House Pricing Prediction Model

Patricia Attah

Contents

Introduction

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this Kaggle competition's dataset proves that much more influences price negotiations than the number of bedrooms or the presence of a white-picket fence. With 1460 houses in the dataset and 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, the goal of this project is to predict the final price of each home.

Data Description

The data in this analysis is from Kaggle's House Prices: Advanced Regressions Techniques competition. The full training dataset, test dataset, and explanation of variables is available here:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques

- There are 1460 houses in the dataset with 79 explanatory variables and 1 response variable (SalePrice).
- The first analysis uses two explanatory, Neighborhood and Above grade/ground living area (GrLivArea), in relationship to sale price.
- The second analysis focuses on variable selection from all the explanatory variables to predict the SalePrice. The output of this analysis will be submitted to Kaggle for scoring.

Analysis Question #1

Problem Statement

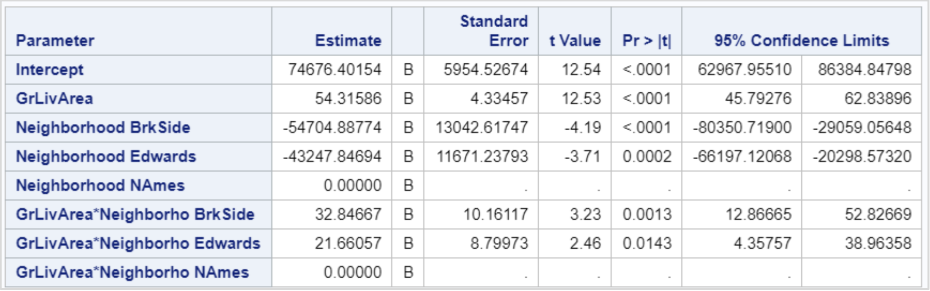
Century 21 Ames only sells houses in the NAmes, Edwards and BrkSide neighborhoods and would like to get an estimate of how the SalePrice of the house is related to the square footage of the living area of the house (GrLIvArea) and if the SalesPrice (and its relationship to square footage) depends on which neighborhood the house is located in.

Build and Fit the Model

Predicted Sale Price = $\boldsymbol{\beta_0}$ ***+*** $\boldsymbol{\beta_1(GrLivArea)}$ ***+*** $\boldsymbol{\beta_2 Neigh_{BrkSide}}$ ***+*** $\boldsymbol{\beta_3 Neigh_{Edwards}} +$
$\boldsymbol{\beta_4(Neigh_{BrkSide}}$ **** GrLivArea) +*** $\boldsymbol{\beta_5(Neigh_{Edwards}}$ **** GrLivArea)***

Predicted (Sale Price | Neighborhood = NAmes) = β_0 + $\beta_1(\text{GrLivArea})$
Predicted (Sale Price | Neighborhood = BrkSide) = β_0 + $\beta_2 + (\beta_1 + \beta_4(\text{GrLivArea}))$
Predicted (Sale Price | Neighborhood = Edwards) = β_0 + $\beta_3 + (\beta_1 + \beta_5(\text{GrLivArea}))$

Predicted (Sale Price | Neighborhood = NAmes) = 74,676 + 54.32(GrLivArea)
Predicted (Sale Price | Neighborhood = BrkSide) = 19,971 +87.17(GrLivArea)
Predicted (Sale Price | Neighborhood = Edwards) = 31,429 +75.98(GrLivArea)



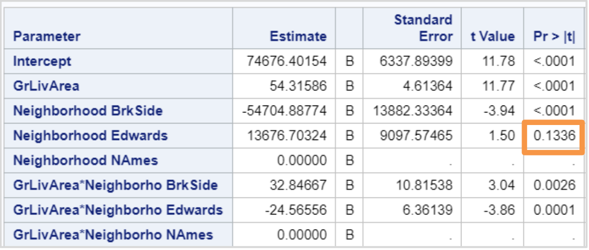
Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	74676.40154	B	5954.52674	12.54	<.0001	62967.95510	86384.84798
GrLivArea	54.31586	B	4.33457	12.53	<.0001	45.79276	62.83896
Neighborhood BrkSide	-54704.88774	B	13042.61747	-4.19	<.0001	-80350.71900	-29059.05648
Neighborhood Edwards	-43247.84694	B	11671.23793	-3.71	0.0002	-66197.12068	-20298.57320
Neighborhood NAmes	0.00000	B	.	.	.	.	.
GrLivArea*Neighborho BrkSide	32.84667	B	10.16117	3.23	0.0013	12.86665	52.82669
GrLivArea*Neighborho Edwards	21.66057	B	8.79973	2.46	0.0143	4.35757	38.96358
GrLivArea*Neighborho NAmes	0.00000	B	.	.	.	.	.

Checking Assumptions

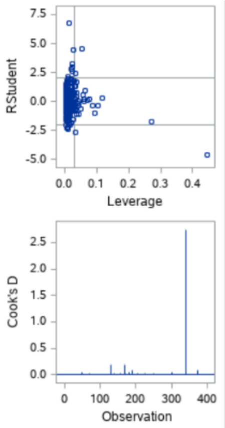
Addressing Outliers

There are two outliers in the dataset in the Edwards neighborhood. Both houses list over 4600 square feet of above ground living area with unusually low sales prices. Upon further investigation, both homes are listed with a sales condition of "partial." These observations have been excluded from the analysis.

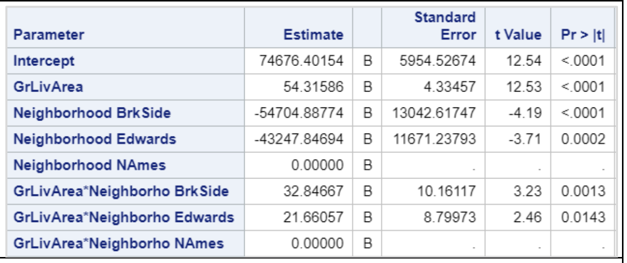
With Outliers



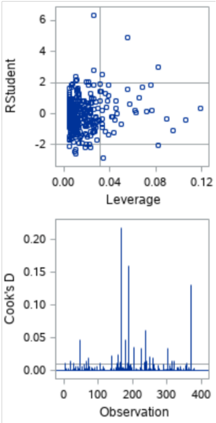
Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	74676.40154	B	6337.89399	11.78	<.0001
GrLivArea	54.31586	B	4.61364	11.77	<.0001
Neighborhood BrkSide	-54704.88774	B	13882.33364	-3.94	<.0001
Neighborhood Edwards	13676.70324	B	9097.57465	1.50	0.1336
Neighborhood NAmes	0.00000	B	.	.	.
GrLivArea*Neighborho BrkSide	32.84667	B	10.81538	3.04	0.0026
GrLivArea*Neighborho Edwards	-24.56556	B	6.36139	-3.86	0.0001
GrLivArea*Neighborho NAmes	0.00000	B	.	.	.



Without Outliers



Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	74676.40154	B	5954.52674	12.54	<.0001
GrLivArea	54.31586	B	4.33457	12.53	<.0001
Neighborhood BrkSide	-54704.88774	B	13042.61747	-4.19	<.0001
Neighborhood Edwards	-43247.84694	B	11671.23793	-3.71	0.0002
Neighborhood NAmes	0.00000	B	.	.	.
GrLivArea*Neighborho BrkSide	32.84667	B	10.16117	3.23	0.0013
GrLivArea*Neighborho Edwards	21.66057	B	8.79973	2.46	0.0143
GrLivArea*Neighborho NAmes	0.00000	B	.	.	.

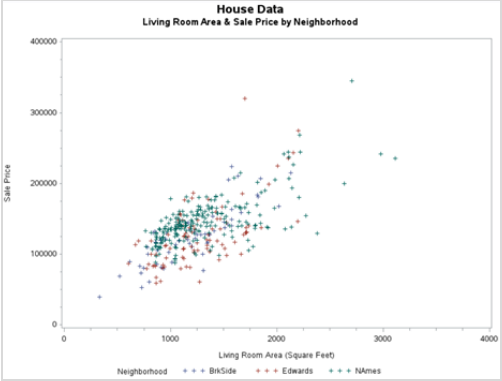


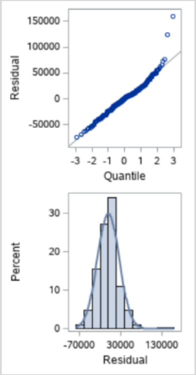
- **Linearity:** Checking pairwise scatter plots indicates a strong linear trend between GrLivArea and Sales Prices.

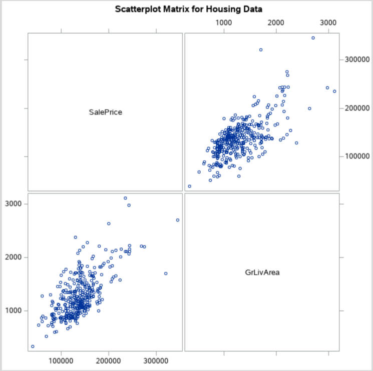
- **Constant Variance:** There is little evidence from the residual plots of heteroscedasticity.

- **Normality:** Judging from scatter plot, q-q plot, and histogram of residuals, there is not strong evidence against normality.

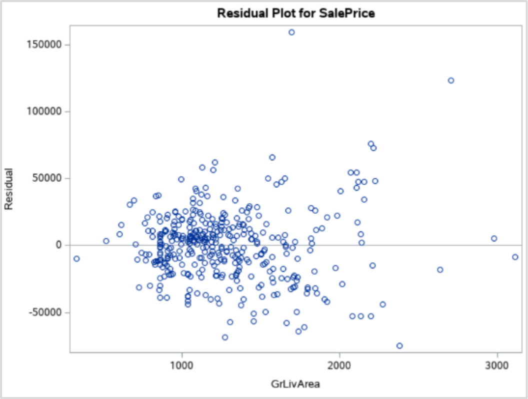
- **Independence:** The samples are from 381 houses after removing the two outliers. We will assume the observations are independent.

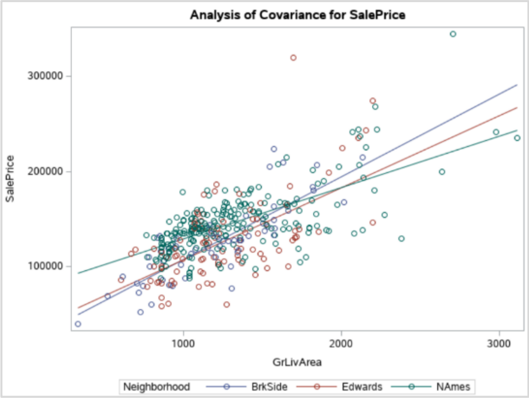






Residual Plots





Comparing Competing Models

See Appendix A

Interpretation

For every 100 square foot increase in living area, the increase in mean estimated sales price is $5,430 for houses in North Ames (p-value < 0.0001). While the mean sale prices of houses in Brookside is estimated to be $54,704 less than mean sale prices in the North Ames, for every one hundred square foot increase in living area in Brookside, the mean sale price is estimated to be $3,285 more than North Ames (p-value = 0.0013).   The mean sale prices of houses in Edwards is estimated to be $43,248 less than mean sale prices in the North Ames, but for every one hundred square foot increase in living area, the mean sale price is estimated to be $2,166 more than North Ames (p-value = 0.0143).

Confidence Intervals

95% confidence interval for the increase in sale price from North Ames to Brookside ($1,287, $5,283) when the living area increases 100 square feet.

95% confidence interval for the increase in sale price from North Ames to Edwards ($436, $3,896) when the living area increases 100 square feet.

Conclusion

The evidence suggests that the sales price increases for additional living area in the Brookside and Edwards neighborhoods compared to additional living area in the North Ames area. Because the sales prices are significantly higher in NAmes than Brkside (p-value = < 0.001) as well as Edwards (p-value = 0.0002), a variable other than living area may be associated with the overall estimated difference in mean prices.

Analysis Question #2

Problem Statement

With 1460 houses in the dataset and 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, the goal of this project is to predict the final price of each home.

Model Selection

This analysis includes the following variable selection techniques for the models: Stepwise, Forward, Backward, and Custom.

Checking Assumptions

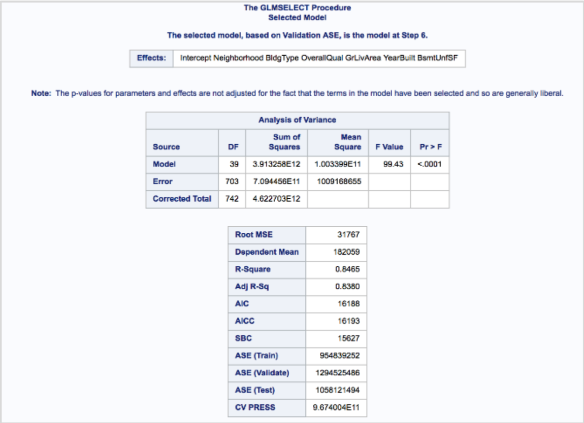
See Appendix B.

Comparing Competing Models

Predictive Models	**Adjusted R2**	**CV PRESS**	**Kaggle Score**
Forward	0.8380	9.67 E11	0.16847
Backward	0.8419	8.86 E11	0.19454
Stepwise	0.8186	9.72 E11	0.20957
CUSTOM	0.7892	1.03 E12	0.19188

Forward selection model variables:

Neighborhood BldgType OverallQual GrLivArea YearBuilt BsmtUnfSF



The GLMSELECT Procedure
Selected Model

The selected model, based on Validation ASE, is the model at Step 6.

Effects:	Intercept Neighborhood BldgType OverallQual GrLivArea YearBuilt BsmtUnfSF

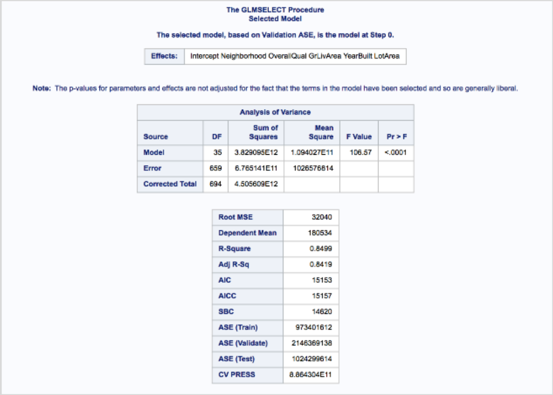
Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	39	3.913258E12	1.003399E11	99.43	<.0001
Error	703	7.094456E11	1009168655		
Corrected Total	742	4.622703E12			

Root MSE	31767
Dependent Mean	182059
R-Square	0.8465
Adj R-Sq	0.8380
AIC	16188
AICC	16193
SBC	15627
ASE (Train)	954839252
ASE (Validate)	1294525486
ASE (Test)	1058121494
CV PRESS	9.674004E11

Backward selection model variables:

Neighborhood OverallQual GrLivArea YearBuilt Lot



The GLMSELECT Procedure
Selected Model

The selected model, based on Validation ASE, is the model at Step 0.

Effects:	Intercept Neighborhood OverallQual GrLivArea YearBuilt LotArea

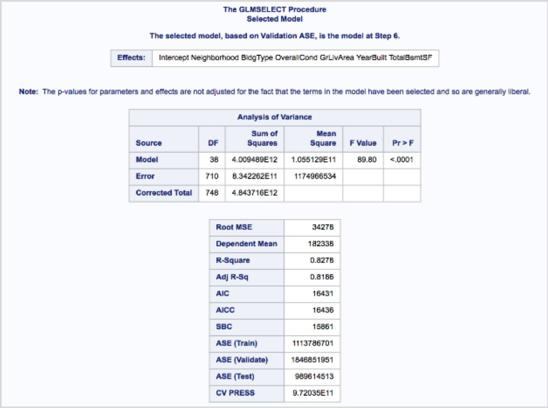
Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	35	3.829095E12	1.094027E11	106.57	<.0001
Error	659	6.765141E11	1026576814		
Corrected Total	694	4.505609E12			

Root MSE	32040
Dependent Mean	180534
R-Square	0.8499
Adj R-Sq	0.8419
AIC	15153
AICC	15157
SBC	14620
ASE (Train)	973401612
ASE (Validate)	2146369138
ASE (Test)	1024299614
CV PRESS	8.864304E11

Stepwise selection model variables:

Neighborhood BldgType OverallCond GrLivArea YearBuilt TotalBsmtSF



The GLMSELECT Procedure
Selected Model

The selected model, based on Validation ASE, is the model at Step 6.

Effects:	Intercept Neighborhood BldgType OverallCond GrLivArea YearBuilt TotalBsmtSF

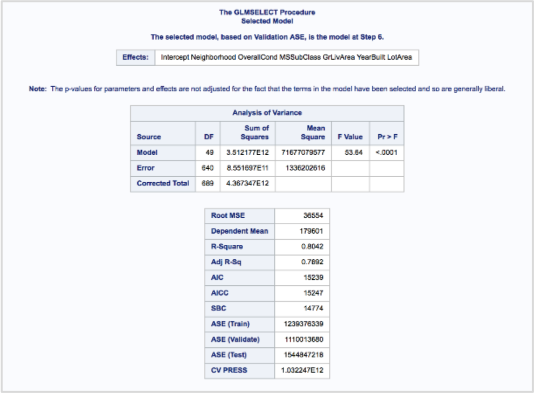
Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	38	4.009489E12	1.055129E11	89.80	<.0001
Error	710	8.342262E11	1174966534		
Corrected Total	748	4.843716E12			

Root MSE	34278
Dependent Mean	182338
R-Square	0.8278
Adj R-Sq	0.8186
AIC	16431
AICC	16436
SBC	15861
ASE (Train)	1113786701
ASE (Validate)	1846851951
ASE (Test)	989614513
CV PRESS	9.72035E11

Custom Selection model variables:

Neighborhood OverallCond MSSubClass GrLivArea YearBuilt LotArea



The GLMSELECT Procedure
Selected Model

The selected model, based on Validation ASE, is the model at Step 6.

Effects:	Intercept Neighborhood OverallCond MSSubClass GrLivArea YearBuilt LotArea

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	49	3.512177E12	71677079577	53.64	<.0001
Error	640	8.551697E11	1336202616		
Corrected Total	689	4.367347E12			

Root MSE	36554
Dependent Mean	179601
R-Square	0.8042
Adj R-Sq	0.7692
AIC	15239
AICC	15247
SBC	14774
ASE (Train)	1239376339
ASE (Validate)	1110013680
ASE (Test)	1544847218
CV PRESS	1.032247E12

Conclusion

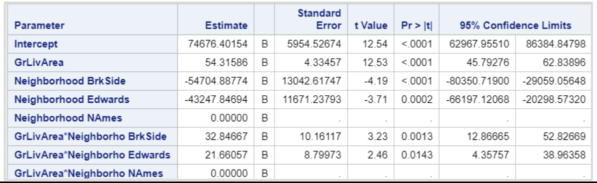
Forward selection with the variables below gave us the best score for SalePrice predictions: Neighborhood BldgType OverallQual GrLivArea YearBuilt BsmtUnfSF.

Appendices

Appendix A – Analysis 1

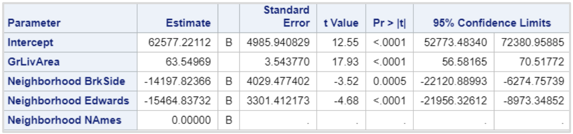
Comparing Competing Models

GrLivArea and Neighborhood With Interactions



Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	74676.40154	B	5954.52674	12.54	<.0001	62967.95510	86384.84798
GrLivArea	54.31586	B	4.33457	12.53	<.0001	45.79276	62.83896
Neighborhood BrkSide	-54704.88774	B	13042.61747	-4.19	<.0001	-80350.71900	-29059.05648
Neighborhood Edwards	-43247.84694	B	11671.23793	-3.71	0.0002	-66197.12068	-20298.57320
Neighborhood NAmes	0.00000	B	.	.	.	.	.
GrLivArea*Neighborho BrkSide	32.84667	B	10.16117	3.23	0.0013	12.86665	52.82669
GrLivArea*Neighborho Edwards	21.66057	B	8.79973	2.46	0.0143	4.35757	38.96358
GrLivArea*Neighborho NAmes	0.00000	B	.	.	.	.	.

GrLivArea and Neighborhood Without Interactions

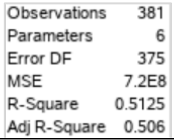


Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	62577.22112	B	4985.940829	12.55	<.0001	52773.48340	72380.95885
GrLivArea	63.54969		3.543770	17.93	<.0001	56.58165	70.51772
Neighborhood BrkSide	-14197.82366	B	4029.477402	-3.52	0.0005	-22120.88993	-6274.75739
Neighborhood Edwards	-15464.83732	B	3301.412173	-4.68	<.0001	-21956.32612	-8973.34852
Neighborhood NAmes	0.00000	B	.	.	.	.	.

Adj R^2

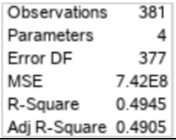
Adj R^2 is slightly better with interactions

GrLivArea and Neighborhood With Interactions



Observations	381
Parameters	6
Error DF	375
MSE	7.2E8
R-Square	0.5125
Adj R-Square	0.506

GrLivArea and Neighborhood Without Interactions

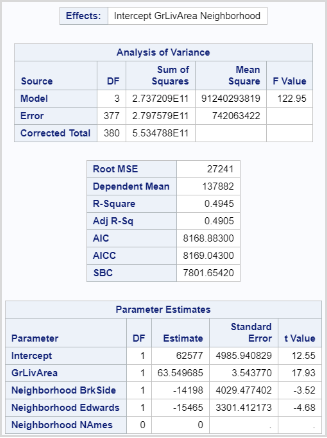


Observations	381
Parameters	4
Error DF	377
MSE	7.42E8
R-Square	0.4945
Adj R-Square	0.4905

Internal CV Press

No difference in variable selection with these variables

Forward



Effects:	Intercept GrLivArea Neighborhood

Analysis of Variance

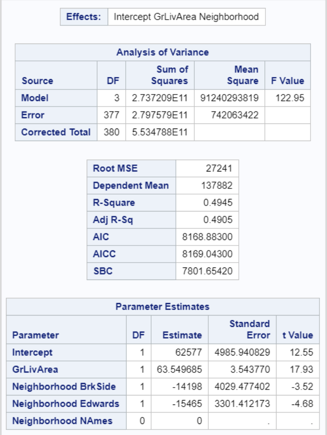
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	2.737209E11	91240293819	122.95
Error	377	2.797579E11	742063422	
Corrected Total	380	5.534788E11		

Root MSE	27241
Dependent Mean	137882
R-Square	0.4945
Adj R-Sq	0.4905
AIC	8168.88300
AICC	8169.04300
SBC	7801.65420

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	62577	4985.940829	12.55
GrLivArea	1	63.549685	3.543770	17.93
Neighborhood BrkSide	1	-14198	4029.477402	-3.52
Neighborhood Edwards	1	-15465	3301.412173	-4.68
Neighborhood NAmes	0	0	.	.

Backward



Effects:	Intercept GrLivArea Neighborhood

Analysis of Variance

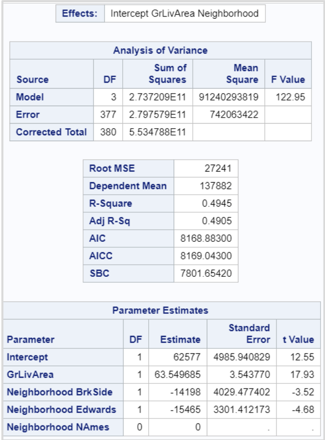
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	2.737209E11	91240293819	122.95
Error	377	2.797579E11	742063422	
Corrected Total	380	5.534788E11		

Root MSE	27241
Dependent Mean	137882
R-Square	0.4945
Adj R-Sq	0.4905
AIC	8168.88300
AICC	8169.04300
SBC	7801.65420

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	62577	4985.940829	12.55
GrLivArea	1	63.549685	3.543770	17.93
Neighborhood BrkSide	1	-14198	4029.477402	-3.52
Neighborhood Edwards	1	-15465	3301.412173	-4.68
Neighborhood NAmes	0	0	.	.

Stepwise



Effects:	Intercept GrLivArea Neighborhood

Analysis of Variance

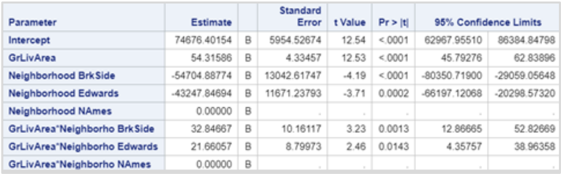
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	2.737209E11	91240293819	122.95
Error	377	2.797579E11	742063422	
Corrected Total	380	5.534788E11		

Root MSE	27241
Dependent Mean	137882
R-Square	0.4945
Adj R-Sq	0.4905
AIC	8168.88300
AICC	8169.04300
SBC	7801.65420

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	62577	4985.940829	12.55
GrLivArea	1	63.549685	3.543770	17.93
Neighborhood BrkSide	1	-14198	4029.477402	-3.52
Neighborhood Edwards	1	-15465	3301.412173	-4.68
Neighborhood NAmes	0	0	.	.

Parameters & Estimates



Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	74676.40154	B	5954.52674	12.54	<.0001	62967.95510	86384.84798
GrLivArea	54.31586	B	4.33457	12.53	<.0001	45.79276	62.83896
Neighborhood BrkSide	-54704.88774	B	13042.61747	-4.19	<.0001	-80350.71900	-29059.05648
Neighborhood Edwards	-43247.84694	B	11671.23793	-3.71	0.0002	-66197.12068	-20298.57320
Neighborhood NAmes	0.00000	B	.	.	.	.	.
GrLivArea*Neighborho BrkSide	32.84667	B	10.16117	3.23	0.0013	12.86665	52.82669
GrLivArea*Neighborho Edwards	21.66057	B	8.79973	2.46	0.0143	4.35757	38.96358
GrLivArea*Neighborho NAmes	0.00000	B	.	.	.	.	.

```
/* Analysis #1 Code */
/* Import data and sort it*/
proc import OUT=WORK.TR
DATAFILE= "/home/u47487140/sasuser.v94/Bridge/train.csv"
DBMS=CSV REPLACE;
GETNAMES=YES;
DATAROW=2;
RUN;

/* Subset the data */
data tr2;
set WORK.TR;
keep Neighborhood SalePrice;
where(Neighborhood in ('NAmes','BrkSide','Edwards'));
if GrLivArea > 4600 then delete;
run;

/* Scatterplot */
title1 "House Data";
title2 "Living Room Area & Sale Price by Neighborhood";
axis1 label=(angle=90 "Sale Price") minor=(n=3);
axis2 label=("Living Room Area (Square Feet)") minor=(n=3);
proc gplot data = tr2;
plot SalePrice * GrLivArea = Neighborhood /vaxis=axis1 haxis=axis2;
run;
quit;

/* Matrix */
proc sgscatter data=tr2;
  title "Scatterplot Matrix for Housing Data";
  matrix SalePrice GrLivArea;
run;
title;

/* Proc GLM with Interactions */
proc glm data = tr2 plot = all;
class Neighborhood (ref='NAmes');
model SalePrice = GrLivArea | Neighborhood / solution clparm;
run;

/* Proc GLM without Interactions */
proc glm data = tr2 plot = all;
class Neighborhood (ref='NAmes');
model SalePrice = GrLivArea Neighborhood / solution clparm;
run;

/* P value on 2 and 375 df */
data pval;
pvalue = 1-PROBF(6.89, 2, 375);
run;

/* Forward Selection */
proc glmselect data = tr2;
class Neighborhood;
model saleprice_log = grlivarea_log Neighborhood / selection = forward;
run;

/* Backward */
proc glmselect data = tr2;
class Neighborhood;
model saleprice_log = grlivarea_log Neighborhood / selection = backward;
run;

/* Stepwise */
proc glmselect data = tr2;
class Neighborhood;
model SalePrice = GrLivArea Neighborhood / selection = stepwise;
run;
```

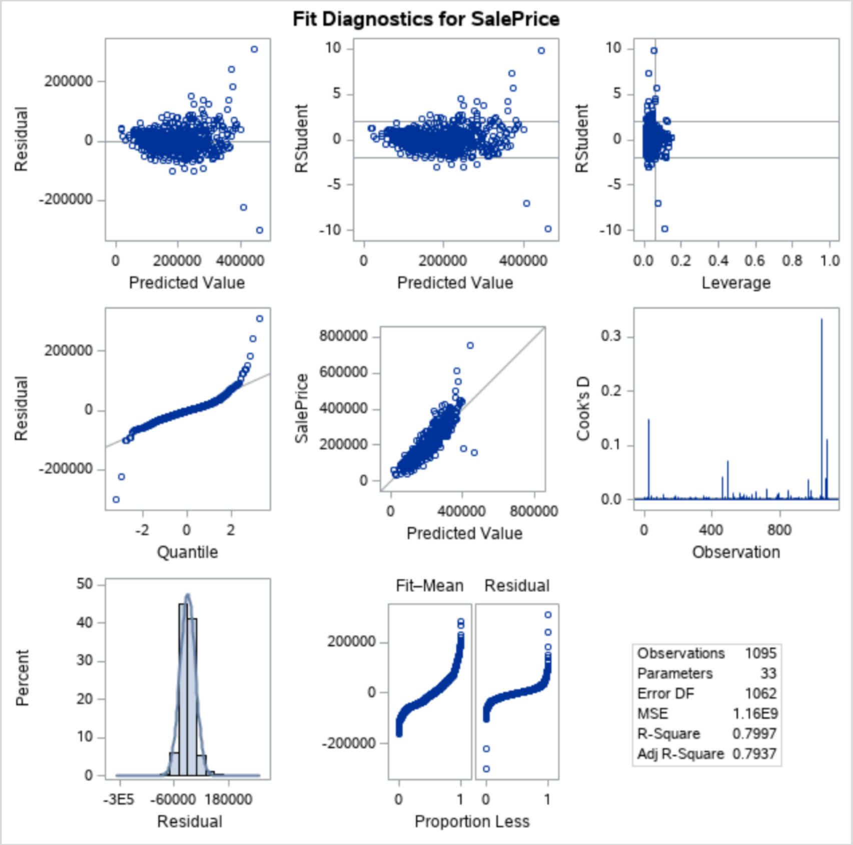
Appendix B – Analysis 2

- **Linearity:** Checking pairwise scatter plots indicates some linear trend between Sales Prices and the continuous variables.

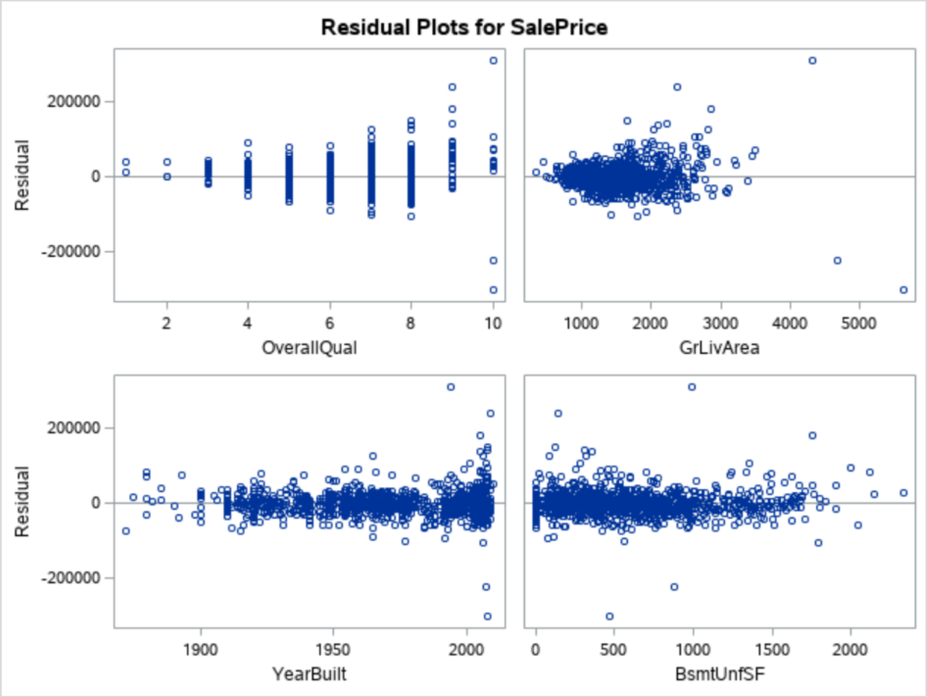
- **Constant Variance:** There is some evidence from the residual plots of heteroscedasticity.

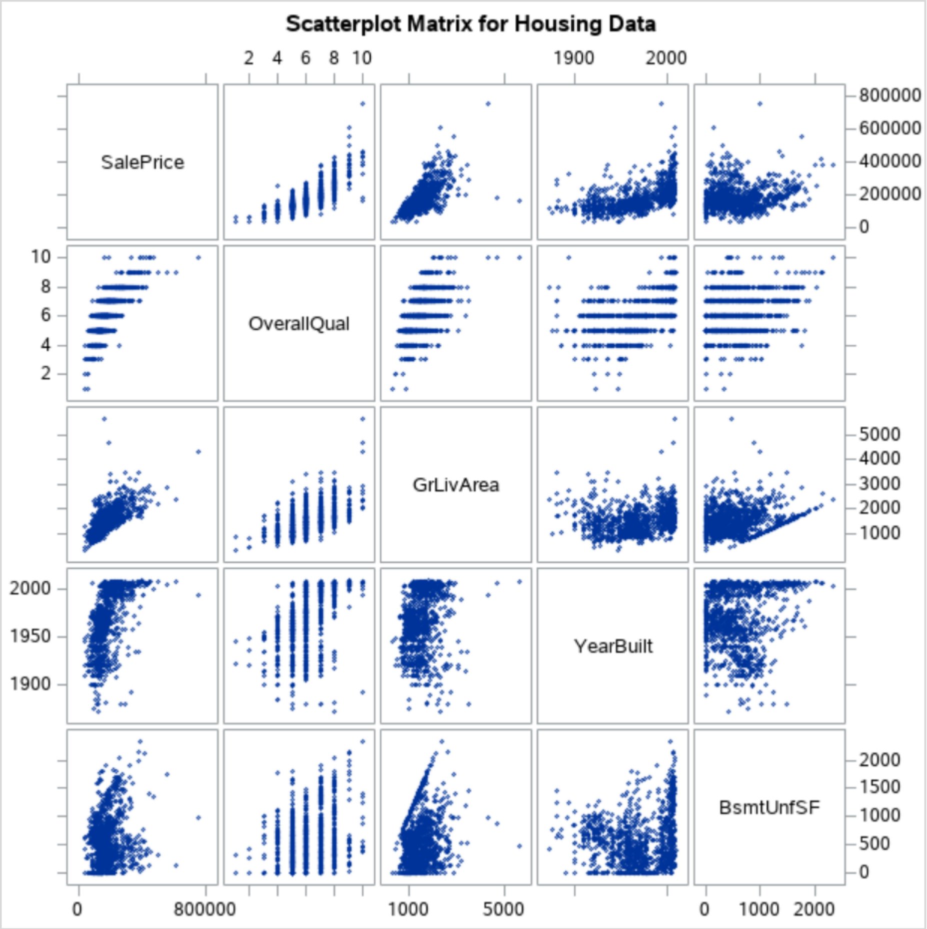
- **Normality:** Judging from scatter plot, q-q plot, and histogram of residuals, there is not strong evidence against normality.

- **Independence:** The samples are from 1460 houses. We will assume the observations are independent.



Observations	1095
Parameters	33
Error DF	1062
MSE	1.16E9
R-Square	0.7997
Adj R-Square	0.7937


Residual Plots for SalePrice
Residual
200000
0
-200000
2
4
6
8
10
OverallQual
1000
2000
3000
4000
5000
GrLivArea
Residual
200000
0
-200000
1900
1950
2000
YearBuilt
0
500
1000
1500
2000
BsmtUnfSF


Scatterplot Matrix for Housing Data
2
4
6
8
10
1900
2000
800000
600000
400000
200000
0
SalePrice
10
8
6
4
2
OverallQual
GrLivArea
5000
4000
3000
2000
1000
2000
1950
1900
YearBuilt
BsmtUnfSF
2000
1500
1000
500
0
0
800000
1000
5000
0
1000
2000