Ames, Iowa Housing Price Analysis

Rachel Liercke, Joshua Turk

Introduction

Century 21 Ames has hired J&R Analysis to analyze home sale data in order to gain insight into the housing market as well as answer a couple questions of interest. Specifically, Century 21 wants to know about the relationship between the living area of a house and its sale price and whether this relationship changes with respect to different neighborhoods that Century 21 operates in. After this, Century 21 wants a model built to predict sales price based off of any predictor variables in the dataset.

Data Description

The dataset used for analysis comes from Kaggle and is a collection of 79 explanatory variables that describe the various details of homes sold in Ames, Iowa from 2006 to 2010. This data set contains a total of 1460 observations from the entirety of Ames. Century 21 Ames primarily works with homes found in the North Ames, Edwards, and BrookSide neighborhoods which account for 383 observations in the dataset. The variable of interest we will be building several models to predict is the SalePrice of the home. To read more about the data set and specific explanatory variables please see:
https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Analysis Question 1

Problem:
Century 21 in Ames, Iowa has asked J&R Analysis to determine if the sale price of a house is affected by the square footage of the living area of a house and if this is affected by the Neighborhood. The neighborhoods of interest for this question are: North Ames, Edwards, and Brookside.
Build and Fit of Model:
To answer this question, we built a model that looked at the relationship between Sale Price and the square footage of the living area with its interaction with the respective neighborhood associated with it.
Assumptions:
The assumptions of a linear regression model are as follows: linearity, normality, constant variance, and independence. We will address these assumptions using the Fit Diagnostics plot in Appendix D.
We can tell the data follows a linear model by looking at the Residual v Quantile plot (second row, first column). If the graph follows a diagonal line then the model will be a good linear fit. This plot shows that the linearity assumption is met.
Normality is met when the histogram of the data follows a normal distribution. The Percent v Residual plot (third row, first column) shows that this data is normally distributed and follows the normal bell curve.
Constant variance can be shown in the Residual v Predicted and the RStudent v Predicted plots (first row, first and second columns). If the data shows a random plot of points with no trends, then the data will have constant variance. Our plots below show that there is no pattern in the Residuals so this assumption is met.

Independence refers to independence of variables and independence of individual data points. Since this data is based off of the houses sold by Century 21 in Ames, we will have to assume that independence is met.
We do have two highly influential observations of large houses sold in the Edwards neighborhood for relatively low prices. While these points are highly influential, we see no reason to believe they are not a part of the population of homes sold in Ames that we are studying, therefore these outliers will be left in the analysis. These points primarily affect the indicator term for the Edwards neighborhood.

Models:
In order to compare the neighborhoods and sale price with the square footage of the living area, we looked at 4 models. We will refer to these models by their number when comparing later in the paper.

1. Sale Price = GrLivArea * Neighborhood
2. Sale Price = log(GrLivArea) * Neighborhood
3. log(Sale Price) = GrLivArea * Neighborhood
4. log(Sale Price) = log(GrLivArea) * Neighborhood

The assumptions were examined for each model, as well as other factors such as Adjusted R^2 and CV Press. The plots of each model assumption are seen, respectively in Appendices A-D.

Model Number	Adj R^2	Assumptions Met?
1	0.44	No - linearity not met
2	.4587	No - linearity not met
3	.4589	Yes -all
4	.5056	Yes - all

Parameters:
We chose the 4th model of log(Sale Price) = log(GrLivArea) * Neighborhood as it produced the highest Adj R^2 and lowest CV Press. The estimate of the sales price ends up with two different models based on the neighborhoods. There is no significant difference in the price of a house in Edwards and North Ames but there is a difference in Brookside.
Our models are:
North Ames and Edwards-
Predicted Log(SalePrice) = 8.0065 + 0.5197*log(GrLivArea)
Brookside-
Predicted Log(SalePrice) = 8.4927 + .8197*log(GrLivArea)

North Ames and Edwards: A doubling of the square footage of a living area is associated with a multiplicative change of 2^.5197 = 1.433 increase in the median Sale Price for the North Ames and

Edwards neighborhoods. We are 95% certain that the multiplicative increase is in the range of (1.3274, 1.5483).

Brookside: A doubling of the square footage of a living area is associated with a multiplicative change of 2^.8197 = 1.765 increase in the median Sale Price for the Brookside neighborhood. We are 95% certain that the true multiplicative increase is in the range (1.5586, 1.9986).

R Shiny App

https://josh-turk.shinyapps.io/StatProject-AmesHousing/

Analysis Question 2

Problem:
Century 21 Ames is looking to find the best model to accurately predict Sale Price of a house using the data they have collected. They want J&R to come up with four models using techniques covered in 6371 and give them the best model of the four.

Model Selection:

Predictive Model	Adjusted R^2	CV PRESS	Kaggle Score
Forward	0.8098	1.853E12	.17502
Backward	0.8137	1.920E12	.17502
Stepwise	0.8098	1.843E12	.17502
Custom	0.8661	20.45173	.16353

Assumptions:
Using the techniques that we have learned has resulted in all three (Forward, Backward, and Stepwise) models having the same predictors. This will allow us to look at the assumptions using the same plots as seen in Appendix I.
Normality is highly violated in the histogram plot because the data is left skewed. Constant variance is not met because both residual plots have a U-shaped pattern. Linearity is not met because the data doesn't follow a linear trend.

The custom model shows that the data looks more normal compared to the forward, backward, and stepwise model. There is a slight tail on this which may be due to some outliers.Constant variance is met based on the residual plots having a random pattern. We will proceed with caution on the linearity assumption as it is not 100% met.

Again we have two highly influential outliers where large houses were sold for significantly less than would otherwise be predicted. We have decided to leave these observations in as we have no reason to believe that these houses are not a part of the population of interest.

Conclusion:

In conclusion, after running a thorough analysis, we have decided to use the custom model containing: Lot Area, Overall Quality, Year Built, Basement Finished Square feet, Living Area, Fireplaces, and Neighborhood. This model had the smallest Adjusted R^2, CV Press, and kaggle score. Out of all the models, this one will provide us with the best estimate of Sale Price in the Ames neighborhoods.

Github links:

https://rachellliercke.github.io/

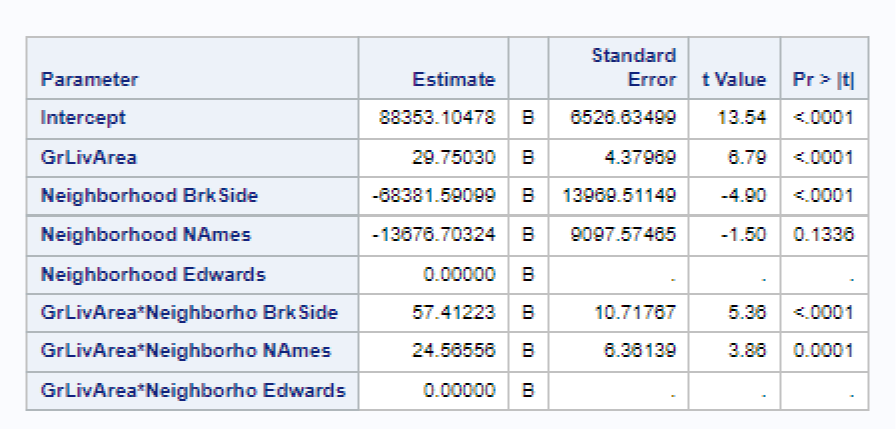
https://josh-turk.github.io/

Data Source:

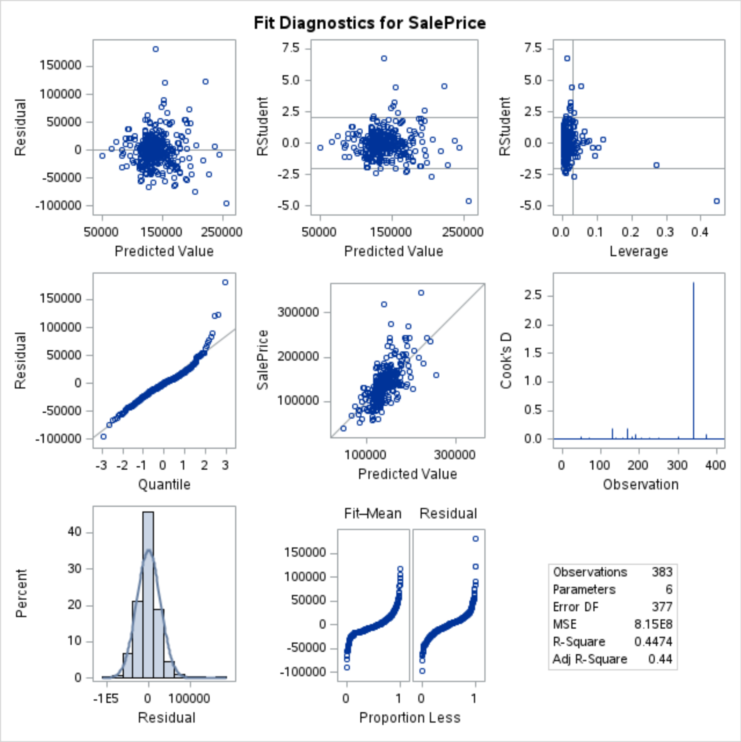
https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

Appendix

Appendix A - Model 1 Stats

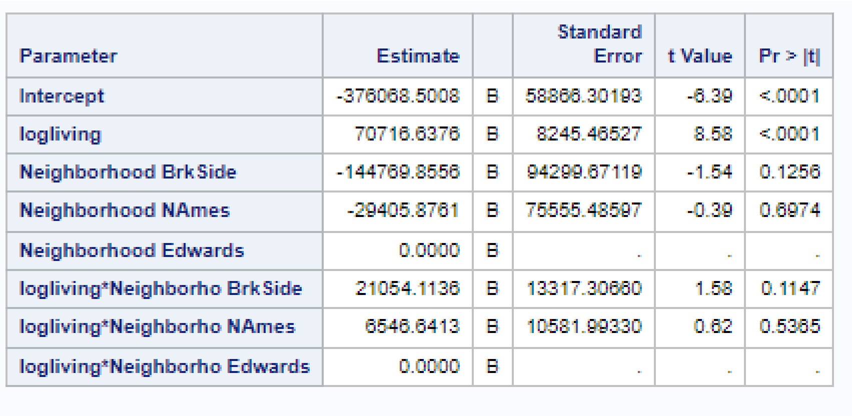


Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	88353.10478	B	6526.63499	13.54	<.0001
GrLivArea	29.75030	B	4.37969	6.79	<.0001
Neighborhood BrkSide	-68381.59099	B	13969.51149	-4.90	<.0001
Neighborhood NAmes	-13676.70324	B	9097.57465	-1.50	0.1336
Neighborhood Edwards	0.00000	B	.	.	.
GrLivArea*Neighborho BrkSide	57.41223	B	10.71767	5.36	<.0001
GrLivArea*Neighborho NAmes	24.56556	B	6.36139	3.86	0.0001
GrLivArea*Neighborho Edwards	0.00000	B	.	.	.

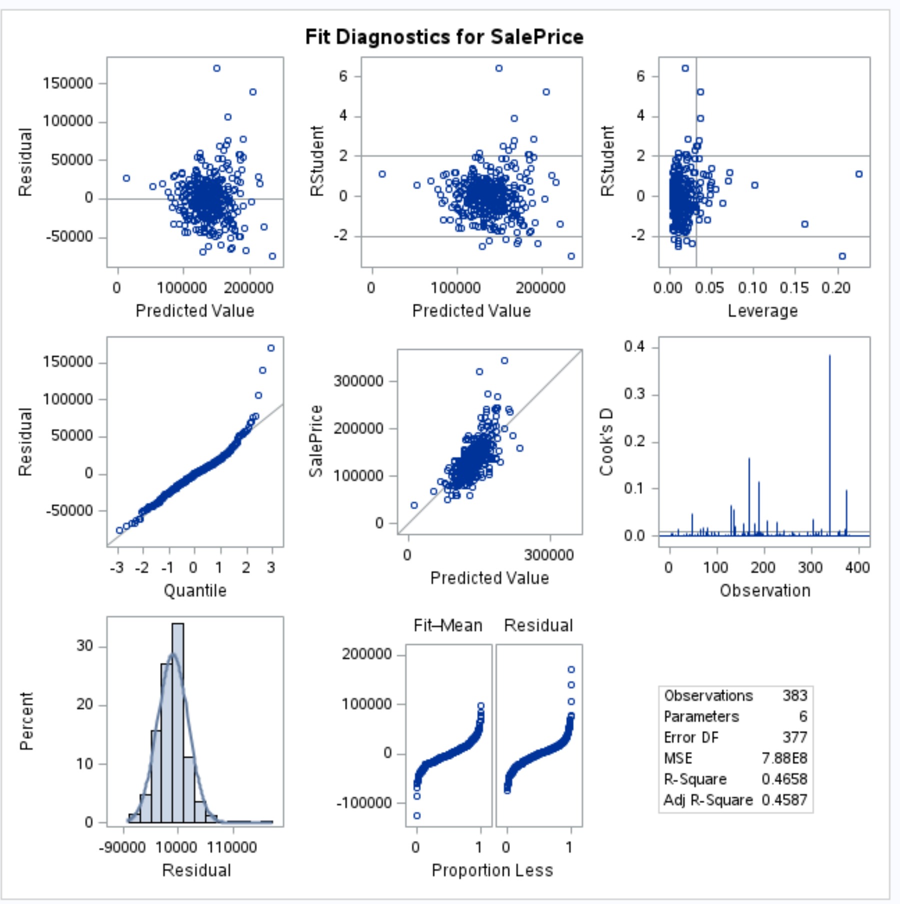


Observations	383
Parameters	6
Error DF	377
MSE	8.15E8
R-Square	0.4474
Adj R-Square	0.44

Appendix B - Model 2 Stats

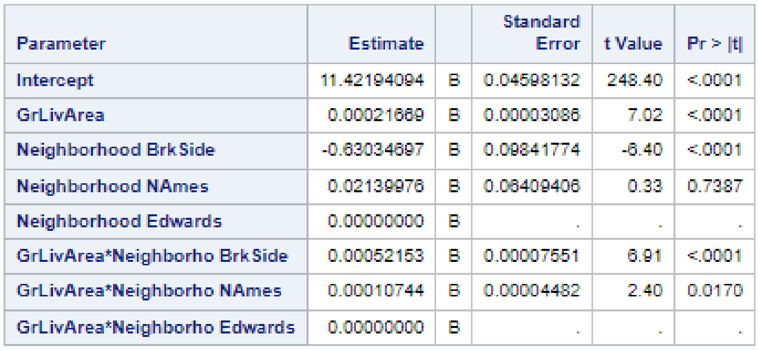


Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-376068.5008	B	58866.30193	-6.39	<.0001
logliving	70716.6376	B	8245.46527	8.58	<.0001
Neighborhood BrkSide	-144769.8556	B	94299.67119	-1.54	0.1256
Neighborhood NAmes	-29405.8761	B	75555.48597	-0.39	0.6974
Neighborhood Edwards	0.0000	B	.	.	.
logliving*Neighborho BrkSide	21054.1136	B	13317.30660	1.58	0.1147
logliving*Neighborho NAmes	6546.6413	B	10581.99330	0.62	0.5365
logliving*Neighborho Edwards	0.0000	B	.	.	.

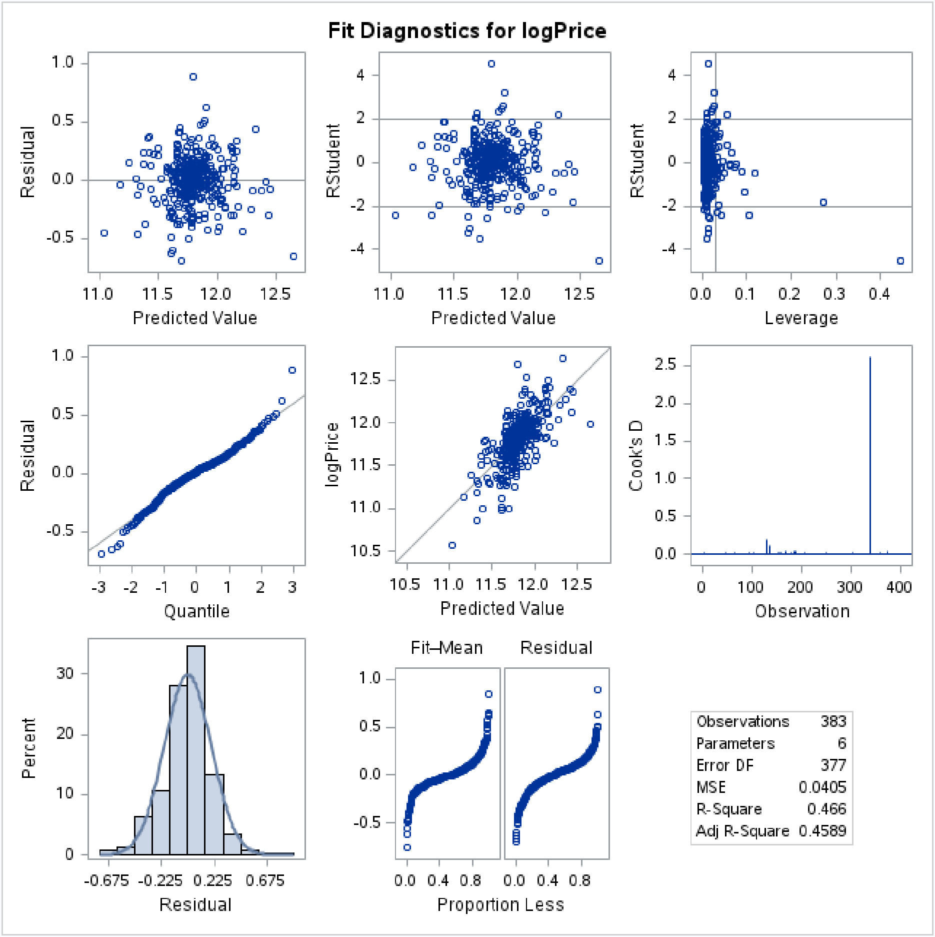


Observations	383
Parameters	6
Error DF	377
MSE	7.88E8
R-Square	0.4658
Adj R-Square	0.4587

Appendix C - Model 3 Stats



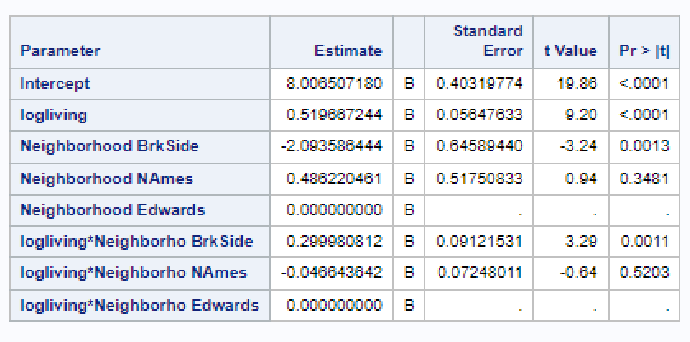
Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	11.42194094	B	0.04598132	248.40	<.0001
GrLivArea	0.00021669	B	0.00003086	7.02	<.0001
Neighborhood BrkSide	-0.63034697	B	0.09841774	-6.40	<.0001
Neighborhood NAmes	0.02139976	B	0.06409406	0.33	0.7387
Neighborhood Edwards	0.00000000	B	.	.	.
GrLivArea*Neighborho BrkSide	0.00052153	B	0.00007551	6.91	<.0001
GrLivArea*Neighborho NAmes	0.00010744	B	0.00004482	2.40	0.0170
GrLivArea*Neighborho Edwards	0.00000000	B	.	.	.



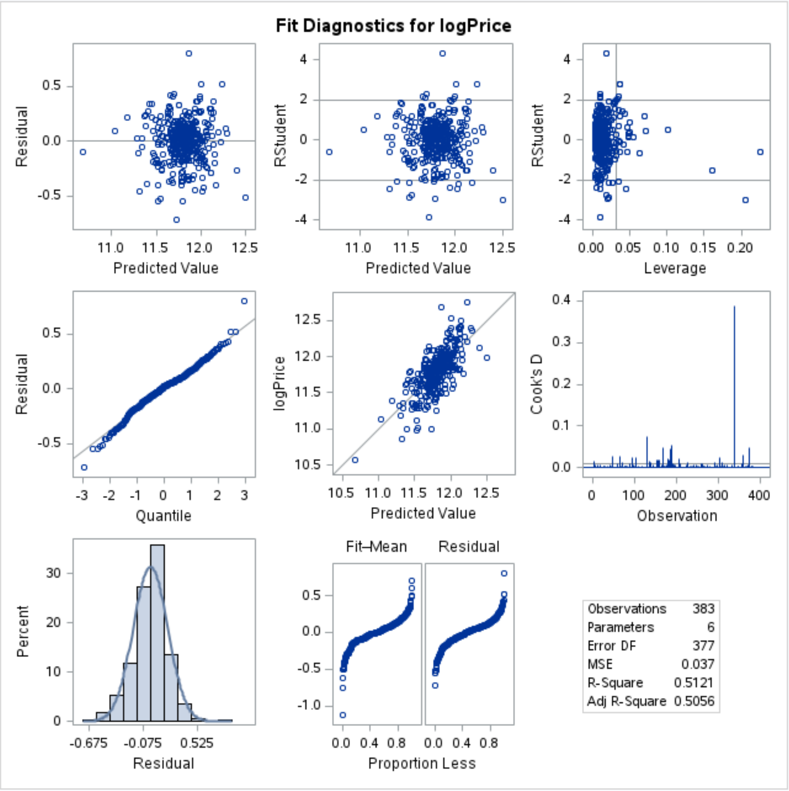
Fit Diagnostics for logPrice

Observations	383
Parameters	6
Error DF	377
MSE	0.0405
R-Square	0.466
Adj R-Square	0.4589

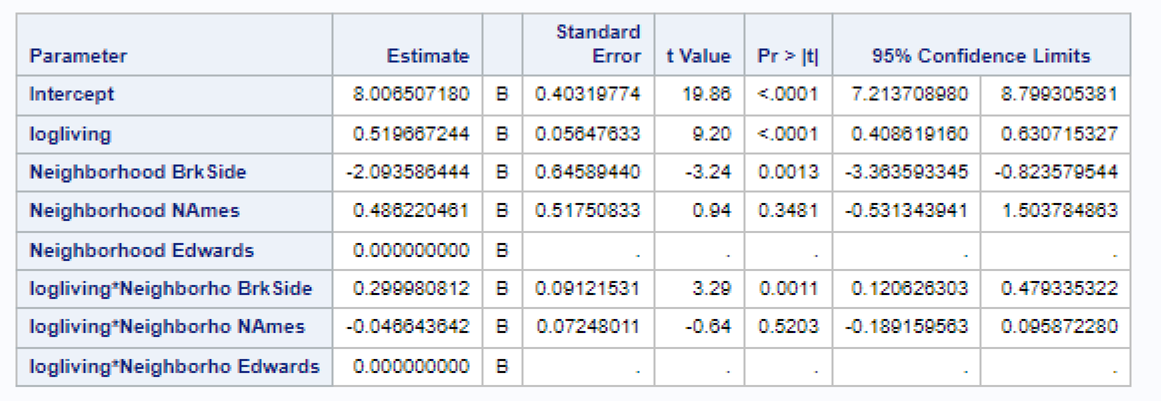
Appendix D - Model 4 Stats



Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	8.006507180	B	0.40319774	19.86	<.0001
logliving	0.519667244	B	0.05647633	9.20	<.0001
Neighborhood BrkSide	-2.093586444	B	0.64589440	-3.24	0.0013
Neighborhood NAmes	0.486220461	B	0.51750833	0.94	0.3481
Neighborhood Edwards	0.000000000	B	.	.	.
logliving*Neighborho BrkSide	0.299980812	B	0.09121531	3.29	0.0011
logliving*Neighborho NAmes	-0.046643642	B	0.07248011	-0.64	0.5203
logliving*Neighborho Edwards	0.000000000	B	.	.	.

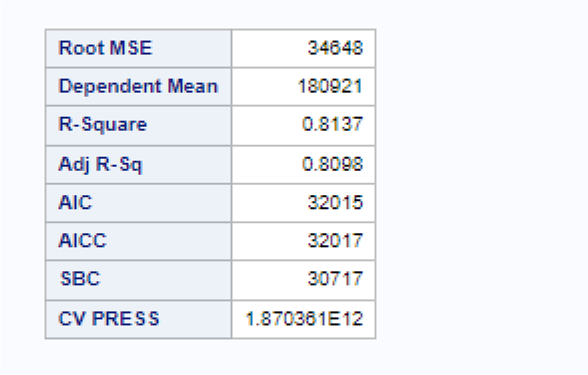

Fit Diagnostics for logPrice

Observations	383
Parameters	6
Error DF	377
MSE	0.037
R-Square	0.5121
Adj R-Square	0.5056

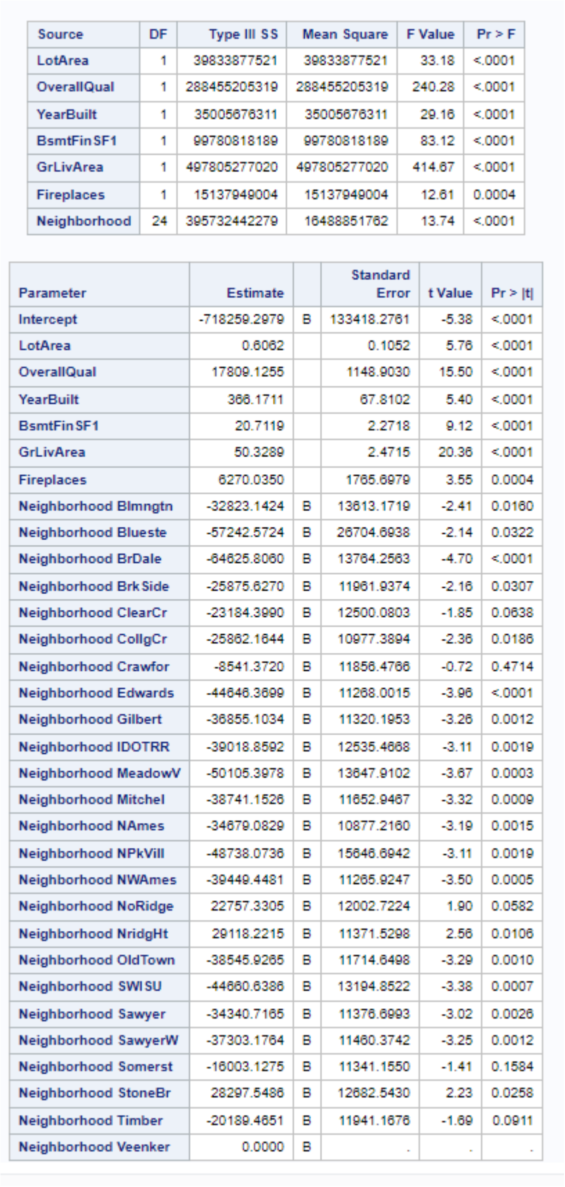


Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	8.006507180	B	0.40319774	19.86	<.0001	7.213708980	8.799305381
logliving	0.519667244	B	0.05647633	9.20	<.0001	0.408619160	0.630715327
Neighborhood BrkSide	-2.093586444	B	0.64589440	-3.24	0.0013	-3.363593345	-0.823579544
Neighborhood NAmes	0.486220461	B	0.51750833	0.94	0.3481	-0.531343941	1.503784863
Neighborhood Edwards	0.000000000	B	.	.	.	.	.
logliving*Neighborho BrkSide	0.299980812	B	0.09121531	3.29	0.0011	0.120626303	0.479335322
logliving*Neighborho NAmes	-0.046643642	B	0.07248011	-0.64	0.5203	-0.189159563	0.095872280
logliving*Neighborho Edwards	0.000000000	B	.	.	.	.	.

Appendix E- Forward Model



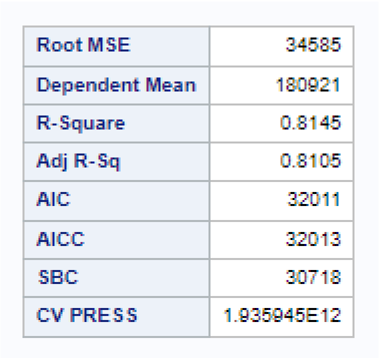
Root MSE	34648
Dependent Mean	180921
R-Square	0.8137
Adj R-Sq	0.8098
AIC	32015
AICC	32017
SBC	30717
CV PRESS	1.870361E12



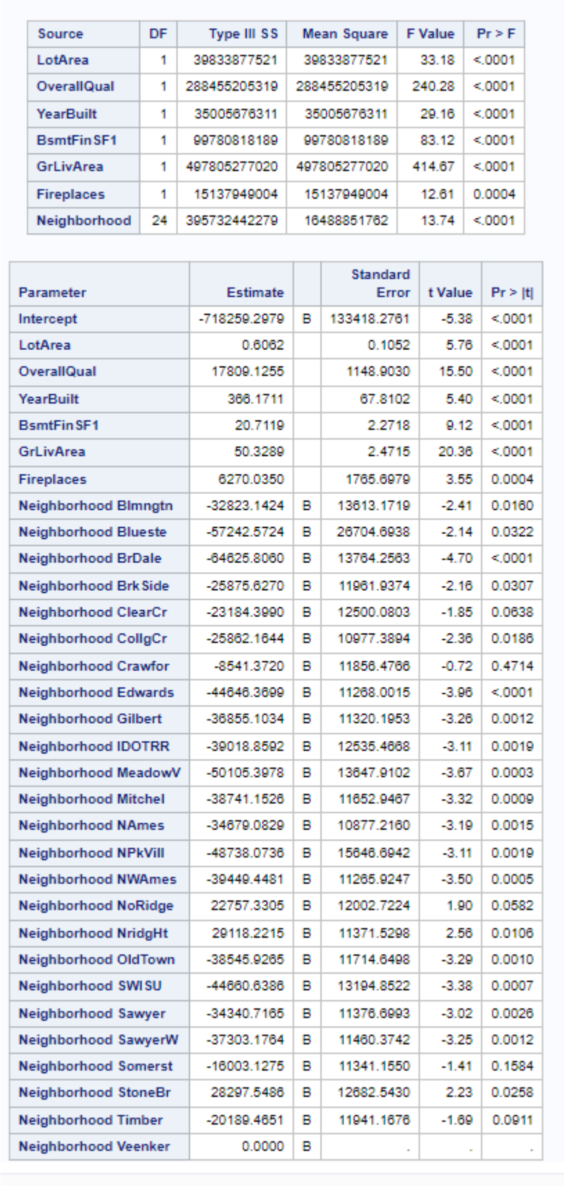
Source	DF	Type III SS	Mean Square	F Value	Pr > F
LotArea	1	39833877521	39833877521	33.18	<.0001
OverallQual	1	288455205319	288455205319	240.28	<.0001
YearBuilt	1	35005676311	35005676311	29.16	<.0001
BsmtFinSF1	1	99780818189	99780818189	83.12	<.0001
GrLivArea	1	497805277020	497805277020	414.67	<.0001
Fireplaces	1	15137949004	15137949004	12.61	0.0004
Neighborhood	24	395732442279	16488851762	13.74	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-718259.2979	B	133418.2761	-5.38	<.0001
LotArea	0.6062		0.1052	5.76	<.0001
OverallQual	17809.1255		1148.9030	15.50	<.0001
YearBuilt	366.1711		67.8102	5.40	<.0001
BsmtFinSF1	20.7119		2.2718	9.12	<.0001
GrLivArea	50.3289		2.4715	20.36	<.0001
Fireplaces	6270.0350		1765.6979	3.55	0.0004
Neighborhood Blmngtn	-32823.1424	B	13613.1719	-2.41	0.0160
Neighborhood Blueste	-57242.5724	B	26704.6938	-2.14	0.0322
Neighborhood BrDale	-64625.8060	B	13764.2563	-4.70	<.0001
Neighborhood BrkSide	-25875.6270	B	11961.9374	-2.16	0.0307
Neighborhood ClearCr	-23184.3990	B	12500.0803	-1.85	0.0638
Neighborhood CollgCr	-25862.1644	B	10977.3894	-2.36	0.0186
Neighborhood Crawfor	-8541.3720	B	11856.4766	-0.72	0.4714
Neighborhood Edwards	-44646.3699	B	11268.0015	-3.96	<.0001
Neighborhood Gilbert	-36855.1034	B	11320.1953	-3.26	0.0012
Neighborhood IDOTRR	-39018.8592	B	12535.4668	-3.11	0.0019
Neighborhood MeadowV	-50105.3978	B	13647.9102	-3.67	0.0003
Neighborhood Mitchel	-38741.1526	B	11652.9467	-3.32	0.0009
Neighborhood NAmes	-34679.0829	B	10877.2160	-3.19	0.0015
Neighborhood NPkVill	-48738.0736	B	15646.6942	-3.11	0.0019
Neighborhood NWAmes	-39449.4481	B	11265.9247	-3.50	0.0005
Neighborhood NoRidge	22757.3305	B	12002.7224	1.90	0.0582
Neighborhood NridgHt	29118.2215	B	11371.5298	2.56	0.0106
Neighborhood OldTown	-38545.9265	B	11714.6498	-3.29	0.0010
Neighborhood SWISU	-44660.6386	B	13194.8522	-3.38	0.0007
Neighborhood Sawyer	-34340.7165	B	11376.6993	-3.02	0.0026
Neighborhood SawyerW	-37303.1764	B	11460.3742	-3.25	0.0012
Neighborhood Somerst	-16003.1275	B	11341.1550	-1.41	0.1584
Neighborhood StoneBr	28297.5486	B	12682.5430	2.23	0.0258
Neighborhood Timber	-20189.4651	B	11941.1676	-1.69	0.0911
Neighborhood Veenker	0.0000	B	.	.	.

Appendix F - Backwards Model



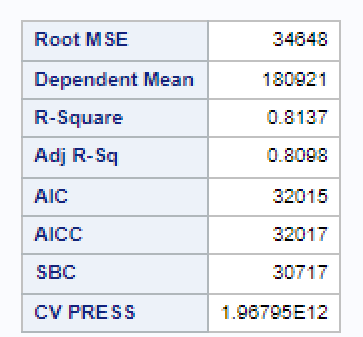
Root MSE	34585
Dependent Mean	180921
R-Square	0.8145
Adj R-Sq	0.8105
AIC	32011
AICC	32013
SBC	30718
CV PRESS	1.935945E12



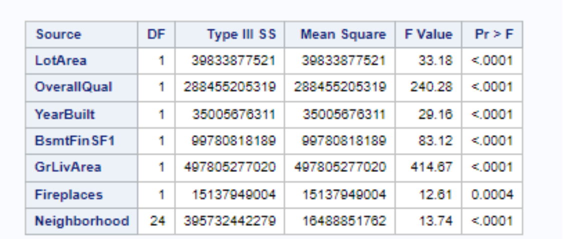
Source	DF	Type III SS	Mean Square	F Value	Pr > F
LotArea	1	39833877521	39833877521	33.18	<.0001
OverallQual	1	288455205319	288455205319	240.28	<.0001
YearBuilt	1	35005676311	35005676311	29.16	<.0001
BsmtFinSF1	1	99780818189	99780818189	83.12	<.0001
GrLivArea	1	497805277020	497805277020	414.67	<.0001
Fireplaces	1	15137949004	15137949004	12.61	0.0004
Neighborhood	24	395732442279	16488851762	13.74	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-718259.2979	B	133418.2761	-5.38	<.0001
LotArea	0.6062		0.1052	5.76	<.0001
OverallQual	17809.1255		1148.9030	15.50	<.0001
YearBuilt	366.1711		67.8102	5.40	<.0001
BsmtFinSF1	20.7119		2.2718	9.12	<.0001
GrLivArea	50.3289		2.4715	20.36	<.0001
Fireplaces	6270.0350		1765.6979	3.55	0.0004
Neighborhood Blmngtn	-32823.1424	B	13613.1719	-2.41	0.0160
Neighborhood Blueste	-57242.5724	B	26704.6938	-2.14	0.0322
Neighborhood BrDale	-64625.8060	B	13764.2563	-4.70	<.0001
Neighborhood BrkSide	-25875.6270	B	11961.9374	-2.16	0.0307
Neighborhood ClearCr	-23184.3990	B	12500.0803	-1.85	0.0638
Neighborhood CollgCr	-25862.1644	B	10977.3894	-2.36	0.0186
Neighborhood Crawfor	-8541.3720	B	11856.4766	-0.72	0.4714
Neighborhood Edwards	-44646.3699	B	11268.0015	-3.96	<.0001
Neighborhood Gilbert	-36855.1034	B	11320.1953	-3.26	0.0012
Neighborhood IDOTRR	-39018.8592	B	12535.4668	-3.11	0.0019
Neighborhood MeadowV	-50105.3978	B	13647.9102	-3.67	0.0003
Neighborhood Mitchel	-38741.1526	B	11652.9467	-3.32	0.0009
Neighborhood NAmes	-34679.0829	B	10877.2160	-3.19	0.0015
Neighborhood NPkVill	-48738.0736	B	15646.6942	-3.11	0.0019
Neighborhood NWAmes	-39449.4481	B	11265.9247	-3.50	0.0005
Neighborhood NoRidge	22757.3305	B	12002.7224	1.90	0.0582
Neighborhood NridgHt	29118.2215	B	11371.5298	2.56	0.0106
Neighborhood OldTown	-38545.9265	B	11714.6498	-3.29	0.0010
Neighborhood SWISU	-44660.6386	B	13194.8522	-3.38	0.0007
Neighborhood Sawyer	-34340.7165	B	11376.6993	-3.02	0.0026
Neighborhood SawyerW	-37303.1764	B	11460.3742	-3.25	0.0012
Neighborhood Somerst	-16003.1275	B	11341.1550	-1.41	0.1584
Neighborhood StoneBr	28297.5486	B	12682.5430	2.23	0.0258
Neighborhood Timber	-20189.4651	B	11941.1676	-1.69	0.0911
Neighborhood Veenker	0.0000	B	.	.	.

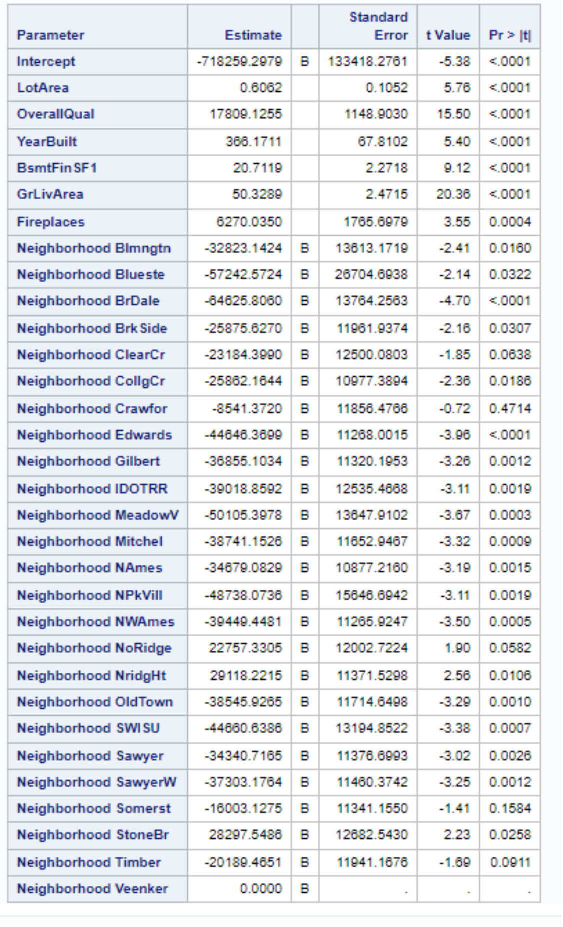
Appendix G - Stepwise Model



Root MSE	34648
Dependent Mean	180921
R-Square	0.8137
Adj R-Sq	0.8098
AIC	32015
AICC	32017
SBC	30717
CV PRESS	1.96795E12

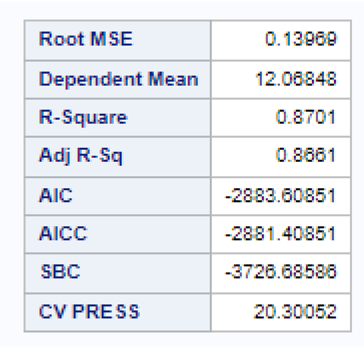


Source	DF	Type III SS	Mean Square	F Value	Pr > F
LotArea	1	39833877521	39833877521	33.18	<.0001
OverallQual	1	288455205319	288455205319	240.28	<.0001
YearBuilt	1	35005676311	35005676311	29.16	<.0001
BsmtFinSF1	1	99780818189	99780818189	83.12	<.0001
GrLivArea	1	497805277020	497805277020	414.67	<.0001
Fireplaces	1	15137949004	15137949004	12.61	0.0004
Neighborhood	24	395732442279	16488851762	13.74	<.0001

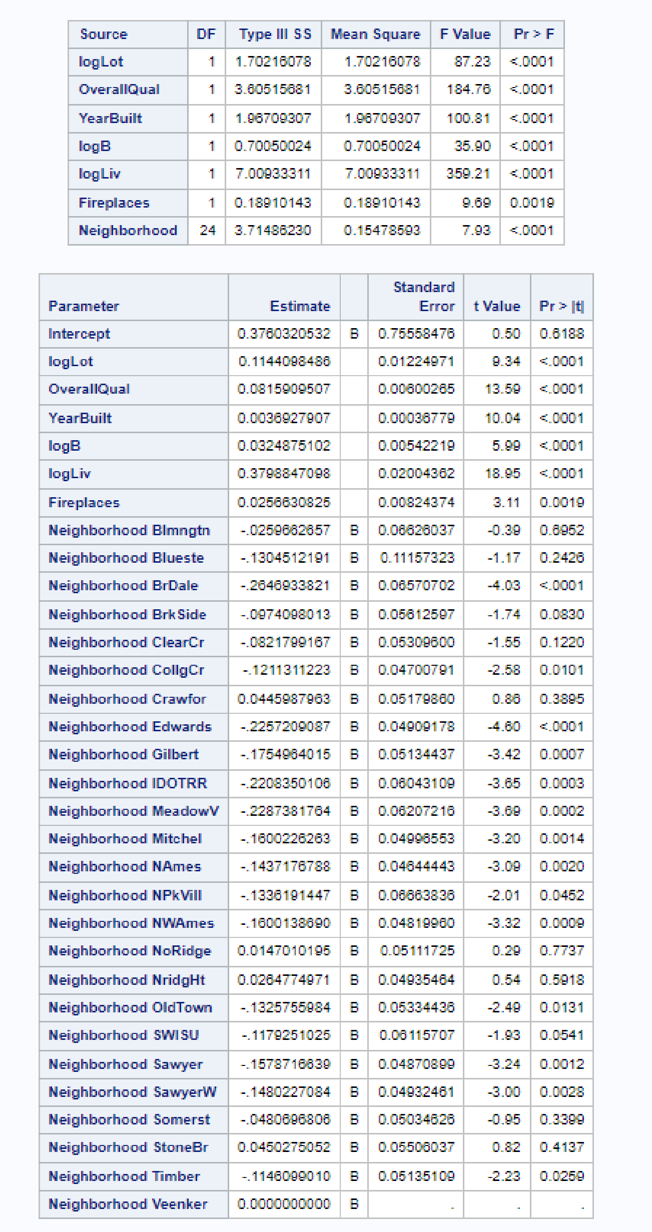


Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-718259.2979	B	133418.2761	-5.38	<.0001
LotArea	0.6062		0.1052	5.76	<.0001
OverallQual	17809.1255		1148.9030	15.50	<.0001
YearBuilt	366.1711		67.8102	5.40	<.0001
BsmtFinSF1	20.7119		2.2718	9.12	<.0001
GrLivArea	50.3289		2.4715	20.36	<.0001
Fireplaces	6270.0350		1765.6979	3.55	0.0004
Neighborhood Blmngtn	-32823.1424	B	13613.1719	-2.41	0.0160
Neighborhood Blueste	-57242.5724	B	26704.6938	-2.14	0.0322
Neighborhood BrDale	-64625.8060	B	13764.2563	-4.70	<.0001
Neighborhood BrkSide	-25875.6270	B	11961.9374	-2.16	0.0307
Neighborhood ClearCr	-23184.3990	B	12500.0803	-1.85	0.0638
Neighborhood CollgCr	-25862.1644	B	10977.3894	-2.36	0.0186
Neighborhood Crawfor	-8541.3720	B	11856.4766	-0.72	0.4714
Neighborhood Edwards	-44646.3699	B	11268.0015	-3.96	<.0001
Neighborhood Gilbert	-36855.1034	B	11320.1953	-3.26	0.0012
Neighborhood IDOTRR	-39018.8592	B	12535.4668	-3.11	0.0019
Neighborhood MeadowV	-50105.3978	B	13647.9102	-3.67	0.0003
Neighborhood Mitchel	-38741.1526	B	11652.9467	-3.32	0.0009
Neighborhood NAmes	-34679.0829	B	10877.2160	-3.19	0.0015
Neighborhood NPkVill	-48738.0736	B	15646.6942	-3.11	0.0019
Neighborhood NWAmes	-39449.4481	B	11265.9247	-3.50	0.0005
Neighborhood NoRidge	22757.3305	B	12002.7224	1.90	0.0582
Neighborhood NridgHt	29118.2215	B	11371.5298	2.56	0.0106
Neighborhood OldTown	-38545.9265	B	11714.6498	-3.29	0.0010
Neighborhood SWISU	-44660.6386	B	13194.8522	-3.38	0.0007
Neighborhood Sawyer	-34340.7165	B	11376.6993	-3.02	0.0026
Neighborhood SawyerW	-37303.1764	B	11460.3742	-3.25	0.0012
Neighborhood Somerst	-16003.1275	B	11341.1550	-1.41	0.1584
Neighborhood StoneBr	28297.5486	B	12682.5430	2.23	0.0258
Neighborhood Timber	-20189.4651	B	11941.1676	-1.69	0.0911
Neighborhood Veenker	0.0000	B	.	.	.

Appendix H - Custom Model



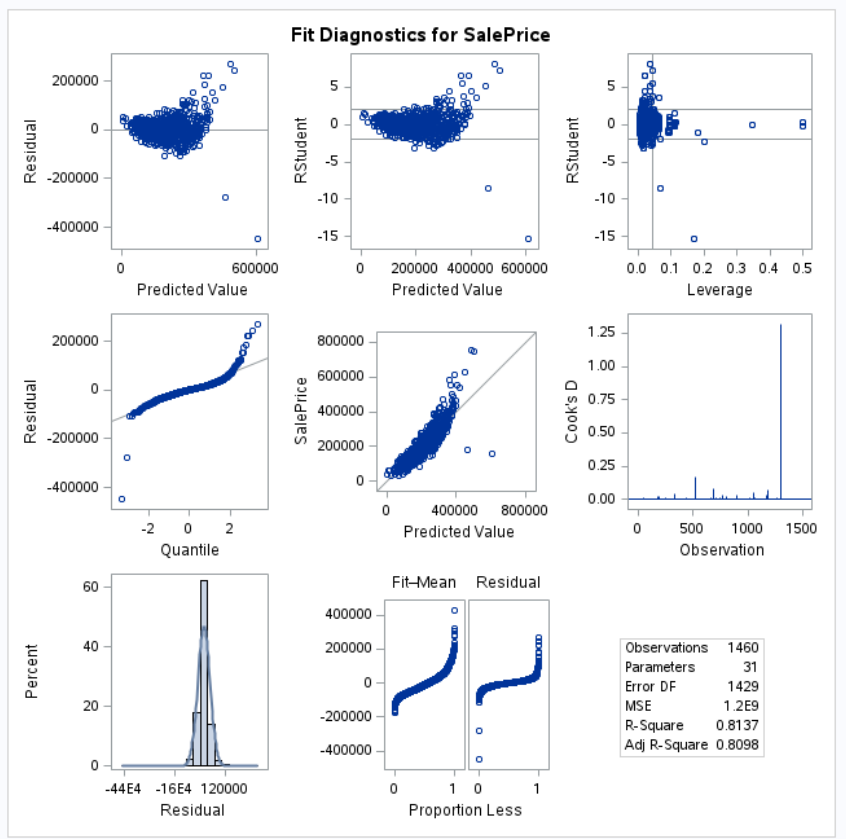
Root MSE	0.13969
Dependent Mean	12.06848
R-Square	0.8701
Adj R-Sq	0.8661
AIC	-2883.60851
AICC	-2881.40851
SBC	-3726.68586
CV PRESS	20.30052



Source	DF	Type III SS	Mean Square	F Value	Pr > F
logLot	1	1.70216078	1.70216078	87.23	<.0001
OverallQual	1	3.60515681	3.60515681	184.76	<.0001
YearBuilt	1	1.96709307	1.96709307	100.81	<.0001
logB	1	0.70050024	0.70050024	35.90	<.0001
logLiv	1	7.00933311	7.00933311	359.21	<.0001
Fireplaces	1	0.18910143	0.18910143	9.69	0.0019
Neighborhood	24	3.71486230	0.15478593	7.93	<.0001

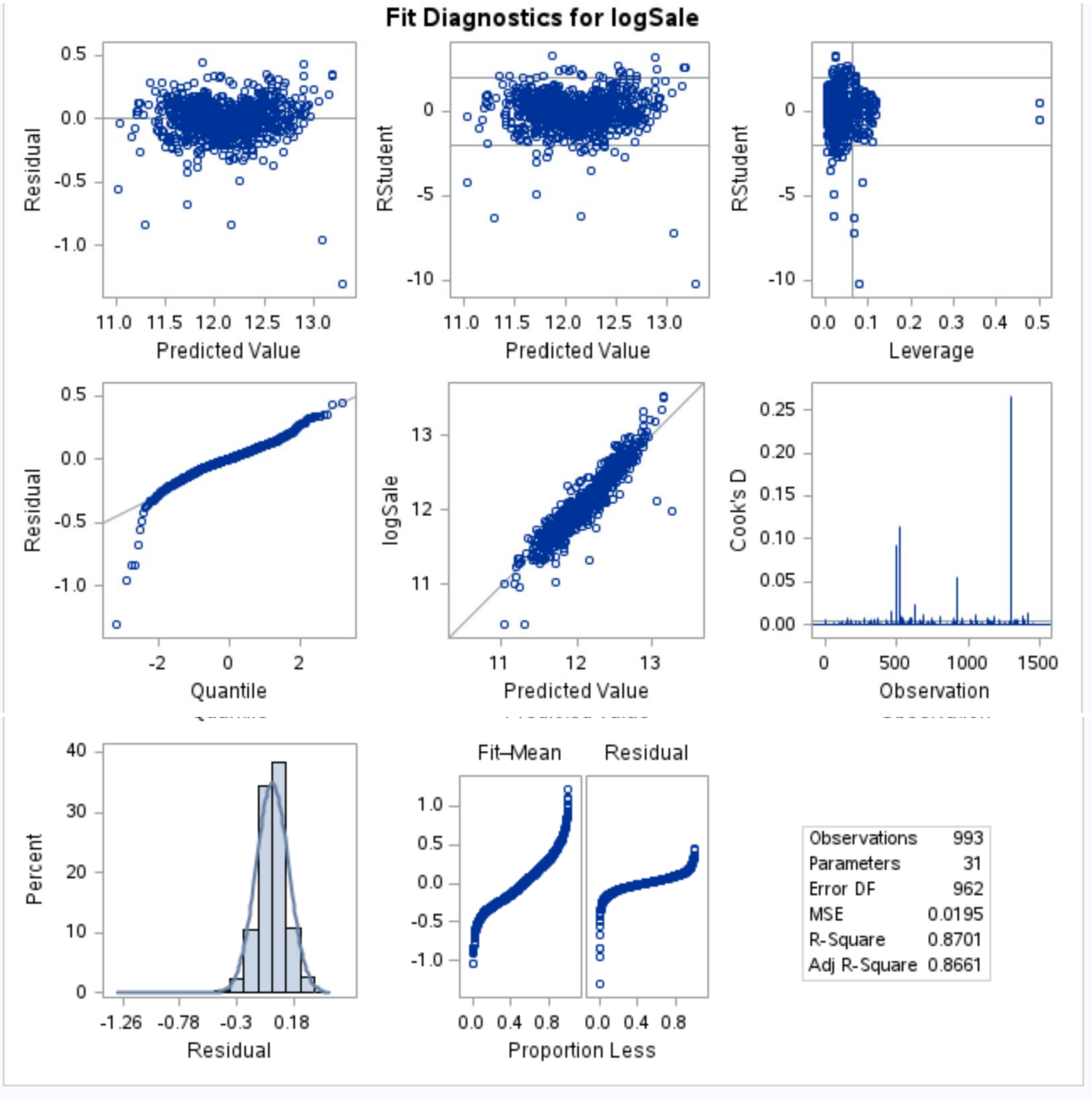
Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	0.3760320532	B	0.75558476	0.50	0.6188
logLot	0.1144098486		0.01224971	9.34	<.0001
OverallQual	0.0815909507		0.00600265	13.59	<.0001
YearBuilt	0.0036927907		0.00036779	10.04	<.0001
logB	0.0324875102		0.00542219	5.99	<.0001
logLiv	0.3798847098		0.02004362	18.95	<.0001
Fireplaces	0.0256630825		0.00824374	3.11	0.0019
Neighborhood Blmngtn	-.0259662657	B	0.06626037	-0.39	0.6952
Neighborhood Blueste	-.1304512191	B	0.11157323	-1.17	0.2426
Neighborhood BrDale	-.2646933821	B	0.06570702	-4.03	<.0001
Neighborhood BrkSide	-.0974098013	B	0.05612597	-1.74	0.0830
Neighborhood ClearCr	-.0821799167	B	0.05309600	-1.55	0.1220
Neighborhood CollgCr	-.1211311223	B	0.04700791	-2.58	0.0101
Neighborhood Crawfor	0.0445987963	B	0.05179860	0.86	0.3895
Neighborhood Edwards	-.2257209087	B	0.04909178	-4.60	<.0001
Neighborhood Gilbert	-.1754964015	B	0.05134437	-3.42	0.0007
Neighborhood IDOTRR	-.2208350106	B	0.06043109	-3.65	0.0003
Neighborhood MeadowV	-.2287381764	B	0.06207216	-3.69	0.0002
Neighborhood Mitchel	-.1600226263	B	0.04996553	-3.20	0.0014
Neighborhood NAmes	-.1437176788	B	0.04644443	-3.09	0.0020
Neighborhood NPkVill	-.1336191447	B	0.06663836	-2.01	0.0452
Neighborhood NWAmes	-.1600138690	B	0.04819960	-3.32	0.0009
Neighborhood NoRidge	0.0147010195	B	0.05111725	0.29	0.7737
Neighborhood NridgHt	0.0264774971	B	0.04935464	0.54	0.5918
Neighborhood OldTown	-.1325755984	B	0.05334436	-2.49	0.0131
Neighborhood SWISU	-.1179251025	B	0.06115707	-1.93	0.0541
Neighborhood Sawyer	-.1578716639	B	0.04870899	-3.24	0.0012
Neighborhood SawyerW	-.1480227084	B	0.04932461	-3.00	0.0028
Neighborhood Somerst	-.0480696806	B	0.05034626	-0.95	0.3399
Neighborhood StoneBr	0.0450275052	B	0.05506037	0.82	0.4137
Neighborhood Timber	-.1146099010	B	0.05135109	-2.23	0.0259
Neighborhood Veenker	0.0000000000	B	.	.	.

Appendix I - Assumptions of Forward, Backward and Stepwise model



Observations	1460
Parameters	31
Error DF	1429
MSE	1.2E9
R-Square	0.8137
Adj R-Square	0.8098

Appendix J - Custom Model Assumptions



Observations	993
Parameters	31
Error DF	962
MSE	0.0195
R-Square	0.8701
Adj R-Square	0.8661

Appendix K - Code:

Analysis Q1:

```
*create log variables;
data Ames2;
set Ames;
logliving = log(GrLivArea);
logPrice = log(SalePrice);

*Model 1;
proc glm data = Ames2 plots=all;
class Neighborhood (ref = "Edwards");
model SalePrice = GrLivArea | Neighborhood / solution;
run;

*Model 2;
proc glm data = Ames2 plots=all;
class Neighborhood (ref = "Edwards");
model SalePrice = logliving | Neighborhood / solution;
run;

*Model 3;
proc glm data = Ames2 plots=all;
class Neighborhood (ref = "Edwards");
model logPrice = GrLivArea | Neighborhood / solution;
run;

*Model 4;
**double log is best;
proc glm data = Ames2 plots=all;
class Neighborhood (ref = "Edwards");
model logPrice = logliving | Neighborhood / solution clparm;
run;
```

Analysis Q2:

```
proc reg data = import1;
model SalePrice = MSSubClass  LotArea OverallQual OverallCond YearBuilt YearRemodAdd
 BsmtFinSF1 BsmtUnfSF TotalBsmtSF  fstFlrSF sndFlrSF
 LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
KitchenAbvGr TotRmsAbvGrd
 Fireplaces GarageCars GarageArea WoodDeckSF
 MiscVal MoSold  / selection = stepwise slentry = 0.05 adjrsq;
```

```
run;

/* backward */
proc glm data = import1;
class Neighborhood BldgType;
model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt BsmtFinSF1
GrLivArea BsmtFullBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces
GarageCars ;
run;



/* stepwise */
proc reg data = import1;
model SalePrice = MSSubClass  LotArea OverallQual OverallCond YearBuilt YearRemodAdd
 BsmtFinSF1 BsmtUnfSF TotalBsmtSF  fstFlrSF sndFlrSF
 LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
KitchenAbvGr TotRmsAbvGrd
 Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch
 PoolArea MiscVal MoSold / selection = stepwise slentry = 0.1 slstay = 0.1  adjrsq;
run;



proc glmselect data = import1;
class Neighborhood MSZoning BldgType HouseStyle RoofStyle GarageType SaleCondition;
model SalePrice = MSSubClass  LotArea OverallQual OverallCond YearBuilt YearRemodAdd
 BsmtFinSF1 BsmtUnfSF TotalBsmtSF  fstFlrSF sndFlrSF
 LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
KitchenAbvGr TotRmsAbvGrd
 Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
 PoolArea MiscVal MoSold Neighborhood MSZoning BldgType HouseStyle RoofStyle GarageType
SaleCondition
 / selection = Stepwise(stop = CV) cvmethod = random(10) stats = adjrsq;
run;


proc glmselect data = import1;
class Neighborhood MSZoning BldgType HouseStyle RoofStyle GarageType SaleCondition;
model SalePrice = MSSubClass  LotArea OverallQual OverallCond YearBuilt YearRemodAdd
 BsmtFinSF1 BsmtUnfSF TotalBsmtSF  fstFlrSF sndFlrSF
 LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
KitchenAbvGr TotRmsAbvGrd
 Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
 PoolArea MiscVal MoSold Neighborhood MSZoning BldgType HouseStyle RoofStyle GarageType
SaleCondition
 / selection = Backward(stop = CV) cvmethod = random(10) stats = adjrsq;
run;

proc glmselect data = import1;
class Neighborhood MSZoning BldgType HouseStyle RoofStyle GarageType SaleCondition;
model SalePrice = MSSubClass  LotArea OverallQual OverallCond YearBuilt YearRemodAdd
 BsmtFinSF1 BsmtUnfSF TotalBsmtSF  fstFlrSF sndFlrSF
 LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
KitchenAbvGr TotRmsAbvGrd
 Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
 PoolArea MiscVal MoSold Neighborhood MSZoning BldgType HouseStyle RoofStyle GarageType
SaleCondition
 / selection = Forward(stop = CV) cvmethod = random(10) stats = adjrsq;
run;


proc glm data = import1 plots=all;
model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt
BsmtFinSF1 GrLivArea BedroomAbvGr GarageCars / solution;
run;

data Ames4;
set import1;
logSale = log(SalePrice);
run;

proc print data = Ames4;
run;



proc glm data = Ames4 plots=all;
model logSale = MSSubClass LotArea OverallQual OverallCond YearBuilt
BsmtFinSF1 GrLivArea BedroomAbvGr GarageCars / solution;
run;



data importSec;
set import1;
logLot = log(LotArea);
logSale = log(SalePrice);
```

```
logB = log(BsmtFinSF1);
logLiv = log(GrLivArea);
run;

proc reg data = importSec;
model logSale =  logLot OverallQual YearBuilt logB
GrLivArea BedroomAbvGr Fireplaces
GarageCars  / vif tol;
run;


proc glm data = importSec plots=all;
class Neighborhood;
model logSale =  logLot OverallQual YearBuilt logB
GrLivArea BedroomAbvGr Fireplaces Neighborhood/ solution;
run;



proc glmselect data = import1;
class Neighborhood;
model SalePrice = LotArea OverallQual YearBuilt BsmtFinSF1
GrLivArea BedroomAbvGr Fireplaces Neighborhood
 / selection = Forward(stop = CV) cvmethod = random(10) stats = adjrsq;
run;


proc glmselect data = import1;
class Neighborhood;
model SalePrice = LotArea OverallQual YearBuilt BsmtFinSF1
GrLivArea BedroomAbvGr Fireplaces Neighborhood
 / selection = Backward(stop = CV) cvmethod = random(10) stats = adjrsq;
run;


proc glmselect data = import1;
class Neighborhood;
model SalePrice = LotArea OverallQual YearBuilt BsmtFinSF1
GrLivArea BedroomAbvGr Fireplaces Neighborhood
 / selection = Stepwise(stop = CV) cvmethod = random(10) stats = adjrsq;
run;
```

```
*Forward model with p-vals;
proc glm data = import1 plots = ALL;
class Neighborhood;
model SalePrice = LotArea OverallQual YearBuilt BsmtFinSF1
GrLivArea Fireplaces Neighborhood / solution;
run;


proc glmselect data = importSec;
class Neighborhood;
model logSale =  logLot OverallQual YearBuilt logB
logLiv Fireplaces Neighborhood/ selection = Stepwise(stop=CV) cvmethod = random(10) stats = adjrsq;
run;


proc glm data = importSec plots=all;
class Neighborhood;
model logSale =  logLot OverallQual YearBuilt logB
logLiv Fireplaces Neighborhood/ solution;
run;
```