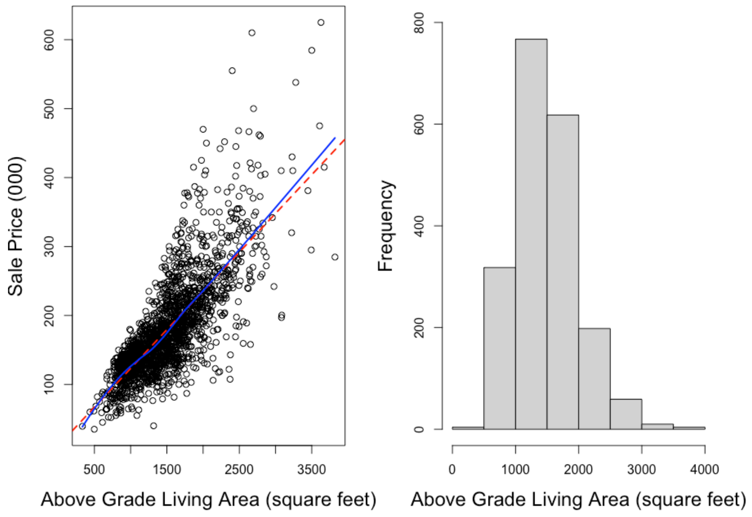
Table 3: Quantile summary.

Variable	0%	25%	50%	75%	100%
GrLivArea	334	1111.25	1444	1755	3820

The above grade living area mean in Table 2 is 1491 square feet and the median is 1444 square feet. The GrLivArea data appear to be positively skewed, which explains why the mean is greater than the median. This implies that there are outliers, or unusual values, for homes with higher above grade living area. The middle 50% of the homes have an above grade living area between 1111 and 1755 square feet.

A scatterplot of sale price versus above grade living area is displayed in Figure 1a. The loess smoother line and the regression line have a strong agreement between the two lines in the scatterplot. Both lines have a positive slope. A heteroscedastic, non-constant variance, is displayed as a widening cone of plotted values as sale price and above grade living area increases. The histogram in Figure 1b indicates that most homes have less than 2000 square feet above grade living area, with the majority between 1000 and 1500 square feet.

Figure 1: (a) Scatterplot of sale price versus above grade living area. The dashed line was estimated using ordinary least squares or OLS. The solid line is a loess smoother. (b) Histogram of above grade living area.



(a) (b)

As TotRmsAbvGrd is a continuous variable, a basic descriptive statistical and quantile summary of the variable was conducted as described above. The total rooms above grade mean in Table 4 is 6 rooms and the median is 6.44 rooms. The TotRmsAbvGrd data appear to be positively skewed, which explains why the mean is greater than the median. This implies that there are outliers, or unusual values, for homes with higher total rooms above grade. The middle 50% of the homes have total rooms above grade between five and seven rooms.

Table 4: Descriptive statistics of TotRmsAbvGrd variable.

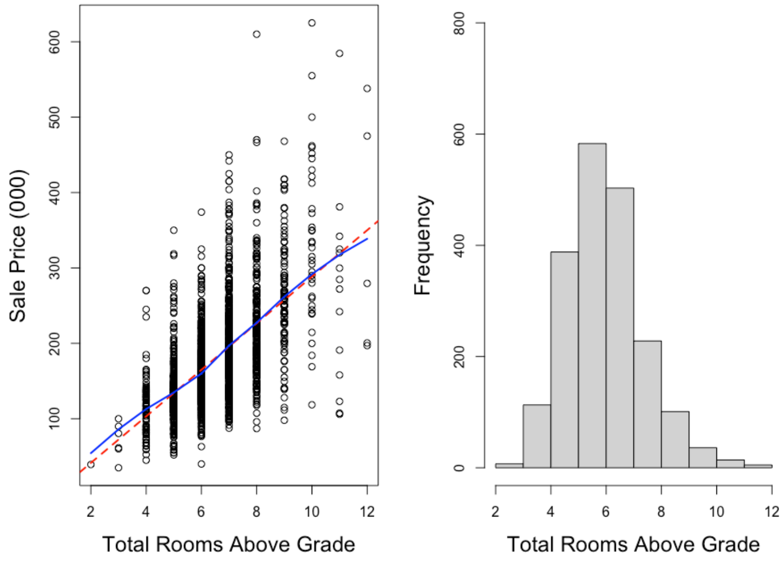
Variable	Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
TotRmsAbvGrd	2	5	6	6.44	7	12

Table 5: Quantile summary.

Variable	0%	25%	50%	75%	100%
TotRmsAbvGrd	2	5	6	7	12

A scatterplot of sale price versus total rooms above grade is displayed in Figure 2a. The loess smoother line and the regression line have a strong agreement between the two lines in the scatterplot. Both lines have a positive slope. A heteroscedastic, non-constant variance, is displayed as a widening cone of plotted values as sale price and total rooms above grade increases. The histogram in Figure 2b indicates that most homes have seven or less total rooms above grade, with the majority of homes between five and seven total rooms.

Figure 2: (a) Scatterplot of sale price versus total rooms above grade. The dashed line was estimated using ordinary least squares or OLS. The solid line is a loess smoother. (b) Histogram of total rooms above grade.



(a) (b)

Table 6: Count of homes by TotRmsAbvGrd variable.

Rooms	2	3	4	5	6	7	8	9	10	11	12
Total	1	6	113	388	583	503	228	101	36	14	5

A count of the home by total rooms above grade is provided in Table 6. There are 19 homes with 11 or more rooms, and seven homes with 3 or less rooms. There are 1474 homes that have between five and seven total rooms above grade.

Simple Linear Regression Models:

Each of the two predictor variables were independently fit in simple linear regression models with sale price. The model summaries and parameters are provided and discussed. The goodness-of-fit of each model was assessed with the coefficient of determination (R^2) and two diagnostic plots, a Q-Q plot and a plot of the model residuals versus the predictor variable.

In a Q-Q plot, a set of equally spaced quantiles is calculated based upon the sample size of the data set. The observed and ordered (from lowest to highest value) data set are plotted against the calculated quantiles. The data points, if normally distributed, will plot approximately along a straight line drawn through the first and third quantiles. Variations from a normal distribution will be detected by data points that drift away from the theoretical quantile line.

The residuals, or the estimated errors, are calculated by subtracting the expected (or predicted) value of Y obtained from the linear equation for the linear regression model from the observed value of Y. The residuals are plotted against the predictor variable. Residuals should be independent, normally distributed and have a constant variance. Residuals should also have a mean of zero.

The coefficient of determination is the proportion of variability of the response variable explained by the predictor variable. The R^2 will have a value between zero and one. A value of zero would indicate that there is no relationship between the response and predictor variables. A value of 1 would indicate that 100% of the variability of the response variable is explained by the predictor variable.

Model #1 Gr Liv Area:

Model #1 is a simple linear regression model of sale price as a function of above grade living area.

Table 7: Residuals for Model #1 Gr Liv Area.

Minimum	1st Qu.	Median	3rd Qu.	Maximum
-169919	-25015	-1264	20235	298649

Table 8: Coefficients for Model #1 Gr Liv Area.

	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	10908.990	3253.481	3.353	0.000814
GrLivArea	112.357	2.073	54.209	< 2e-16

Model 1: E(Sale Price) = $b_0 + b_1$GrLivArea

E(Sale Price) = 10908.990 + 112.357GrLivArea

Table 9: Model #1 summary.

Residual standard error	45140 on 1976 degrees of freedom
Multiple R-squared	0.5979
Adjusted R-squared	0.5977
F-statistic	2939 on 1 and 1976 DF
p-value	< 2.2e-16

From the model, sale price is expected to increase by $112 for each square foot increase in above grade living area. The estimated intercept of 10908.990 is the expected value of sale price when the above grade living area equals zero. A home with zero square footage is not meaningful, but this value could represent an empty lot with no structure.

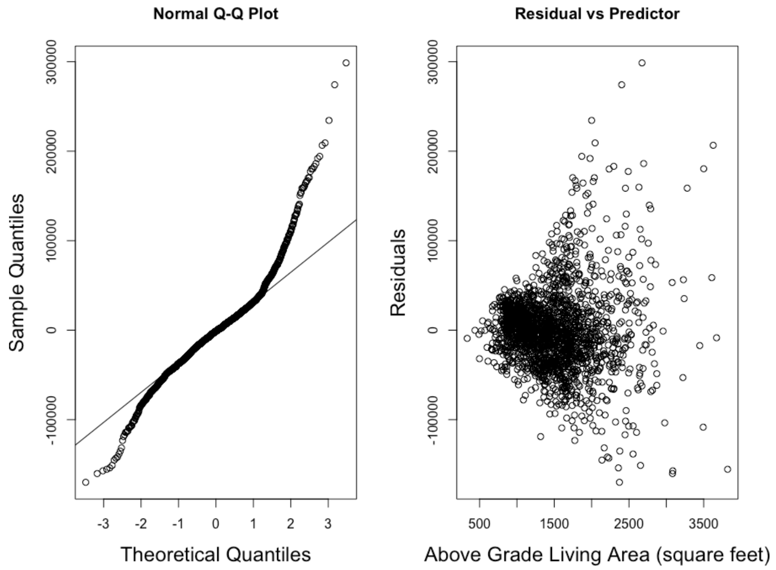
The residual standard error is 45140. If model #1 was used to predict the sale price from above grade living area square footage, we can expect to be accurate to within approximately ±90280 dollars at a 95% confidence level.

The multiple R-squared value is 0.5979, which indicates that 59.7% of the variation in sale price (about its mean) can be explained by the variable above grade living area. The adjusted R-squared value is 0.5977 and takes into account the number of predictors for the model. The adjusted R-square value will always be lower than the R-squared value. As this is a simple linear regression model with only one variable, both the adjusted R-square and the multiple R-square are close in value.

The linear association between sale price and above grade living area is statistically significant at the 5% significance level given the t value of 54.209, based upon the results of the hypothesis tests for the model coefficients (Pr(> |t|)). The F-statistic of 2939 is statistically significant at the 95% confidence level given the p-value < 2.2e-16. From this information, it can be concluded that the above grade living area predictor is linearly associated with sale price.

The Q-Q plot of model #1 indicates a departure of values from a normal distribution for lower and higher sale priced homes. The residual versus predictor plot of model #1 demonstrates a heteroscedastic, non-constant variance, displayed as a widening cone of plotted values as the above grade living area increases.

Figure 3: Goodness-of-fit diagnostic plots. (a) Q-Q plot of model #1. (b) Residual versus predictor plot of model #1.



(a) (b)

Model #2 Tot Rms Abv Grd:

Model #2 is a simple linear regression model of sale price as a function of the total rooms above grade.

Table 10: Residuals for Model #2 Tot Rms Abv Grd.

Minimum	1st Qu.	Median	3rd Qu.	Maximum
-213205	-31727	-8040	22460	383378

Table 11: Coefficients for Model #2 Tot Rms Abv Grd.

	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	-20263.4	5891.8	-3.439	0.000595
TotRmsAbvGrd	30860.7	893.6	34.535	< 2e-16

$$\text{Model 2:} \qquad \text{E(Sale Price)} = b_0 + b_1\text{TotRmsAbvGrd}$$

$$\text{E(Sale Price)} = -20263.4 + 30860.7\text{TotRmsAbvGrd}$$

Table 12: Model #2 summary.

Residual standard error	56210 on 1976 degrees of freedom
Multiple R-squared	0.3764
Adjusted R-squared	0.3761
F-statistic	1193 on 1 and 1976 DF
p-value	< 2.2e-16

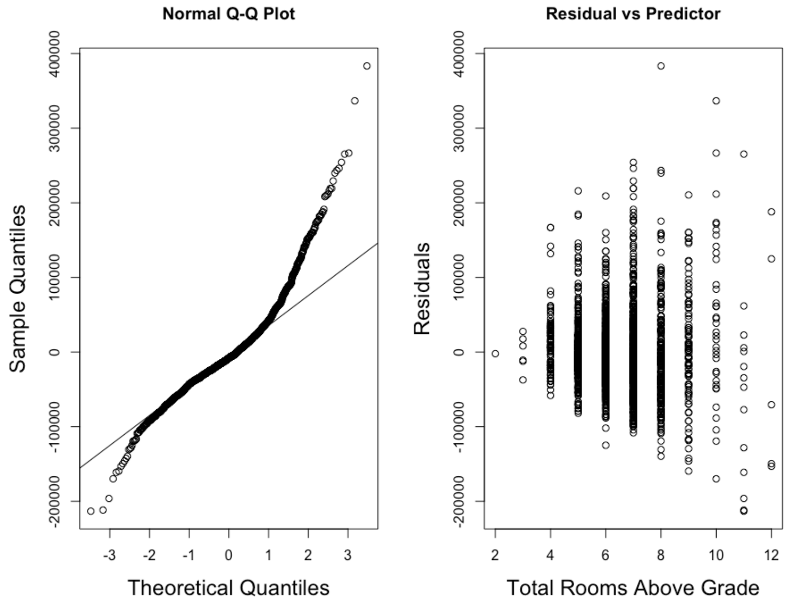
From model #2, sale price is expected to increase by $30860.70 for each increase in total rooms above grade. The estimated intercept of -20263.4 is the expected value of sale price when the total rooms above grade equals zero. A home with zero rooms above grade is not meaningful, and the negative value of the intercept has no practical interpretation.

The residual standard error is 56210. If model #2 was used to predict the sale price from total rooms above grade, we can expect to be accurate to within approximately ±112420 dollars at a 95% confidence level. The multiple R-squared value is 0.3764, which indicates that 37.6% of the variation in sale price (about its mean) can be explained by the variable total rooms above grade. The adjusted R-squared value is 0.3761 and takes into account the number of predictors for the model. The adjusted R-square value will always be lower than the R-squared value. As this is a simple linear regression model with only one variable, both the adjusted R-square and the multiple R-square are close in value.

The linear association between sale price and total rooms above grade is statistically significant at the 5% significance level given the t value of 34.535, based upon the results of the hypothesis tests for the model coefficients (Pr(> |t|)). The F-statistic of 1193 is statistically significant at the 95% confidence level given the p-value < 2.2e-16. From this information, it can be concluded that the total rooms above grade predictor is linearly associated with sale price.

The Q-Q plot of model #2 indicates a departure of values from a normal distribution for lower and higher sale priced homes. The residual versus predictor plot of model #2 demonstrates a heteroscedastic, non-constant variance, displayed as a widening cone of plotted values as the total rooms above grade increases.

Figure 4: Goodness-of-fit diagnostic plots. (a) Q-Q plot of model #2. (b) Residual versus predictor plot of model #2.



(a) (b)

Multiple Linear Regression Model – Model #3:

Model #3 is a multiple linear regression model of sale price as a function of the above grade living area and the total rooms above grade.

Table 13: Residuals for Model #3.

Minimum	1st Qu.	Median	3rd Qu.	Maximum
-160645	-25520	-994	19557	292197

Table 14: Coefficients for Model #3.

	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	25223.641	4907.596	5.140	3.02e-7
GrLivArea	124.435	3.732	33.341	< 2e-16
TotRmsAbvGrd	-5019.937	1292.043	-3.885	0.000106

Model 3: E(Sale Price) = b_0 + b_1GrLivArea + b_2TotRmsAbvGrd

E(Sale Price) = 25223.641 + 124.436GrLivArea -5019.937TotRmsAbvGrd

Table 15: Model #3 summary.

Residual standard error	44980 on 1975 degrees of freedom
Multiple R-squared	0.601
Adjusted R-squared	0.6006
F-statistic	1487 on 2 and 1975 DF
p-value	< 2.2e-16

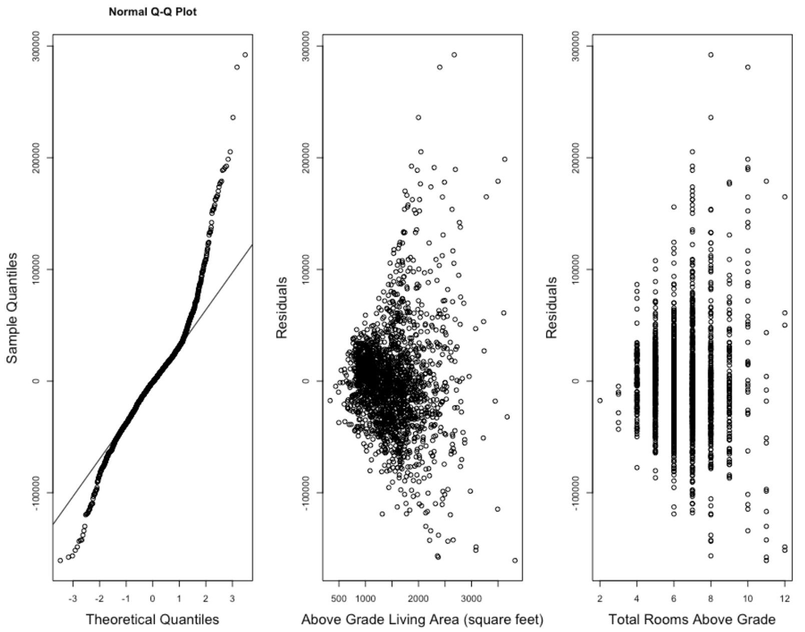
From the model, sale price is expected to increase by $124.435 for each square foot increase in above grade living area when the total rooms above grade is held constant. The sale price is expected to decrease by $5019.937 for each increase in the total rooms above grade when the above grade living area remains constant. The estimated intercept of 25223.641 is the expected value of sale price when the above grade living area and total rooms above grade equal zero. A home with zero square footage and total rooms is not meaningful, but this value could represent an empty lot with no structure.

The residual standard error is 44980. If model #3 was used to predict the sale price from above grade living area square footage and the total rooms above grade, we can expect to be accurate to within approximately ±89960 dollars at a 95% confidence level. The multiple R-squared value is 0.601, which indicates that 60.1% of the variation in sale price (about its mean) can be explained by a multiple linear regression association between the above grade living area and the total rooms above grade. The adjusted R-squared value is 0.6006 and takes into account the number of predictors for the model. The adjusted R-square value will always be lower than the R-squared value. This is a multiple linear regression model with two variables, and both the adjusted R-square and the multiple R-square are close in value. This indicates that both predictor variables contribute towards explaining the variation in the sale price and the model is not overfit.

The linear association between sale price and total rooms above grade, holding above grade living area constant, and between sale price and above grade living area, holding the total rooms above grade constant, is statistically significant at the 5% significance level given the respective t values of -3.885 and 33.341, based upon the results of the hypothesis tests for the model coefficients (Pr(> |t|)). The F-statistic of 1487 is statistically significant at the 95% confidence level given the p-value < 2.2e-16. From this information, it can be concluded that the above grade living area and total rooms above grade predictors are linearly associated with sale price.

The Q-Q plot of model #3 in Figure 5a indicates a departure of values from a normal distribution for lower and higher sale priced homes. The residual versus predictor plots of model #3 in Figures 5b and 5c demonstrate a heteroscedastic, non-constant variance, displayed as a widening column of plotted values as the above grade living area and total rooms above grade increase.

Figure 5: Goodness-of-fit diagnostic plots. (a) Q-Q plot of model #3. (b) Residual versus above grade living area predictor plot of model #3. (c) Residual versus total rooms above grade predictor plot of model #3.



(a) (b) (c)

As noted in Table 16, the multiple linear regression model #3 has a higher R-squared value compared to the two simple linear regression models, although the increase is only 0.4% over model #1. The residual standard error for the predictor variables is smaller for model #3 compared to models #1 and #2. An analysis of the residuals for all models indicates a heteroscedastic variance and a departure from a normal distribution. A transformation of variables may be useful.

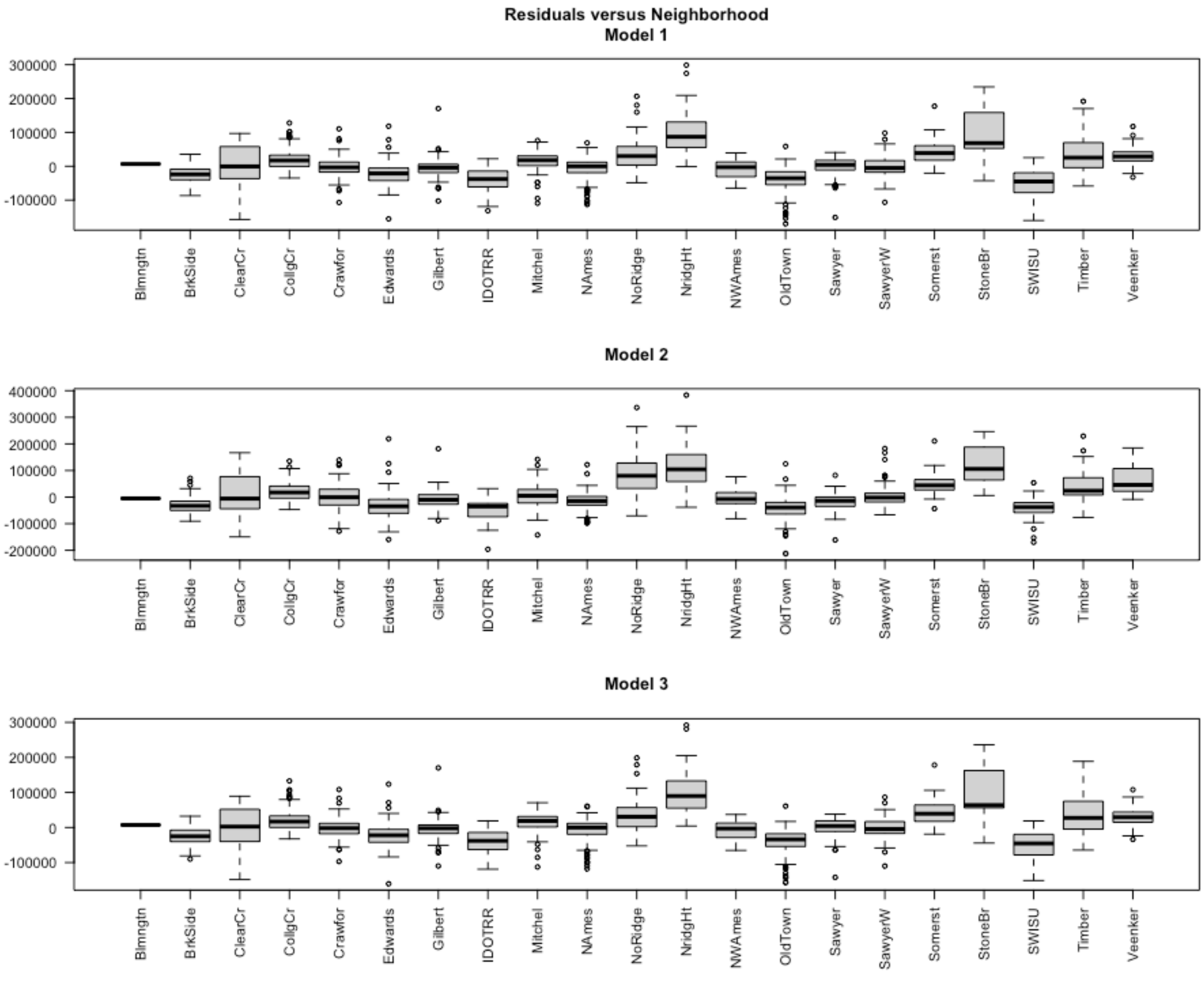
Table 16: Comparison of Model Output for Models #1, #2 and #3.

	Model 1	Model 2	Model 3
Residual Standard Error	45140 on 1976 DF	56210 on 1976 DF	44980 on 1975 DF
Multiple R-squared	0.5979	0.3764	0.601
Adjusted R-squared	0.5977	0.3761	0.6006
F-statistic	2939 on 1 and 1976 DF	1193 on 1 and 1976 DF	1487 on 2 and 1975 DF
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16

Neighborhood Accuracy:

A boxplot of the model residuals by neighborhood was made to determine which neighborhoods were better fit by the model. This was conducted for models #1, 2 and 3.

Figure 6: Boxplot of Residuals by Neighborhood for Models #1, 2 and 3.



Neighborhoods were consistently over- and under-predicted. Neighborhoods that were over-predicted in each model were NridgHt and StoneBr, while IDOTRR, OldTown and SWISU were consistently under-predicted. The mean absolute error (MAE) and mean sale price per square foot were calculated for each neighborhood and then plotted in a scatterplot.

Table 17: MAE and Mean Sale Price Per Square Foot by Neighborhood.

Neighborhood	MAE	Mean Sale Price Per Sq ft
Blmngtn	7256.59	126.29937
BrkSide	27946.44	103.93649
ClearCr	43240.20	124.97980
CollgCr	25010.84	134.43971
Crawfor	22503.85	118.23056
Edwards	31657.97	104.82935
Gilbert	18667.35	117.77165
IDOTRR	41137.91	94.89429
Mitchel	26165.82	135.66889
NAmes	19983.30	119.85863
NoRidge	43568.04	131.90394
NridgHt	98425.82	165.66915
NWAmes	22669.93	116.77809
OldTown	41534.06	95.31745
Sawyer	20805.84	125.01947
SawyerW	21799.37	119.82202
Somerst	44538.98	144.29737
StoneBr	102879.18	165.07479
SWISU	54405.70	93.10390
Timber	50512.92	142.66587
Veenker	38451.89	138.96907

An examination of Figure 7 reveals that the two neighborhoods that were over-predicted have the highest sale price per square foot and also the highest MAE. The three neighborhoods that were under-predicted have the lowest sale price per square foot. The MAE generally decreases as the sale price per square foot increases up to $120 per square foot. Then the MAE generally increases until the maximum is reached with StoneBr. Nineteen of the twenty-one neighborhoods have a MAE equal to or less than 54405.70, with the exceptions of NridgHt and StoneBr.

Figure 7: Scatterplot of MAE versus Mean Sale Price per Square Foot by Neighborhood.

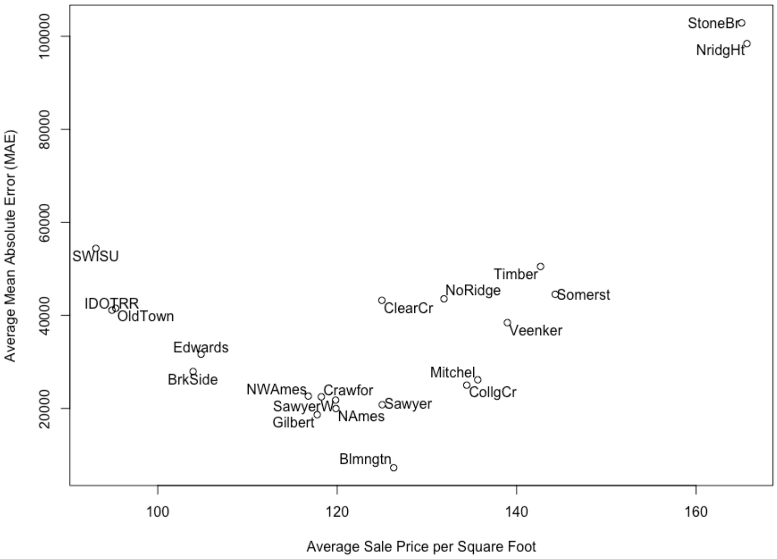


Table 18: Neighborhood Groups Based on Sale Price per Square Foot.

Group	Neighborhood	Number of Homes
1	IDOTRR, OldTown, SWISU	268
2	BrkSide, Edwards	224
3	NWAmes, Crawfor, Gilbert, NAmes, Sawyer, SawyerW, Blmngtn, ClearCr, NoRidge, Veenker, CollgCr, Mitchel	1293
4	Timber, Somerst	114
5	StoneBr, NridgHt	79

Based upon this information, the neighborhoods were placed into 5 groups based on sale price per square foot (Table 18). Model #3 was then refit with the neighborhood groups and named Model #4. In order to determine the difference between the neighborhood group 1 with the lowest averages of sale price per square foot, neighborhood groups 2 through 5 were used in the model.

Table 19: Residuals for Model #4.

Minimum	1st Qu.	Median	3rd Qu.	Maximum
-132405	-18188	-432	15701	232055

Table 20: Coefficients for Model #4.

	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	16964.30	4189.12	4.050	5.33e-5
GrLivArea	111.31	2.86	38.920	< 2e-16
TotRmsAbvGrd	-7358.15	983.63	-7.481	1.11e-13
Neighborhood2	15107.48	3098.53	4.876	1.17e-06

Neighborhood3	45838.10	2299.56	19.933	< 2e-16
Neighborhood4	88861.34	3862.07	23.009	< 2e-16
Neighborhood5	152572.19	4530.70	33.675	< 2e-16

Model 4: E(Sale Price) = b_0 + b_1GrLivArea + b_2TotRmsAbvGrd + b_3Neighborhood2 + b_4Neighborhood3 + b_5Neighborhood4 + b_6Neighborhood5

E(Sale Price) = 16964.30 + 111.31GrLivArea -7358.15TotRmsAbvGrd + 15107.48Neighborhood2 + 45838.10Neighborhood3 + 88861.34Neighborhood4 + 152572.19Neighborhood5

Table 21: Model #4 summary.

Residual standard error	34160 on 1971 degrees of freedom
Multiple R-squared	0.7703
Adjusted R-squared	0.7696
F-statistic	1101 on 6 and 1971 DF
p-value	< 2.2e-16

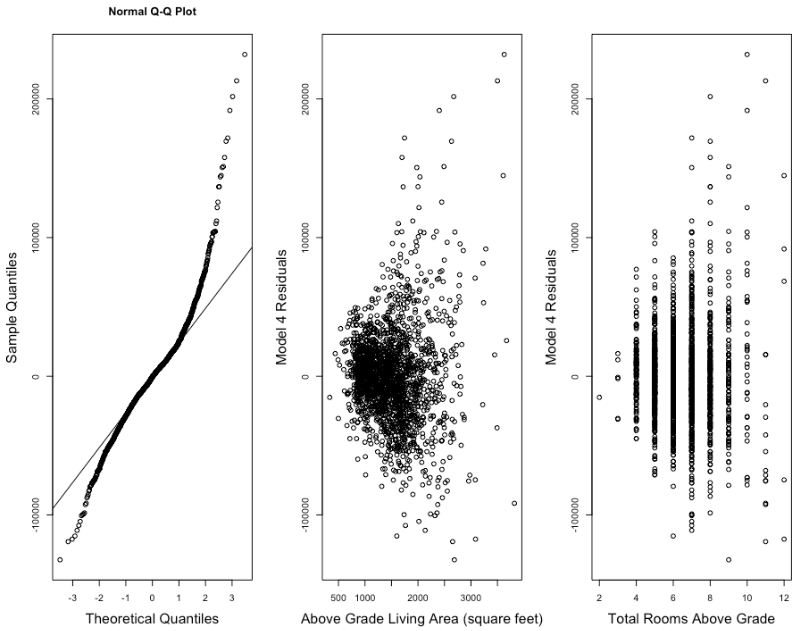
From the model, sale price is expected to increase by $111.31 for each square foot increase in above grade living area when the total rooms above grade is held constant and the neighborhood group is group 1. The sale price is expected to decrease by $7358.15 for each increase in the total rooms above grade when the above grade living area remains constant and the neighborhood group is group 1. The sale price is expected to increase $15107.48 if the house is located in neighborhood group 2, and the above grade living area and the total rooms above grade remain constant. Similarly, the sale price would increase $45838.10 for a home in neighborhood group 3, $88861.34 for a home in neighborhood group 4, and $152572.19 for a home in neighborhood group 5 when above grade living area and total rooms above grade are held constant. The estimated intercept of 16964.30 is the expected value of sale price when the above grade living area and total rooms above grade equal zero, and the location is in neighborhood group 1. A home with zero square footage and total rooms is not meaningful, but this value could represent an empty lot with no structure.

The residual standard error is 34160. If model #4 was used to predict the sale price from above grade living area square footage, total rooms above grade and neighborhood group, we can expect to be accurate to within approximately ±68320 dollars at a 95% confidence level. The multiple R-squared value is 0.7703, which indicates that 77.0% of the variation in sale price (about its mean) can be explained by a multiple linear regression association between the above grade living area, total rooms above grade, and neighborhood group. The adjusted R-squared value is 0.7696 and takes into account the number of predictors for the model. The adjusted R-square value will always be lower than the R-squared value. This is a multiple linear regression model with six variables, and both the adjusted R-square and the multiple R-square are close in value. This indicates that all the predictor variables contribute towards explaining the variation in the sale price and the model is not overfit.

The linear association between sale price and each predictor variable while holding all others constant is statistically significant at the 5% significance level given the respective t values found in Table 20, based upon the results of the hypothesis tests for the model coefficients (Pr(> |t|)). The F-statistic of 1101 is statistically significant at the 95% confidence level given the p-value < 2.2e-16. From this information, it can be concluded that the above grade living area, total rooms above grade, and neighborhood group predictors are linearly associated with sale price.

The Q-Q plot of model #4 in Figure 8a indicates a departure of values from a normal distribution for lower and higher sale priced homes. The residual versus predictor plots of model #4 in Figures 8b and 8c demonstrate a heteroscedastic, non-constant variance, displayed as a widening column of plotted values as the above grade living area and total rooms above grade increase.

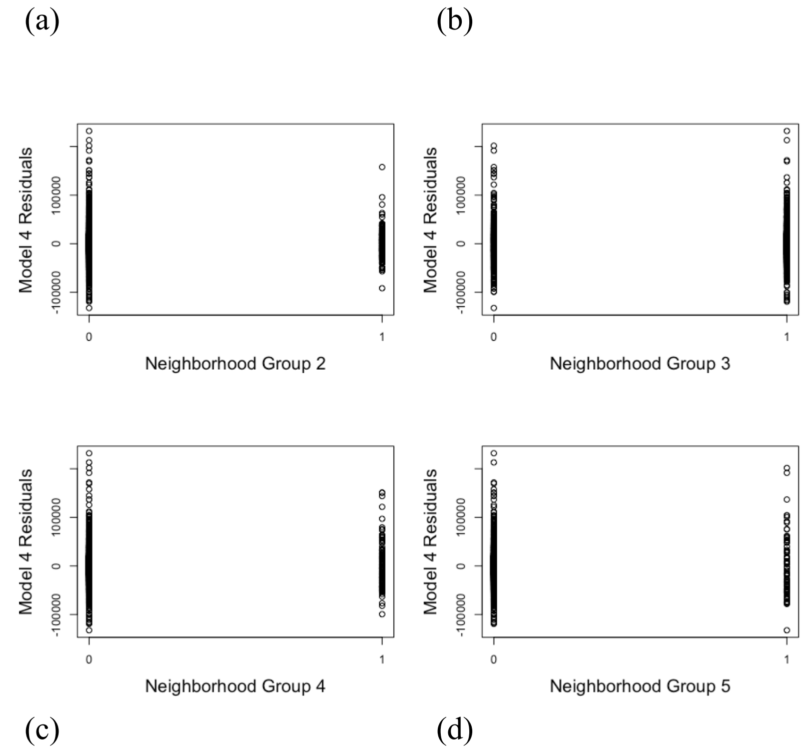
Figure 8: Goodness-of-fit diagnostic plots. (a) Q-Q plot of model #4. (b) Residual versus above grade living area predictor plot of model #4. (c) Residual versus total rooms above grade predictor plot of model #4.



(a) (b) (c)

The residual versus predictor plots of model #4 in Figures 9a - 9d demonstrate a heteroscedastic, non-constant variance, displayed as a narrowing cone of plotted values for neighborhood groups 2, 4, and 5. A homoscedastic variance is observed with the plotted residuals for neighborhood groups 3, and it is noted that this neighborhood group contained 1293 homes whereas all the other neighborhood groups contained 268 homes or less.

Figure 9: Goodness-of-fit diagnostic plots. (a) Residual versus neighborhood group 2 predictor plot of model #4. (b) Residual versus neighborhood group 3 predictor plot of model #4. (c) Residual versus neighborhood group 4 predictor plot of model #4. (d) Residual versus neighborhood group 5 predictor plot of model #4.



Comparison of the MAE by neighborhood for model #3 and model #4 indicate a reduced MAE for all neighborhoods except for two. The reduction ranged from -2 to -96 percent. Based upon the MAE, model #4 is better than model #3.

Table 22: MAE by Neighborhood for Model #3 and Model #4.

Neighborhood	MAE Model #3	MAE Model #4	% Difference
Blmngtn	7256.59	322.1079	-96
BrkSide	27946.44	16988.7065	-39
ClearCr	43240.20	37813.2100	-13
CollgCr	25010.84	21932.3785	-12
Crawfor	22503.85	23290.5200	3
Edwards	31657.97	23493.3416	-26
Gilbert	18667.35	16953.5154	-9
IDOTRR	41137.91	23083.9301	-44
Mitchel	26165.82	20921.8384	-20
NAmes	19983.30	19674.7068	-2
NoRidge	43568.04	51930.3364	19
NridgHt	98425.82	47692.8683	-52
NWAmes	22669.93	18620.3897	-18
OldTown	41534.06	21878.1836	-47

Sawyer	20805.84	18356.3222	-12
SawyerW	21799.37	19113.6295	-12
Somerst	44538.98	26674.4251	-40
StoneBr	102879.18	65276.3329	-37
SWISU	54405.70	28024.8773	-48
Timber	50512.92	48804.4285	-3
Veenker	38451.89	36942.3153	-4

Sale Price Versus Log Sale Price as the Response:

Two models were fit using the same set of predictor variables, but the response variables would be sale price and the log transformation of sale price. Each model used the following predictor variables: GrLivArea, TotalBsmtSF (total square feet of basement area), BsmtFinSF1 (basement finished in square feet), Garage Area, Year Built, and Neighborhood group 5. The model summaries and parameters are provided and discussed. The goodness-of-fit of each model were assessed with two diagnostic plots, a Q-Q plot and a plot of the model residuals versus the predictor variable.

Model #5:

Model #5 is a multiple linear regression model of sale price as a function of the above grade living area, total square feet of basement area, basement finished in square feet, garage area, year built, and neighborhood group 5.

Table 23: Residuals for Model #5.

Minimum	1st Qu.	Median	3rd Qu.	Maximum
-126039	-15778	-1084	13153	223310

Table 24: Coefficients for Model #5.

	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	-974762.381	47861.203	-20.37	< 2e-16
GrLivArea	76.396	1.492	51.20	< 2e-16
TotalBsmtSF	30.795	2.101	14.66	< 2e-16
BsmtFinSF1	22.255	1.775	12.54	< 2e-16
GarageArea	43.862	4.036	10.87	< 2e-16
YearBuilt	495.542	24.842	19.95	< 2e-16
Neighborhood5	59119.289	3447.371	17.15	< 2e-16

Model 5: $$\text{E(Sale Price)} = b_0 + b_1\text{GrLivArea} + b_2\text{TotalBsmtSF} + b_3\text{BsmtFinSF1} + b_4\text{GarageArea} + b_5\text{YearBuilt} + b_6\text{Neighborhood5}$$

$$\text{E(Sale Price)} = -974762.381 + 76.396\text{GrLivArea} + 30.795\text{TotalBsmtSF} + 22.255\text{BsmtFinSF1} + 43.862\text{GarageArea} + 495.542\text{YearBuilt} + 59119.289\text{Neighborhood5}$$

Table 25: Model #5 summary.

Residual standard error	27540 on 1972 degrees of freedom
Multiple R-squared	0.8507
Adjusted R-squared	0.8503
F-statistic	1872 on 6 and 1971 DF
p-value	< 2.2e-16

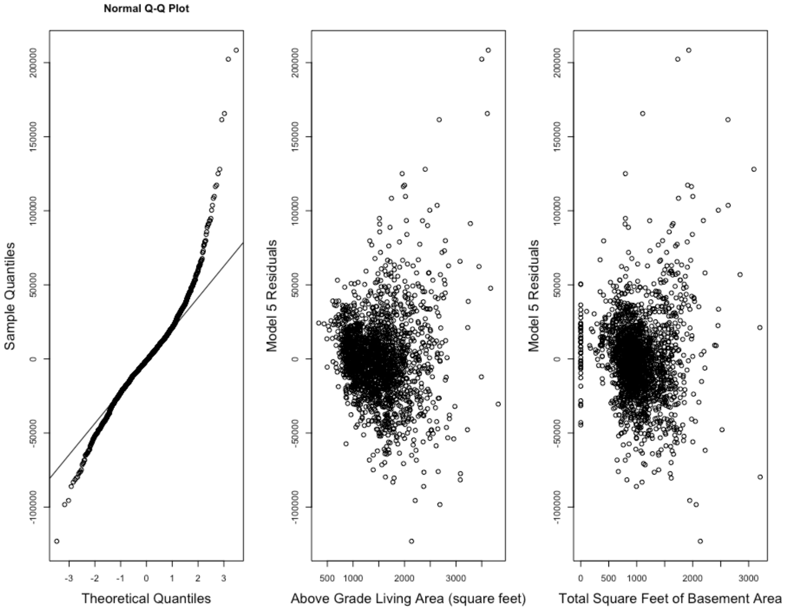
From the model, sale price is expected to increase by $76.396 for each square foot increase in above grade living area when all other variables are held constant. The sale price is expected to increase by $30.795 for each square foot increase in total basement area when all other variables are held constant. The sale price is expected to increase $22.255 for each square foot increase in finished basement when all other variables are held constant. The sale price is expected to increase $43.862 for each square foot increase in garage area when all other variables are held constant. The sale price is expected to increase by $495.542 for every year after 1872 when all other variables are held constant. The estimated intercept of -974762.381 is the expected value of sale price when all other variables are zero. The negative value of the intercept has no practical interpretation.

The residual standard error is 27540. If model #5 was used to predict the sale price with the six predictor variables, we can expect to be accurate to within approximately ±55080 dollars at a 95% confidence level. The multiple R-squared value is 0.8507, which indicates that 85.0% of the variation in sale price (about its mean) can be explained by a multiple linear regression association between the six predictor variables. The adjusted R-squared value is 0.8503 and takes into account the number of predictors for the model. The adjusted R-square value will always be lower than the R-squared value. This is a multiple linear regression model with six variables, and both the adjusted R-square and the multiple R-square are close in value. This indicates that all the predictor variables contribute towards explaining the variation in the sale price and the model is not overfit.

The linear association between sale price and each predictor variable while holding all others constant is statistically significant at the 5% significance level given the respective t values found in Table 24, based upon the results of the hypothesis tests for the model coefficients (Pr(> |t|)). The F-statistic of 1872 is statistically significant at the 95% confidence level given the p-value < 2.2e-16. From this information, it can be concluded that the above grade living area, total square feet of basement area, basement finished in square feet, garage area, year built, and neighborhood group 5 predictors are linearly associated with sale price.

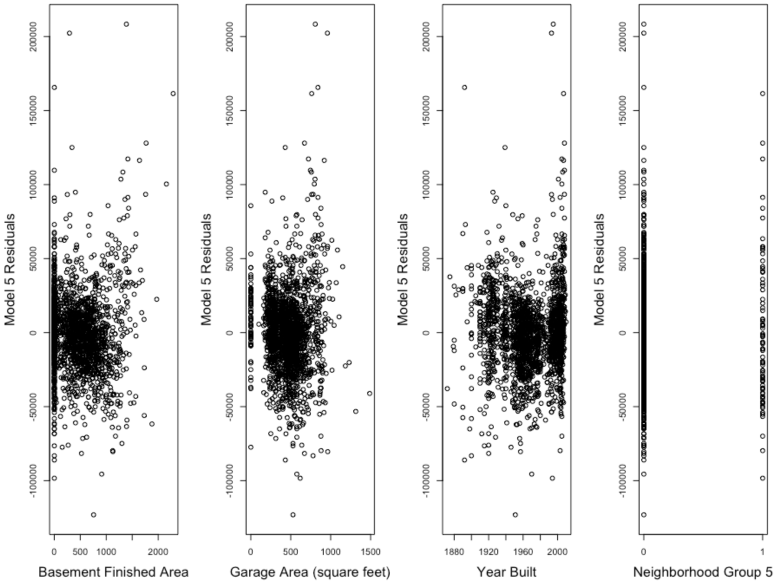
The Q-Q plot of model #5 in Figure 10a indicates a departure of values from a normal distribution for lower and higher sale priced homes. The residual versus predictor plots of model #5 in Figures 10b – 10c demonstrate a heteroscedastic, non-constant variance, displayed as a widening cone of plotted values as the above grade living area and total square feet of basement area increase. A homoscedastic variance is observed with the plotted residuals in Figure 11 for basement finished in square feet, garage area, year built, and neighborhood group 5 predictors.

Figure 10: Goodness-of-fit diagnostic plots. (a) Q-Q plot of model #5. (b) Residual versus above grade living area predictor plot of model #5. (c) Residual versus total square feet of basement area predictor plot of model #5.



(a) (b) (c)

Figure 11: Goodness-of-fit diagnostic plots. (a) Residual versus basement finished area predictor plot of model #5. (b) Residual versus garage area predictor plot of model #5. (c) Residual versus year built predictor plot of model #5. (d) Residual versus neighborhood group 5 predictor plot of model #5.



(a) (b) (c) (d)

Model #6:

Model #6 is a multiple linear regression model of log transformed sale price as a function of the above grade living area, total square feet of basement area, basement finished in square feet, garage area, year built, and neighborhood group 5.

Table 26: Residuals for Model #6.

Minimum	1st Qu.	Median	3rd Qu.	Maximum

Table 27: Coefficients for Model #6.

	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	4.066854728	0.246887985	16.472	< 2e-16
GrLivArea	0.000398098	0.000007697	51.724	< 2e-16
TotalBsmtSF	0.000153627	0.000010837	14.176	< 2e-16
BsmtFinSF1	0.000097038	0.000009158	10.596	< 2e-16
GarageArea	0.000255566	0.000020817	12.277	< 2e-16
YearBuilt	0.003576733	0.000128143	27.912	< 2e-16
Neighborhood5	0.109173076	0.017782971	6.139	1e-09

Model 6: $$E(\log(\text{Sale Price})) = b_0 + b_1\text{GrLivArea} + b_2\text{TotalBsmtSF} + b_3\text{BsmtFinSF1} + b_4\text{GarageArea} + b_5\text{YearBuilt} + b_6\text{Neighborhood5}$$

$$E(\log(\text{Sale Price})) = 4.066854728 + 0.000398098\text{GrLivArea} + 0.000153627\text{TotalBsmtSF} + 0.000097038\text{BsmtFinSF1} + 0.000255566\text{GarageArea} + 0.003576733\text{YearBuilt} + 0.109173076\text{Neighborhood5}$$

Table 28: Model #6 summary.

Residual standard error	0.145 on 1971 degrees of freedom
Multiple R-squared	0.8529
Adjusted R-squared	0.8525
F-statistic	1905 on 6 and 1971 DF
p-value	< 2.2e-16

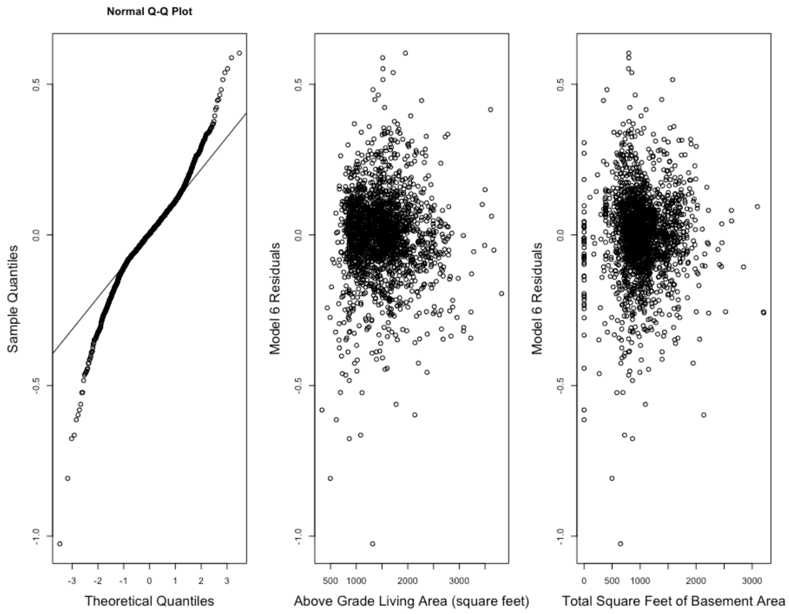
From the model, the untransformed sale price is expected to be 0.03% higher ($\exp^{0.000398098}$) for each square foot increase in above grade living area when all other variables are held constant. The sale price is expected to be 0.01% higher ($\exp^{0.000153627}$) for each square foot increase in total basement area when all other variables are held constant. The sale price is expected to be 0.009% higher ($\exp^{0.000097038}$) for each square foot increase in finished basement when all other variables are held constant. The sale price is expected to be 0.02% higher ($\exp^{0.000255566}$) for each square foot increase in garage area when all other variables are held constant. The sale price is expected to be 0.35% higher ($\exp^{0.003576733}$) for every year after 1872 when all other variables are held constant. The estimated intercept is 4.066854728.

The residual standard error is 0.145. The multiple R-squared value is 0.8529, which indicates that 85.2% of the variation in sale price (about its mean) can be explained by a multiple linear regression association between the six predictor variables. The adjusted R-squared value is 0.8525 and takes into account the number of predictors for the model. The adjusted R-square value will always be lower than the R-squared value. This is a multiple linear regression model with six variables, and both the adjusted R-square and the multiple R-square are close in value. This indicates that all the predictor variables contribute towards explaining the variation in the sale price and the model is not overfit.

The linear association between sale price and each predictor variable while holding all others constant is statistically significant at the 5% significance level given the respective t values found in Table 27, based upon the results of the hypothesis tests for the model coefficients (Pr(> |t|)). The F-statistic of 1905 is statistically significant at the 95% confidence level given the p-value < 2.2e-16. From this information, it can be concluded that the above grade living area, total square feet of basement area, basement finished in square feet, garage area, year built, and neighborhood group 5 predictors are linearly associated with log transformed sale price.

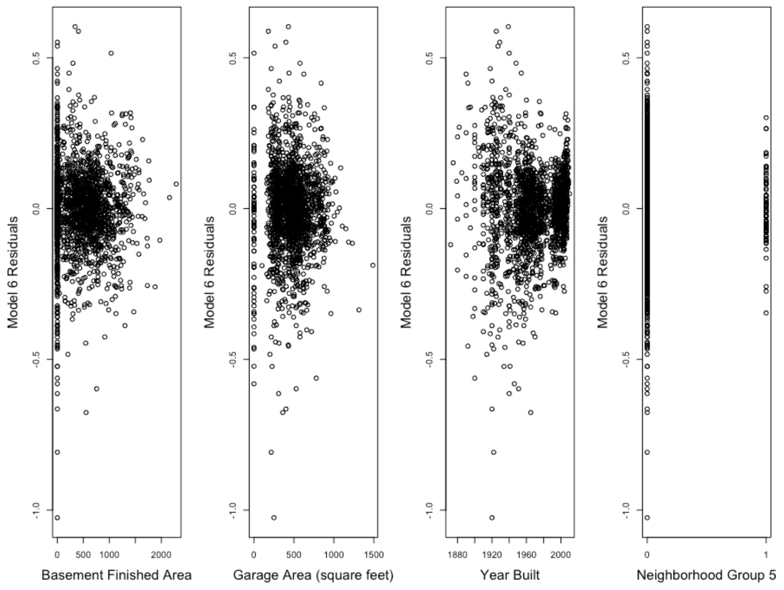
The Q-Q plot of model #6 in Figure 12a indicates a departure of values from a normal distribution for lower and higher sale priced homes. The residual versus predictor plots of model #6 in Figure 11d demonstrate a heteroscedastic, non-constant variance, displayed as a narrowing cone of plotted values for the neighborhood group 5 predictor. as the above grade living area and total square feet of basement area increase. Generally, a homoscedastic variance is observed with the plotted residuals in Figures 12b – 12c and 13a – 13c for the above grade living area, total square feet of basement area, basement finished in square feet, garage area, year built predictors.

Figure 12: Goodness-of-fit diagnostic plots. (a) Q-Q plot of model #6. (b) Residual versus above grade living area predictor plot of model #6. (c) Residual versus total square feet of basement area predictor plot of model #6.



(a) (b) (c)

Figure 13: Goodness-of-fit diagnostic plots. (a) Residual versus basement finished area predictor plot of model #6. (b) Residual versus garage area predictor plot of model #6. (c) Residual versus year built predictor plot of model #6. (d) Residual versus neighborhood group 5 predictor plot of model #6.



(a) (b) (c) (d)

Table 29: Comparison of Model Output for Models #5 and #6.

	Model 5	Model 6
Residual Standard Error	27540 on 1971 DF	0.142 on 1971 DF
Multiple R-squared	0.8507	0.8529
Adjusted R-squared	0.8503	0.8525
F-statistic	1872 on 6 and 1971 DF	1905 on 6 and 1971 DF
p-value	< 2e-16	< 2e-16

Conclusion:

Model #3 indicates that 60.1% of the variation in sale price (about its mean) can be explained by a multiple linear regression association between the above grade living area and the total rooms above grade. However, an assessment of the goodness-of-fit indicated that the residuals failed the zero mean assumption due to the heteroscedastic variance. Model #4 refit Model #3 taking into account the variability of neighborhood average sale price per square foot. An assessment of the goodness-of-fit indicated that the residuals failed the zero mean assumption due to the heteroscedastic variance for Model #4, also.

Model #6 indicates that 85.2% of the variation in log transformed sale price (about its mean) can be explained by a multiple linear regression association as a function as above grade living area, total square feet of basement area, basement finished in square feet, garage area, year built, and neighborhood group 5. An improvement in the assessment of the goodness-of-fit was observed for the residuals in Figures 12 and 13, although the normality regression assumption was noted due to a departure of values from a normal distribution for lower and higher sale priced homes. The goodness-of-fit was improved with the log transformed sale price compared to the same model without the log transformation.

References:

Benoit, K. (2011, March 17). Linear Regression Models with Logarithmic Transformations. Retrieved from http://kenbenoit.net/assets/courses/ME104/logmodels2.pdf

Black, K. (2017). Business Statistics for Contemporary Decision Making, Ninth Edition. Hoboken, NJ: John Wiley & Sons, Inc.

De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project, Journal of Statistics Education, 19:3. Retrieved from http://amstat.tandfonline.com/doi/abs/10.1080/10691898.2011.11889627#.WciC-WinFTE

De Cock, D. Ames Housing Data Documentation. Retrieved from https://ww2.amstat.org/publications/jse/v19n3/decock/DataDocumentation.txt

Pardoe, I. (2012). Applied Regression Modeling, Second Edition. Hoboken, NJ: John Wiley & Sons, Inc.

Stowell, S. (2014). Using R for Statistics. New York, NY: Apress.

Weisberg, S. (2014). Applied Linear Regression, Fourth Edition. Hoboken, NJ: John Wiley & Sons, Inc.

Code:

```
# Jennifer Wanat
# Fall 2017
# Ames_assignment3_final.R


install.packages("maptools")
require(maptools)

path.name <- "~/Desktop/R/"
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)

# Show the header of the data frame;
head(ames.df)

# Show the structure of the data frame;
str(ames.df)

#Creating a waterfall of drop conditions
ames.df$dropCondition <- ifelse(ames.df$BldgType!='1Fam','01: Not SFR',
                    ifelse(ames.df$SaleCondition!='Normal','02: Non-Normal Sale',
                        ifelse(ames.df$Street!='Pave','03: Street Not Paved',
                            ifelse(ames.df$GrLivArea >4000,'04: LT 4000 SqFt',
                                ifelse(ames.df$LotArea >100000,'05: Lot 100000 SqFt',
                                    ifelse(ames.df$BedroomAbvGr <1, '06: No Bedrooms',
                                        ifelse(ames.df$FullBath <1, '07: No Full Baths',
                                            ifelse(ames.df$PoolArea >0, '08: Pool',
                                        '99: Eligible Sample')
                            )))))))

table(ames.df$dropCondition)

# Save the table
waterfall <- table(ames.df$dropCondition);

# Format the table as a column matrix for presentation;
as.matrix(waterfall,8,1)

# 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;
table(eligible.population$dropCondition)
```

```
#check the structure of the data frame
str(eligible.population)

# Save the R data frame as an .RData object
saveRDS(eligible.population,file='/Users/jmwanat/Documents/Northwestern classes/MSPA
410/410 R/ames_assignment3.RData')


#***********Section*************
#EDA
#EDA of grade above living area
table(eligible.population$GrLivArea!="NA")
table(eligible.population$GrLivArea!=0)
summary(eligible.population$GrLivArea)
quantile(eligible.population$GrLivArea)
#describe(eligible.population$GrLivArea)

#EDA of total rooms above grade (does not include bathrooms)
table(eligible.population$TotRmsAbvGrd!="NA")
table(eligible.population$TotRmsAbvGrd!=0)
summary(eligible.population$TotRmsAbvGrd)
quantile(eligible.population$TotRmsAbvGrd)
table(eligible.population$TotRmsAbvGrd)


#scatterplot and histogram of above grade living area
par(mfrow = c(1,2))
plot(eligible.population$GrLivArea, eligible.population$SalePrice/1000,
    ylab = "Sale Price (000)",
    xlab = "Above Grade Living Area (square feet)",
    cex.lab = 1.5)
abline(lm(eligible.population$SalePrice/1000 ~ eligible.population$GrLivArea), col =
"red", lwd = 2, lty = 2)
GrLivArea.loess <-
loess(eligible.population$SalePrice/1000~eligible.population$GrLivArea)
GrLivArea.predict <- predict(GrLivArea.loess)
lines(eligible.population$GrLivArea[order(eligible.population$GrLivArea)],
GrLivArea.predict[order(eligible.population$GrLivArea)],
     col = "blue", lwd = 2, lty = 1)
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010")
hist(eligible.population$GrLivArea, xlab = "Above Grade Living Area (square feet)", col =
"lightgrey",
    ylim = c(0,800),
    main = "",
    cex.lab = 1.5)
```

```
par(mfrow = c(1,1))

#scatterplot and histogram of total rooms above grade
par(mfrow = c(1,2))
plot(eligible.population$TotRmsAbvGrd, eligible.population$SalePrice/1000,
    ylab = "Sale Price (000)",
    xlab = "Total Rooms Above Grade",
    cex.lab = 1.5)
abline(lm(eligible.population$SalePrice/1000 ~ eligible.population$TotRmsAbvGrd), col =
"red", lwd = 2, lty = 2)
TotRmsAbvGrd.loess <-
loess(eligible.population$SalePrice/1000~eligible.population$TotRmsAbvGrd)
TotRmsAbvGrd.predict <- predict(TotRmsAbvGrd.loess)
lines(eligible.population$TotRmsAbvGrd[order(eligible.population$TotRmsAbvGrd)],
TotRmsAbvGrd.predict[order(eligible.population$TotRmsAbvGrd)],
     col = "blue", lwd = 2, lty = 1)
#title(main = "Residential Properties sold in Ames, IA from 2006 - 2010")
hist(eligible.population$TotRmsAbvGrd, xlab = "Total Rooms Above Grade", col =
"lightgrey",
    ylim = c(0,800),
    main = "",
    cex.lab = 1.5)
par(mfrow = c(1,1))



#***********Section*************
# Fit a linear regression model with R
#Sale price as a function of Above Grade Living Area
model.1 <- lm(SalePrice ~ GrLivArea, data=eligible.population)

# Panel the plots
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(model.1)
par(mfrow = c(1,1))

# Hint:  We need to check the assumptions of normality and homoscedasticity;
# (1) QQ Plot
# (2) Scatterplot of residuals versus predictor

par(mfrow = c(1,2))
# Use the Base R functon qqplot() to assess the normality of the residuals
qqnorm(model.1$residuals, cex.lab = 1.5)
qqline(model.1$residuals)
# Make a scatterplot
plot(eligible.population$GrLivArea,model.1$residuals,
par(mfrow = c(1,3))
# Use the Base R functon qqplot() to assess the normality of the residuals
qqnorm(model.3$residuals, cex.lab = 1.5)
qqline(model.3$residuals)
# Make a scatterplot
plot(eligible.population$GrLivArea, model.3$residuals,
    ylab = "Residuals",
    xlab = "Above Grade Living Area (square feet)",
    cex.lab = 1.5)
plot(eligible.population$TotRmsAbvGrd, model.3$residuals,
    ylab = "Residuals",
    xlab = "Total Rooms Above Grade",
    cex.lab = 1.5)
par(mfrow = c(1,1))

summary(model.3)



#***********Section*************
#*************Neighborhood Accuracy*******************
par(mfrow = c(3,1))
#boxplot of model1 residuals versus neighborhood
boxplot(model.1$residuals ~ eligible.population$Neighborhood, col = "lightgrey", las = 2)
title(main = "Residuals versus Neighborhood\nModel 1")
#boxplot of model2 residuals versus neighborhood
boxplot(model.2$residuals ~ eligible.population$Neighborhood, col = "lightgrey", las = 2)
title(main = "Model 2")
#boxplot of model3 residuals versus neighborhood
boxplot(model.3$residuals ~ eligible.population$Neighborhood, col = "lightgrey", las = 2)
title(main = "Model 3")
par(mfrow = c(1,1))


#Create sale price per square foot variable
eligible.population$PricebySqft <-
eligible.population$SalePrice/eligible.population$GrLivArea

#Compute the mean sale price per square foot by neighborhood
avgSalePricePerSqft <- aggregate(eligible.population$PricebySqft,
by=list(Neighborhood=eligible.population$Neighborhood), FUN=mean)
#Change the column names
colnames(avgSalePricePerSqft) <- c('Neighborhood','AvgSalePricePerSqft')

#Compute the mean MAE for each neighborhood
```

```
avgMAE <- aggregate((abs(model.3$residuals)),
by=list(Neighborhood=eligible.population$Neighborhood), FUN=mean)
colnames(avgMAE) <- c('Neighborhood','AvgMAE')

#plot average MAE vs average price/sqft for each neighborhood
plot(avgSalePricePerSqft$AvgSalePricePerSqft, avgMAE$AvgMAE,
    xlab = "Average Sale Price per Square Foot",
    ylab = "Average Mean Absolute Error (MAE)")
pointLabel(avgSalePricePerSqft$AvgSalePricePerSqft, avgMAE$AvgMAE, labels =
avgMAE$Neighborhood)
#text(avgSalePricePerSqft$AvgSalePricePerSqft, avgMAE$AvgMAE,
avgMAE$Neighborhood,
#    cex = 0.5, pos = 4, col = "blue")


#table of Neighborhood counts
table(eligible.population$Neighborhood)

# Let's create a family of indicator variables for neighborhoods by price per square foot;
eligible.population$Neighborhood1 <- ifelse((eligible.population$Neighborhood ==
c("IDOTRR")|
                                  (eligible.population$Neighborhood == c("OldTown")|
                                    (eligible.population$Neighborhood == c("SWISU")))), 1,0);

eligible.population$Neighborhood2 <- ifelse((eligible.population$Neighborhood ==
c("BrkSide")|
                                  (eligible.population$Neighborhood == c("Edwards"))), 1,0);

eligible.population$Neighborhood3 <- ifelse((eligible.population$Neighborhood ==
c("NWAmes")|
                                  (eligible.population$Neighborhood == c("Crawfor")|
                                    (eligible.population$Neighborhood == c("Gilbert")|
                                      (eligible.population$Neighborhood == c("NAmes")|
                                        (eligible.population$Neighborhood == c("Sawyer")|
                                          (eligible.population$Neighborhood == c("SawyerW")|
                                            (eligible.population$Neighborhood == c("Blmngtn")|
                                              (eligible.population$Neighborhood == c("ClearCr")|
                                                (eligible.population$Neighborhood ==
c("NoRidge")|
                                                  (eligible.population$Neighborhood ==
c("Veenker")|
                                                    (eligible.population$Neighborhood ==
c("CollgCr")|
                                                      (eligible.population$Neighborhood ==
c("Mitchel")
                                                       ))))))))))), 1,0);
```

```
eligible.population$Neighborhood4 <- ifelse((eligible.population$Neighborhood ==
c("Timber")|
                                    (eligible.population$Neighborhood == c("Somerst"))), 1,0);

eligible.population$Neighborhood5 <- ifelse((eligible.population$Neighborhood ==
c("StoneBr")|
                                   (eligible.population$Neighborhood == c("NridgHt"))), 1,0);

table(eligible.population$Neighborhood1)
table(eligible.population$Neighborhood2)
table(eligible.population$Neighborhood3)
table(eligible.population$Neighborhood4)
table(eligible.population$Neighborhood5)


#Sale price as a function of Above Grade Living Area and TotRmsAbvGrd and
Neighborhood
model.4 <- lm(SalePrice ~ GrLivArea + TotRmsAbvGrd +
             Neighborhood2 + Neighborhood3 + Neighborhood4 + Neighborhood5,
data=eligible.population)

# Panel the plots
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(model.4)
par(mfrow = c(1,1))

par(mfrow = c(1,3))
# Use the Base R functon qqplot() to assess the normality of the residuals
qqnorm(model.4$residuals, cex.lab = 1.5)
qqline(model.4$residuals)
#scatterplots
plot(eligible.population$GrLivArea, model.4$residuals,
    ylab = "Model 4 Residuals",
    xlab = "Above Grade Living Area (square feet)",
    cex.lab = 1.5)
plot(eligible.population$TotRmsAbvGrd, model.4$residuals,
    ylab = "Model 4 Residuals",
    xlab = "Total Rooms Above Grade",
    cex.lab = 1.5)
par(mfrow = c(1,1))

par(mfrow = c(2,2))
plot(eligible.population$Neighborhood2, model.4$residuals,
    ylab = "Model 4 Residuals",
    xlab = "Neighborhood Group 2",
    lab = c(1,5,1), cex.lab = 1.5)
plot(eligible.population$Neighborhood3, model.4$residuals,
    ylab = "Model 4 Residuals",
    xlab = "Neighborhood Group 3",
    lab = c(1,5,1), cex.lab = 1.5)
plot(eligible.population$Neighborhood4, model.4$residuals,
    ylab = "Model 4 Residuals",
    xlab = "Neighborhood Group 4",
    lab = c(1,5,1), cex.lab = 1.5)
plot(eligible.population$Neighborhood5, model.4$residuals,
    ylab = "Model 4 Residuals",
    xlab = "Neighborhood Group 5",
    lab = c(1,5,1), cex.lab = 1.5)
par(mfrow = c(1,1))

summary(model.4)


#Compute the mean MAE for each neighborhood
avgMAE.4 <- aggregate((abs(model.4$residuals)),
by=list(Neighborhood=eligible.population$Neighborhood), FUN=mean)
colnames(avgMAE.4) <- c('Neighborhood','AvgMAE')




#***********Section*************
#Sale price as a function of multiple predictor variables
model.5 <- lm(SalePrice ~ GrLivArea + TotalBsmtSF + BsmtFinSF1 + GarageArea +
YearBuilt + Neighborhood5, data=eligible.population)

# Panel the plots
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(model.5)
par(mfrow = c(1,1))

par(mfrow = c(1,3))
# Use the Base R functon qqplot() to assess the normality of the residuals
qqnorm(model.5$residuals, cex.lab = 1.5)
qqline(model.5$residuals)
#scatterplots
plot(eligible.population$GrLivArea, model.5$residuals,
    ylab = "Model 5 Residuals",
    xlab = "Above Grade Living Area (square feet)",
    cex.lab = 1.5)
```

```
plot(eligible.population$TotalBsmtSF, model.5$residuals,
    ylab = "Model 5 Residuals",
    xlab = "Total Square Feet of Basement Area",
    cex.lab = 1.5)
par(mfrow = c(1,1))

par(mfrow = c(1,4))
plot(eligible.population$BsmtFinSF1, model.5$residuals,
    ylab = "Model 5 Residuals",
    xlab = "Basement Finished Area",
    cex.lab = 1.5)
plot(eligible.population$GarageArea, model.5$residuals,
    ylab = "Model 5 Residuals",
    xlab = "Garage Area (square feet)",
    cex.lab = 1.5)
plot(eligible.population$YearBuilt, model.5$residuals,
    ylab = "Model 5 Residuals",
    xlab = "Year Built",
    cex.lab = 1.5)
plot(eligible.population$Neighborhood5, model.5$residuals,
    ylab = "Model 5 Residuals",
    xlab = "Neighborhood Group 5",
    lab = c(1,5,1), cex.lab = 1.5)
par(mfrow = c(1,1))

summary(model.5)




#log(Sale price) as a function of multiple predictor variables
model.6 <- lm(log(SalePrice) ~ GrLivArea + TotalBsmtSF + BsmtFinSF1 + GarageArea +
YearBuilt + Neighborhood5, data=eligible.population)

# Panel the plots
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(model.6)
par(mfrow = c(1,1))

par(mfrow = c(1,3))
# Use the Base R functon qqplot() to assess the normality of the residuals
qqnorm(model.6$residuals, cex.lab = 1.5)
qqline(model.6$residuals)
#scatterplots
plot(eligible.population$GrLivArea, model.6$residuals,
    ylab = "Model 6 Residuals",
    xlab = "Above Grade Living Area (square feet)",
    cex.lab = 1.5)
plot(eligible.population$TotalBsmtSF, model.6$residuals,
    ylab = "Model 6 Residuals",
    xlab = "Total Square Feet of Basement Area",
    cex.lab = 1.5)
par(mfrow = c(1,1))

par(mfrow = c(1,4))
plot(eligible.population$BsmtFinSF1, model.6$residuals,
    ylab = "Model 6 Residuals",
    xlab = "Basement Finished Area",
    cex.lab = 1.5)
plot(eligible.population$GarageArea, model.6$residuals,
    ylab = "Model 6 Residuals",
    xlab = "Garage Area (square feet)",
    cex.lab = 1.5)
plot(eligible.population$YearBuilt, model.6$residuals,
    ylab = "Model 6 Residuals",
    xlab = "Year Built",
    cex.lab = 1.5)
plot(eligible.population$Neighborhood5, model.6$residuals,
    ylab = "Model 6 Residuals",
    xlab = "Neighborhood Group 5",
    lab = c(1,5,1), cex.lab = 1.5)
par(mfrow = c(1,1))

summary(model.6)
```