STAT 6371 PROJECT

An analysis exploring relationships between home sales prices, square footage, and neighborhood locations & creating a predictive model for sales prices of homes in all of Ames, Iowa, neighborhoods.

Analysis presented by:

Joshua Hudson - Hudson Data Science Website (hud44.github.io)

Akib Hossain – https://github.com/flyche213

TABLE OF CONTENTS

INTRODUCTION

Century 21 in Ames, Iowa, reached out to our team to gather evidence of a relationship between the sale prices of houses and square footage and to help determine if that is dependent on the neighborhoods where each house is located.

Additionally, Century 21 requested that our team build a predictive model for sale prices of houses in all of the neighborhoods in Ames, Iowa.

To gather evidence for both analysis questions, our team was directed to procure data from the following source: https://www.kaggle.com/c/house-prices-advanced-regression-techniques. See the Data Description section of Appendix A for complete details regarding this data for each analysis in this report.

ANALYSIS 1

Problem Statement

Century 21 reached out to our team to gather evidence of a relationship between the sale prices of houses and their square footage. Additionally, Century 21 requested our team to gather evidence to determine if the relationship between home sale prices and square footage is dependent on the neighborhoods where each house is located.

Multiple Linear Regression Model

Figure 1 is a scatter plot that shows the apparent linear associations between sales prices and square footage for houses sold in each of the three neighborhoods included within the scope of this analysis.

To formally model these relationships, our team selected to use Multiple Linear Regression (MLR) to construct a model of the association between the average home sale price and square footage and determine if this association is dependent on which of the neighborhoods each house is located. This MLR model can be shorthandedly described by the following mathematical equation:

$$\mu\{SP \mid sf, BS, ED\} = \beta_0 + \beta_1 * sf + \beta_2 * BS + \beta_3 * ED + \beta_4 * (sf \times BS) + \beta_5 * (sf \times ED)$$

$$SP = Sales\ Price;\ sf = Square\ Footage;\ BS = BrkSide\ Neighborhood;\ ED = Edwards\ Neighborhood;\ NA = NAmes\ Neighborhood$$

NA is missing from this model as it is used as this model's point of reference (0 point)

MLR Model Assumptions

In order for our team to be confident in this MLR model, we explored the data to ensure it met the following basic assumptions for MLR.

1. **Normality:**
   Judging from the histogram (Figure 2) of residuals and q-q plot (Figure 3), there is some evidence of long tailed distribution due to some outliers in the data that must be investigated.

2. **Linearity:**
   Judging from the scatter plot in Figure 4, there is strong visual evidence of linearity between sale prices of houses and square footage between and within each neighborhood of interest.

3. **Equality of Variance:**
   Judging from the residual plot in Figure 5, though there is some evidence of heteroscedasticity, it is not strong enough to reject this assumption.

4. **Independence:**
   For the purposes of this study, we will assume each observation is independent of each other WITHIN each neighborhood of interest. It is understood that there is risk of clustering dependence BETWEEN neighborhoods of interest.

Some outliers were removed in order to obtain the best fitted model. See MLR Assumptions section in Appendix B for complete details as to the logic for removing these outliers.

Conclusion

In conclusion, there is strong evidence that a relationship between sale prices of houses and square footage exists, and this relationship is dependent upon which of the three neighborhoods in Ames, Iowa (NAmes, BrkSide, and Edwards) the house is located.

Key Takeaways

1. For houses in the NAmes neighborhood, our team is 95% confident that every increase of one hundred square feet is associated with an average home sale price increase between $4,150.50 and $5,761.75. Our best estimate for this associated increase is $4,956.13/100ft² (p-value < 0.0001).

2. For houses in the BrkSide neighborhood, our team is 95% confident that every increase of one hundred square feet is associated with an average home sale price increase between $1,911.34 and $5,608.92 higher than that of NAmes. Our best estimate for this associated increase for homes sold in BrkSide is $3,760.13/100ft² higher than those sold in NAmes (p-value < 0.0001).

3. For houses in the Edwards neighborhood, our team is 95% confident that every increase of one hundred square feet is associated with an average home sale price increase between $446.55 and $3,672. 87 higher than that of NAmes. Our best estimate for this associated increase for homes sold in Edwards is $2,059.71/100ft² higher than those sold in NAmes (p-value = 0.0125).

See Fitted Full MLR Model section of Appendix B for complete details regarding the full MLR model. Additionally, the Full MLR model was compared to models selected through more advanced selection techniques. For more details about these comparison models, see Comparing Competing Models section of Appendix B.

It should be noted that this data was not procured from a randomized study. Therefore, it would be inappropriate for our team to conclude that neither the square footage nor the neighborhood the house is located in during the time of the sale drives sale prices of houses. Additionally, the data is observational in nature and samples were not randomly drawn. Therefore, it would be inappropriate to infer these results to the rest of the homes in Ames, Iowa. However, the results are interesting and may be used to support effective decision making.

To visualize the differences in the associations between Sale Prices and Square footage for each of the three neighborhoods, visit: C21HomeSalePriceApp.knit (shinyapps.io).

ANALYSIS 2

Problem Statement

Century 21 in Ames, Iowa, reached out to our team to build a predictive model for sales prices of homes in all of neighborhoods in Ames, Iowa.

Model Assumptions for Selected Variables

In order to determine which variables would provide the best model for predicting the sales prices of houses, our team explored the data and used our domain knowledge and common sense to select preliminary variables to use for the predictive model. For complete details regarding this variable selection process, see the Manual Variable Selection section in Appendix C.

In order for our team to be confident in this MLR model, we explored the data to ensure it met the following basic assumptions for regression.

1. **Normality:**
   Judging from the histogram of residuals and q-q plot (**Figure 11**), there is some evidence of long tailed distribution due to some outliers in the data that must be investigated.

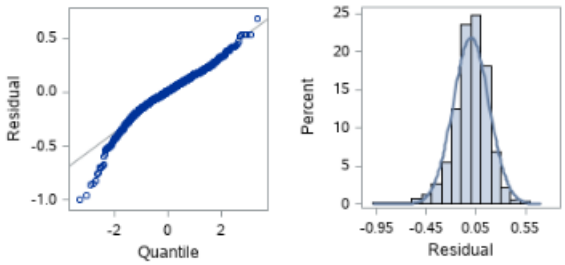

Figure 11

2. **Linearity:**
   Judging from the scatter plot in **Figure 12**, there is strong visual evidence of linearity between sale prices of houses and square footage between and within each neighborhood of interest.

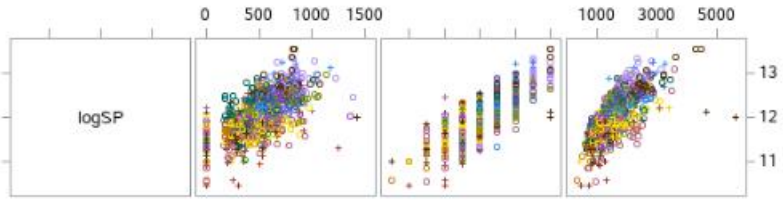

Figure 12

3. **Equality of Variance:**
   Judging from the residual plot in **Figure 13**, though there is some evidence of heteroscedasticity, it is not strong enough to reject this assumption.

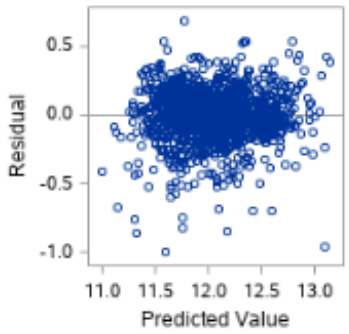

Figure 13

4. **Independence:**
   For the purposes of this study, we will assume each observation and variable is independent of each other WITHIN each neighborhood of interest. It is understood that there is risk of clustering dependence BETWEEN neighborhoods of interest.

As stated in the normality and linearity assumption analysis, the following outliers were investigated as they appeared to be heavily influential on the fitted model due to being unusually distant from the other explanatory values (See the RStudent and Cook's D plots found in **Figure 14** for additional evidence).

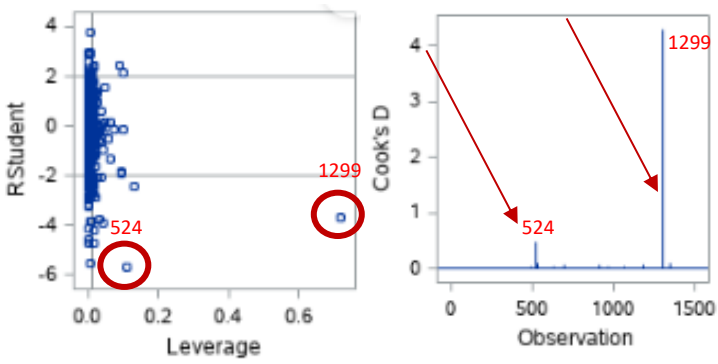

Figure 14

The outlying values were investigated in the dataset, but there was no real explanation identified. It is likely the sales price or square footage values were recorded in error and/or the neighborhood was mis-recorded for the observation. Though our team cannot account for the true root cause of the outlying data points, we ultimately decided to remove these values. We determined that the risk of overfitting our model by including potentially erroneous outlying explanatory values outweighed the risk of underfitting our model by excluding true/non-erroneous outlying explanatory values.

With these outlying explanatory values removed, visual exploration of the data reveals stronger evidence that the remaining data meets the basic assumptions for multiple linear regression and thus, would create a strong predictive model. See **Figure 15** for visual evidence of the met MLR assumptions post outlier removal.

With all this in mind, our team chose to proceed with caution in developing our predictive model with the highly influential outliers removed from the training data at the risk of marginal underfitting.

Model Selection

After strong visual associations were identified, our team then deployed the following advanced regression techniques for the final predictive model. **Table 1** shows each techniques predictive statistics:

Predictive Models	Adjusted R2	CV PRESS	Kaggle Score
Backward	0.8512	25.77800	0.18492
Forward	0.8450	26.98540	0.18905
Stepwise	0.8425	27.27974	0.18606
No Elimination	0.8655	1232.34038	0.92042

Table 2

As **Table 2** shows Backward Elimination produced the best predictive model with the variables our team manually selected. Below is the predictive model that Backward Elimination produces:

$$\boldsymbol{Predicted\ LogSP = \beta_0 + \beta_1 * gSF + \beta_2 * OQual + \beta_3 * lSF + \beta_4 * (gSF \times lSF) + \beta_5 * (gSF \times OQual \times lSF) + \beta_{i=1-25} * (NH_{i=1-25})}$$

$$\boldsymbol{Predicted\ Sale\ Price = e^{LogSP}}$$

Conclusion

As can be seen in **Table 2**, the explanatory variables in the backward elimination prediction model account for more than 85% of the variation of logged sale prices for houses sold in all of the neighborhoods in Ames, Iowa. To use this model, all one would need to obtain is a home's general living area square footage, garage square footage, and overall quality rating for a house located in any of the 25 neighborhoods in Ames, Iowa. Armed with that information, one could quite easily obtain the homes predicted sale price estimate with reasonable confidence.

Appendix A:

Data Description

To gather evidence for both analysis questions, our team was directed to procure data from the following source: https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Analysis Question 1:

The data set our team used to construct our fitted model from this source is as follows:

- train.csv – 80 Variables with 1,460 observations
    - This data was used to develop our model
    - Observations were reduced to only include the observations for homes located in the 3 neighborhoods of interest
        i.e. - 1,460 observations were reduced to 383 observations from this data set

From this set, the variables of interest are as follows:

1. Response variable:
    - Sale Price (SP) – The sales price of the homes included in the study
        - The column title in this set is "SalePrice"

2. Explanatory variables:
    - Square Footage (SF) – The square footage of the homes included in the study
        - The column title in this set is "GrLIvArea"

    - Neighborhood (NH) – The neighborhoods* of homes included in the study
        - The column title in this set is "Neighborhood"
        **Note: The only neighborhoods of interest for Century 21 in this study are NAmes, Edwards and BrkSide in Ames, Iowa.*

Analysis Question 2:

The data set our team used to construct our fitted model from this source is as follows:

- train.csv – 80 Variables with 1,460 observations
    - This data was used to develop our model

From this set, the variables of interest are as follows:

- Response – ***logSP***: log(**SalePrice**):
- Explanatory variables
    1. ***GarageArea (gSF)****:* Square footage of the home garage
    2. ***OverallQual (OQual)****:* Home quality score (1-10; 10 assumed to be highest)
    3. ***GrLivArea (lSF)****:* Square footage of the total home living area
    4. ***Neighborhood (NH)****:* The neighborhood of where the home is located

Appendix B: Analysis 1 Reference Materials

Scatterplot: Sale Prices vs. Square Footage

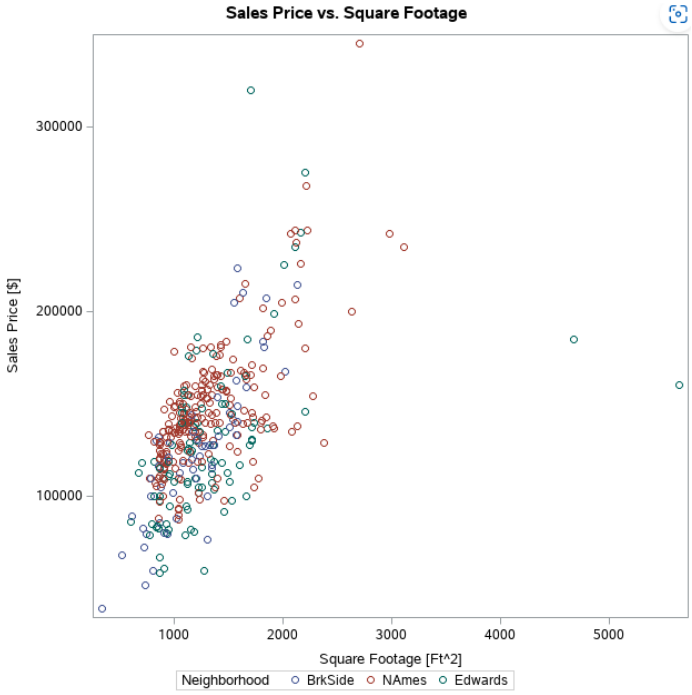

Figure 1

(Back)

Normality & QQ Plots with Outliers

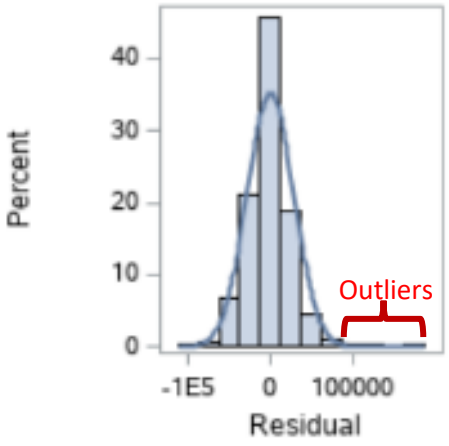

Figure 2

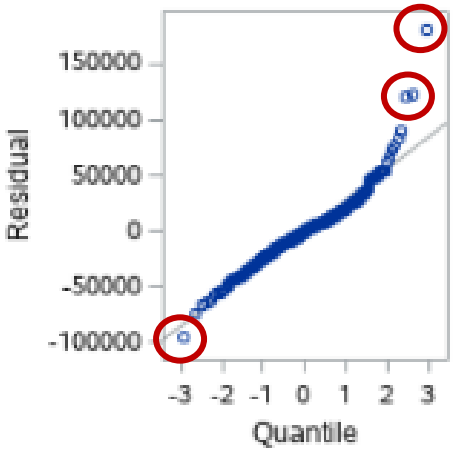

Figure 3

(Back)

Linearity Plot with Outliers

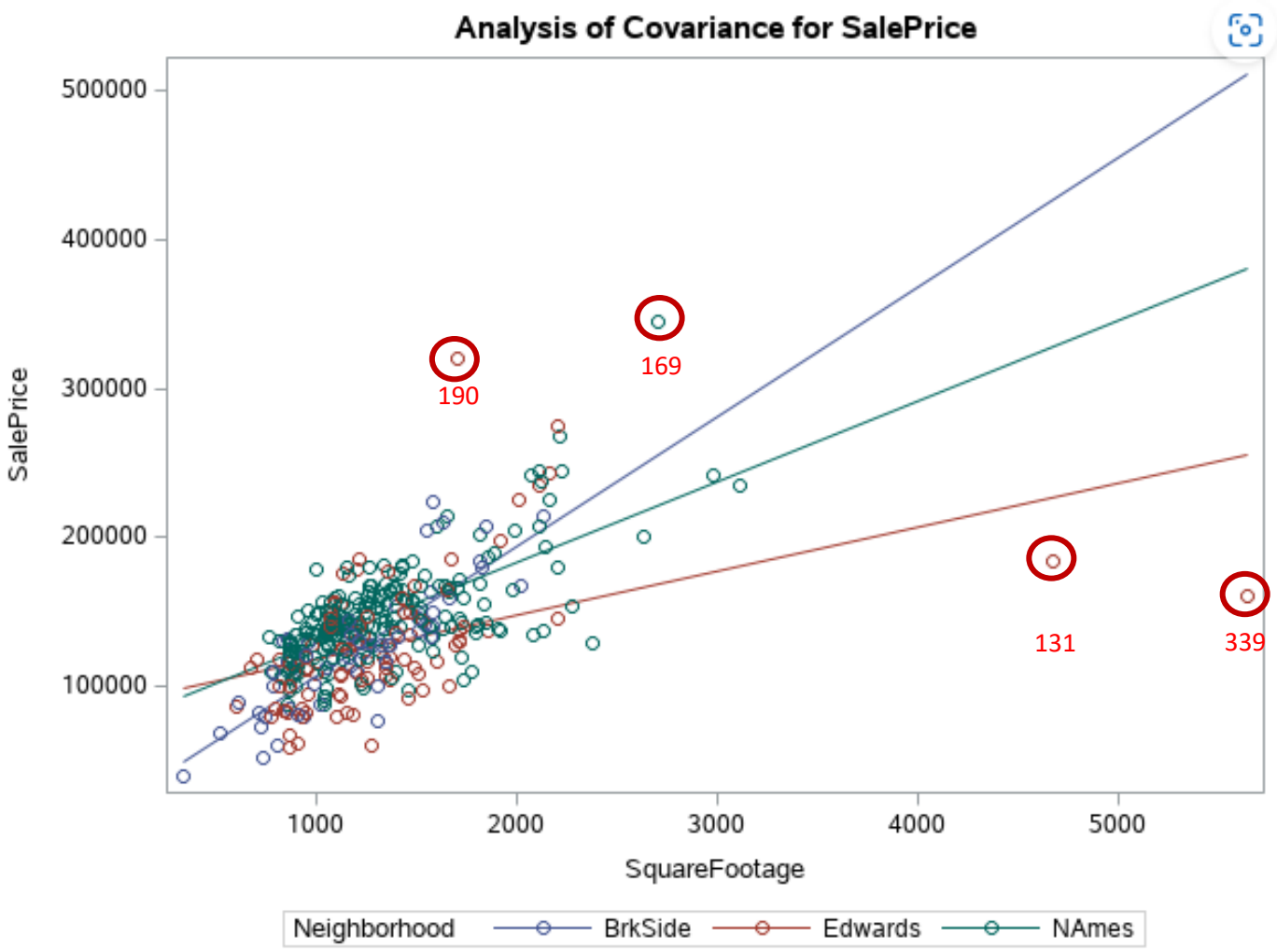

Figure 4

(Back)

Residual Plot with Outliers

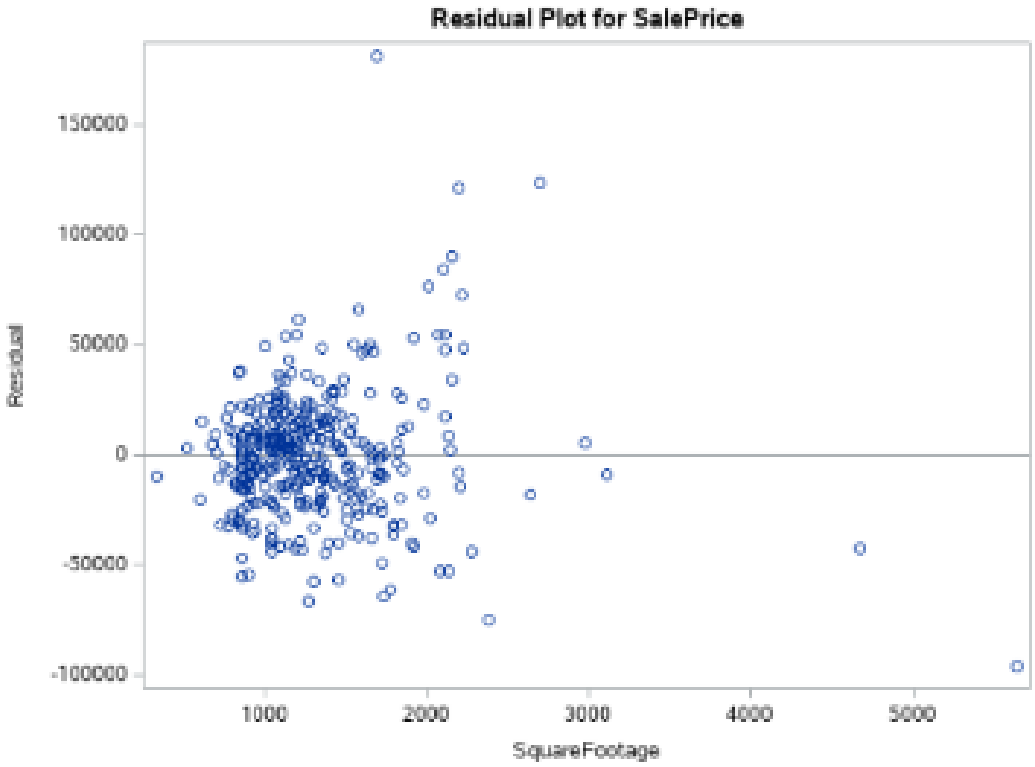

Figure 5

(Back)

MLR Assumptions (Outlier Removal)

As stated in the normality and linearity assumption analysis, the following outliers were investigated as they appeared to be heavily influential on the fitted model due to being unusually distant from the other explanatory values (See the RStudent and Cook's D plots found in **Figure 6** for additional evidence).

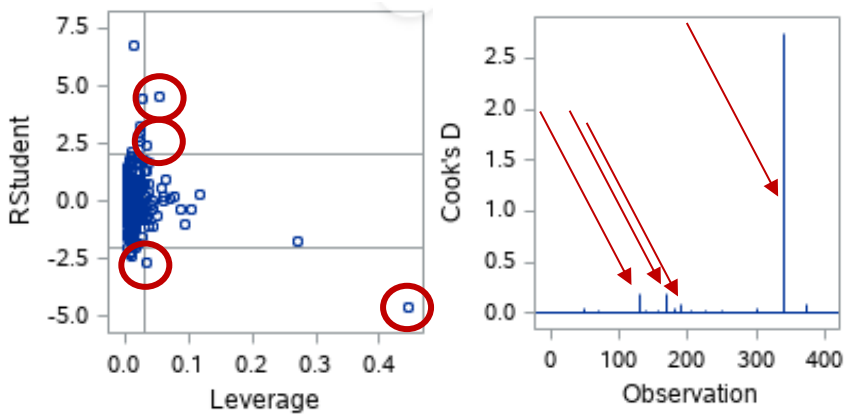

Figure 6

(Back)

The outlying values were investigated in the dataset, but there was no real explanation identified. It is likely the sales price or square footage values were recorded in error and/or the neighborhood was mis-recorded for the observation. Though our team cannot account for the true root cause of the outlying data points, we ultimately decided to remove these values. We determined that the risk of overfitting our model by including potentially erroneous outlying explanatory values outweighed the risk of underfitting our model by excluding true/non-erroneous outlying explanatory values.

With these outlying explanatory values removed, visual exploration of the data reveals stronger evidence that the remaining data meets the basic assumptions for multiple linear regression. See **Figure 7** for visual evidence of the met MLR assumptions post outlier removal.

As **Figure 7** reveals, the strength supporting each assumption for MLR increased after the removal of the outlying explanatory variables. Therefore, our team chose to proceed with caution in developing our MLR with the highly influential outliers removed from the training data at the risk of marginal underfitting.

MLR Fit Diagnostics w/ outliers removed

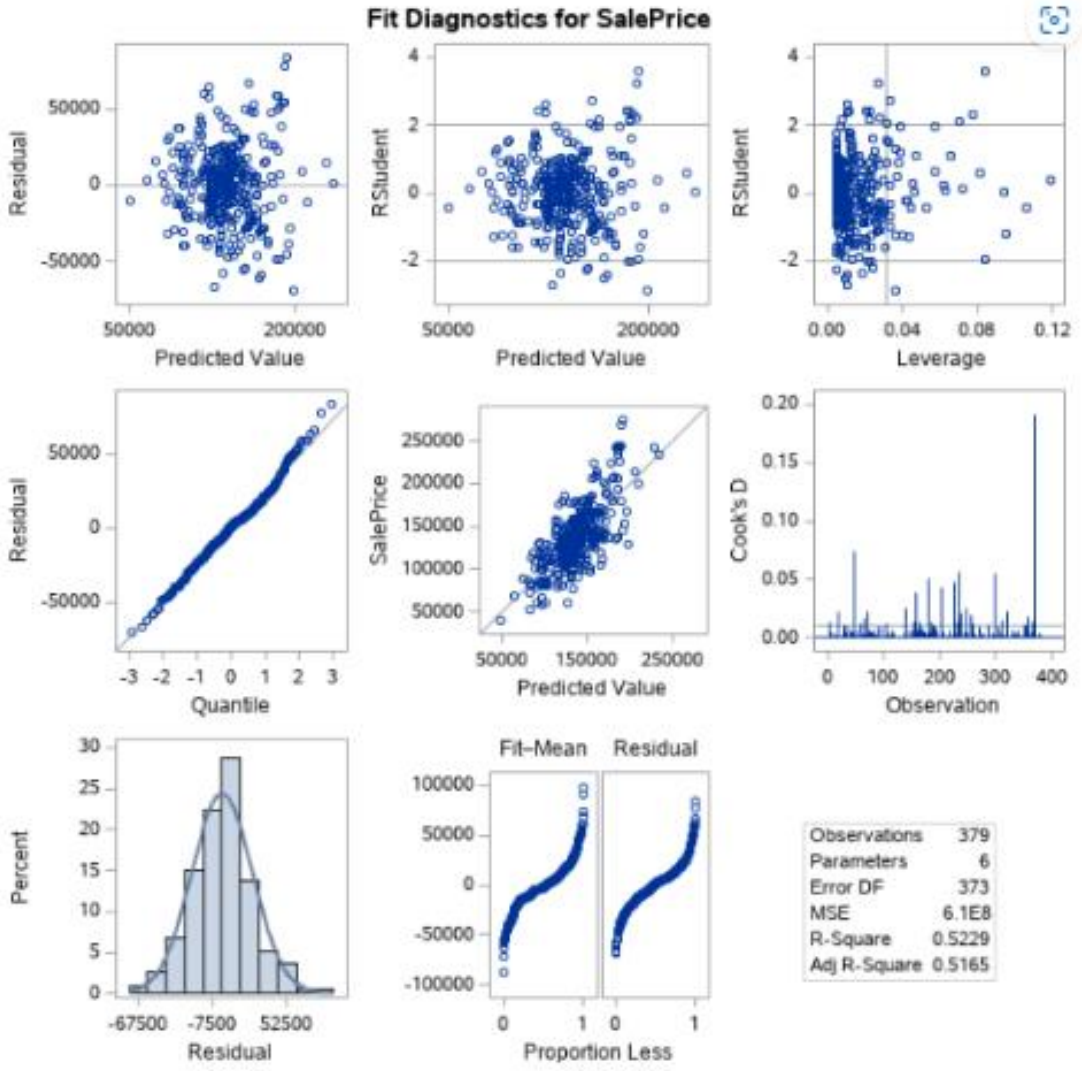

Figure 7

Comparing Competing Models

In selecting the appropriate model, we needed to gather evidence that the association between sale prices of houses and square footage differs between the three neighborhoods.

To do this, our team deployed an extra sum of squares F-test to compare the Full Model* and of a Parallel Model**.

Note 1: Full Model in this context refers to an MLR model that results in separate magnitudes of association between sales price & square footage due to the neighborhood where the house is located.

****Note 2: Parallel Model in this context refers to an MLR model that does not result in separate magnitudes of association between sales price & square footage due to the neighborhood where the house is located.***

Below is the output of the least squares fit to the regression of the Full Model (w/ interactions):

$$\boldsymbol{Average\ Sales\ Price = Square\ Footage + NEGIHBORHOODS + (Square\ Footage \times NEGIHBORHOODS)}$$

$$\boldsymbol{\mu\{sf, BS, ED\} = \beta_0 + sf * \beta_1 + BS * \beta_2 + ED * \beta_3 + (sf \times BS) * \beta_4 + (sf \times ED) * \beta_5}$$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	249429640074	49885928015	81.76	<.0001
Error	373	227584871181	610147107.72		
Corrected Total	378	477014511255			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.522897	18.04909	24701.16	136855.4

A p-value < 0.0001 is strong evidence that at least one of the slopes in the Full Model differs from zero. Namely, there is evidence that the Full Model explains over 50% of the variation of the response.

Below is the output of the least squares fit to the regression of the Parallel Model (w/out interactions):

$$\boldsymbol{Average\ Sales\ Price = Square\ Footage + NEGIHBORHOODS}$$

$$\boldsymbol{\mu\{sf, BS, ED\} = \beta_0 + sf * \beta_1 + BS * \beta_2 + ED * \beta_3}$$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	237947769241	79315923080	124.41	<.0001
Error	375	239066742014	637511312.04		

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Corrected Total	378	477014511255			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.498827	18.44939	25248.99	136855.4

Again, a p-value < 0.0001 is strong evidence that at least one of the slopes in the Parallel Model differs from zero. Namely, there is evidence that the Parallel Model explains nearly 50% of the variation of the response.

Below is the output of the extra sum of squares F-test for comparing the Full Model to the Parallel Model:

	Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
	Model	2	11481870833	5740935417	9.4091	0.0001
Full Model	Error	373	227584871181	610147107.72		
Parallel Model	Corrected Total	375	239066742014			

*The small p-value (*0.0001*) in this Full Model vs. Parallel Model comparison test indicates that the association between sale prices of houses and square footage does differ between the three neighborhoods.*

Our team deduced from the extra sum of squares F-test that it was inappropriate to exclude the interaction terms that the Full Model provides. Therefore, our team proceeded with confidence that there is indeed a relationship between living area and sale price with respect to the NAmes, Edwards, and BrkSide neighborhoods. Our task was to fit a model that best describes this relationship.

To determine if the fitted Full Model best describes the relationship between living area and sale price with respect to neighborhoods of interest, our team deployed advanced regression techniques to see if all of the interactions truly provide an optimal model for sale price prediction. Our team used the forward & backward elimination techniques and stepwise regression techniques for optimal parameter selection. We compared each of these model's Adj R^2 and Internal CV Press***** and their logical practicalities. The following are the results of this comparison:

Note 1: Generally, a high** Adj R^2 **and low CV Press indicate the "best" fitted model. Some exceptions to this general rule apply.

1. **Forward Selection Model:**

$$\mu\{SP \mid sf, BS, ED\} = \beta_0 + \beta_2 * BS + \beta_3 * ED + \beta_4 * (sf \times BS) + \beta_5 * (sf \times ED)$$

$$sf \sim \textbf{Square Footage was removed}$$

- Adj R^2 = 0.52

- Internal CV Press = 1.648327E11

2. **Backward Elimination Model**

$$\mu\{SP \mid sf, BS, ED\} = \beta_0 + \beta_2 * BS + \beta_3 * ED + \beta_4 * (sf \times BS) + \beta_5 * (sf \times ED)$$

$$sf \sim Square\ Footage\ was\ removed$$

- Adj R^2 = 0.54
- Internal CV Press = 1.720894E11

3. **Stepwise Regression Model**

$$\mu\{SP \mid sf, BS, ED\} = \beta_0 + \beta_2 * BS + \beta_3 * ED + \beta_4 * (sf \times BS) + \beta_5 * (sf \times ED)$$

$$sf \sim Square\ Footage\ was\ removed$$

- Adj R^2 = 0.52
- Internal CV Press = 1.660324E11

4. **No Elimination Model**

$$\mu\{SP \mid sf, BS, ED\} = \beta_0 + \beta_1 * sf + \beta_2 * BS + \beta_3 * ED + \beta_4 * (sf \times BS) + \beta_5 * (sf \times ED)$$

$$No\ parameters\ were\ removed$$

- Adj R^2 = 0.50
- Internal CV Press = 1.685967E11

Though both the Adj R^2 and Internal CV Press values of the forward, backward, and stepwise elimination models was more optimal than the regression model without variables removed, the forward, backward, and stepwise elimination models removed square footage as an explanatory variable even though it was included in the explanatory variables of interactions with neighborhoods. It is inappropriate to exclude individual explanatory variables when there is a significant interaction of that variable with another. Therefore, our team chose to cautiously proceed with the Full Model (No variables excluded) to best explain the relationship between living area and sale price with respect to neighborhoods of interest.

Full Model Parameter Estimates Table

Parameter	Estimate	SE	t-val	p-val	95% Confidence Limits	
Intercept	80325.71	5592.04	14.36	<.0001	69329.84	91321.58
Square Footage	49.56	4.10	12.10	<.0001	41.51	57.62
BrkSide Neighborhood	-60354.20	12060.03	-5.00	<.0001	-84068.38	-36640.02
Edward Neighborhood	-43225.29	10837.82	-3.99	<.0001	-64536.17	-21914.41
Square Footage * BrkSide Neighborhood	37.60	9.40	4.00	<.0001	19.11	56.09

Parameter	Estimate	SE	t-val	p-val	95% Confidence Limits	
Square Footage * Edward Neighborhood	20.60	8.20	2.51	0.0125	4.47	36.73

Table 1

(Back)

Fitted Full MLR Model

Table 1 displays the parameter estimates for each explanatory variable's slope using NAmes as the reference neighborhood:

Using the parameter estimates provided in **Table 1**, the fitted Full Model for average sale prices of houses given square footage and neighborhood location can be expressed as follows:

$$\mu\{SP \mid sf, BS, ED\} = \$80,325.71 + 49.56 * sf - \$60,354.20 * BS - \$43,225.29 * ED + 37.60 * (sf \times BS) + 20.60 * (sf \times ED)$$

The following is a breakdown of how the Full Model estimates the association between the average sale prices for houses and square footage when neighborhood location for each home is accounted for:

1. **NAmes Neighborhood Model**

$$\mu\{SP \mid sf, NA\} = 80,325.71 + 49.56 * sf$$

For houses in the NAmes neighborhood, our team is 95% confident that every increase of one hundred square feet is associated with an average home sale price increase between $4,150.50 and $5,761.75. Our best estimate for this associated increase is $4,956.13/100ft² (p-value < 0.0001).

The following are 95% CI levels for each of the fitted model parameters.
Intercept: ($69,329.83, $91,321.58)
Slope: ($4,150.50/100ft², $5,761.75/100ft²)

2. **BrkSIde Neighborhood Model**

$$\mu\{SP \mid sf, BS\} = 19,971.51 + 87.16 * sf$$

For houses in the BrkSide neighborhood, our team is 95% confident that every increase of one hundred square feet is associated with an average home sale price increase between $7,052.22 and $10,380.29. Our best estimate for this associated increase is $8,716.25/100ft² (p-value < 0.0001).

The following are 95% CI levels for each of the fitted model parameters.
Intercept: ($0, $ 40,982.30)
Slope: ($7,052.22 /100ft², $10,380.29/ft²)

3. **Edwards Neighborhood Model**

$$\mu\{SP \mid sf, ED\} = 37,100.42 + 70.16 * sf$$

For houses in the Edwards neighborhood, our team is 95% confident that every increase of one hundred square feet is associated with an average home sale price increase between $5,618.25 and $8,413.43. Our best estimate for this associated increase is $7,015.84/100ft² (p-value < 0.0001).

The following are 95% CI levels for each of the fitted model parameters.
Intercept: ($18,845.44, $55,355.40)
Slope: ($5,618.25 /100ft², $8,413.43 /100ft²)

(back)

APPENDIX C: ANALYSIS 2 REFERENCE MATERIAL

Manual Variable Selection

In order to determine which variables would provide the best model for predicting the sales prices of houses, our team explored the data as follows:

1. Reviewed the SalePrice distribution to determine if it approximates a normal distribution. **Figure 8** below shows a clear right skew. This can be easily addressed by log-transforming the SalePrice variable. As **Figure 9** below shows, the log transformed distribution of SalePrice approximates a normal distribution much more closely.

2. Now that the distribution of **SalePrice** was addressed, our team reviewed all available explanatory variables and their values and filtered out those deemed by common sense and domain knowledge to be irrelevant to sale prices of houses. Examples of unused variables are as follows:
    - **BsmtCond**, **BsmtFullBath**, **WoodDeckSF**, etc.

3. Our team then plotted the remaining continuous variables to help visually discern any immediate strong associations. See **Figure 10** below for this visual analysis.
    - After further examination of the multi-scatter plot in **Figure 10**, it appears that **GrLivArea** is dependent on **firstFlSF**, **SecondFlrSF**, and **TotalBsmtSF** variable. This is logical as **GrLivArea** (total square footage of the home living space) would include first floor, second floor, and basement square footage. With this understanding, our team decided to only include **GrLivArea** in the model to simply it AND to prevent multi-variable dependency.

4. After this exploration of the data and manual variable selections, our team ended up with the following explanatory variable selections to be compared using advanced regression techniques:
    - Response – ***logSP***: log(**SalePrice**):
    - Explanatory variables
        1. ***GarageArea (gSF)****:* Square footage of the home garage
        2. ***OverallQual (OQual)****:* Home quality score (1-10; 10 assumed to be highest)
        3. ***GrLivArea (lSF)****:* Square footage of the total home living area
        4. ***Neighborhood (NH)****:* The neighborhood of where the home is located

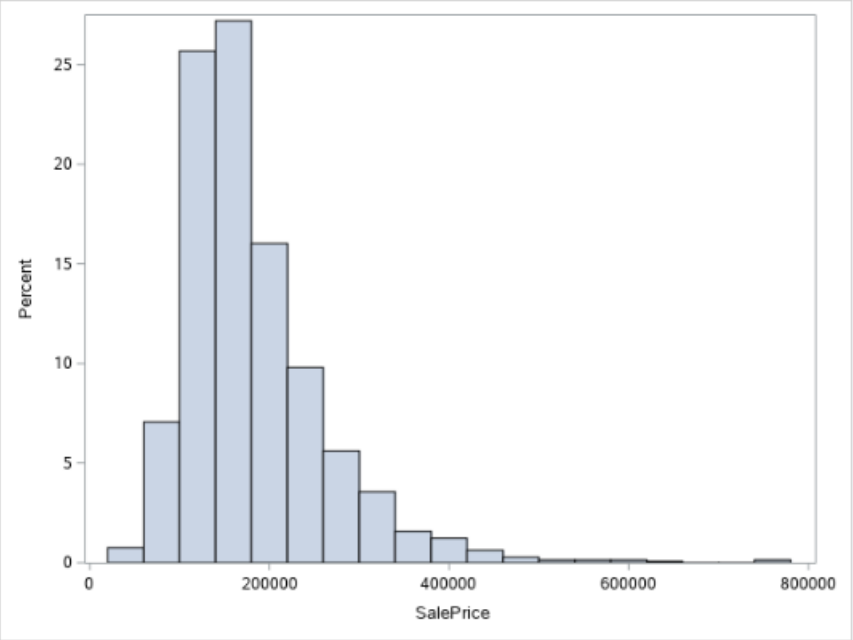

Figure 8

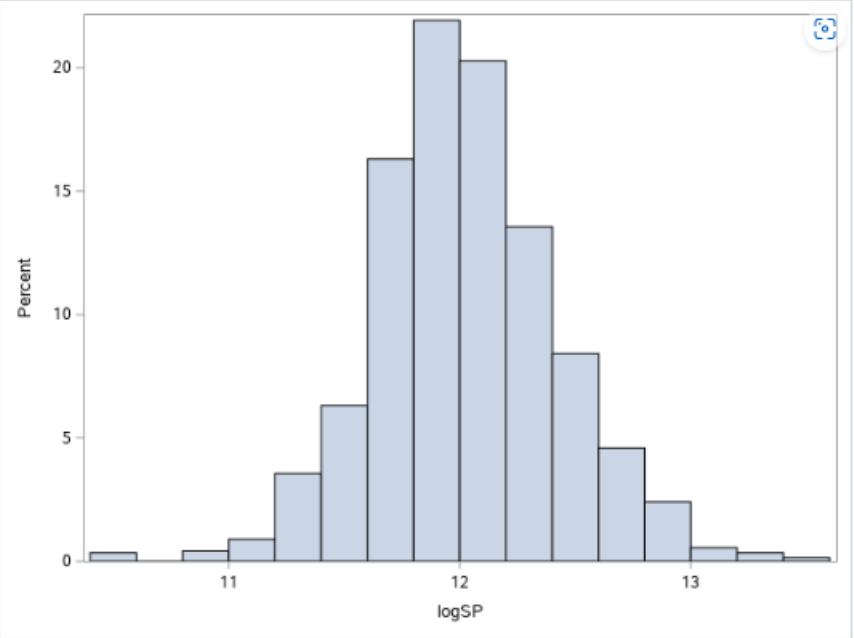

Figure 9

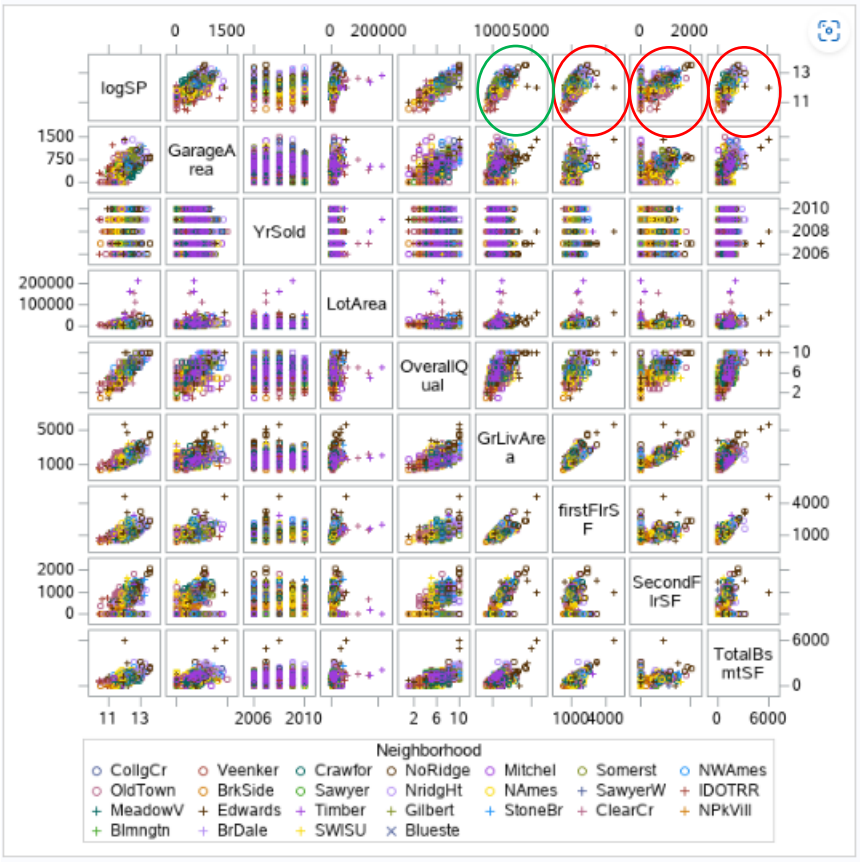

Figure 10

(Back)

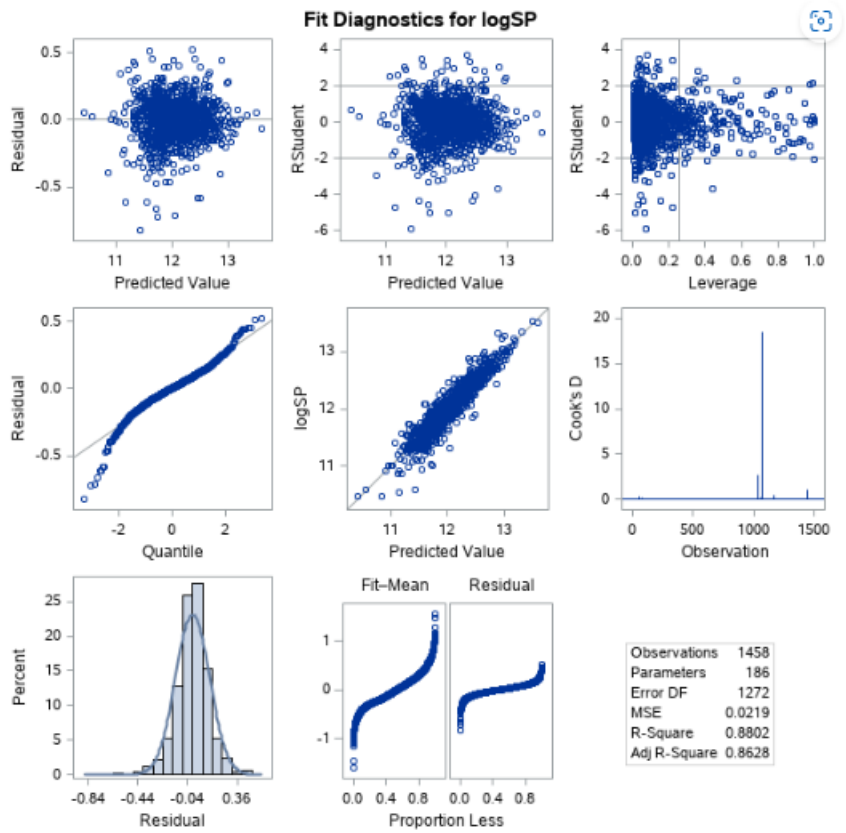
Fit Diagnostics for logSP
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
logSP
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 1458
Parameters 186
Error DF 1272
MSE 0.0219
R-Square 0.8802
Adj R-Square 0.8628

Figure 15

APPENDIX D: SAS CODE

Analysis 1 Code:

```
/* Importing the data*/

FILENAME REFFILE '/home/u60034905/Project/test.csv';


PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=test;

	GETNAMES=YES;

RUN;


FILENAME REFFILE '/home/u60034905/Project/train.csv';


PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=train;

	GETNAMES=YES;

RUN;


/* Exploring the relationships between Sale Price and Square Footage and dependance on Neighborhood
*/

data train2;

	/* filter training data for only three neighborhoods of interest */

	set train;


	if Neighborhood IN("NAmes", "Edwards", "BrkSide");

	logSalePrice=LOG(SalePrice);

	SquareFootage=GrLivArea;

	logSquareFootage=LOG(GrLivArea);

run;


/* Scatter Plot of data */

proc sgscatter data=train2;
```

```
	compare y=SalePrice x=(SquareFootage) / group=Neighborhood;
	title 'Sales Price vs. Square Footage w/ Outliers';
	label SquareFootage='Square Footage [Ft^2]';
	label SalePrice='Sales Price [$]';
run;

/*Full Model (with interaction between Explainatory variables) Prior to removing outliers*/
proc glm data=train2 plots=ALL;
	class Neighborhood;
	model SalePrice=SquareFootage | Neighborhood / solution clparm;
	run;

	/*Remove influential outliers*/
data train2_outlier_removed;
	set train2;

	if _n_=339 then

		/*high leverage and influence on edwards neighborhood*/
		delete;

	if _n_=131 then

		/*high leverage and influence on edwards neighborhood*/
		delete;

	if _n_=190 then

		/*high influence on edwards neighborhood*/
		delete;
```

```
	if _n_=169 then

		/*high leverage and influence on NAmes neighborhood*/
		delete;
	*/
run;

	/* Scatter Plot of data */
proc sgscatter data=train2_outlier_removed;
	compare y=SalePrice x=(SquareFootage) / group=Neighborhood;
	title 'Sales Price vs. Square Footage w/o Outliers';
	label SquareFootage='Square Footage [Ft^2]';
	label SalePrice='Sales Price [$]';
run;

/*Full Model (with interaction between Explainatory variables)*/
proc glm data=train2_outlier_removed plots=all;
	class Neighborhood;
	model SalePrice=SquareFootage | Neighborhood / solution clparm;
	run;

	/*Parallel Model (with no interaction between Explainatory variables)*/
proc glm data=train2_outlier_removed plots=all;
	class Neighborhood;
	model SalePrice=SquareFootage Neighborhood / solution clparm;
	run;

	/*Reduced Model (assuming parallel model with equal intercepts and no interaction between Explainatory variables)*/
```

```
proc glm data=train2_outlier_removed plots=fitplot;
	model SalePrice=SquareFootage / solution clparm;
	run;

	/*Full Model Opitmized*/
proc glmselect data=train2_outlier_removed seed=635422552;
	class Neighborhood;
	partition fraction(test=0.3);
	model SalePrice=SquareFootage|Neighborhood / selection=forward stats=adjrsq
		stats=press;
run;

proc glmselect data=train2_outlier_removed seed=695793964;
	class Neighborhood;
	partition fraction(test=0.3);
	model SalePrice=SquareFootage|Neighborhood / selection=backward stats=adjrsq
		stats=press;
run;

proc glmselect data=train2_outlier_removed seed=739101952;
	class Neighborhood;
	partition fraction(test=0.3);
	model SalePrice=SquareFootage|Neighborhood / selection=stepwise stats=adjrsq
		stats=press;
run;

proc glmselect data=train2_outlier_removed seed=512192299;
	class Neighborhood;
	partition fraction(test=0.3);
	model SalePrice=SquareFootage|Neighborhood / selection=none stats=adjrsq
```

```
                stats=press;
run;

/*Changing the reference neighborhoods to obtain 95% CIs for each parameter*/
proc glm data=train2_outlier_removed plots=all;
        class Neighborhood (ref="NAmes");
        ;
        model SalePrice=SquareFootage | Neighborhood / solution clparm;
        run;

proc glm data=train2_outlier_removed plots=all;
        class Neighborhood (ref="BrkSide");
        ;
        model SalePrice=SquareFootage | Neighborhood / solution clparm;
        run;

proc glm data=train2_outlier_removed plots=all;
        class Neighborhood (ref="Edwards");
        ;
        model SalePrice=SquareFootage | Neighborhood / solution clparm;
        run;
```

Analysis 2 Code:

```
/* Importing the data
FILENAME REFFILE '/home/u60034905/Project/test.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=test;
GETNAMES=YES;
RUN;

FILENAME REFFILE '/home/u60034905/Project/train.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=train;
GETNAMES=YES;
```

```
RUN;
/*Prepare test data to be appended to the training data for Kaggle
data = test;
set test;
SalePrice = .;
logSP = .;
run;

/*Append the data
data train2;
set train test;
run;
/*Plot overall SalePrice data to see if it itself
appears to approximate a normally distribution*/
proc sgplot data=train2;
       histogram SalePrice;
run;

/*Log the sales price as there is evidence of right skew*/
data train2_logSP;
       set train2;
       logSP=log(SalePrice);
run;

/*Plot logged SalePrice data to see if it itself
appears to approximate a normally distribution*/
proc sgplot data=train2_logSP;
       histogram logSP;

       /*Logged data appears to approach a
       normal distribution*/
run;

/*Explore the data after common sense (based on
domain knowledge) variable selection*/
proc sgscatter data=train2_logSP;
       matrix logSP GarageArea YrSold LotArea OverallQual GrLivArea
firstFlrSF
              SecondFlrSF TotalBsmtSF / group=Neighborhood;
run;

proc sgscatter data=train2_logSP;
       matrix logSP GarageArea OverallQual GrLivArea /
group=Neighborhood;
run;

/*Explore the full model and residual analysis*/
proc glm data=train2_logSP plots=all;
```

```
	model logSP=GarageArea | OverallQual | GrLivArea / solution
clparm;
	run;

	/*Outlier removal process*/
data train2__logSP_outlier_removed;
	set train2_logSP;

	if _n_=1299 then

		/*high leverage and influence on General Living area*/
		delete;

	if _n_=524 then

		/*high leverage and influence on General Living area*/
		delete;
run;

/*GarageArea, OverallQual,  GrLivArea categorized by neighborhood
appear to be correlated
to Logged Sales Price. We will explre these associations more*/

/*forward selection*/
proc glmselect data=train2__logSP_outlier_removed seed=635422552;
	partition fraction(test=0.3);
	class neighborhood;
	model logSP=GarageArea | OverallQual | GrLivArea | neighborhood
/ selection=forward
		stats=adjrsq stats=press;
	output out=resultsf p=Predict;
run;

data resultsff;
	set resultsf;

	if Predict > 0 then
		SalePrice=exp(Predict);

	if Predict < 0 then
		SalePrice=10000;
	keep id SalePrice;
	where id > 1460;
run;

proc means data = resultsff;
var SalePrice;
run;
```

```
/*proc export data=resultsff
    outfile="/home/u60034905/Project/resultsff.csv"
    dbms=csv;
run;

/*backward elimination*/
proc glmselect data=train2__logSP_outlier_removed seed=635422552;
       partition fraction(test=0.3);
       class neighborhood;
       model logSP=GarageArea | OverallQual | GrLivArea |
neighborhood/selection=backward
              stats=adjrsq stats=press;
       output out=resultsb p=Predict;
run;

data resultsbb;
       set resultsb;

       if Predict > 0 then
              SalePrice=exp(Predict);

       if Predict < 0 then
              SalePrice=10000;
       keep id SalePrice;
       where id > 1460;
run;

proc means data = resultsbb;
var SalePrice;
run;

/*proc export data=resultsbb
    outfile="/home/u60034905/Project/resultsbb.csv"
    dbms=csv;
run;

/*stepwise elimination*/
proc glmselect data=train2__logSP_outlier_removed seed=635422552;
       partition fraction(test=0.3);
       class neighborhood;
       model logSP=GarageArea | OverallQual | GrLivArea |
neighborhood/ selection= stepwise
              stats=adjrsq stats=press;
       output out=resultss p=Predict;
run;

data resultsss;
```

```
	set resultss;

	if Predict > 0 then
		SalePrice=exp(Predict);

	if Predict < 0 then
		SalePrice=10000;
	keep id SalePrice;
	where id > 1460;
run;

proc means data = resultsss;
var SalePrice;
run;

/*proc export data=resultsss
    outfile="/home/u60034905/Project/resultsss.csv"
    dbms=csv;
run;

/*No elimination (Custom Model1)*/
proc glmselect data=train2__logSP_outlier_removed seed=635422552;
	partition fraction(test=0.3);
	class neighborhood;
	model logSP=GarageArea | OverallQual | GrLivArea |
neighborhood/ selection= none
		stats=adjrsq stats= press;
	output out=resultsn p=Predict;
run;

data resultsnn;
	set resultsn;

	if Predict > 0 then
		SalePrice=exp(Predict);

	if Predict < 0 then
		SalePrice=10000;
	keep id SalePrice;
	where id > 1460;
run;

proc means data = resultsnn;
var SalePrice;
run;

/*
proc export data=resultsnn
```

```
    outfile="/home/u60034905/Project/resultsnn.csv"
    dbms=csv;
run;

/*Explore the full model MLR*/
proc glm data=train2__logSP_outlier_removed plots=all;
       class neighborhood;
       model logSP=GarageArea | OverallQual | GrLivArea | neighborhood
/ cli solution clparm;
       output out=resultsm p=Predict;

       run;

data resultsmm;
       set resultsm;

       if Predict > 0 then
              SalePrice=exp(Predict);

       if Predict < 0 then
              SalePrice=10000;
       keep id SalePrice;
       where id > 1460;
run;

proc means data = resultsmm;
var SalePrice;
run;

/*proc export data=resultsmm
    outfile="/home/u60034905/Project/resultsmm.csv"
    dbms=csv;
run;*/
```