Summary

Century 21 Ames only sells home in NAmes, Edwards, and BrkSide neighborhoods and would like to get an estimate of how the SalePrice of each 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. This portion of the report will build a model with the provided dataset and provide our conclusion that quantifies the relationship between living area and sale price with respect to the three neighborhoods.

Descriptive Statistics

- File name: test.csv
- Observations
    - Sample: 383
- Variables
    - Sample: 81
- Averages by neighborhoods and all three combined

Neighborhood	Number of Properties	Avg. Living Area	Avg. Sale Price
NAmes	225	1310.31	$145,847.08
Edwards	100	1340.04	$128,219.70
BrkSide	58	1203.07	$124,834.05
All Three	383	1301.83	$138,062.50

Assumptions

Using the 383 observations for the three neighborhoods, our goal is to select a model that best fits the relationship between living area and sale price. This section will provide details on the selection process, the assumptions, and the model.

Selection Process

This section will demonstrate the various models ran, results of each, and our decision on which model is best based on how it satisfies the following assumptions:

1. Normality
2. Linearity
3. Equal Variance
4. Independence
5. Outliers

Model

The model assigns SalePrice as the response variable and GrLivArea as the explanatory variable. Based on the scatterplots, there is strong evidence the model will have different slopes and intercepts for each significant parameter.

Predicted SalePrice = ß0 + ß1 (GrLivArea) + ß2 (GrLivArea * NAmes) +
ß3 (GrLivArea * Edwards) + ß4 (GrLivArea * BrkSide)

Summary Table of Models

Vanilla Regression						
Model	**Normality**	**Linearity**	**Equal Variance**	**Independence**	**Outliers**	**R^2**
Original Data	✓			✓		0.447
Log-Linear	✓		✓	✓		0.465
Linear-Log	✓			✓		0.466
Log-Log	✓	✓	✓	✓		0.512
Original Data (Outliers Addressed)	✓	✓	✓	✓	✓	0.523
Log-Linear (Outliers Addressed)	✓	✓	✓	✓	✓	0.525
Linear-Log (Outliers Addressed)	✓		✓	✓	✓	0.0498
Log-Log (Outliers Addressed)	✓		✓	✓	✓	0.053

Assumptions

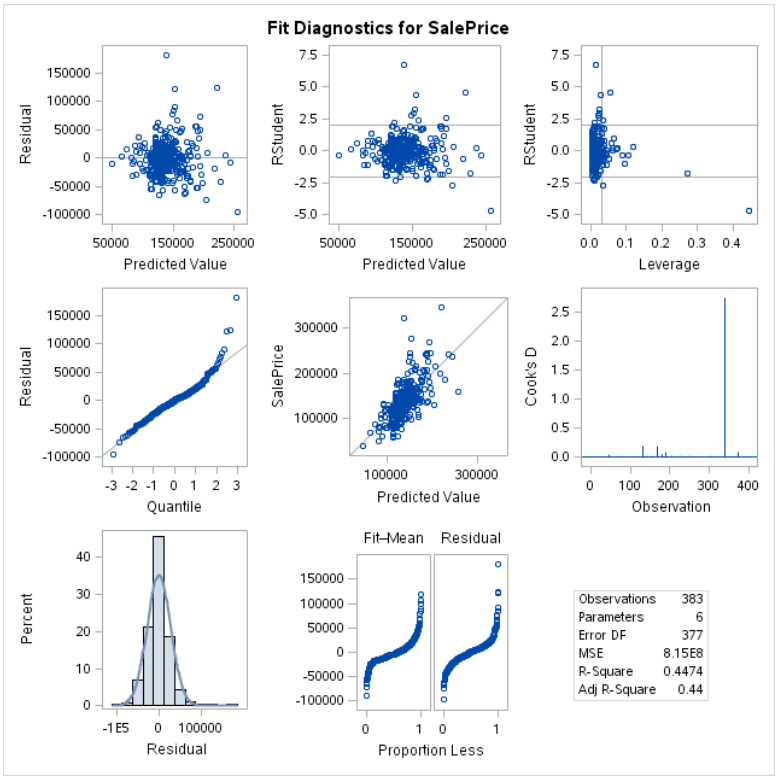
After running 8 different models, there is sufficient evidence that the vanilla regression model using the original data (three neighborhoods) while addressing the outliers is the best fit model. The following addresses the assumptions. **Refer to Addendum 5 for charts and values**.

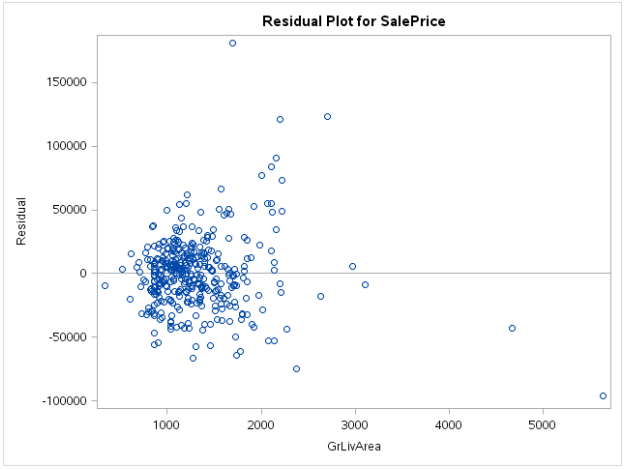
1. **Normality:** Judging by the histogram, the original data (outliers addressed) looks slightly better than the rest and shows evidence of normality.
2. **Linearity:** Judging by the scatterplots, the original data (outliers addressed) looks slightly better than the other two outliers addressed (log-linear and linear-log). There is sufficient evidence for linearity.
3. **Equal Variance:** Judging by the residual scatterplots and QQplot, the original data (outliers addressed) looks better than the rest and shows sufficient evidence for equal variance.
4. **Independence:** We will assume independence, although the data gathering process was not explained.
5. **Outliers:** There are a few outliers that were addressed. Specifically, homes in the Edwards neighborhood with living area greater than 4,000 sq. ft and two homes (one in Edwards and NAmes) with sale price over $300,000.

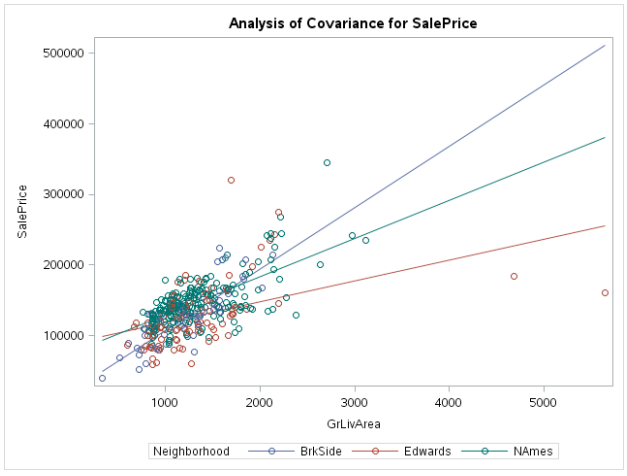
Addendum 1: Vanilla Regression – Raw Data | Model Original

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	.	.	.

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.447376	20.68070	28552.30	138062.5



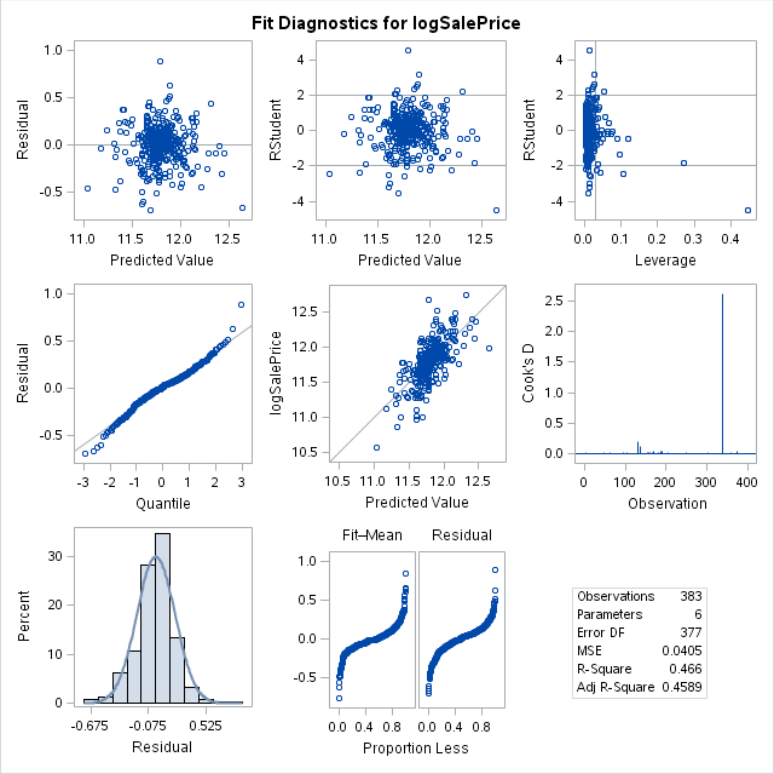


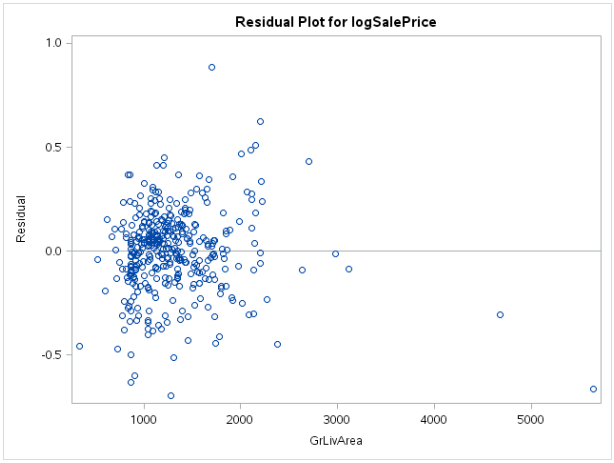


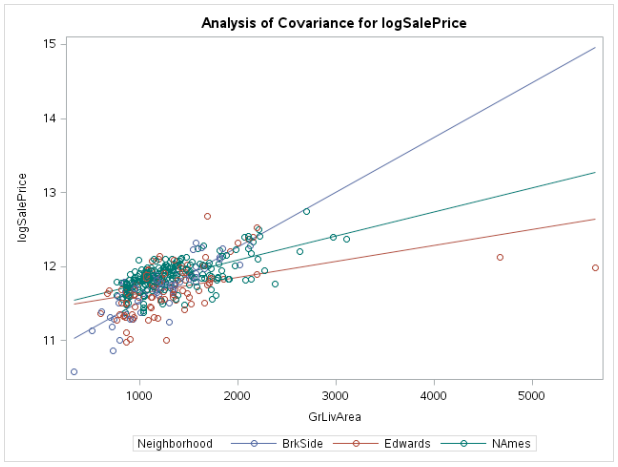
Addendum 2: Vanilla Regression | Model: Log-Linear

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	11.44334070	B	0.04465161	256.28	<.0001
GrLivArea	0.00032412	B	0.00003250	9.97	<.0001
Neighborhood BrkSide	-0.65174673	B	0.09780355	-6.66	<.0001
Neighborhood Edwards	-0.02139976	B	0.06409406	-0.33	0.7387
Neighborhood NAmes	0.00000000	B	.	.	.
GrLivArea*Neighborho BrkSide	0.00041410	B	0.00007620	5.43	<.0001
GrLivArea*Neighborho Edwards	-0.00010744	B	0.00004482	-2.40	0.0170
GrLivArea*Neighborho NAmes	0.00000000	B	.	.	.

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.465985	1.704877	0.201156	11.79887



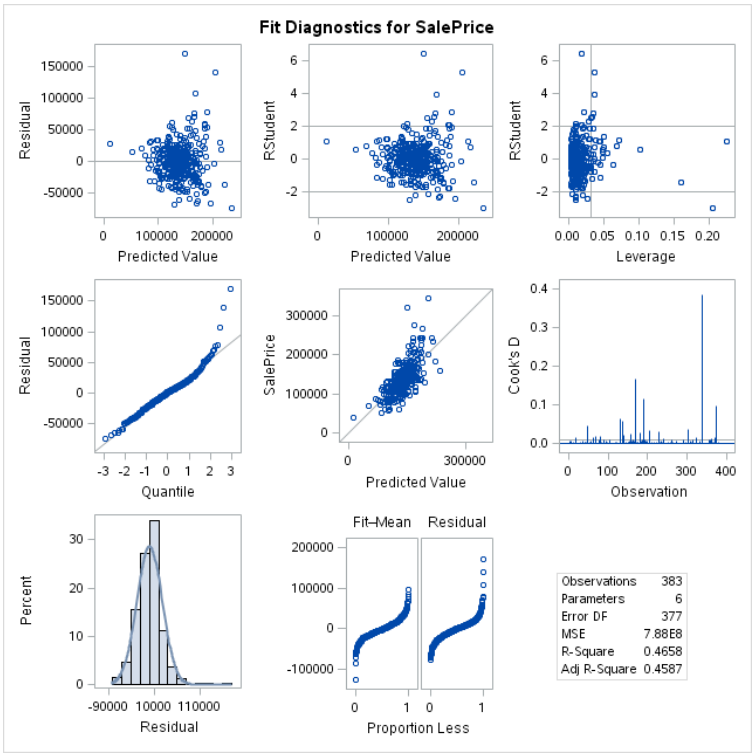


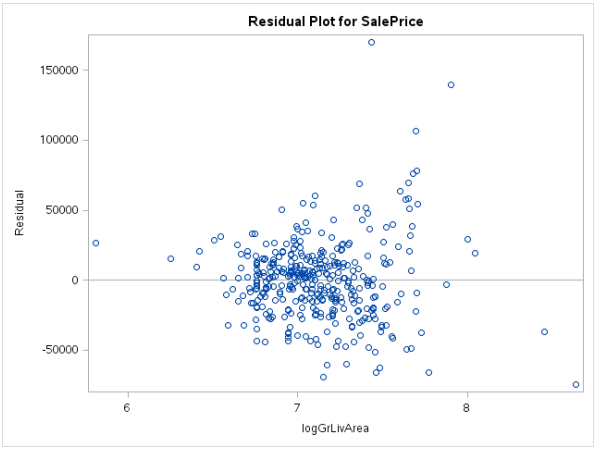


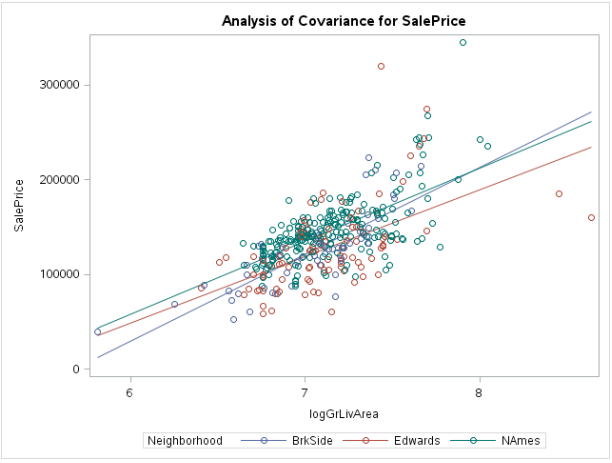
Addendum 3: Vanilla Regression | Model: Linear - Log

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-405474.3769	B	47364.43769	-8.56	<.0001
logGrLivArea	77263.2789	B	6632.56245	11.65	<.0001
Neighborhood BrkSide	-115363.9795	B	87581.82712	-1.32	0.1886
Neighborhood Edwards	29405.8761	B	75555.48597	0.39	0.6974
Neighborhood NAmes	0.0000	B	.	.	.
logGrLivA*Neighborho BrkSide	14507.4723	B	12383.61185	1.17	0.2421
logGrLivA*Neighborho Edwards	-6546.6413	B	10581.99330	-0.62	0.5365
logGrLivA*Neighborho NAmes	0.0000	B	.	.	.

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.465809	20.33287	28072.06	138062.5



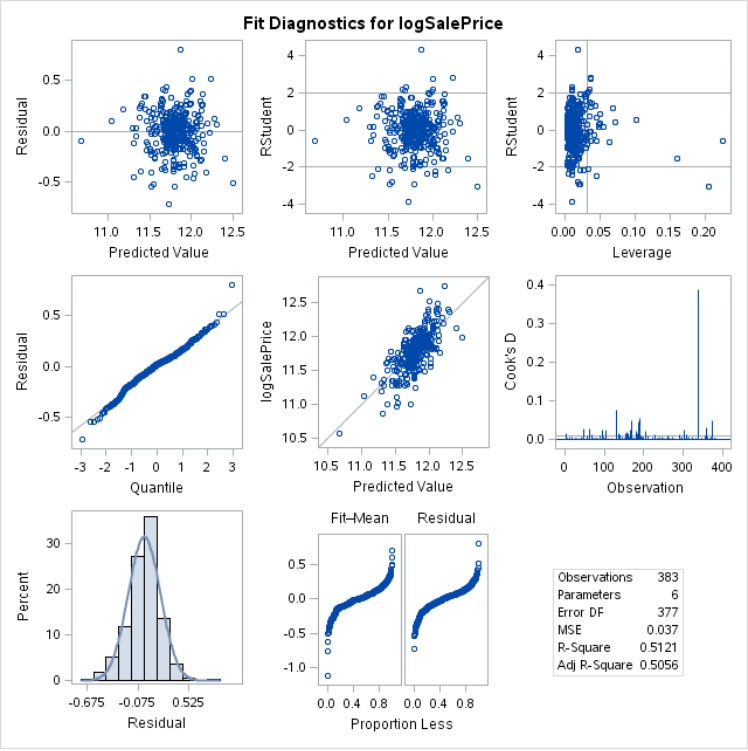


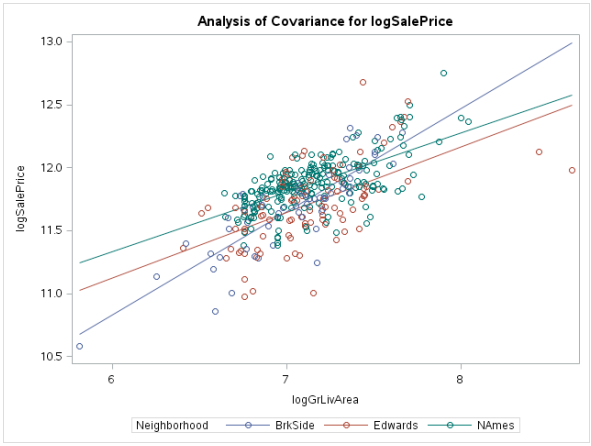


Addendum 4: Vanilla Regression | Model: Log-Log

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	8.492727641	B	0.32441709	26.18	<.0001
logGrLivArea	0.473023602	B	0.04542895	10.41	<.0001
Neighborhood BrkSide	-2.579806905	B	0.59988132	-4.30	<.0001
Neighborhood Edwards	-0.486220461	B	0.51750833	-0.94	0.3481
Neighborhood NAmes	0.000000000	B	.	.	.
logGrLivA*Neighborho BrkSide	0.346624454	B	0.08482008	4.09	<.0001
logGrLivA*Neighborho Edwards	0.046643642	B	0.07248011	0.64	0.5203
logGrLivA*Neighborho NAmes	0.000000000	B	.	.	.

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.512092	1.629617	0.192276	11.79887

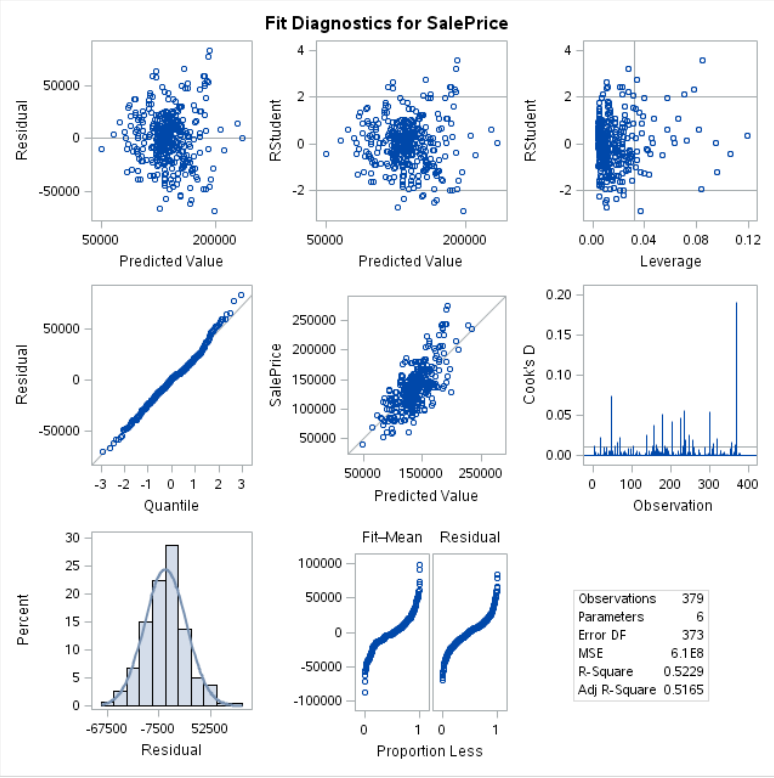


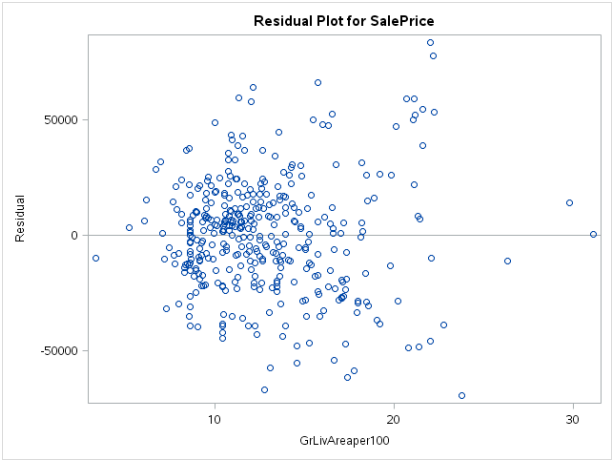


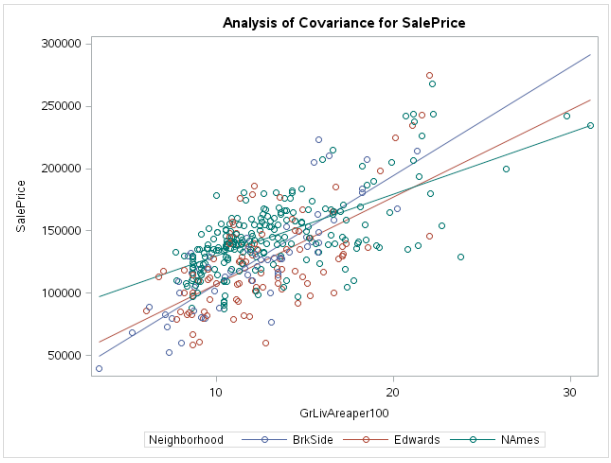
Addendum 5: Vanilla Regression | Model: Outliers Addressed*

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	80325.71230	B	5592.03832	14.36	<.0001
GrLivAreaper100	4956.12477	B	409.70671	12.10	<.0001
Neighborhood BrkSide	-60354.19850	B	12060.03479	-5.00	<.0001
Neighborhood Edwards	-43225.29073	B	10837.81644	-3.99	<.0001
Neighborhood NAmes	0.00000	B	.	.	.
GrLivArea*Neighborho BrkSide	3760.12849	B	940.21789	4.00	<.0001
GrLivArea*Neighborho Edwards	2059.71212	B	820.38610	2.51	0.0125
GrLivArea*Neighborho NAmes	0.00000	B	.	.	.

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



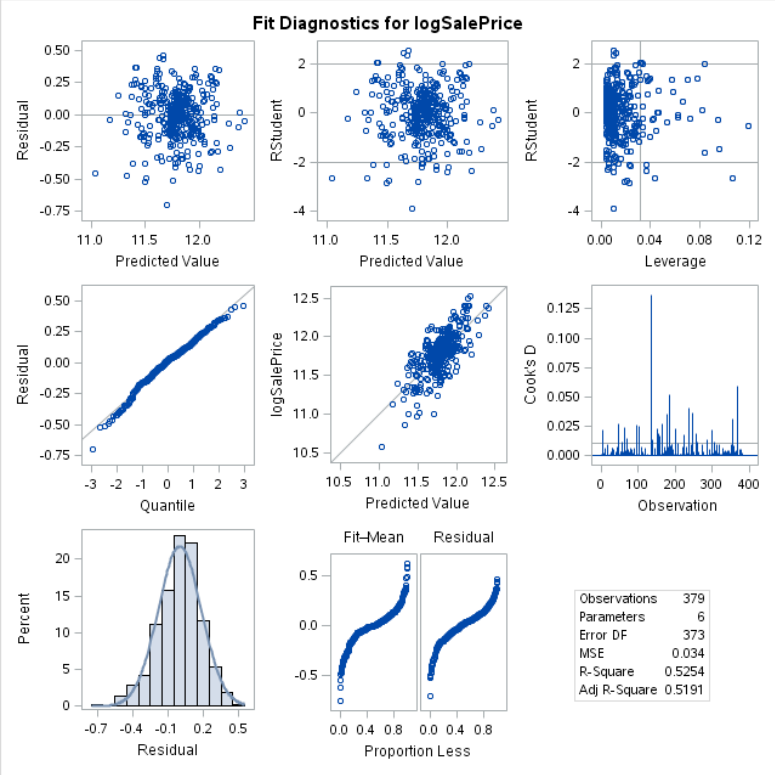


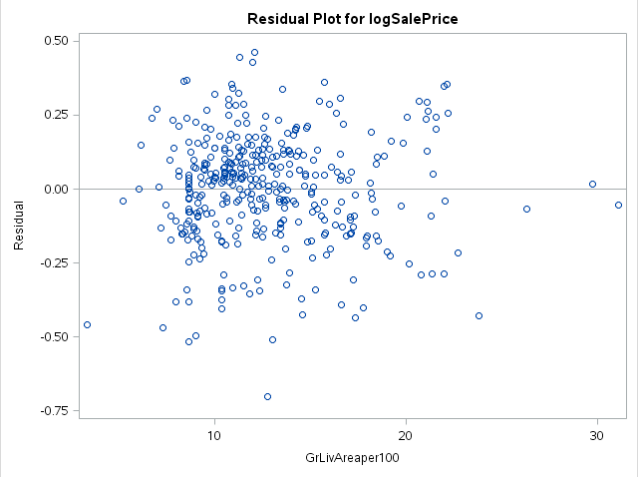


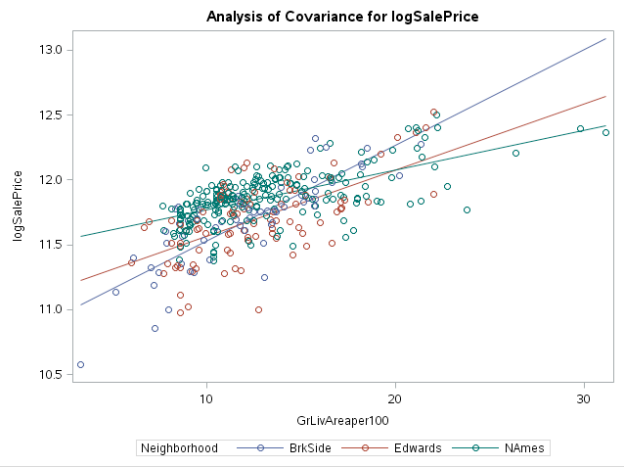
Addendum 6: Vanilla Regression | Model: Log-Linear (Outliers Addressed)

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	11.46308762	B	0.04175292	274.55	<.0001
GrLivAreaper100	0.03075050	B	0.00305907	10.05	<.0001
Neighborhood BrkSide	-0.67149365	B	0.09004618	-7.46	<.0001
Neighborhood Edwards	-0.41144266	B	0.08092049	-5.08	<.0001
Neighborhood NAmes	0.00000000	B	.	.	.
GrLivArea*Neighborho BrkSide	0.04307178	B	0.00702013	6.14	<.0001
GrLivArea*Neighborho Edwards	0.02043149	B	0.00612541	3.34	0.0009
GrLivArea*Neighborho NAmes	0.00000000	B	.	.	.

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.525413	1.563944	0.184431	11.79269



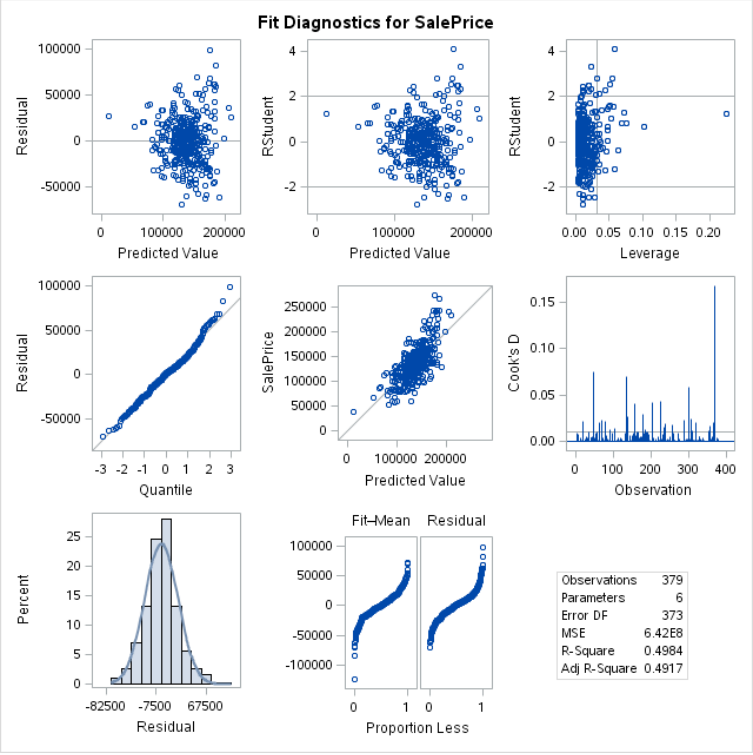


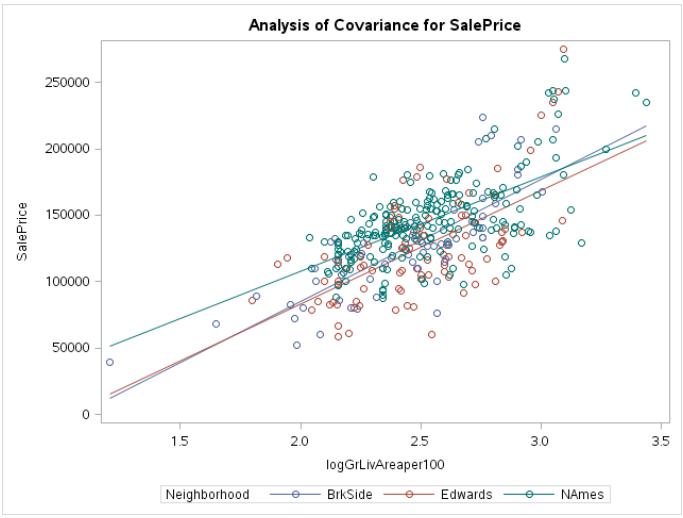


Addendum 7: Vanilla Regression | Model: Linear-Log (Outliers Addressed)

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-34567.92221	B	15470.76545	-2.23	0.0260
logGrLivAreaper100	71042.34430	B	6085.38391	11.67	<.0001
Neighborhood BrkSide	-63650.50684	B	27862.76048	-2.28	0.0229
Neighborhood Edwards	-52636.52333	B	28037.21033	-1.88	0.0612
Neighborhood NAmes	0.00000	B	.	.	.
logGrLivA*Neighborho BrkSide	20728.40689	B	11227.58949	1.85	0.0657
logGrLivA*Neighborho Edwards	14188.73230	B	11130.71521	1.27	0.2032
logGrLivA*Neighborho NAmes	0.00000	B	.	.	.

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.498378	18.50708	25327.93	136855.4





Addendum 8: Vanilla Regression | Model: Log-Log (Outliers Addressed)

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	10.72724642	B	0.11211841	95.68	<.0001
logGrLivAreaper100	0.44987850	B	0.04410147	10.20	<.0001
Neighborhood BrkSide	-1.03970690	B	0.20192461	-5.15	<.0001
Neighborhood Edwards	-0.62585626	B	0.20318887	-3.08	0.0022
Neighborhood NAmes	0.00000000	B	.	.	.
logGrLivA*Neighborho BrkSide	0.36976955	B	0.08136763	4.54	<.0001
logGrLivA*Neighborho Edwards	0.18931410	B	0.08066557	2.35	0.0195
logGrLivA*Neighborho NAmes	0.00000000	B	.	.	.

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.529914	1.556511	0.183554	11.79269