Century 21 Ames: Analyzing Sales Price Relationships in Different Neighborhoods and Predictive Modeling for Future Prices

Southern Methodist University
DS-6371 Statistical Foundations for Data Science
Pejal Rath & Kyle Kuberski
August 6, 2023

Introduction:

We are attempting to provide thorough analysis for Century 21 Ames by detailing two distinct analyses. These analyses are designed to provide valuable insights and aid in overall decision-making processes for the real estate company in Ames, Iowa.

Our first analysis centers on the relationship between house sales prices and square footage of living areas within specific neighborhoods. Furthermore, we seek to estimate the impact of the living areas on sales price considering the influence of each neighborhood. Our goal is to construct a model that accurately quantifies these relationships.

In our second analysis, we focus on developing the most predictive model for sales prices for homes across all neighborhoods in Ames, Iowa. To address this, we implement various cross-validation models as well as custom-built models for the purpose of identifying the most appropriate forecast. The forecast relates to future sales prices that may assist in guiding future strategic decisions for Century 21 Ames.

Data Description:

Our data comes from Kaggle (by Dean De Cock) and is a compilation of many housing statistics from various neighborhoods in Ames, Iowa. There are 1,460 observations across all neighborhoods, and these can be found from the linked website on Kaggle for more information. With respect to this analysis, the main variables used are as follows: GrLivArea, SalePrice, Neighborhood, OverallQual, TotRmsAbvGrd, BsmtExposure, PoolArea, KitchenQual, Fireplaces, GarageCars, BldgType, FullBath, BsmtFinType1, BsmtQual, and SaleCondition

Analysis 1: Relationship Between Home Sales Price and Square Footage in Ames, Iowa

Problem: The real estate company Century 21 Ames seeks to obtain a robust estimate of the relationship between sales prices of houses and the square footage of their living areas. Additionally, the company aims to investigate how this relationship varies across different neighborhoods where the houses are situated.

Approach: We will proceed with two competing models: our first model uses a log-log transformation of the dataset and includes the full scope. Our second model shares the log-log transformation, however, we will be narrowing our data's scope to only include square footage where GrLivArea >= 1000 and GrLivArea <= 3250 and SalePrice is SalePrice >= 75000 and SalePrice <= 150000. We will provide the fitted models and models broken out by neighborhood and here on name these models "Unrestricted Model" and "Restricted Model"

Unrestricted Model (Figure 2.17):

log_SalePrice = 5.9129 + 0.8196 * logGrLiv + 2.0935 * Edwards + 2.5798 * NAmes - 0.2999 * (logGrLiv * Edwards) - 0.3466 * (logGrLiv * NAmes)

Model by Neighborhood:

BrkSide (reference): Predicted SalePrice = 5.9129 + 0.8196 * logGrLiv

Edwards: Predicted SalePrice = 7.0064 + 0.8196 * logGrLiv + 2.0935

NAmes: Predicted SalePrice = 8.4927 + 0.8196 * logGrLiv + 2.5798

Restricted Model (Figure 2.17)**:**

log_SalePrice = 8.4195 + 0.4607 * logGrLiv + 0.7059 * Edwards + 2.2418 * NAmes - 0.1051 * (logGrLiv * Edwards) - 0.3050 * (logGrLiv * NAmes)

Model by Neighborhood:

BrkSide (reference): Predicted SalePrice = 8.4195 + 0.4607 * logGrLiv

Edwards: Predicted SalePrice =8.4195 + 0.4607 * logGrLiv + 0.7059 - 0.1051 * (logGrLiv * Edwards)

NAmes: Predicted SalePrice = 8.4195 + 0.4607 * logGrLiv + 2.2418 - 0.3050 * (logGrLiv * NAmes)

Assumptions (Figure 2.13-2.15)**:**

Unrestricted Model:

Normality: Judging from the scatter plot, Q-Q plot, and the histogram of residuals we see a normal distribution with some residual outliers; however, we will move forward with the assumption of normality.

Linear Trend: We see a positive linearity across our pairwise plots for each individual neighborhood.

Equal SD: There is some evidence of heteroscedasticity from our residual plots, where a few outliers are slightly out of range for comfort, however, we will proceed with caution and attempt competing models to confirm this assumption further.

Independence: We will assume the observations are independent.

Restricted Model:

Normality: For the restricted data, we see similar normality to our first model (unrestricted data). We will assume normality.

Linear Trend: We see a positive linearity across our pairwise plots for each individual neighborhood. The Edwards neighborhood shows an outlier; however, this is less influential than the original dataset.

Equal SD: There is some evidence of heteroscedasticity from our residual plots again, however, it appears to be less in strength (judging from our residual plots and DFBetas).

Independence: We will assume the observations are independent.

Comparing Competing Models (Figures 2.15/2.17) **:**

Stat	Unrestricted Model	Restricted Model
R^2	0.5121	0.3098
Adj R^2	0.5056	0.2965
CV Press	16.9386*	6.2902*

Parameters, Estimates, Interpretations, and Confidence Intervals:

Our model parameters and estimates (Figure 2.18) show overall significant interactions between each neighborhood.

The intercept estimate is 5.912920736. This represents the estimated baseline value of the log(SalePrice) when all other predictor variables are zero. In this case, when logGrLiv, logGrLiv*Neighborhood interaction terms, and the neighborhood indicators (Edwards and NAmes) are zero, the estimated log(SalePrice) is approximately 5.91. Our 95% confidence interval for the intercept ranges from approximately ($e^{4.92}$,$e^{6.91}$) = (137.002, 1002.25) where we can be 95% confident the true value of SalePrice falls within this range.

The estimate for logGrLivArea is 0.819648056. This indicates that for each one-unit increase in the natural logarithm of GrLivArea (living area square footage), the estimated log(SalePrice) is expected to increase by approximately 0.82 units, assuming all other variables are held constant. We can be 95% confident the true value GrLivArea lies within the range of ($e^{0.68}$,$e^{0.96}$) = (1.974, 2.612) where for each one-unit increase in GrLivArea, the estimated SalePrice is expected to increase by 1.974 to 2.612 units holding other variables constant.

The estimates for Edwards and NAmes neighborhoods are 2.093586444 and 2.579806905, respectively. These estimates represent the difference in the estimated log(SalePrice) between the Edwards and NAmes neighborhoods compared to the reference neighborhood (BrkSide). The log(SalePrice) is expected to be higher in the Edwards neighborhood by approximately 2.09 units and in the NAmes neighborhood by approximately 2.58 units compared to the BrkSide neighborhood. Our 95% confidence interval is approximately ($e^{0.82}$,$e^{3.36}$) = (2.27, 28.79) for the range of possible differences in log(SalePrice) between Edwards and NAmes compared to the reference neighborhood (BrkSide).

The estimates for the interaction terms logGrLivNeighborhood Edwards and log(GrLivArea)*Neighborhood NAmes are -0.299980812 and -0.346624454, respectively. These estimates indicate that the relationship between log(GrLivArea)and log(SalePrice) varies depending on the neighborhood. The negative values suggest that the effect of log(GrLivArea)on log(SalePrice) is less pronounced in the Edwards and NAmes neighborhoods compared to the BrkSide neighborhood. Our 95% confidence interval is between ($e^{-0.51}$,$e^{-0.18}$) = (0.6004, 0.835) for the effect of GrLivArea on SalePrice in the Edwards and NAmes neighborhoods.

Conclusion:

The analysis conducted between the unrestricted and restricted data involved exploring the relationship between the log-transformed SalePrice and the predictors, specifically log-transformed GrLivArea and the categorical variable Neighborhood (with interaction terms). In the unrestricted model, we found that logGrLiv had a significant positive effect on log(SalePrice), indicating that an increase in living area is associated with a higher sale price. Additionally, the neighborhoods Edwards and NAmes had significantly higher log(SalePrice) compared to the reference neighborhood BrkSide. When comparing the two models, the unrestricted model provided a better fit to the data, as evidenced by a higher adjusted R-squared value. However, the restricted model offered valuable insights into the relationship between logGrLiv and log(SalePrice) specific to each neighborhood. The interaction terms revealed that the effect of logGrLiv on log(SalePrice) varied depending on the neighborhood, with Edwards and NAmes showing a weaker relationship compared to BrkSide.

Shiny App displaying Living Area v Sales Price per these three neighborhoods!

Analysis 2: Predictive Models for Home Sale Prices in Ames, Iowa

The objective of this analysis is to develop predictive models for home sale prices in Ames, Iowa, utilizing techniques covered in the course. The analysis involves constructing four distinct models: forward selection, backward elimination, stepwise selection, and a custom model. To facilitate accurate predictions, data preprocessing involves addressing missing values by filling NA values in categorical columns with the most common category and NA values in continuous columns with the average value.

Model Selection

- **Forward Selection**

The forward selection method involves systematically adding variables based on their contribution to model performance. The model equation is provided:

SalePrice = OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + PoolArea + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath + BsmtFinType1 + BsmtQual + SaleCondition

This model achieved an adjusted R-squared of 0.875, CVPRESS of 1.4671E12, and AIC of 31458. Please see **Figure 2.3 Forward Selection**

- **Backward Elimination**

The backward elimination technique entails iteratively removing variables with minimal contribution to model performance. The model equation is provided:

SalePrice = OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + PoolArea + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath + BsmtFinType1 + BsmtQual

This model achieved an adjusted R-squared of 0.871, CVPRESS of 1.6316E12, and AIC of 31495. Please see **Figure 2.4 Backward Selection**

- **Stepwise Selection**

Stepwise selection combines elements of forward and backward techniques. The model equation is provided:

SalePrice = OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + PoolArea + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath + BsmtFinType1 + BsmtQual

This model achieved an adjusted R-squared of 0.871, CVPRESS of 1.5379E12, and AIC of 31495. Please see **Figure 2.5 Stepwise Selection**

- **Custom Model**

The custom model involves manual variable selection, yielding the following equation:

SalePrice = Neighborhood + TotRmsAbvGrd + BsmtExposure + KitchenQual + Fireplaces + GarageCars + FireplaceQu + OverallQual * FullBath

This model achieved an adjusted R-squared of 0.851, CVPRESS of 2.1535E12, and AIC of 31713. Please see **Figure 2.6 Custom Selection**

Comparing the competing models based on their adjusted R-squared, CVPRESS, and AIC scores, the forward selection model emerges as the most promising for predicting future sale prices of homes in Ames, Iowa. It exhibits the highest adjusted R-squared and relatively low CVPRESS and AIC scores, which are R-squared of 0.875, CVPRESS of 1.4671E12, and AIC of 31458, suggesting favorable predictive capabilities.

Model Assumption Check

Residual plots indicate reasonably normal distribution of residuals, suggesting adherence to the assumption of normality. However, the presence of high leverage points and notable Cook's D values suggests potential outliers, warranting further scrutiny and potential mitigation. Please see **Figure 2.1 Fit Diagnostics for SalePrice with High Leverage and Cook's D.**

In order to enhance the robustness and reliability of the predictive models, careful consideration of high leverage points was undertaken. High leverage points can significantly influence model outcomes and result in inflated values of metrics like Cook's D and leverage. Therefore, a crucial step in refining the models involved the identification and removal of these high leverage points. Please see **Figure 2.2 Fit Diagnostics for SalePrice with Low Leverage and Cook's D.**

Following the removal of high leverage points, the models were re-evaluated to assess their performance under the new conditions. The selected variables were updated for each type of selection method, yielding the following model specifications:

- **Forward Selection** (After Removing High Leverage Points)

SalePrice = OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath + BsmtQual + BsmtFullBath + CentralAir + HalfBath + FireplaceQu + GarageFinish

This model achieved Adjusted R-squared of 0.868, CVPRESS of 6.2289E11 and AIC: 23059. Please see **Figure 2.7 Forward Selection without High Leverage**

- **Backward Elimination** (After Removing High Leverage Points)

SalePrice = OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath + BsmtQual + BsmtFinType1 + SaleCondition

This model achieved Adjusted R-squared of 0.859, CVPRESS of 6.5093E11 and AIC of 23120. Please see **Figure 2.8 Backward Selection without High Leverage**

- **Stepwise Selection** (After Removing High Leverage Points):

SalePrice = OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath + BsmtQual + BsmtFullBath + CentralAir + HalfBath + FireplaceQu + GarageFinish

This model achieved Adjusted R-squared of 0.868, CVPRESS of 6.2938E11, and AIC of 23059. Please see **Figure 2.9 Stepwise Selection without High Leverage**

- **Custom Model** (After Removing High Leverage Points):

SalePrice = OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath + BsmtQual + BsmtFullBath + CentralAir + HalfBath + FireplaceQu + GarageFinish

This model achieved Adjusted R-squared of 0.866, CVPRESS of 6.3086E11, and AIC of 23074. Please see **Figure 2.10 Custom Selection without High Leverage**

Comparative Analysis

Upon revisiting the models after addressing the issue of high leverage points, it is evident that the models' performance remains largely consistent with their earlier counterparts. The adjusted R-squared values continue to reflect the models' goodness of fit, while the CVPRESS scores provide insight into their prediction accuracy. Considering the AIC values, which capture the models' complexity, it is apparent that the forward selection model maintains a balance between fit and complexity, resulting in a favorable predictive performance.

After submitting the models for public score evaluation, the following results were obtained:

- Forward selection achieved a public score of 0.16482.
- Forward selection, after removing high leverage points, achieved a slightly higher public score of 0.16681.
- Backward selection achieved a public score of 0.16861.
- Backward selection, after removing high leverage points, achieved a lower public score of 0.16729.
- Stepwise selection achieved a public score of 0.16861.
- Stepwise selection, after removing high leverage points, achieved a slightly lower public score of 0.16681.
- Custom selection achieved a higher public score of 0.18012.
- Custom selection, after removing high leverage points, achieved the same lower public score of 0.16681.

Model Selection with High Leverage Point

Predictive Models	Adjusted R2	CV PRESS	Kaggle Score
Forward	0.875	1.4671 E12	0.16482
Backward	0.871	1.6316 E12	0.16861
Stepwise	0.871	1.5379 E12	0.16861
CUSTOM	0.851	2.1535 E12	0.18012

Model Selection without High Leverage Point

Predictive Models	Adjusted R2	CV PRESS	Kaggle Score
Forward	0.875	1.4671 E12	0.16681
Backward	0.871	1.6316 E12	0.16729
Stepwise	0.871	1.5379 E12	0.16681
CUSTOM	0.851	2.1535 E12	0.16681

Considering that a lower public score indicates better predictive performance, the forward selection model without high leverage points adjustment emerges as the most promising choice among the tested models. It achieves the lowest public score, signifying superior predictive accuracy while maintaining an optimal balance between model fit and complexity. This outcome aligns with our earlier assessment and underscores the forward selection model's robustness and suitability for predicting future sale prices of homes in Ames, Iowa.

Conclusion

Upon comprehensive evaluation of the models with high leverage points addressed, the forward selection model emerges as the most suitable choice for predicting future sale prices of homes in Ames, Iowa. This model effectively combines an elevated adjusted R-squared, CVPRESS, and a competitive AIC score. However, it is crucial to acknowledge that the existence of outliers and high leverage points can impact model robustness. Therefore, further diagnostic analysis and sensitivity tests are recommended to validate the selected model's robustness and applicability.

Appendix

Github Page Links:

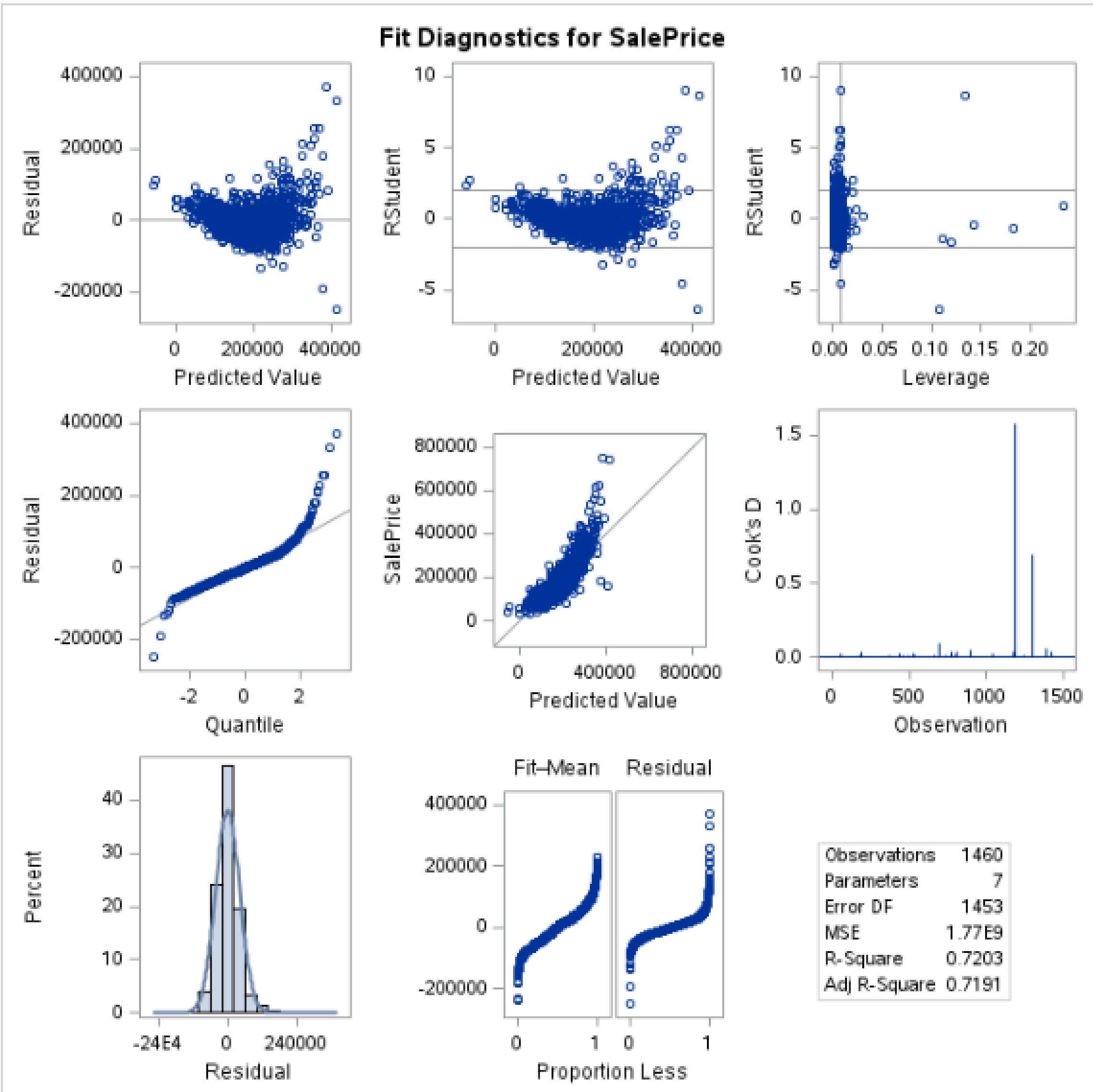
Github Repo (Code, datasets, report)

Kyle Kuberski: Github.io

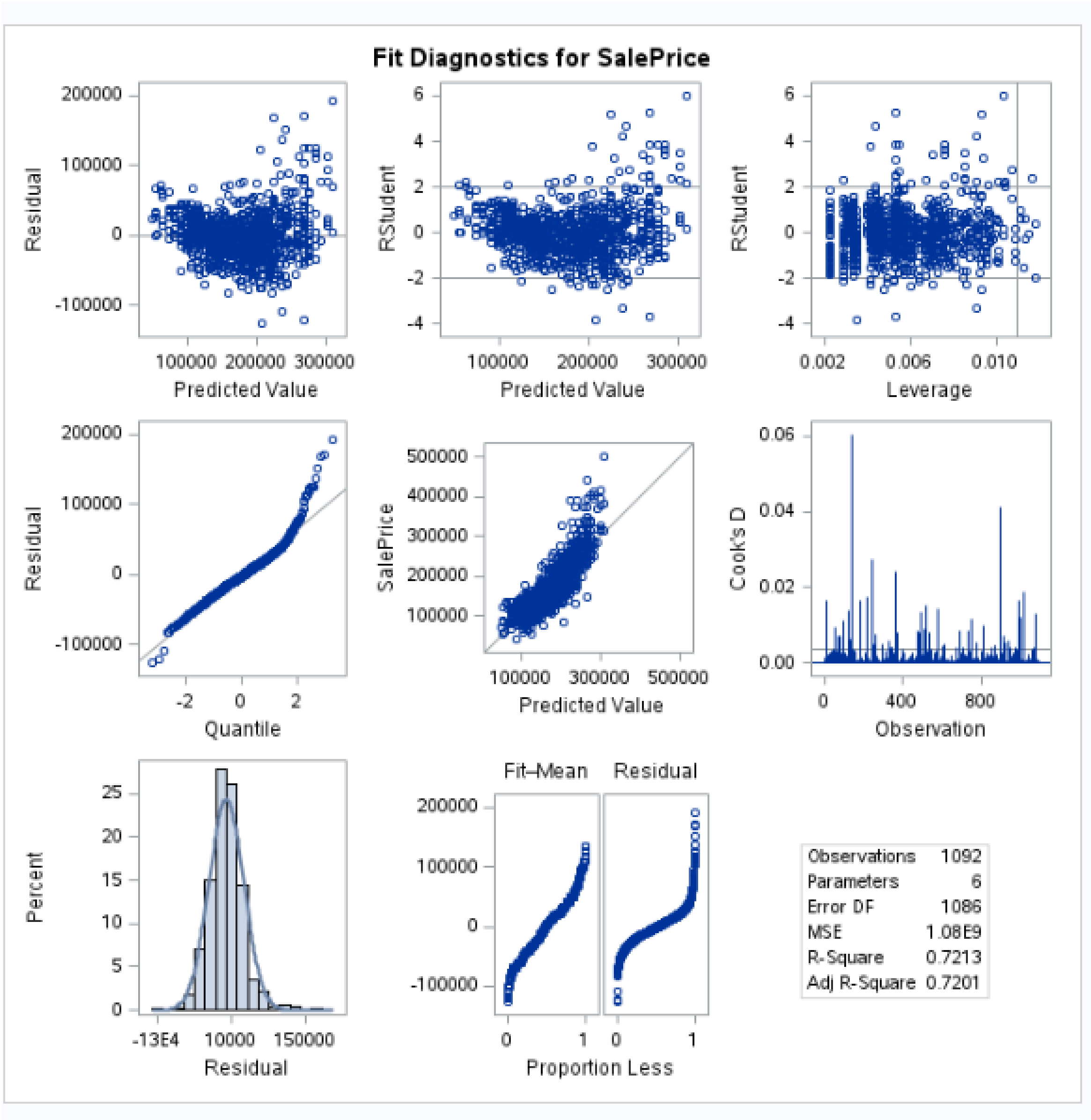
Pejal Rath

Figures:

- Figure 2.1 Fit Diagnostics for SalePrice with High Leverage and Cook's D



- Figure 2.2 Fit Diagnostics for SalePrice with Low Leverage and Cook's D



- Figure 2.3 Forward Selection

The GLMSELECT Procedure

Data Set	WORK.MYDATA
Dependent Variable	SalePrice
Selection Method	Forward
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	694623671

Number of Observations Read	1460
Number of Observations Used	1460

Dimensions	
Number of Effects	74
Number of Parameters	7598

The GLMSELECT Procedure

Forward Selection Summary						
Step	Effect Entered	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0	Intercept	1	1	0.0000	32952.0292	9.21859E12
1	OverallQual	2	10	0.6822	31334.8280	2.99646E12
2	Neighborhood	3	34	0.7497	31136.8041	2.37274E12
3	TotRmsAbvGrd	4	45	0.7843	30988.6375	2.11564E12
4	BsmtExposure	5	48	0.8007	30891.6431	1.95479E12
5	PoolArea	6	55	0.8177	30805.7461	1.9511E12
6	KitchenQual	7	58	0.8281	30738.3682	1.85461E12
7	Fireplaces	8	61	0.8367	30682.0167	1.79159E12
8	GarageCars	9	65	0.8455	30625.7547	1.69506E12
9	BldgType	10	69	0.8522	30586.0097	1.63659E12
10	FullBath	11	72	0.8591	30535.7083	1.57127E12
11	BsmtFinType1	12	77	0.8663	30489.5141	1.5124E12
12	BsmtQual	13	80	0.8710	30456.2403	1.47823E12
13	SaleCondition	14	85	0.8747*	30445.3285*	1.46714E12*
* Optimal Value of Criterion						

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details				
Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	Condition2	1.53803E12	>	1.46714E12

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 13).

Effects:	Intercept Neighborhood BldgType OverallQual BsmtQual BsmtExposure BsmtFinType1 FullBath KitchenQual TotRmsAbvGrd Fireplaces GarageCars PoolArea SaleCondition

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	84	8.120349E12	96670815744	122.22
Error	1375	1.087563E12	790954772	
Corrected Total	1459	9.207911E12		

Root MSE	28124
Dependent Mean	180921
R-Square	0.8819
Adj R-Sq	0.8747
AIC	31458
AICC	31469
SBC	30445
CV PRESS	1.467144E12

- Figure 2.4 Backward Selection

Dimensions	
Number of Effects	15
Number of Parameters	108

The GLMSELECT Procedure

Data Set	WORK.MYDATA
Dependent Variable	SalePrice
Selection Method	Backward
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	556753002

Number of Observations Read	1460
Number of Observations Used	1460

The GLMSELECT Procedure

Backward Selection Summary

Step	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0		15	92	0.8798*	30427.5477*	1.69853E12
1	Condition2	14	85	0.8747	30445.3285	1.65872E12
2	SaleCondition	13	80	0.8710	30456.2403	1.63165E12*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Removal	BsmtQual	1.67644E12	>	1.63165E12

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 2).

Effects:	Intercept OverallQual Neighborhood TotRmsAbvGrd BsmtExposure PoolArea KitchenQual Fireplaces GarageCars BldgType FullBath BsmtFinType1 BsmtQual

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	79	8.084505E12	1.023355E11	125.71
Error	1380	1.123407E12	814062833	
Corrected Total	1459	9.207911E12		

Root MSE	28532
Dependent Mean	180921
R-Square	0.8780
Adj R-Sq	0.8710
AIC	31495
AICC	31505
SBC	30456
CV PRESS	1.631646E12

- Figure 2.5 Stepwise Selection

Dimensions	
Number of Effects	74
Number of Parameters	7598

The GLMSELECT Procedure

Stepwise Selection Summary							
Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0	Intercept		1	1	0.0000	32952.0292	9.2177E12
1	OverallQual		2	10	0.6822	31334.8280	2.98663E12
2	Neighborhood		3	34	0.7497	31136.8041	2.41751E12
3	TotRmsAbvGrd		4	45	0.7843	30988.6375	2.18174E12
4	BsmtExposure		5	48	0.8007	30891.6431	2.02077E12
5	PoolArea		6	55	0.8177	30805.7461	2.01841E12
6	KitchenQual		7	58	0.8281	30738.3682	1.91639E12
7	Fireplaces		8	61	0.8367	30682.0167	1.84953E12
8	GarageCars		9	65	0.8455	30625.7547	1.75832E12
9	BldgType		10	69	0.8522	30586.0097	1.70964E12
10	FullBath		11	72	0.8591	30535.7083	1.62452E12
11	BsmtFinType1		12	77	0.8663	30489.5141	1.57691E12
12	BsmtQual		13	80	0.8710*	30456.2403*	1.53796E12*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details				
Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	SaleCondition	1.55258E12	>	1.53796E12
Removal	BsmtQual	1.57691E12	>	1.53796E12

The GLMSELECT Procedure

Data Set	WORK.MYDATA
Dependent Variable	SalePrice
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	955180600

Number of Observations Read	1460
Number of Observations Used	1460

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 12).

Effects:	Intercept Neighborhood BldgType OverallQual BsmtQual BsmtExposure BsmtFinType1 FullBath KitchenQual TotRmsAbvGrd Fireplaces GarageCars PoolArea

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	79	8.084505E12	1.023355E11	125.71
Error	1380	1.123407E12	814062833	
Corrected Total	1459	9.207911E12		

Root MSE	28532
Dependent Mean	180921
R-Square	0.8780
Adj R-Sq	0.8710
AIC	31495
AICC	31505
SBC	30456
CV PRESS	1.537963E12

- Figure 2.6 Custom Selection

The GLMSELECT Procedure

Data Set	WORK.MYDATA
Dependent Variable	SalePrice
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	694623671

Number of Observations Read	1460
Number of Observations Used	1460

The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0	Intercept		1	1	0.0000	32952.0292	9.21859E12
1	OverallQual*FullBath		2	30	0.7231	31259.2484	4.18263E12
2	Neighborhood		3	54	0.7793	31078.3577	3.33987E12
3	GarageCars		4	58	0.8001	30958.8276	2.96575E12
4	Fireplaces		5	61	0.8155	30860.2172	2.76229E12
5	KitchenQual		6	64	0.8247	30804.2998	2.55191E12
6	BsmtExposure		7	67	0.8324	30757.1673	2.41384E12
7	TotRmsAbvGrd		8	77	0.8453	30702.8711	2.2182E12
8	BldgType		9	81	0.8514*	30668.8503*	2.1536E12*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	BsmtFullBath	2.21477E12	>	2.1536E12
Removal	BldgType	2.2182E12	>	2.1536E12

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 8).

Effects:	Intercept Neighborhood TotRmsAbvGrd BsmtExposure KitchenQual Fireplaces GarageCars OverallQual*FullBath BldgType

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	80	7.914868E12	98935851769	105.51
Error	1379	1.293043E12	937667290	
Corrected Total	1459	9.207911E12		

Root MSE	30621
Dependent Mean	180921
R-Square	0.8596
Adj R-Sq	0.8514
AIC	31703
AICC	31713
SBC	30669
CV PRESS	2.153598E12

- Figure 2.7 Forward Selection without High Leverage

The GLMSELECT Procedure

Data Set	WORK.MYDATA
Dependent Variable	SalePrice
Selection Method	Forward
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	694623671

Number of Observations Read	1093
Number of Observations Used	1092

The GLMSELECT Procedure

Forward Selection Summary

Step	Effect Entered	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0	Intercept	1	1	0.0000	24112.9934	4.23314E12
1	OverallQual	2	7	0.6610	22967.6337	1.45048E12
2	GarageCars	3	10	0.7130	22803.6577	1.2366E12
3	Neighborhood	4	34	0.7706	22702.5982	1.01448E12
4	TotRmsAbvGrd	5	41	0.7978	22606.3763	9.01317E11
5	BsmtFullBath	6	43	0.8198	22492.7655	8.08546E11
6	KitchenQual	7	46	0.8289	22454.1586	7.70308E11
7	BsmtQual	8	49	0.8365	22422.3554	7.40739E11
8	Fireplaces	9	51	0.8432	22388.1336	7.12199E11
9	BsmtExposure	10	54	0.8484	22369.0356	6.97577E11
10	BldgType	11	58	0.8550	22344.2800	6.72097E11
11	FullBath	12	59	0.8588	22321.7252	6.56995E11
12	CentralAir	13	60	0.8607	22312.4845	6.45445E11
13	HalfBath	14	62	0.8629	22306.7412	6.34829E11
14	FireplaceQu	15	66	0.8660	22305.9604	6.28477E11
15	GarageFinish	16	68	0.8676*	22304.8507*	6.22889E11*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	ExterQual	6.26958E11	>	6.22889E11

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 15).

Effects:	Intercept Neighborhood BldgType OverallQual BsmtQual BsmtExposure CentralAir BsmtFullBath FullBath HalfBath KitchenQual TotRmsAbvGrd Fireplaces FireplaceQu GarageFinish GarageCars

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	67	3.69544E12	55155826332	107.67
Error	1024	5.245569E11	512262644	
Corrected Total	1091	4.219997E12		

Root MSE	22633
Dependent Mean	174761
R-Square	0.8757
Adj R-Sq	0.8676
AIC	23059
AICC	23069
SBC	22305
CV PRESS	6.228889E11

- Figure 2.8 Backward Selection without High Leverage

The GLMSELECT Procedure

Data Set	WORK.MYDATA
Dependent Variable	SalePrice
Selection Method	Backward
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	556753002

Number of Observations Read	1093
Number of Observations Used	1092

The GLMSELECT Procedure

Backward Selection Summary

Step	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0		15	70	0.8597	22379.9678	6.53937E11
	PoolArea	14	70	0.8597	22379.9678	6.53937E11
1	Condition2	13	67	0.8598*	22380.9872*	6.50939E11*

* Optimal Value of Criterion

Note: Effects dropped at step 0 are redundant.

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Removal	SaleCondition	6.62595E11	>	6.50939E11

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 1).

Effects:	Intercept OverallQual Neighborhood TotRmsAbvGrd BsmtExposure KitchenQual Fireplaces GarageCars BldgType FullBath BsmtFinType1 BsmtQual SaleCondition

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	66	3.66422E12	55518487280	102.39
Error	1025	5.557772E11	542221611	
Corrected Total	1091	4.219997E12		

Root MSE	23286
Dependent Mean	174761
R-Square	0.8683
Adj R-Sq	0.8598
AIC	23120
AICC	23129
SBC	22381
CV PRESS	6.50939E11

- Figure 2.9 Stepwise Selection without High Leverage

The GLMSELECT Procedure

Data Set	WORK.MYDATA
Dependent Variable	SalePrice
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	955180600

Number of Observations Read	1093
Number of Observations Used	1092

The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0	Intercept		1	1	0.0000	24112.9834	4.22871E12
1	OverallQual		2	7	0.6610	22967.6337	1.47323E12
2	GarageCars		3	10	0.7130	22803.6577	1.26391E12
3	Neighborhood		4	34	0.7706	22702.5982	1.04466E12
4	TotRmsAbvGrd		5	41	0.7978	22606.3763	9.25249E11
5	BsmtFullBath		6	43	0.8198	22492.7655	8.32141E11
6	KitchenQual		7	46	0.8289	22454.1586	7.89811E11
7	BsmtQual		8	49	0.8365	22422.3554	7.58716E11
8	Fireplaces		9	51	0.8432	22388.1336	7.30626E11
9	BsmtExposure		10	54	0.8484	22369.0356	7.0127E11
10	BldgType		11	58	0.8550	22344.2800	6.73014E11
11	FullBath		12	59	0.8588	22321.7252	6.59294E11
12	CentralAir		13	60	0.8607	22312.4845	6.50694E11
13	HalfBath		14	62	0.8629	22306.7412	6.42175E11
14	FireplaceQu		15	66	0.8660	22305.9604	6.34681E11
15	GarageFinish		16	68	0.8676*	22304.8507*	6.29378E11*

* Optimal Value of Criterion

Selection stopped as adding or dropping any effect does not improve the selection criterion.

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 15).

Effects:	Intercept Neighborhood BldgType OverallQual BsmtQual BsmtExposure CentralAir BsmtFullBath FullBath HalfBath KitchenQual TotRmsAbvGrd Fireplaces FireplaceQu GarageFinish GarageCars

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	67	3.69544E12	55155826332	107.67
Error	1024	5.245569E11	512262644	
Corrected Total	1091	4.219997E12		

Root MSE	22633
Dependent Mean	174761
R-Square	0.8757
Adj R-Sq	0.8676
AIC	23059
AICC	23069
SBC	22305
CV PRESS	6.293784E11

- Figure 2.10 Custom Selection without High Leverage

The GLMSELECT Procedure

Data Set	WORK.MYDATA
Dependent Variable	SalePrice
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	694623671

Number of Observations Read	1093
Number of Observations Used	1092

The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0	Intercept		1	1	0.0000	24112.9934	4.23314E12
1	OverallQual*FullBath		2	10	0.6918	22881.4463	1.32166E12
2	Neighborhood		3	34	0.7580	22760.9760	1.06084E12
3	BsmtFullBath		4	36	0.7810	22664.0034	9.63665E11
4	BldgType*HalfBath		5	45	0.8097	22563.8254	8.53429E11
5	GarageCars		6	48	0.8272	22476.5957	7.77442E11
6	Fireplaces		7	50	0.8338	22445.8298	7.4794E11
7	KitchenQual		8	53	0.8401	22421.8473	7.20935E11
8	BsmtExposure		9	56	0.8471	22390.8217	7.02332E11
9	TotRmsAbvGrd		10	63	0.8553	22371.7246	6.69557E11
10	BsmtQual		11	66	0.8608	22347.5586	6.48749E11
11	CentralAir		12	67	0.8629	22336.6547	6.37249E11
12	FireplaceQu		13	71	0.8661*	22334.9944*	6.30866E11*

* Optimal Value of Criterion

Selection stopped as adding or dropping any effect does not improve the selection criterion.

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 12).

Effects:	Intercept Neighborhood TotRmsAbvGrd BsmtExposure KitchenQual Fireplaces GarageCars BsmtQual BsmtFullBath CentralAir FireplaceQu OverallQual*FullBath BldgType*HalfBath

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	70	3.691023E12	52728906666	101.77
Error	1021	5.289738E11	518093874	
Corrected Total	1091	4.219997E12		

Root MSE	22762
Dependent Mean	174761
R-Square	0.8747
Adj R-Sq	0.8661
AIC	23074
AICC	23085
SBC	22335
CV PRESS	6.308657E11

- Figure 2.11 R-Code used to predict the SalePrice

```r
#Predict Formula and Write to CSV Files
#Forward 1
fit_forward1=lm(SalePrice ~ OverallQual+Neighborhood+TotRmsAbvGrd+BsmtExposure+PoolArea+KitchenQual+ Fireplaces+ GarageCars+ BldgType+ FullBath+ BsmtFinType1+
BsmtQual+ SaleCondition, data = df_subset)
predicted_sale_price_forward1 <- predict(fit_forward1, newdata = df_test)
df_f1 <- data.frame(Id = df_test$Id, SalePrice = predicted_sale_price_forward1)
write.csv(df_f1, file = "C:/Users/pejal/OneDrive/Desktop/SMU/Classes/Summer 2023/Stat/project/forward.csv", row.names = FALSE)

#Forward 1 RH
fit_forward1_RH=lm(SalePrice ~ OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath + BsmtQual +
BsmtFullBath + CentralAir + HalfBath + FireplaceQu + GarageFinish, data = data_filtered)
predicted_sale_price_forward1_RH <- predict(fit_forward1_RH, newdata = df_test)
df_f1_RH <- data.frame(Id = df_test$Id, SalePrice = predicted_sale_price_forward1_RH)
write.csv(df_f1_RH, file = "C:/Users/pejal/OneDrive/Desktop/SMU/Classes/Summer 2023/Stat/project/forward_RH.csv", row.names = FALSE)

#Backward 2
fit_backward1=lm(SalePrice ~OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + PoolArea + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath +
BsmtFinType1 + BsmtQual, data = df_subset)
predicted_sale_price_backward1 <- predict(fit_backward1, newdata = df_test)
df_b1 <- data.frame(Id = df_test$Id, SalePrice = predicted_sale_price_backward1)
write.csv(df_b1, file = "C:/Users/pejal/OneDrive/Desktop/SMU/Classes/Summer 2023/Stat/project/backward.csv", row.names = FALSE)

#Backward 2 RH
fit_backward1_RH=lm(SalePrice ~ OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath + BsmtQual
+ BsmtFinType1 + SaleCondition, data = data_filtered)
predicted_sale_price_backward1_RH <- predict(fit_backward1_RH, newdata = df_test)
df_b1_RH <- data.frame(Id = df_test$Id, SalePrice = predicted_sale_price_backward1_RH)
write.csv(df_b1_RH, file = "C:/Users/pejal/OneDrive/Desktop/SMU/Classes/Summer 2023/Stat/project/backward_RH.csv", row.names = FALSE)

#Stepwise 3
fit_stepwise1=lm(SalePrice ~ OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + PoolArea + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath +
BsmtFinType1 + BsmtQual, data = df_subset)
predicted_sale_price_stepwise1 <- predict(fit_stepwise1, newdata = df_test)
df_s1 <- data.frame(Id = df_test$Id, SalePrice = predicted_sale_price_stepwise1)
write.csv(df_s1, file = "C:/Users/pejal/OneDrive/Desktop/SMU/Classes/Summer 2023/Stat/project/stepwise.csv", row.names = FALSE)

#Stepwise 3 RH
fit_stepwise1_RH=lm(SalePrice ~ OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath + BsmtQual
+ BsmtFullBath + CentralAir + HalfBath + FireplaceQu + GarageFinish, data = data_filtered)
predicted_sale_price_stepwise1_RH <- predict(fit_stepwise1_RH, newdata = df_test)
df_s1_RH <- data.frame(Id = df_test$Id, SalePrice = predicted_sale_price_stepwise1_RH)
write.csv(df_s1_RH, file = "C:/Users/pejal/OneDrive/Desktop/SMU/Classes/Summer 2023/Stat/project/stepwise_RH.csv", row.names = FALSE)

#Custom 4

fit_1=lm(SalePrice ~ Neighborhood + TotRmsAbvGrd + BsmtExposure + KitchenQual + Fireplaces + GarageCars + FireplaceQu + OverallQual * FullBath, data = df_subset)
predicted_sale_price_1 <- predict(fit_1, newdata = df_test)
df_c1 <- data.frame(Id = df_test$Id, SalePrice = predicted_sale_price_1)
write.csv(df_c1, file = "C:/Users/pejal/OneDrive/Desktop/SMU/Classes/Summer 2023/Stat/project/custom.csv", row.names = FALSE)

#Custom 4 RH
fit_1_RH=lm(SalePrice ~ OverallQual + Neighborhood + TotRmsAbvGrd + BsmtExposure + KitchenQual + Fireplaces + GarageCars + BldgType + FullBath + BsmtQual +
BsmtFullBath + CentralAir + HalfBath + FireplaceQu + GarageFinish, data = data_filtered)
predicted_sale_price_1_RH <- predict(fit_1_RH, newdata = df_test)
df_c1_RH <- data.frame(Id = df_test$Id, SalePrice = predicted_sale_price_1_RH)
write.csv(df_c1_RH, file = "C:/Users/pejal/OneDrive/Desktop/SMU/Classes/Summer 2023/Stat/project/custom_RH.csv", row.names = FALSE)
```

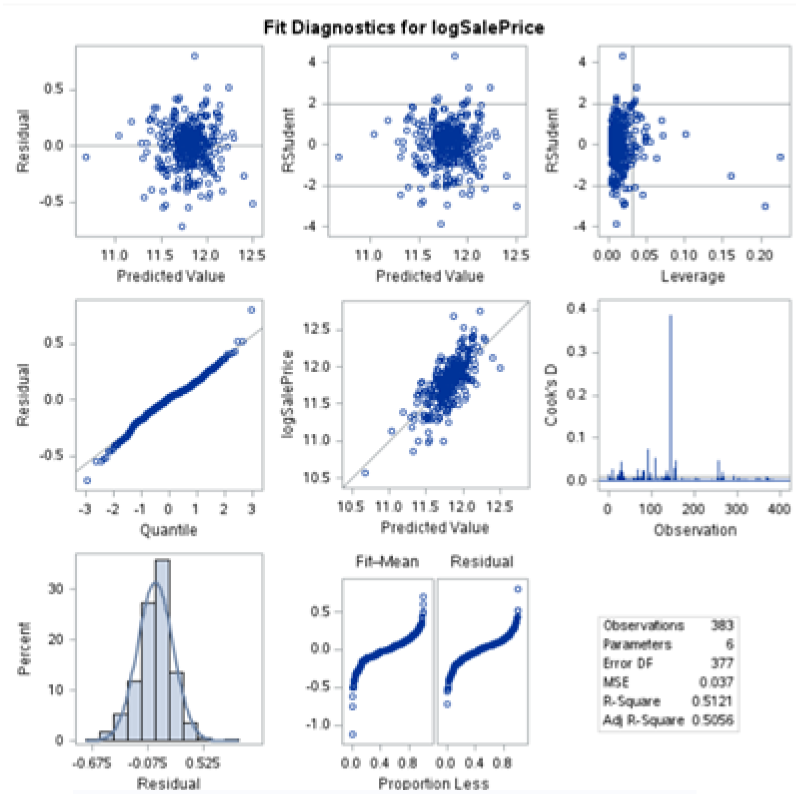
- Figure 2.12 SAS-Code used to Select Variables

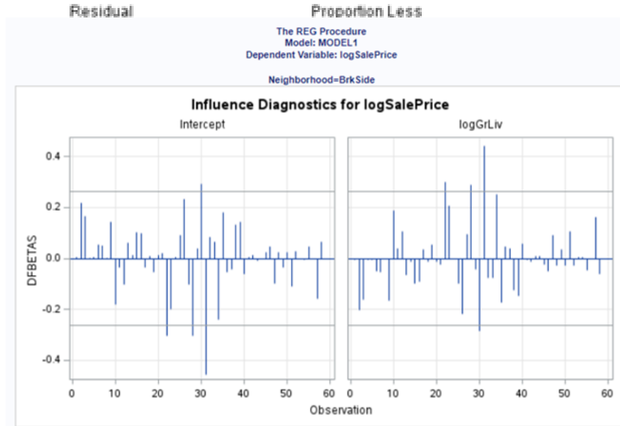
```
/*Forward*/
proc glmselect data=mydata seed=694623671;
    class MSSubClass MSZoning LotFrontage LotArea Street LotShape LandContour
        LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle
        OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st
        Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation BsmtQual
        BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2
        BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical X1stFlrSF
        X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
        BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces
        FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars GarageArea
        GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch
        ScreenPorch PoolArea MiscVal MoSold YrSold SaleType SaleCondition;
    model SalePrice=MSSubClass MSZoning LotFrontage LotArea Street LotShape
        LandContour LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
        HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
        Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
        BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
        BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
        X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
        HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
        Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
        GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
        EnclosedPorch ScreenPorch PoolArea MiscVal MoSold YrSold SaleType
        SaleCondition / selection=Forward(stop=CV) cvmethod=random(5) stats=adjrsq;
run;
```

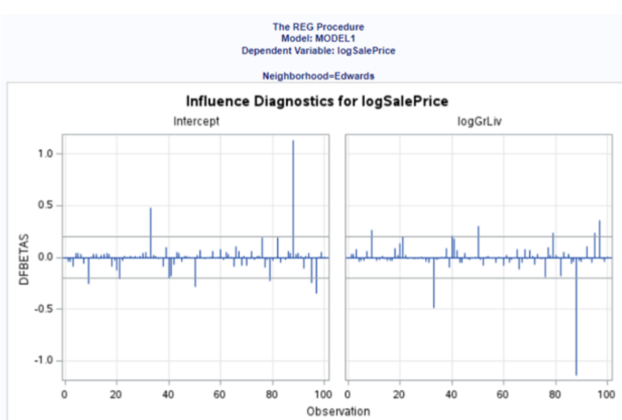
```
/*Backward*/
proc glmselect data=mydata seed=556753002;
    class MSSubClass MSZoning LotFrontage LotArea Street LotShape LandContour
        LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle
        OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st
        Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation BsmtQual
        BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2
        BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical X1stFlrSF
        X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
        BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces
        FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars GarageArea
        GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch
        ScreenPorch PoolArea MiscVal MoSold YrSold SaleType SaleCondition;
    model SalePrice=OverallQual Neighborhood TotRmsAbvGrd BsmtExposure PoolArea
        KitchenQual Fireplaces GarageCars BldgType FullBath BsmtFinType1 BsmtQual
        SaleCondition Condition2/ selection=Backward(stop=CV) cvmethod=random(5)
        stats=adjrsq;
run;
```

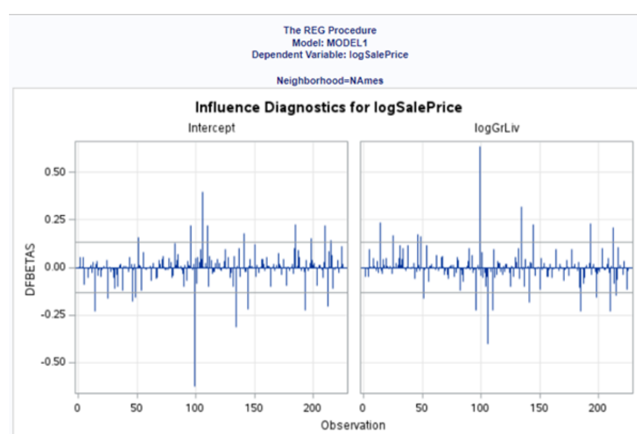
```
/*Stepwise*/
proc glmselect data=mydata seed=955180600;
   class MSSubClass MSZoning LotFrontage LotArea Street LotShape LandContour
      LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle
      OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st
      Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation BsmtQual
      BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2
      BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical X1stFlrSF
      X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
      BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces
      FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars GarageArea
      GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch
      ScreenPorch PoolArea MiscVal MoSold YrSold SaleType SaleCondition;
   model SalePrice=MSSubClass MSZoning LotFrontage LotArea Street LotShape
      LandContour LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
      HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
      Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
      BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
      BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
      X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
      HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
      Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
      GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
      EnclosedPorch ScreenPorch PoolArea MiscVal MoSold YrSold SaleType
      SaleCondition / selection=Stepwise(stop=CV) cvmethod=random(5) stats=adjrsq;
run;

/*Custome*/
proc glmselect data=mydata seed=694623671;
   class OverallQual Neighborhood TotRmsAbvGrd BsmtExposure KitchenQual
      Fireplaces GarageCars BldgType FullBath BsmtQual BsmtFullBath CentralAir
      HalfBath FireplaceQu GarageFinish;
   model SalePrice=FireplaceQu BsmtQual BsmtFullBath CentralAir BsmtExposure
      Fireplaces GarageCars OverallQual | Neighborhood | TotRmsAbvGrd | KitchenQual | BldgType | FullBath | HalfBath | GarageFinish
      / selection=Forward(stop=CV) cvmethod=random(5) stats=adjrsq;
run;
```

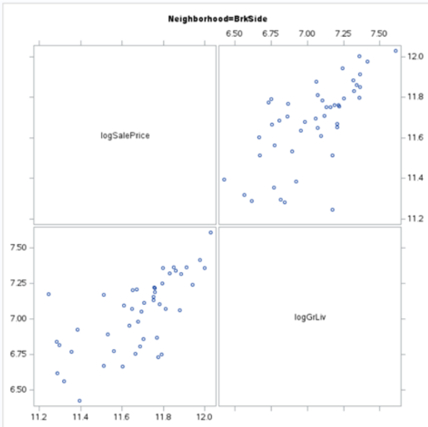
- Figure 2.13 Assumption Plots, Residuals, Influential plots, and Leverage for Unrestricted data

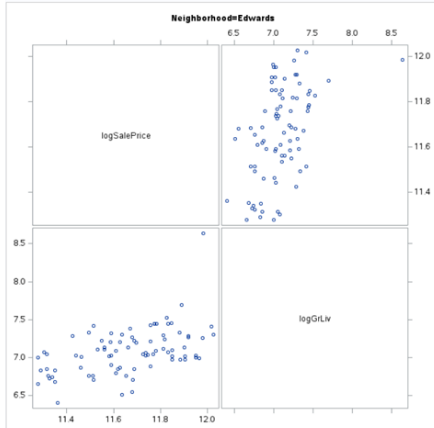


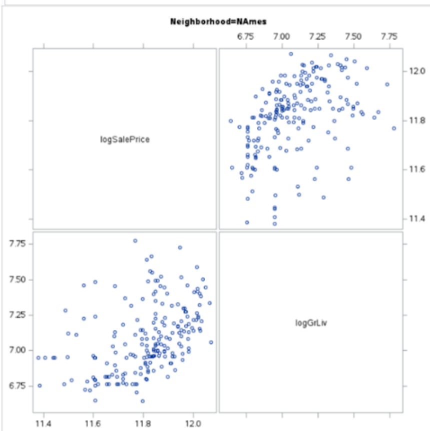


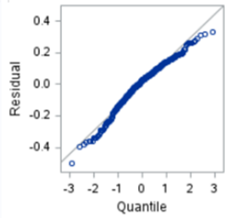


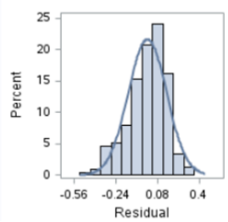


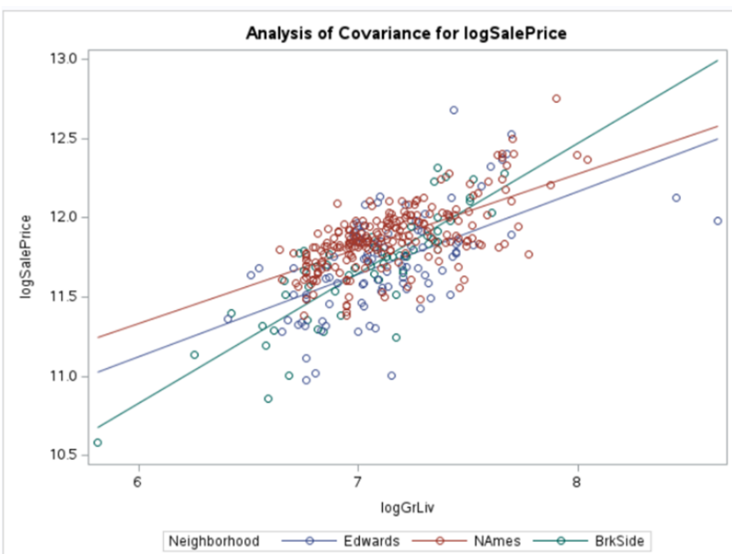

Neighborhood=BrkSide
logSalePrice
logGrLiv


Neighborhood=Edwards
logSalePrice
logGrLiv

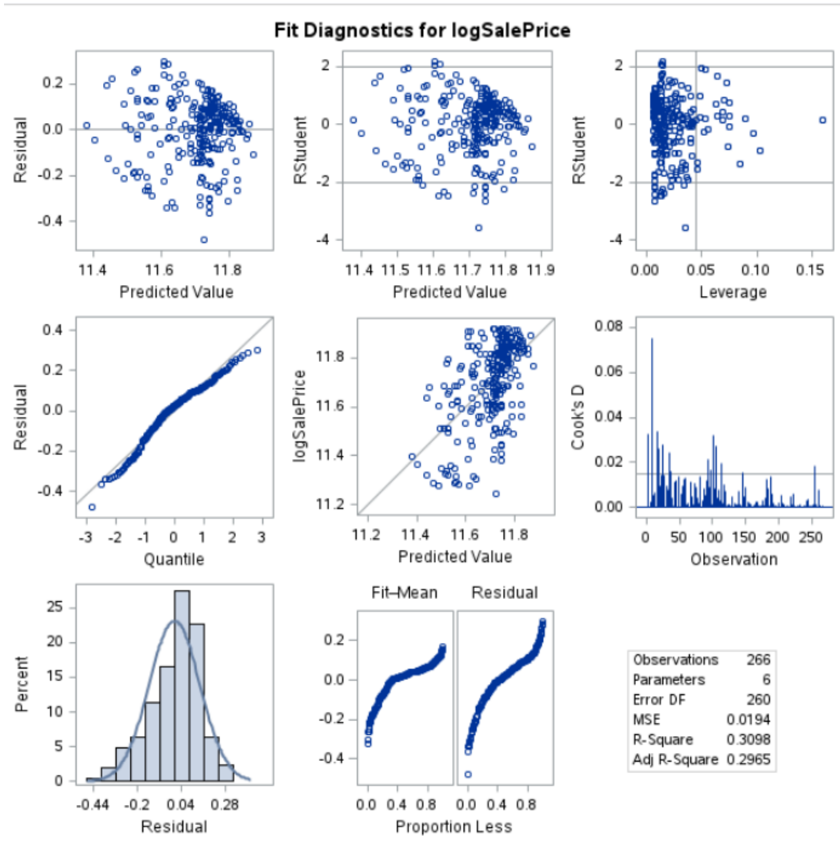

Neighborhood=NAmes
logSalePrice
logGrLiv

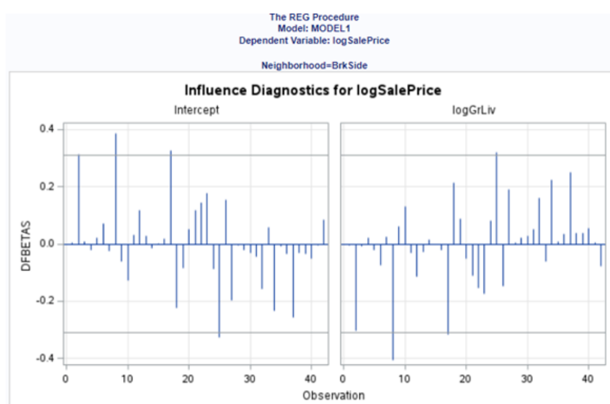

Residual
Quantile

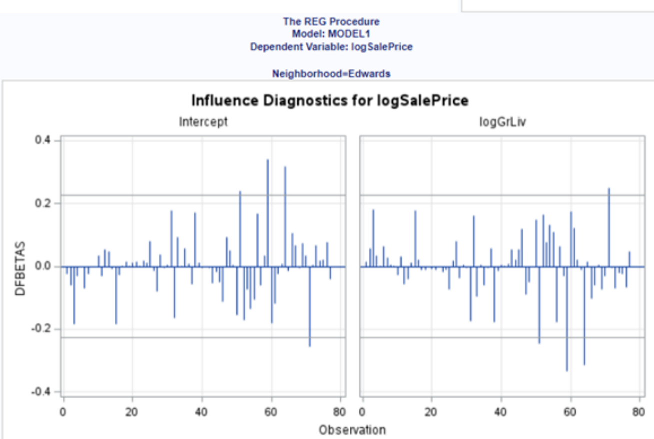

Percent
Residual

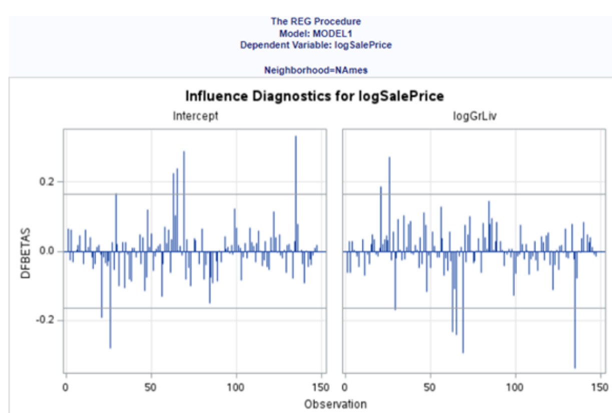

Analysis of Covariance for logSalePrice
logSalePrice
logGrLiv
Neighborhood
Edwards
NAmes
BrkSide

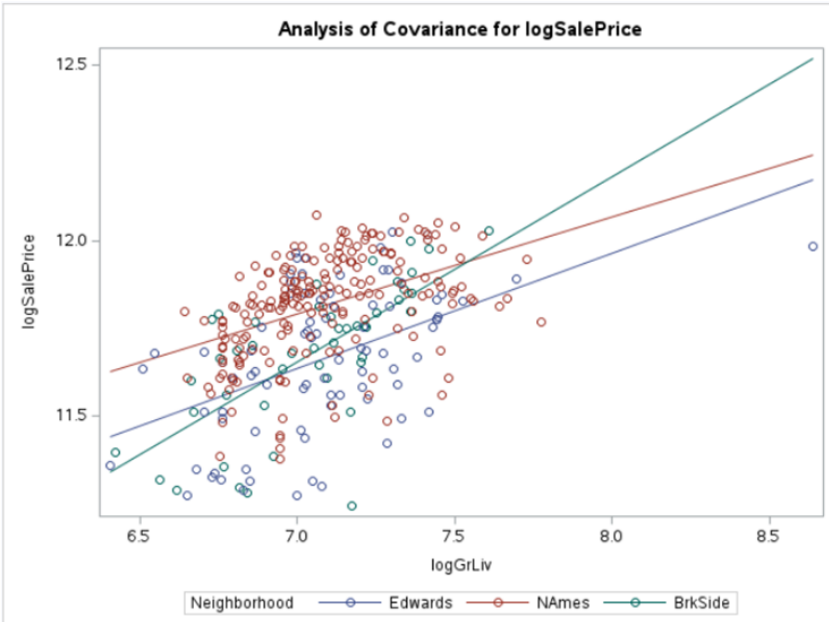
- Figure 2.14 Assumption Plots, Residuals, Influential plots, and Leverage for Restricted data











- Figure 2.15 GLMSELECT procedure for R^2, Adj R-Sq, CV PRESS

Unrestricted Model	Restricted Model

Unrestricted Model

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 1).

Effects:	Intercept logGrLiv

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	12.00843	12.00843	276.32
Error	381	16.55790	0.04346	
Corrected Total	382	28.56633		

Root MSE	0.20847
Dependent Mean	11.79887
R-Square	0.4204
Adj R-Sq	0.4188
AIC	-814.06885
AICC	-814.00553
SBC	-1191.17278
CV PRESS	16.93860

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	7.753378	0.243603	31.83
logGrLiv	1	0.568241	0.034185	16.62

Restricted Model

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 1).

Effects:	Intercept logGrLiv

Analysis of Variance

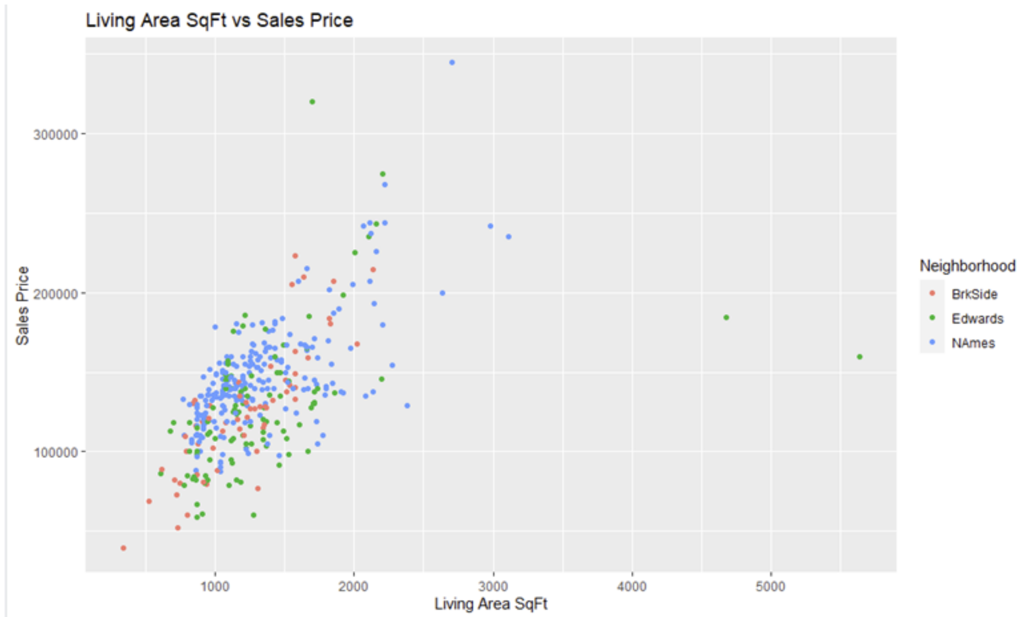
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	1.08372	1.08372	45.95
Error	264	6.22698	0.02359	
Corrected Total	265	7.31070		

Root MSE	0.15358
Dependent Mean	11.70489
R-Square	0.1482
Adj R-Sq	0.1450
AIC	-726.72499
AICC	-726.63339
SBC	-987.55800
CV PRESS	6.29020

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	9.869850	0.270885	36.44
logGrLiv	1	0.260750	0.038468	6.78

- Figure 2.16 General Scatter for un-altered data LivArea v Sale Price



- Figure 2.17 Estimates, SE, t-Value, P-val, 95% Confidence Limits

Unrestricted:

The GLM Procedure

Dependent Variable: logSalePrice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	14.62857557	2.92571511	79.14	<.0001
Error	377	13.93775037	0.03697016		
Corrected Total	382	28.56632594			

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

Source	DF	Type I SS	Mean Square	F Value	Pr > F
logGrLiv	1	12.00843049	12.00843049	324.81	<.0001
Neighborhood	2	1.98063548	0.99031774	26.79	<.0001
logGrLiv*Neighborhoo	2	0.63950960	0.31975480	8.65	0.0002

Source	DF	Type III SS	Mean Square	F Value	Pr > F
logGrLiv	1	11.69403942	11.69403942	316.31	<.0001
Neighborhood	2	0.69497286	0.34748643	9.40	0.0001
logGrLiv*Neighborhoo	2	0.63950960	0.31975480	8.65	0.0002

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	5.912920736	B	0.50459008	11.72	<.0001	4.920757171	6.905084301
logGrLiv	0.819648056	B	0.07162860	11.44	<.0001	0.678806432	0.960489681
Neighborhood Edwards	2.093586444	B	0.64589440	3.24	0.0013	0.823579544	3.363593345
Neighborhood NAmes	2.579806905	B	0.59988132	4.30	<.0001	1.400274428	3.759339383
Neighborhood BrkSide	0.000000000	B	.	.	.	.	.
logGrLiv*Neighborhoo Edwards	-0.299980812	B	0.09121531	-3.29	0.0011	-0.479335322	-0.120626303
logGrLiv*Neighborhoo NAmes	-0.346624454	B	0.08482008	-4.09	<.0001	-0.513404171	-0.179844737
logGrLiv*Neighborhoo BrkSide	0.000000000	B	.	.	.	.	.

Restricted:

The GLM Procedure

Dependent Variable: logSalePrice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2.26465889	0.45293178	23.34	<.0001
Error	260	5.04603839	0.01940784		
Corrected Total	265	7.31069729			

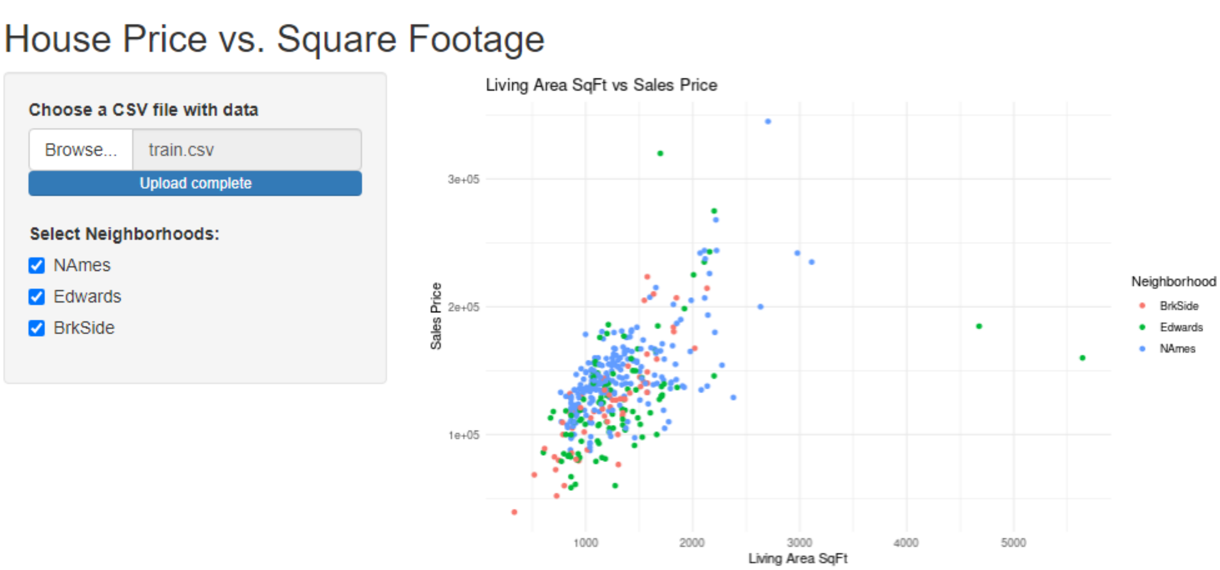
R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.309773	1.190204	0.139312	11.70489

Source	DF	Type I SS	Mean Square	F Value	Pr > F
logGrLiv	1	1.08371964	1.08371964	55.84	<.0001
Neighborhood	2	0.94607408	0.47303704	24.37	<.0001
logGrLiv*Neighborhoo	2	0.23486518	0.11743259	6.05	0.0027

Source	DF	Type III SS	Mean Square	F Value	Pr > F
logGrLiv	1	1.31812125	1.31812125	67.92	<.0001
Neighborhood	2	0.26588747	0.13294374	6.85	0.0013
logGrLiv*Neighborhoo	2	0.23486518	0.11743259	6.05	0.0027

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	8.419475736	B	0.61887516	13.60	<.0001	7.200830091	9.638121382
logGrLiv	0.460710853	B	0.08829387	5.22	<.0001	0.286848736	0.634572971
Neighborhood Edwards	0.705909089	B	0.75471400	0.94	0.3505	-0.780220902	2.192039080
Neighborhood NAmes	2.241837333	B	0.70784081	3.17	0.0017	0.848006749	3.635667916
Neighborhood BrkSide	0.000000000	B	.	.	.	.	.
logGrLiv*Neighborhoo Edwards	-0.105125439	B	0.10736939	-0.98	0.3284	-0.316549718	0.106298841
logGrLiv*Neighborhoo NAmes	-0.304952911	B	0.10089622	-3.02	0.0028	-0.503630688	-0.106275134
logGrLiv*Neighborhoo BrkSide	0.000000000	B	.	.	.	.	.

- Figure 2.18 Shiny App



- Figure 2.19 SAS Code for Analysis 1

```
/* Filtering the data to include only needed neighborhoods*/
data filtered_houses;
    set houses;
    where Neighborhood in ("NAmes", "Edwards", "BrkSide");
run;

/* Start by plotting the data */
proc sort data=filtered_houses;
by Neighborhood;
run;

proc sgplot data=filtered_houses;
    title "Scatter Plot of SalePrice vs GrLivArea by Neighborhood";
    scatter x=GrLivArea y=SalePrice / group=Neighborhood;
    xaxis label="Living Area SqFt";
    yaxis label="Sales Price";
    by Neighborhood;
run;

proc sgscatter data = filtered_houses;
by Neighborhood;
matrix SalePrice GrLivArea;
run;

data loghouses;
set filtered_houses;
logGrLiv = log(GrLivArea);
logSalePrice = log(SalePrice);
;

proc sgscatter data = loghouses;
by Neighborhood;
matrix logSalePrice logGrLiv;
run;


proc glm data = loghouses plots=all;
class Neighborhood (REF = "BrkSide");
model logSalePrice = logGrLiv | Neighborhood / solution clparm;
run;


/*Unrestricted MODEL*/
proc glm data = loghouses plots=all;
class Neighborhood (REF = "BrkSide");
model logSalePrice = logGrLiv Neighborhood logGrLiv*Neighborhood / solution clparm;
run;

/*Conf/Pred Plots */
proc reg data=loghouses outest=cooks;
  by Neighborhood;
  model logSalePrice = logGrLiv / stb clb;
run;
```

```
run;
/*Influential Points DFBeta plots */
proc reg data=loghouses plots(only)=DFBetas;
  by Neighborhood;
  model logSalePrice = logGrLiv / stb clb;
run;
/* CV Press non-restricted*/
proc glmselect data= loghouses;
class Neighborhood;
model logSalePrice = logGrLiv
/selection = forward(stop=CV) cvmethod=random(5) stats= adjrsq;
run;



/*Set Restriction on dataset*/
data restricted_data;
  set loghouses;
  where GrLivArea >= 1000 and GrLivArea <= 3250;
  where SalePrice >= 75000 and SalePrice <= 150000;
run;

data loghouses;
set filtered_houses;
logGrLiv = log(GrLivArea);
logSalePrice = log(SalePrice);
;
/*Restricted Model*/
proc glm data = restricted_data plots=all;
class Neighborhood (REF = "BrkSide");
model logSalePrice = logGrLiv Neighborhood logGrLiv*Neighborhood / solution clparm;
run;
/*Influential Points DFBeta plots (restricted)*/
proc reg data=restricted_data plots(only)=DFBetas;
  by Neighborhood;
  model logSalePrice = logGrLiv / stb clb;
run;
/*Scatter for restricted data*/
proc sgscatter data = restricted_data;
by Neighborhood;
matrix logSalePrice logGrLiv;
run;

/* CV Press restricted*/
proc glmselect data= restricted_data;
class Neighborhood;
model logSalePrice = logGrLiv
/selection = forward(stop=CV) cvmethod=random(5) stats= adjrsq;
run;
```

- Figure 2.20 RShiny Code for Analysis 1

```r
library(shiny)
library(ggplot2)
library(readr)

ui <- fluidPage(
  titlePanel("House Price vs. Square Footage"),
  sidebarLayout(
    sidebarPanel(
      fileInput("datafile", "Choose a CSV file with data"),
      checkboxGroupInput("neighborhoods", "Select Neighborhoods:",
                         choices = c("NAmes", "Edwards", "BrkSide"),
                         selected = c("NAmes", "Edwards", "BrkSide")
      )
    ),
    mainPanel(
      plotOutput("scatterplot")
    )
  )
)

server <- function(input, output) {
  data <- reactive({
    req(input$datafile)
    read_csv(input$datafile$datapath)
  })

  output$scatterplot <- renderPlot({
    req(data())
    filtered_data <- subset(data(), Neighborhood %in% input$neighborhoods)
    ggplot(filtered_data, aes(x = GrLivArea, y = SalePrice, color = Neighborhood)) +
      geom_point() +
      xlab("Living Area SqFt") +
      ylab("Sales Price") +
      ggtitle("Living Area SqFt vs Sales Price") +
      theme_minimal()
  })
}

shinyApp(ui, server)
```