KAGGLE PROJECT (MSDS 6371)

Daniel Chang & Garret Drake

Introduction

We have been asked by Century 21 Ames (a real estate company) in Ames Iowa to get an estimate of the sale price of a house based on the square footage of the living area and to see the sales price (and relationship to square footage) depend on which neighborhood the house is located in for the NAmes, Edwards, and BrkSide neighborhoods. Therefore Century 21 would like an estimate (or estimates if it varies by neighborhood) as well as confidence intervals for our estimate(s).

Data Description

The Ames Housing dataset was compiled by Dean De Cock, and is available to download via Kaggle.com. While the entire training data set examines 1460 observations of 79 different variables of home ownership in Ames, Iowa, for example, square footage, lot size, number of bathrooms, number of bedrooms, etc, more information about all the variables can be found on the Kaggle website.

For the first analysis we focused on what our client, Century 21, is interested in, which includes the sales price of a home, square footage, and the three neighborhoods they sell in, which are the NAmes, Edwards, and BrkSide neighborhoods. After filtering out these neighborhoods from the original 1460 observations, we were left with 383 observations with no missing values. There did appear to be some significant outliers that were ultimately from the data set and noted below because there did not appear to be a transformation that successfully helped improve the assumptions of regression.

For the second analysis, the same Ames housing dataset from kaggle.com was used. Again, the entire data set examines a training data set with 1460 observations and a test data set of 1459 both with 79 different variables of home ownership in Ames, Iowa. However, for this analysis, we conducted four separate types of regression stepwise, forward, and backward, and built a custom model. Further details of variable specifics are listed below in the Analysis of question 2 section.

Analysis Question 1:

Restatement of Problem

Century 21 Ames (a real estate company) in Ames, Iowa has commissioned us get an estimate of the sale price of a house for the three neighborhoods they currently sell in, NAmes, Edwards, and BrkSide, based on its square footage of living area, and to see if the sales price (and relationship to square footage) depends on which neighborhood the house is located in.

Build and Fit the Model

The first step was to examine a scatter plot of SalePrice vs GrLivArea by neighborhood [see Appendix, Figure 1.2]. The results from this appear to demonstrate a positive linear relationship between the square footage living area and sale price, however, there do appear to be some clear outliers that we will attempt to deal with in the modeling stages.

1. **First tentative model:**

$$\text{Model 1: } \mu(\text{SalePrice}) = b_0 + b_1(\text{GrLIvArea})$$

The following observations are made after viewing Appendix Figures 1.2 and 1.3 to check the assumptions of regression.

- Linearity: There appears to be a linear trend
- Normality: Based on the histogram of residuals this appears relatively normal
- Equal standard deviations: QQ Plot appears mostly linear, while there is a significant amount of clustering within the residual plot, likely due to outliers
- Independence: Given that they are looking at specific neighborhoods there could be a possible clustering effect, but we will assume independence, although not much is known about how these houses were selected.
- Outliers: There appear to be 5 outliers with studentized residuals greater than 3 and Cook's D greater than 5.
- The Adjusted R-Square = 0.3406

We checked various model transformations such as log-linear, linear-log, and log-log, however these did not appear to improve residual plots of assumptions therefore, the 5 outliers mentioned above were removed in the subsequent analysis [see Appendix Figure 1.4].

2. **Second tentative model: rerun the first model with the outliers removed:**

The following observations are made after viewing Appendix Figures 1.5 and 1.6 to check the assumptions of regression.

- Linearity: There appears to be a positive linear trend
- Normality: Based on the histogram of residuals this appears relatively normal and improved with outliers removed

- Equal standard deviations: QQ Plot appears mostly linear, there has been improvement in the residual plot (more randomly distributed, but still some clustering)
- Independence: Same assumption as above
- The Adjusted R-Square = 0.4408

3. **Third tentative model including the Neighborhood variables with interaction effects**

$$\text{Model 3: } \mu(\text{SalePrice}) = b_0 + b_1(\text{GrLIvArea}) + b_2(\text{GrLivArea*Neighborhood})$$

The following observations are made after viewing Appendix Figures 1.8, 1.9, 1.10, and 1.11 to check the assumptions of regression.

- Linearity: Appears to be a positive linear trend within each neighborhood
- Normality: Based on the histogram of residuals this appears relatively normal and improved with outliers removed and neighborhood interactions added
- Equal standard deviations: QQ Plot appears linear, there has been substantial improvement in the residual plot (more randomly distributed)
- Independence: Same assumption as above
- The Adjusted R-Square = 0.5131

This appears to be the best fitting model, nor does there appear to be any need for transformation of data, so this model was ultimately selected. Since this model has interaction effects for the neighborhoods a separate regression for each neighborhood is written below for ease of interpretation from the SAS output in Appendix Figure 1.12.

- Regression model for BrkSide neighborhood:

  μ(SalePrice|BrkSide) = 19971.51 +87.16*GrLivArea
- Regression model for Edwards neighborhood:

  μ(SalePrice|Edwards) = 45,110.28 + 63.04*GrLivArea
- Regression model for NAmes neighborhood:

  μ(SalePrice|NAmes) = 80,325.71 +49.56*GrLivArea

Conclusion and Interpretation

This model suggests that the linear regression fitted, $\mu(\text{SalePrice}) = b_0 + b_1(\text{GrLIvArea}) + b_2(\text{GrLivArea*Neighborhood})$, is a good fit based on significant F-test =80.46, df(5,372), and p-value <.0001. R-square = 0.5196, meaning that 51.96% of the variability of sale price can be explained by the living area square footage.

- Interpretation of BrkSide model: For every 100 sq.ft increase in living space (GrLivArea) in the BrkSide neighborhood there is an estimated increase in mean sale price of $8,716., with a 95% confidence interval between $7,078.04 and $10,354.66.
- Interpretation of Edwards model: For every 100 sq.ft increase in living space (GrLivArea) in the Edwards neighborhood there is an estimated increase in mean sale price of $6,304., with a 95% confidence interval between $2,491.37 and $10,117.12
- Interpretation of NAmes model: For every 100 sq.ft increase in living space (GrLivArea) in the NAmes neighborhood there is an estimated increase in mean sale price of $4,956., with a 95% confidence interval between $1,497.80 and $8,414.44.

Scope: While there is a positive correlation between sale price, square footage and the neighborhoods, no causal inferences can be made since this is an observational study. Additionally, there is no mention of random sampling so caution should be used in generalizing results beyond this population.

Rshiny App

R Shiny app of Price vs Living Area by Each Neighborhood

Analysis Question 2:

Restatement of Problem

In this analysis, we will build a predictive model for the sale prices of individual residential property in all neighborhoods in Ames, Iowa. To do this, we will use multiple linear regression to evaluate all variables in the dataset to build a good model that does this accurately. To select the variables, we will use Stepwise, Forward, Backward and Custom process selection as part of our analysis and compare the parameters(adjusted R-squared, internal CV Press and Kaggle Score) of these different models.

Cleaning/Pre-Processing Data

First, we clean our data by removing variables with lots of NA. For us, this would only be GarageYrBlt. We also cleaned up any columns with misspelled names to make the two datasets match. Next, we removed the 5 outliers that we discovered. Lastly, we keep any variables that we feel are pivotal to helping us determine SalePrice and log the SalePrice. Please refer to Figure 1.14 in the Appendix.

With the remaining variables, we examined the correlation and scatter plots to see if there are any linear relationships between the numerical variables with SalePrice. This leaves us with: OverallQual, OverallCond, YearBuilt, YearRemodAdd, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF GrLivArea FullBath HalfBath BedroomAbvGr TotRmsAbvGrd Fireplaces GarageCars, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, ScreenPorch PoolArea, and YrSold. Please refer to Figure 1.15 in the Appendix.

Next, we selected these categorical variables as candidates for our multiple linear regression model: Neighborhood, MSZoning, LotShape, LotConfig, Condition1, Condition2, BldgType, HouseStyle, RoofStyle, BsmtFinType1, HeatingQc, CentralAir, Electrical, KitchenQual, GarageType, GarageFinish and SaleType.

Building Models

1. **Forward Variable Selection Model (Appendix Figure 1.16)**
    - **Interested Variables:** OverallQual, OverallCond, YearBuilt, YearRemodAdd, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, GrLivArea, Fireplaces, GarageArea, Neighborhood, MSZoning, and BldgType **(Appendix Figure 1.17 and 1.18)**
        i. Adj R-Square = .885 after 14 steps
        ii. For Assumptions: there appears to be linearity, and based on the residual, studentized residual, QQ plot, Cook's D, and histogram all look relatively normal (Appendix Figure 1.19)
        iii. Scatter plots indicate randomly distributed residuals with no patterns.
        iv. Cook's D values are mostly less than 0.20, and some hover around it.
        v. Small violation with the normality from QQ plot and histogram; nothing to worry about.
        vi. We do not see any concern with variance; constant variance.
        vii. No noteworthy outliers in the residual plot.
        viii. **Our Linear Regression Model-Numeric Variables (Appendix Figure 1.19)**:

μ(logSalePrice) = b_0 + b_1(OverallQual) + b_2(OverallCond)+ b_3(YearBuilt) + b_4(YearRemodAdd) + b_5(BsmtFinSF1) + b_6(BsmtFinSF2) + b_7(BsmtUnfSF) + b_8(GrLivArea) + b_9(Fireplaces) + b_{10}(GarageCars) + b_{11}(GarageArea)

B_0 = 1.79, b_1 = 0.07, b_2=0.05, b_3=0.003, b_4=0.001, b_5=0.0002, b_6=0.0001, b_7 = 0.0001, b_8= 0.00002, b_9= 0.06, b_{10}=, b11=0.00002

2. **Backward Variable Selection Model (Appendix Figure 1.20)**
    - **Interested Variables:** OverallQual, OverallCond, Yearbuilt, YearRemodAdd, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, GrLivArea, FullBath, HalfBath, BedroomAbvGr, TolRmsAbvGrd, Fireplaces, GarageCars, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, ScreenPorch, PoolArea, YrSold, Neighborhood, MSZoning, LotShape, LotConfig, Condition1, BldgType, BsmtFinType1, HeatingQC, CentralAir, Electrical, KitchenQual, GarageType, GarageFinish, and SaleType. **(Appendix Figure 1.21)**
        i. Adj R-Square = .8932 after 3 steps
        ii. CV Press= **20.6768**
        iii. For Assumptions: there appears to be linearity, and based on the residual, studentized residual, QQ plot, Cook's D, and histogram all look relatively normal (Appendix Figure 1.22).
        iv. Scatter plots indicate randomly distributed residuals with no patterns, very similar to our forward model.
        v. Cook's D values are all less than 0.20. The highest is about 0.12.
        vi. Small violation with the normality from QQ plot and histogram; nothing to worry about. A slight left-skew.
        vii. We do not see any concern with a variance; constant variance.
        viii. No noteworthy outliers in the residual plot.

ix. **Linear Regression Model-Numeric Variables (Appendix Figure 1.21)**:

μ(logSalePrice) = b_0 + b_1(OverallQual) + b_2(OverallCond)+ b_3(YearBuilt) + b_4(YearRemodAdd) + b_5(BsmtFinSF1) + b_6(BsmtFinSF2) + b_7(BsmtUnfSF) + b_8(GrLivArea) + b_9(Fullbath) + b_{10}(HalfBath) + b_{11}(BedroomAbvGr) + b_{12}(TotRmsAbvGrd) +b_{13}(Fireplaces) + b_{14}(GarageCars) + b_{15}(GarageArea) + b_{17}(WoodDeckSF) +b_{18}(OpenPorchSF) +b_{19}(EnclosedPorch) + b_{20}(ScreenPorch) + b_{21}(PoolArea) + b_{22}(YrSold)

b_0=5.2, b_1= 0.004, b_2= 0.003, b_3 =0.0002, b_4=0.0002, b_5=0.00001, b_6 = 0.00002, b_7 = 0.00001, b_8=0.00001, b_9=0.01, b_{10}=0.009, b_{11}=0.006, b_{12}=0.004, b_{13}=0.006, b_{14}=0.01, b_{15}=0.00003, b_{16}=0.00001, b_{17}=0.00002, b_{18}=0.000005, b_{19}=0.00006, b_{20}=-0.00001, b_{21}=-0.005

3. **Stepwise Variable Selection Model(Appendix Figure 1.23)**
   - **Interested Variable:** OverallQual, OverallCond, YearBuilt, YearRemodAdd, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, GrLivArea, Fireplaces, GarageArea, Neighborhood, MSZoning, and BldgType
     i. Adj R-Square = **.8907** after 13 steps **(Appendix Figure 1.24)**
     ii. CV Press = **20.8527**
     iii. For Assumption: there appears to be linearity, and based on the residual, studentized residual, QQ plot, Cook's D, and histogram all look relatively normal (Appendix Figure 1.25).
     iv. Scatter plots indicate randomly distributed residuals with no patterns, very similar to our forward model.
     v. Cook's D values are mostly less than 0.20, and some hover around it. Similar to our forward model.
     vi. Normality and variance assumptions are very similar to the previous 2 models. There are no noteworthy outliers.
     vii. **Linear Regression Model- Numeric Variables (Appendix Figure 1.24)**:

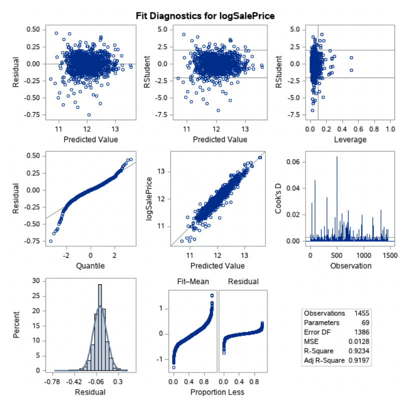
μ(logSalePrice) = b_0 + b_1(OverallQual) + b_2(OverallCond)+ b_3(YearBuilt) + b_4(YearRemodAdd) + b_5(BsmtFinSF1) + b_6(BsmtFinSF2) + b_7(BsmtUnfSF) + b_8(GrLivArea) +b_9(Fireplaces) + b_{10}(GarageArea)

b_0=2.2, b_1= 0.08, b_2= 0.05, b_3 =0.003, b_4=0.001, b_5=0.0002, b_6 = 0.0001, b_7 = 0.00001, b_8=0.00001, b_9=0.05, b_{10}=0.0002

4. **Custom Model(Appendix Figure 1.26):**
   - Since backward selection was our best model, we decided to use the variables associated with the selection process, but we did not use all of them. Using our intuition and logic, we were able to remove some variables we feel are not pivotal to determining the SalePrice from the backward model, such as Electrical, SaleType, GarageFinish, etc.

     We decided to run one more backward variable selection, which suggested we remove GarageType after 1 step. After a short discussion, we decided this was appropriate. Our variables are stated below. **(Appendix Figure 1.27)**

- **Interested Variables:** OverallQual, OverallCond, YearBuilt, YearRemodAdd, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, GrLivArea, FullBath, HalfBath, BedroomAbvGr, TotRmsAbvGrd, Fireplaces, GarageCars, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, ScreenPorch, PoolArea, YrSold, Neighborhood, MSZoning, LotShape, BldgType, HeatingQC, CentralAir, and KitchenQual
- Our assumptions are still the same as the previous 3 models where the normality and variance don't show any departure. In fact, the plots look exactly the same. This time our CookD did not surpass 0.06!



- **(Appendix Figure 1.28)** After doing this, we were able to achieve a CV Press of **20.4266.**
- Adjusted R-square= **.9197**

Summary

Please refer to Appendix Figure 1.29 for how we got our predictions in R.

Appendix Figure 1.30: Forward Selection Model Prediction

Appendix Figure 1.31: Backward Selection Model Prediction

Appendix Figure 1.32: Stepwise Selection Model Prediction

Appendix Figure 1.33: Custom Selection Model Prediction

Models	**Adjusted R-squared**	**CV Press**	**Kaggle Score**
Forward	**.885**	**20.8470**	**0.14255**
Backward	**.8932**	**20.6768**	**0.14206**
Stepwise	**.8907**	**20.8527**	**0.14272**
Custom	**.9197**	**20.4266**	**0.14137**

Appendix:

Analysis 1:

SAS Code

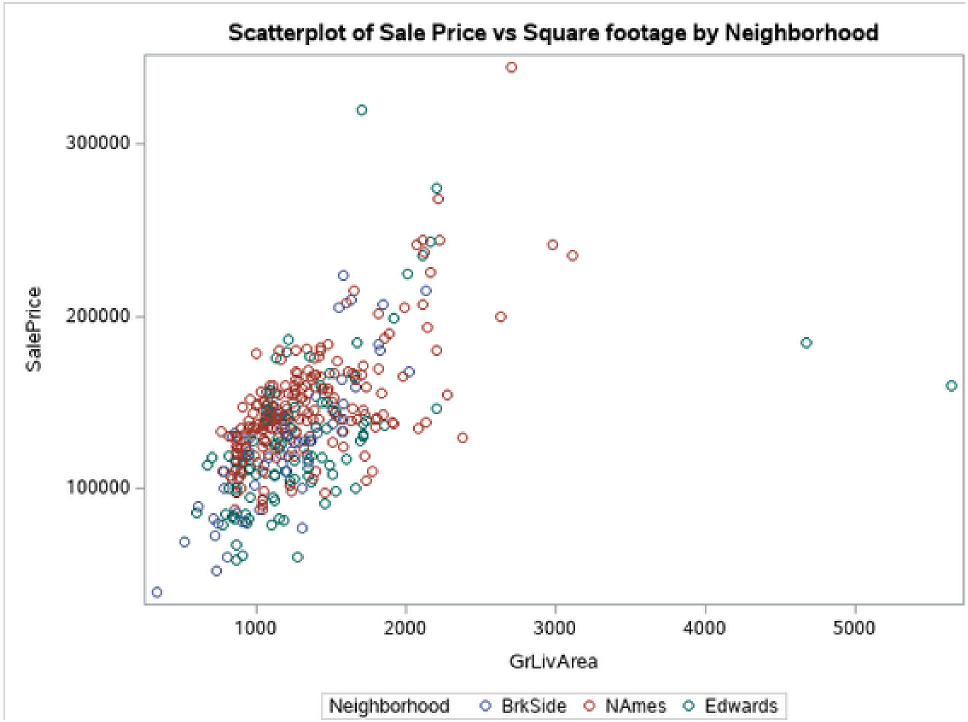
```
/*Question 1*/
/*Filter our dataset and Log Transform*/
data train2;
set train;
where Neighborhood contains "Edwards"
    or Neighborhood contains"NAmes"
    or Neighborhood contains "BrkSide";
run;

proc print data=train2;
run;
```

Figure 1.1

376	1436	20	RL	80	8400	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes
377	1437	20	RL	60	9000	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	NAmes
378	1444	30	RL	NA	8854	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	BrkSide
379	1449	50	RL	70	11767	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Edwards
380	1451	90	RL	60	9000	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	NAmes
381	1453	180	RM	35	3675	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Edwards
382	1459	20	RL	68	9717	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes
383	1460	20	RL	75	9937	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Edwards

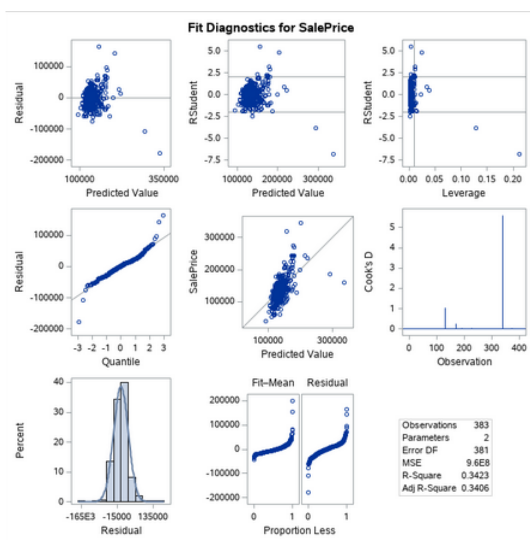
```
/*Scatterplot with Outliers by neighborhood*/
proc sgplot data=train2;
scatter x=GrLivArea y=SalePrice / group = neighborhood;
title 'Scatterplot of Sale Price vs Square footage by Neighborhood';
run;
```

Figure 1.2




```
/* Build Model1 with outliers*/
proc reg data= train2;
model SalePrice = GrLIvArea / vif clb cli clm;
run;
```

Figure 1.3



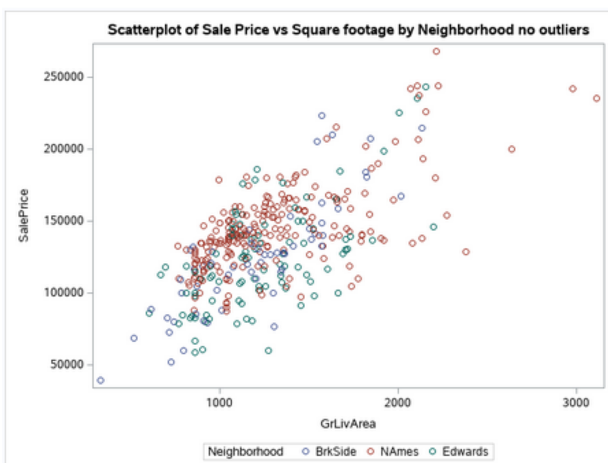
```
/* Identify Cook's D Outliers */
proc glm data=train2 alpha=0.05;
class Neighborhood;
model SalePrice = GrLivArea/ solution clparm;
output out=outliers1 P=Fitted PRESS=PRESS H=HAT
RSTUDENT=SRESID R=RESID DFFITS=DFFITS COOKD=COOKD;
run ;
proc print data=outliers1;

data outliers1;
set outliers1;
where abs(SRESID) > 3 or COOKD > 5;
run;
proc print data=outliers1;
```

Figure 1.4

Obs	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood
1	524	60	RL	13	40094	Pave	NA	IR1	Bnk	AllPub	Inside	Gtl	Edwards
2	643	80	RL	75	13860	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes
3	725	20	RL	86	13286	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Edwards
4	1299	60	RL	31	63887	Pave	NA	IR3	Bnk	AllPub	Corner	Gtl	Edwards
5	1424	80	RL	NA	19690	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	Edwards

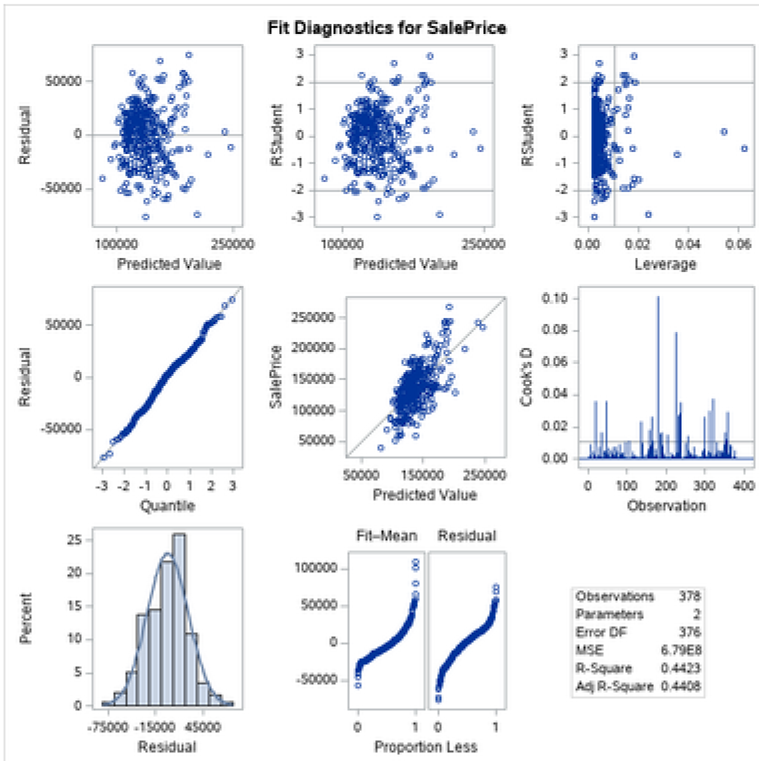
```
/*Plot without Outliers*/
proc sgplot data=train3;
 scatter x=GrLivArea y=SalePrice / group=Neighborhood;
 title 'Scatterplot of Sale Price vs Square footage by Neighborhood no outliers';
run;
```

Figure 1.5



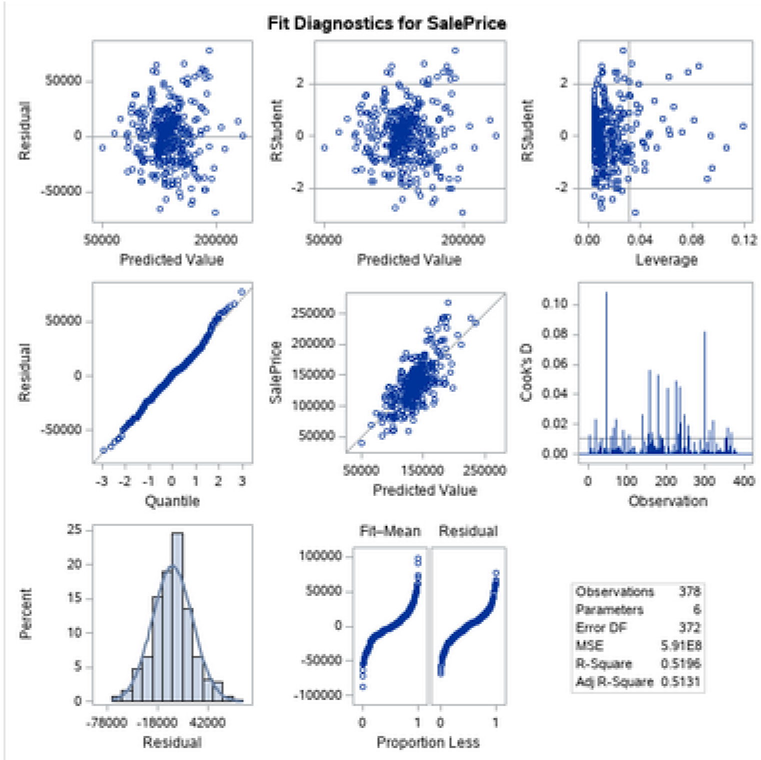
```
/* Run Model 2 Without Outliers */
proc glm data=train3 alpha=0.05 plots = All;
class Neighborhood;
model SalePrice = GrLivArea / solution;
run;
```

Figure 1.6




```
proc glm data= train3 plots = all;
class neighborhood (REF = "BrkSide");
model SalePrice = GrLIvArea|Neighborhood / solution clparm cli;
run;
```

Figure 1.7




```
/*  Model 3 */
/* Scatter plots of three neighborhoods without outliers */
title 'Scatter plot of BrkSide: SalePrice vs GrLIvArea';
PROC sgplot DATA=train3;
where neighborhood = 'BrkSide';
scatter x=GrLIvArea y=SalePrice;
run;

title 'Scatter plot of NAmes: SalePrice vs GrLIvArea';
PROC sgplot DATA=train3;
where neighborhood = 'NAmes';
scatter x=GrLIvArea y=SalePrice;
run;

title 'Scatter plot of Edwards: SalePrice vs GrLIvArea';
PROC sgplot DATA=train3;
where neighborhood = 'Edwards';
scatter x=GrLIvArea y=SalePrice;
run;
```

Figure 1.8

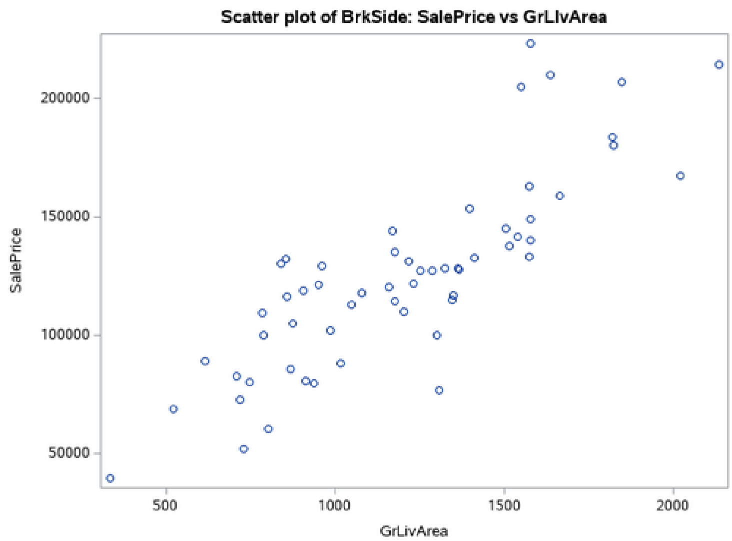


Figure 1.9

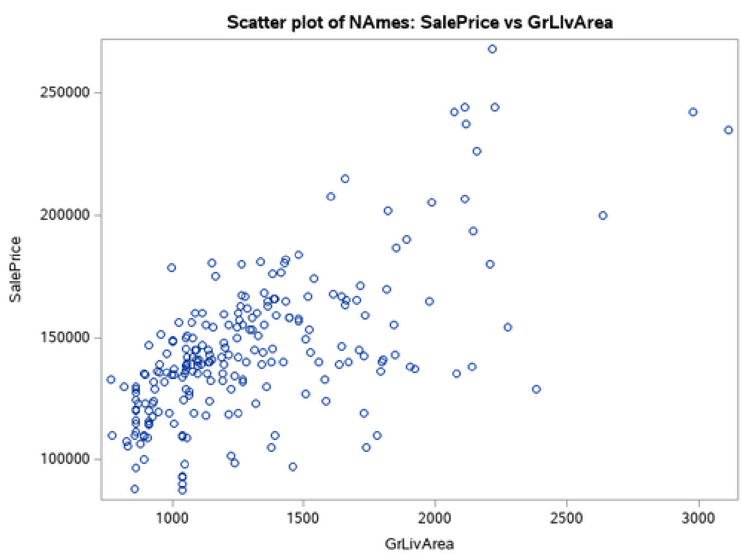
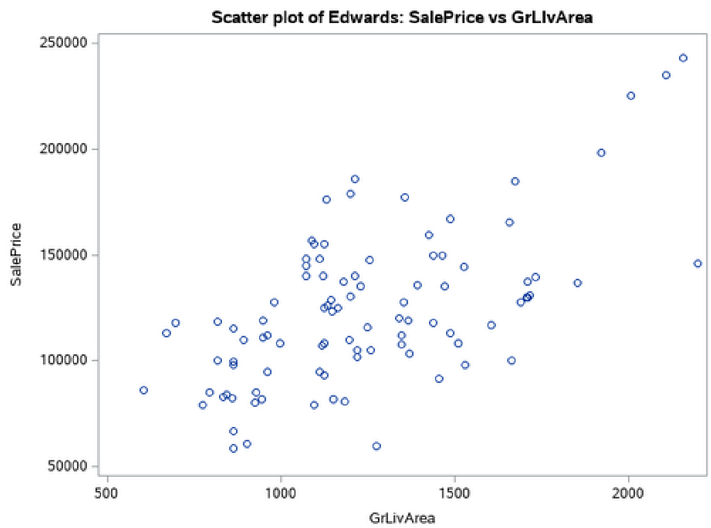


Figure 1.10




```
proc glm data= train3 plots = all;
class neighborhood (REF = "BrkSide");
model SalePrice = GrLIvArea|Neighborhood / solution clparm cli;
run;
```

Figure 1.11

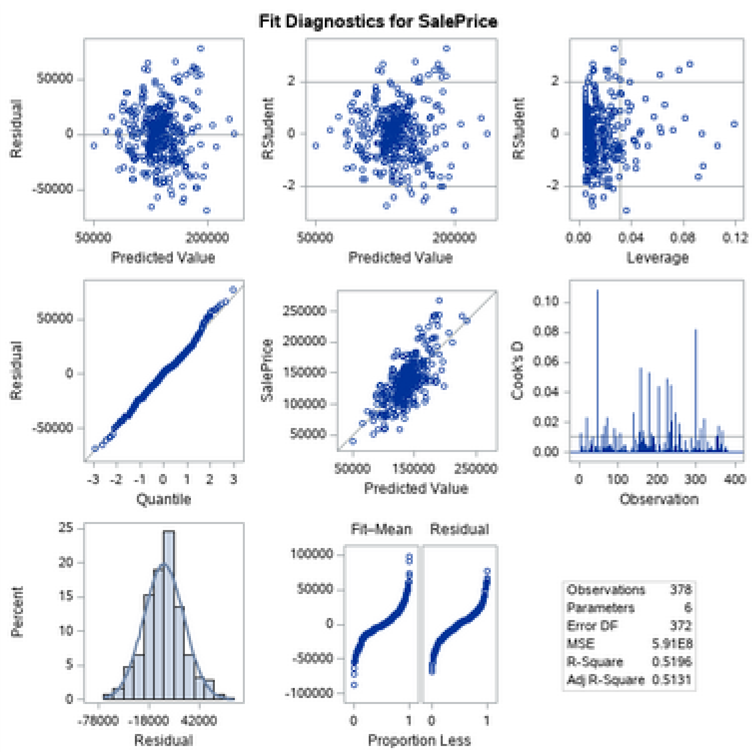


Figure 1.12

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	19971.51379	B	10519.32258	1.90	0.0584	-713.27717	40656.30476
GrLivArea	87.16253	B	8.33119	10.46	<.0001	70.78040	103.54466
Neighborhood Edwards	25138.76985	B	14113.06590	1.78	0.0757	-2612.61964	52890.15934
Neighborhood NAmes	60354.19850	B	11872.81191	5.08	<.0001	37007.95822	83700.43879
Neighborhood BrkSide	0.00000	B	.	.	.	.	.
GrLivArea*Neighborho Edwards	-24.12011	B	11.05931	-2.18	0.0298	-45.86672	-2.37350
GrLivArea*Neighborho NAmes	-37.60128	B	9.25622	-4.06	<.0001	-55.80235	-19.40022
GrLivArea*Neighborho BrkSide	0.00000	B	.	.	.	.	.

Figure 1.13

The GLM Procedure

Dependent Variable: SalePrice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	237906192195	47581238439	80.46	<.0001
Error	372	219982205379	591350014.46		
Corrected Total	377	457888397574			

Analysis 2:

SAS Code

Figure 1.14

```
/* Remove Outliers */
data train;
set train;
where Id ~= 524 AND Id ~= 643 AND Id ~= 725 AND Id ~= 1299 AND Id ~= 1424;
run;

/* Create new dataset with interested variables*/
data train4;
set train;
keep MSSubClass MSZoning LotArea LotShape LandContour LotConfig LandSlope   Neighborhood
Condition1  Condition2  BldgType    HouseStyle  OverallQual OverallCond YearBuilt   YearRemodAdd
RoofStyle   RoofMatl    Exterior1st Exterior2nd MasVnrType  ExterQual   ExterCond   Foundation  BsmtQual
BsmtCond    BsmtExposure    BsmtFinType1    BsmtFinSF1  BsmtFinType2    BsmtFinSF2  BsmtUnfSF   TotalBsmtSF
Heating HeatingQC   CentralAir  Electrical  LowQualFinSF    GrLivArea   BsmtFullBath    BsmtHalfBath    FullBath
HalfBath    BedroomAbvGr    KitchenAbvGr    KitchenQual TotRmsAbvGrd    Fireplaces GarageType
GarageFinish    GarageCars  GarageArea  GarageQual  GarageCond  PavedDrive  WoodDeckSF  OpenPorchSF EnclosedPorch
ScreenPorch PoolArea Fence  MiscFeature MiscVal MoSold  YrSold  SaleType SaleCondition  SalePrice;
run;


/* Log Sale Price to account for normality/linearity*/
data train4;
set train4;
logSalePrice = log(SalePrice);
run;
```

Figure 1.15

```
/* Check for Linear Relationships for Numerical Variables */
proc corr data = train; *check correlations first;
run;

PROC sgscatter DATA=train4;
matrix logSalePrice MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2
BsmtUnfSF;
run;

PROC sgscatter DATA=train4;
matrix logSalePrice LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea;
run;

PROC sgscatter DATA=train4;
matrix logSalePrice WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch PoolArea MiscVal MoSold YrSold;
run;
```

Correlations

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

wQualFinSF	GrLivArea	BsmtFullBath	BsmtHalfBath	FullBath	HalfBath	BedroomAbvGr	KitchenAbvGr	TotRmsAbvGrd	Fireplaces	GarageYrBlt
-0.04412 0.0925 1455	0.00194 0.9410 1455	0.00095 0.9712 1455	-0.01981 0.4502 1455	0.00498 0.8495 1455	0.00711 0.7864 1455	0.03636 0.1657 1455	0.00329 0.9003 1455	0.02447 0.3510 1455	-0.02342 0.3721 1455	-0.00011 0.9968 1374
0.04653 0.0760 1455	0.07584 0.0038 1455	0.00386 0.8832 1455	-0.00227 0.9310 1455	0.13180 <.0001 1455	0.17703 <.0001 1455	-0.02444 0.3515 1455	0.28193 <.0001 1455	0.04004 0.1269 1455	-0.04753 0.0699 1455	0.08615 0.0014 1374
0.00567 0.8290 1455	0.23211 <.0001 1455	0.14745 <.0001 1455	0.05035 0.0548 1455	0.11895 <.0001 1455	0.00772 0.7686 1455	0.11926 <.0001 1455	-0.01657 0.5276 1455	0.17459 <.0001 1455	0.26119 <.0001 1455	-0.03219 0.2330 1374
-0.02991 0.2542 1455	0.58952 <.0001 1455	0.10164 0.0001 1455	-0.03895 0.1375 1455	0.54811 <.0001 1455	0.27121 <.0001 1455	0.10046 0.0001 1455	-0.18399 <.0001 1455	0.41937 <.0001 1455	0.39089 <.0001 1455	0.54670 <.0001 1374
0.02561 0.3289 1455	-0.08086 0.0020 1455	-0.05366 0.0407 1455	0.11817 <.0001 1455	-0.19495 <.0001 1455	-0.06078 0.0204 1455	0.01074 0.6822 1455	-0.08698 0.0009 1455	-0.05777 0.0276 1455	-0.02616 0.3188 1455	-0.32323 <.0001 1374
-0.18378 <.0001 1455	0.19481 <.0001 1455	0.18428 <.0001 1455	-0.03765 0.1511 1455	0.46711 <.0001 1455	0.24197 <.0001 1455	-0.07112 0.0067 1455	-0.17457 <.0001 1455	0.09072 0.0005 1455	0.14529 <.0001 1455	0.82517 <.0001 1374
-0.06228 0.0175 1455	0.28750 <.0001 1455	0.11527 <.0001 1455	-0.01187 0.6510 1455	0.43818 <.0001 1455	0.18179 <.0001 1455	-0.04068 0.1209 1455	-0.14943 <.0001 1455	0.18818 <.0001 1455	0.10919 <.0001 1455	0.64135 <.0001 1374
-0.06916 0.0085 1447	0.37367 <.0001 1447	0.07475 0.0044 1447	0.02824 0.2830 1447	0.27180 <.0001 1447	0.19844 <.0001 1447	0.10415 <.0001 1447	-0.03689 0.1607 1447	0.26995 <.0001 1447	0.24386 <.0001 1447	0.24805 <.0001 1366
-0.06662 0.0110 1455	0.13987 <.0001 1455	0.65766 <.0001 1455	0.07451 0.0045 1455	0.04632 0.0774 1455	-0.01068 0.6839 1455	-0.11691 <.0001 1455	-0.08287 0.0016 1455	0.00816 0.7558 1455	0.23889 <.0001 1455	0.14755 <.0001 1374
0.01469 0.5755 1455	-0.00634 0.8089 1455	0.16055 <.0001 1455	0.07073 0.0070 1455	-0.07555 0.0039 1455	-0.03179 0.2256 1455	-0.01510 0.5649 1455	-0.04097 0.1183 1455	-0.03362 0.2000 1455	0.04924 0.0604 1455	-0.08758 0.0012 1374
0.02821 0.2822 1455	0.24760 <.0001 1455	-0.42465 <.0001 1455	-0.09576 0.0003 1455	0.28874 <.0001 1455	-0.04153 0.1133 1455	0.16666 <.0001 1455	0.03016 0.2502 1455	0.25192 <.0001 1455	0.05251 0.0452 1455	0.19098 <.0001 1374
-0.03361 0.2002 1455	0.40785 <.0001 1455	0.29481 <.0001 1455	0.00300 0.9091 1455	0.32725 <.0001 1455	-0.06789 0.0096 1455	0.05021 0.0555 1455	-0.07017 0.0074 1455	0.26449 <.0001 1455	0.32441 <.0001 1455	0.32776 <.0001 1374
-0.01323 0.6142 1455	0.53155 <.0001 1455	0.23015 <.0001 1455	0.00530 0.8399 1455	0.38084 <.0001 1455	-0.13840 <.0001 1455	0.12796 <.0001 1455	0.07388 0.0048 1455	0.39406 <.0001 1455	0.39882 <.0001 1455	0.23125 <.0001 1374
0.06399 0.0146 1455	0.69516 <.0001 1455	-0.17559 <.0001 1455	-0.02315 0.3775 1455	0.41831 <.0001 1455	0.60897 <.0001 1455	0.50373 <.0001 1455	0.06023 0.0216 1455	0.61448 <.0001 1455	0.19049 <.0001 1455	0.06883 0.0107 1374
1.00000 1455	0.14134 <.0001 1455	-0.04691 0.0737 1455	-0.00594 0.8208 1455	-0.00023 0.9929 1455	-0.02695 0.3042 1455	0.10602 <.0001 1455	0.00744 0.7769 1455	0.13303 <.0001 1455	-0.02072 0.4298 1455	-0.03622 0.1797 1374
0.14134	1.00000	0.01319	-0.01658	0.63822	0.41920	0.53662	0.10656	0.82921	0.45371	0.22925

GarageCars	GarageArea	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch	PoolArea	MiscVal	MoSold	YrSold	SalePrice
0.01660 0.5270 1455	0.01461 0.5777 1455	-0.03477 0.1850 1455	-0.00528 0.8406 1455	0.00346 0.8951 1455	-0.04653 0.0760 1455	0.00176 0.9466 1455	0.03097 0.2378 1455	-0.00611 0.8157 1455	0.02288 0.3832 1455	0.00206 0.9374 1455	-0.02279 0.3849 1455
-0.03955 0.1316 1455	-0.09912 0.0002 1455	-0.01367 0.6024 1455	-0.00799 0.7608 1455	-0.01195 0.6487 1455	-0.04382 0.0948 1455	-0.02598 0.3221 1455	0.00100 0.9697 1455	-0.00766 0.7702 1455	-0.01489 0.5704 1455	-0.02071 0.4299 1455	-0.08474 0.0012 1455
0.15209 <.0001 1455	0.16363 <.0001 1455	0.16666 <.0001 1455	0.05992 0.0223 1455	-0.01613 0.5388 1455	0.02150 0.4126 1455	0.04560 0.0821 1455	0.02796 0.2864 1455	0.03918 0.1352 1455	0.00561 0.8307 1455	-0.01323 0.6141 1455	0.26748 <.0001 1455
0.59998 <.0001 1455	0.55652 <.0001 1455	0.23511 <.0001 1455	0.29760 <.0001 1455	-0.11259 <.0001 1455	0.03133 0.2323 1455	0.06709 0.0105 1455	0.05314 0.0427 1455	-0.03110 0.2358 1455	0.07488 0.0043 1455	-0.02888 0.2710 1455	0.79545 <.0001 1455
-0.18565 <.0001 1455	-0.15041 <.0001 1455	-0.00675 0.7970 1455	-0.03328 0.2046 1455	0.07072 0.0070 1455	0.02562 0.3288 1455	0.05508 0.0357 1455	-0.01689 0.5199 1455	0.06894 0.0085 1455	-0.00508 0.8465 1455	0.04464 0.0888 1455	-0.08052 0.0021 1455
0.53679 <.0001 1455	0.47741 <.0001 1455	0.22435 <.0001 1455	0.18491 <.0001 1455	-0.38708 <.0001 1455	0.03168 0.2272 1455	-0.04981 0.0575 1455	-0.00330 0.9001 1455	-0.03423 0.1919 1455	0.01437 0.5839 1455	-0.01418 0.5890 1455	0.52412 <.0001 1455
0.41946 <.0001 1455	0.36926 <.0001 1455	0.20679 <.0001 1455	0.22498 <.0001 1455	-0.19356 <.0001 1455	0.04560 0.0821 1455	-0.03826 0.1446 1455	0.01008 0.7009 1455	-0.01013 0.6995 1455	0.02360 0.3683 1455	0.03437 0.1901 1455	0.50834 <.0001 1455
0.36250 <.0001 1447	0.36232 <.0001 1447	0.15821 <.0001 1447	0.10655 <.0001 1447	-0.10948 <.0001 1447	0.01962 0.4559 1447	0.06357 0.0156 1447	-0.01501 0.5683 1447	-0.02966 0.2595 1447	-0.00243 0.9263 1447	-0.00811 0.7578 1447	0.48428 <.0001 1447
0.22769 <.0001 1455	0.27195 <.0001 1455	0.20536 <.0001 1455	0.07334 0.0051 1455	-0.10344 <.0001 1455	0.02952 0.2605 1455	0.06919 0.0083 1455	0.07730 0.0032 1455	0.00494 0.8506 1455	-0.00222 0.9324 1455	0.01238 0.6372 1455	0.40736 <.0001 1455
-0.03750 0.1528 1455	-0.01682 0.5215 1455	0.07025 0.0073 1455	0.00582 0.8244 1455	0.03620 0.1676 1455	-0.03011 0.2510 1455	0.08863 0.0007 1455	0.05809 0.0267 1455	0.00485 0.8532 1455	-0.01552 0.5541 1455	0.03174 0.2263 1455	-0.01046 0.6901 1455
0.21411 <.0001 1455	0.18499 <.0001 1455	-0.00609 0.8164 1455	0.12884 <.0001 1455	-0.00242 0.9265 1455	0.02081 0.4277 1455	-0.01250 0.6339 1455	-0.04498 0.0863 1455	-0.02381 0.3640 1455	0.03366 0.1994 1455	-0.04055 0.1221 1455	0.21524 <.0001 1455
0.45151 <.0001 1455	0.47462 <.0001 1455	0.23503 <.0001 1455	0.21637 <.0001 1455	-0.09634 0.0002 1455	0.04126 0.1157 1455	0.09337 0.0004 1455	0.05522 0.0352 1455	-0.01839 0.4833 1455	0.02758 0.2931 1455	-0.01803 0.4919 1455	0.65061 <.0001 1455
0.44858 <.0001 1455	0.47730 <.0001 1455	0.23072 <.0001 1455	0.17590 <.0001 1455	-0.06325 0.0158 1455	0.06008 0.0219 1455	0.09623 0.0002 1455	0.05587 0.0331 1455	-0.02081 0.4276 1455	0.04145 0.1140 1455	-0.01409 0.5913 1455	0.62939 <.0001 1455
0.18217 <.0001 1455	0.13287 <.0001 1455	0.08908 0.0007 1455	0.19700 <.0001 1455	0.06342 0.0155 1455	-0.02407 0.3589 1455	0.04166 0.1122 1455	0.07620 0.0036 1455	0.01655 0.5282 1455	0.03343 0.2025 1455	-0.02724 0.2992 1455	0.32095 <.0001 1455
-0.09432 0.0003 1455	-0.06754 0.0100 1455	-0.02489 0.3427 1455	0.01970 0.4527 1455	0.06095 0.0201 1455	-0.00434 0.8685 1455	0.02669 0.3089 1455	0.07877 0.0026 1455	-0.00383 0.8840 1455	-0.02235 0.3942 1455	-0.02896 0.2695 1455	-0.02529 0.3351 1455
0.47593	0.45718	0.24309	0.29993	0.01402	0.02291	0.10882	0.11393	-0.00139	0.05693	-0.03650	0.73418

Scatterplots

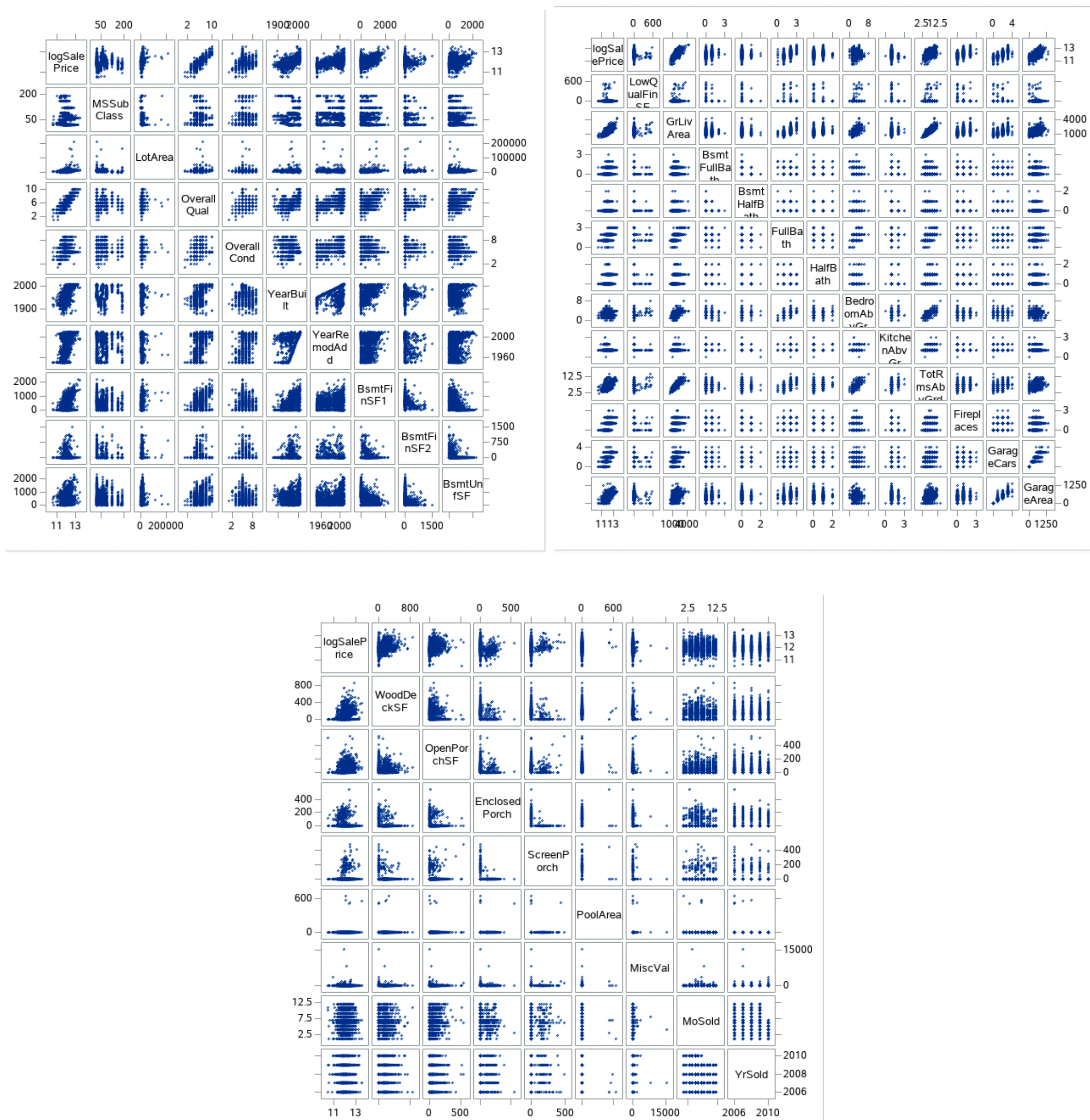


Figure 1.16

```
/* Forward Selection Model*/
/*Agjusted R-Squared: .885 after 14 steps, CV Press: 20.8470, Kaggle*/
proc glmselect data = train4;
class Neighborhood MSZoning LotShape LotConfig Condition1 Condition2 BldgType HouseStyle RoofStyle
BsmtFinType1 HeatingQc CentralAir Electrical KitchenQual GarageType GarageFinish SaleType;
model logSalePrice = OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF GrLivArea
FullBath HalfBath BedroomAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ScreenPorch PoolArea YrSold Neighborhood MSZoning LotShape LotConfig Condition1 Condition2 BldgType HouseStyle
RoofStyle BsmtFinType1 HeatingQc CentralAir Electrical KitchenQual GarageType
GarageFinish SaleType
/selection = Forward(stop = cv) cvmethod = random(5) CVDETAILS stats = adjrsq;
store ForwardTrainModel;
run;
```

Figure 1.17

The GLMSELECT Procedure

Data Set	WORK.TRAIN4
Dependent Variable	logSalePrice
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	112742069

Number of Observations Read	1455
Number of Observations Used	1455

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	-2666.7593	232.6648
1	OverallQual	2	2	0.6743	-4292.6492	75.8743
2	GrLivArea	3	3	0.7627	-4746.9361	55.1294
3	Neighborhood	4	27	0.8331	-5108.4656	39.6983
4	BsmtFinSF1	5	28	0.8586	-5343.5834	33.6068
5	OverallCond	6	29	0.8708	-5468.1415	30.8810
6	YearBuilt	7	30	0.8840	-5619.4611	27.8588
7	GarageArea	8	31	0.8927	-5725.7323	25.7797
8	BsmtUnfSF	9	32	0.8979	-5792.9075	24.7387
9	BsmtFinSF2	10	33	0.9024	-5852.1555	23.5903
10	MSZoning	11	37	0.9075	-5904.8779	22.3702
11	Fireplaces	12	38	0.9100	-5938.1907	21.7878
12	YearRemodAdd	13	39	0.9113	-5953.5284	21.4943
13	BldgType	14	43	0.9137	-5967.8449	21.1219
14	GarageCars	15	44	0.9147*	-5978.3059*	20.8470*
* Optimal Value of Criterion						

Figure 1.18

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 14).

Effects:	Intercept OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF GrLivArea Fireplaces GarageCars GarageArea Neighborhood MSZoning BldgType

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	43	212.40417	4.93963	363.44
Error	1411	19.17726	0.01359	
Corrected Total	1454	231.58143		

Root MSE	0.11658
Dependent Mean	12.02272
R-Square	0.9172
Adj R-Sq	0.9147
AIC	-4753.74736
AICC	-4750.80911
SBC	-5978.30587
CV PRESS	20.84700

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	1140	315	3.7586
2	1162	293	3.9332
3	1144	311	4.9888
4	1186	269	4.4896
5	1188	267	3.6768
Total			20.8470

Figure 1.19

```
/*Run Linear Regression Model for Forward Model: Numeric Only*/
proc glm data = train4 plots=all;
class Neighborhood MSZoning BldgType;
model logSalePrice = OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF GrLivArea
GarageArea Fireplaces /solution;
run;
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	2.196793766	0.42841657	5.13	<.0001
OverallQual	0.080454794	0.00428997	18.75	<.0001
OverallCond	0.053270015	0.00385040	13.83	<.0001
YearBuilt	0.003116234	0.00019590	15.91	<.0001
YearRemodAdd	0.001078945	0.00024504	4.40	<.0001
BsmtFinSF1	0.000249933	0.00001178	21.22	<.0001
BsmtFinSF2	0.000199907	0.00002319	8.62	<.0001
BsmtUnfSF	0.000138964	0.00001187	11.71	<.0001
GrLivArea	0.000301176	0.00000935	32.21	<.0001
GarageArea	0.000217521	0.00002189	9.94	<.0001

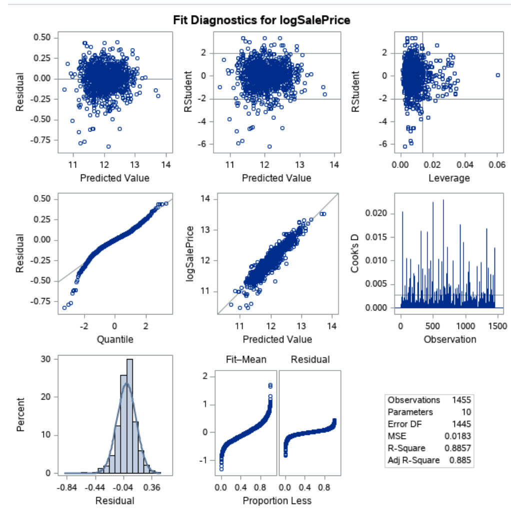


Figure 1.20

```
/* Backward Selection Model*/
/*Agjusted R-Squared: .8932 after 3 steps, CV Press: 20.6768, Kaggle: 0.14206*/
proc glmselect data = train4;
class Neighborhood MSZoning LotShape LotConfig Condition1 Condition2 BldgType HouseStyle RoofStyle
BsmtFinType1 HeatingQc CentralAir Electrical KitchenQual GarageType GarageFinish SaleType;
model logSalePrice = OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF GrLivArea
FullBath HalfBath BedroomAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ScreenPorch PoolArea YrSold Neighborhood MSZoning LotShape LotConfig Condition1 Condition2 BldgType HouseStyle
RoofStyle BsmtFinType1 HeatingQc CentralAir Electrical KitchenQual GarageType
GarageFinish SaleType
/ selection = Backward(stop = cv) cvmethod = random(5) CVDETAILS stats = adjrsq;
store BackwardTrainModel;
run;
```

Figure 1.21

The GLMSELECT Procedure

Backward Selection Summary						
Step	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0		39	120	0.9236	-5665.4862	21.3779
1	HouseStyle	38	113	0.9236*	-5709.8605	21.2294
2	Condition2	37	106	0.9233	-5746.5272	20.8200
3	RoofStyle	36	101	0.9234	-5780.1135*	20.3127*
* 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	Neighborhood	21.5026	>	20.3127

The GLMSELECT Procedure

Backward Selection Summary						
Step	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0		39	120	0.9236	-5665.4862	21.3779
1	HouseStyle	38	113	0.9236*	-5709.8605	21.2294
2	Condition2	37	106	0.9233	-5746.5272	20.8200
3	RoofStyle	36	101	0.9234	-5780.1135*	20.3127*
* 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	Neighborhood	21.5026	>	20.3127

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 3).

Effects:	Intercept OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF GrLivArea FullBath HalfBath BedroomAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch PoolArea YrSold Neighborhood MSZoning LotShape LotConfig Condition1 BldgType BsmtFinType1 HeatingQC CentralAir Electrical KitchenQual GarageType GarageFinish SaleType

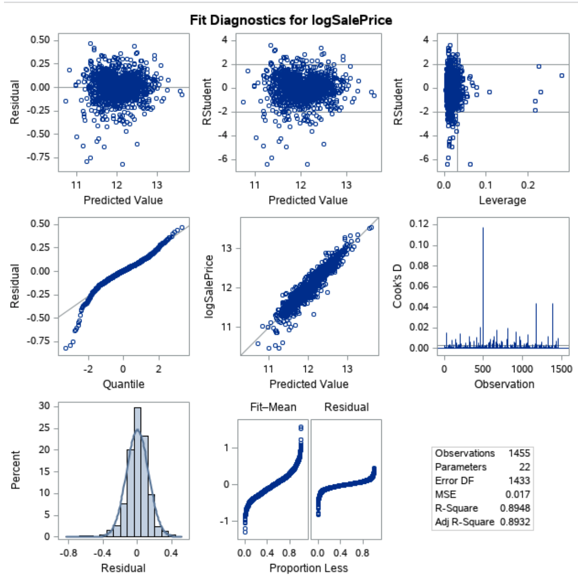
Analysis of Variance				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**
Model	100	215.06038	2.15060	176.25
Error	1354	16.52106	0.01220	
Corrected Total	1454	231.58143		

Root MSE	0.11046
Dependent Mean	12.02272
R-Square	0.9287
Adj R-Sq	0.9234
AIC	-4856.67237
AICC	-4841.13095
SBC	-5780.11349
CV PRESS	20.67681

Cross Validation Details			
	Observations		
Index	**Fitted**	**Left Out**	**CV PRESS**
1	1161	294	4.5183
2	1149	306	5.3683
3	1181	274	3.1787
4	1154	301	3.9180
5	1175	280	3.6935
Total			20.6768

Figure 1.22

```
/*Run Linear Regression Model for Backward Model: Numeric Only*/
proc glm data = train4 plots=all;
class Neighborhood MSZoning BldgType LotShape CentralAir Electrical LotConfig Condition1 BsmtFinType1
HeatingQC KitchenQual GarageType GarageFinish SaleType;
model logSalePrice = OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF GrLivArea
FullBath HalfBath BedroomAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF
EnclosedPorch ScreenPorch PoolArea YrSold/solution;
run;
```



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	12.77302559	5.21694551	2.45	0.0145
OverallQual	0.07157299	0.00431176	16.60	<.0001
OverallCond	0.05240085	0.00378933	13.83	<.0001
YearBuilt	0.00309243	0.00022413	13.80	<.0001
YearRemodAdd	0.00125950	0.00024489	5.14	<.0001
BsmtFinSF1	0.00023853	0.00001271	18.76	<.0001
BsmtFinSF2	0.00018088	0.00002320	7.80	<.0001
BsmtUnfSF	0.00013852	0.00001238	11.19	<.0001
GrLivArea	0.00025191	0.00001718	14.66	<.0001
FullBath	0.00243935	0.01021821	0.24	0.8114
HalfBath	0.01083511	0.00902511	1.20	0.2301
BedroomAbvGr	-0.00554031	0.00622052	-0.89	0.3733
TotRmsAbvGrd	0.00672557	0.00444679	1.51	0.1306
Fireplaces	0.04849351	0.00643878	7.53	<.0001
GarageCars	0.03412374	0.01082150	3.15	0.0016
GarageArea	0.00012291	0.00003646	3.37	0.0008
WoodDeckSF	0.00008919	0.00002983	2.99	0.0028
OpenPorchSF	0.00005648	0.00005702	0.99	0.3221
EnclosedPorch	0.00014112	0.00006290	2.24	0.0250
ScreenPorch	0.00025301	0.00006449	3.92	<.0001
PoolArea	-0.00015386	0.00010583	-1.45	0.1462
YrSold	-0.00540229	0.00259534	-2.08	0.0376

Figure 1.23

```
/* Stepwise Selection Model*/
/*Adjusted R-Squared: .8907 after 13 steps, CV Press: 20.8527, Kaggle: 0.14272*/
proc glmselect data = train4;
class Neighborhood MSZoning LotShape LotConfig Condition1 Condition2 BldgType HouseStyle RoofStyle
BsmtFinType1 HeatingQc CentralAir Electrical KitchenQual GarageType GarageFinish SaleType;
model logSalePrice = OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF GrLivArea
FullBath HalfBath BedroomAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ScreenPorch PoolArea YrSold Neighborhood MSZoning LotShape LotConfig Condition1 Condition2 BldgType HouseStyle
RoofStyle BsmtFinType1 HeatingQc CentralAir Electrical KitchenQual GarageType
GarageFinish SaleType
/ selection = Stepwise(stop = cv) cvmethod = random(5) CVDETAILS stats = adjrsq;
store StepwiseTrainModel;
run;
```

Figure 1.24

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	-2666.7593	231.8212
1	OverallQual		2	2	0.6743	-4292.6492	75.4169
2	GrLivArea		3	3	0.7627	-4746.9361	55.0155
3	Neighborhood		4	27	0.8331	-5108.4656	39.6195
4	BsmtFinSF1		5	28	0.8586	-5343.5834	33.5021
5	OverallCond		6	29	0.8708	-5468.1415	30.6572
6	YearBuilt		7	30	0.8840	-5619.4611	27.6907
7	GarageArea		8	31	0.8927	-5725.7323	25.6577
8	BsmtUnfSF		9	32	0.8979	-5792.9075	24.3155
9	BsmtFinSF2		10	33	0.9024	-5852.1555	23.2106
10	MSZoning		11	37	0.9075	-5904.8779	22.2320
11	Fireplaces		12	38	0.9100	-5938.1907	21.6869
12	YearRemodAdd		13	39	0.9113	-5953.5284	21.3718
13	BldgType		14	43	0.9137*	-5967.8449*	20.8527*
* Optimal Value of Criterion							

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

The GLMSELECT Procedure
Selected Model

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

Effects:	Intercept OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF GrLivArea Fireplaces GarageArea Neighborhood MSZoning BldgType

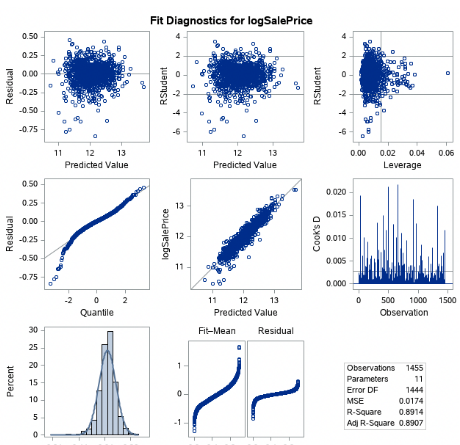
Analysis of Variance				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**
Model	42	212.16888	5.05164	367.44
Error	1412	19.41256	0.01375	
Corrected Total	1454	231.58143		

Root MSE	0.11725
Dependent Mean	12.02272
R-Square	0.9162
Adj R-Sq	0.9137
AIC	-4738.00361
AICC	-4735.19510
SBC	-5967.84488
CV PRESS	20.85269

Cross Validation Details			
	Observations		
Index	**Fitted**	**Left Out**	**CV PRESS**
1	1127	328	4.8335
2	1162	293	4.3304
3	1181	274	2.9367
4	1162	293	4.6440
5	1188	267	4.1082
Total			20.8527

Figure 1.25

```
/*Run Linear Regression Model for Stepwise Model: Numeric Only*/
proc glm data = train4 plots=all;
class Neighborhood MSZoning BldgType;
model logSalePrice = OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF
GrLivArea Fireplaces GarageArea/solution;
run;
```



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1.791906411	0.42026260	4.26	<.0001
OverallQual	0.074474340	0.00423837	17.57	<.0001
OverallCond	0.051891603	0.00375735	13.81	<.0001
YearBuilt	0.003108933	0.00019099	16.28	<.0001
YearRemodAdd	0.001317082	0.00024046	5.48	<.0001
BsmtFinSF1	0.000233747	0.00001163	20.09	<.0001
BsmtFinSF2	0.000183642	0.00002268	8.10	<.0001
BsmtUnfSF	0.000134749	0.00001158	11.64	<.0001
GrLivArea	0.000278044	0.00000949	29.29	<.0001
Fireplaces	0.055358544	0.00634399	8.73	<.0001
GarageArea	0.000222105	0.00002135	10.40	<.0001

Figure 1.26

```
/* Custom Selection Model*/
proc glmselect data = train4;
class Neighborhood MSZoning LotShape BldgType HeatingQC CentralAir  KitchenQual GarageType;
model logSalePrice = OverallQual OverallCond Yearbuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF
GrLivArea FullBath HalfBath BedroomAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF
OpenPorchSF EnclosedPorch ScreenPorch PoolArea YrSold Neighborhood MSZoning LotShape
BldgType HeatingQC CentralAir KitchenQual GarageType/ selection = Backward(stop = cv) cvmethod = random(5) CVDETAILS stats = adjrsq;
store BackwardTrainModel;
run;
```

Figure 1.27

The GLMSELECT Procedure

Backward Selection Summary						
Step	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0		30	71	0.9199*	-5902.6266	20.9330
1	GarageType	29	65	0.9194	-5929.3867*	20.7877*
* Optimal Value of Criterion						

Figure 1.28

```
/*Run Linear Regression Model for Custom Model: Numeric Only*/
proc glm data = train4 plots=all;
class Neighborhood MSZoning LotShape  BldgType  HeatingQC CentralAir KitchenQual LotConfig;
model logSalePrice = OverallQual OverallCond Yearbuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF
GrLivArea FullBath HalfBath BedroomAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF
OpenPorchSF EnclosedPorch ScreenPorch PoolArea YrSold Neighborhood MSZoning LotShape LotConfig
BldgType HeatingQC CentralAir  KitchenQual/solution;
run;
```

The GLMSELECT Procedure
Selected Model

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

Effects:	Intercept OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF GrLivArea Fireplaces GarageArea Neighborhood MSZoning BldgType

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	42	212.16888	5.05164	367.44
Error	1412	19.41256	0.01375	
Corrected Total	1454	231.58143		

Root MSE	0.11725
Dependent Mean	12.02272
R-Square	0.9162
Adj R-Sq	0.9137
AIC	-4738.00361
AICC	-4735.19510
SBC	-5967.84488
CV PRESS	20.85269

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	1127	328	4.8335
2	1162	293	4.3304
3	1181	274	2.9367
4	1162	293	4.6440
5	1188	267	4.1082
Total			20.8527

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	-2666.7593	231.8212
1	OverallQual		2	2	0.6743	-4292.6492	75.4169
2	GrLivArea		3	3	0.7627	-4746.9361	55.0155
3	Neighborhood		4	27	0.8331	-5108.4656	39.6195
4	BsmtFinSF1		5	28	0.8586	-5343.5834	33.5021
5	OverallCond		6	29	0.8708	-5468.1415	30.6572
6	YearBuilt		7	30	0.8840	-5619.4611	27.6907
7	GarageArea		8	31	0.8927	-5725.7323	25.6577
8	BsmtUnfSF		9	32	0.8979	-5792.9075	24.3155
9	BsmtFinSF2		10	33	0.9024	-5852.1555	23.2106
10	MSZoning		11	37	0.9075	-5904.8779	22.2320
11	Fireplaces		12	38	0.9100	-5938.1907	21.6869
12	YearRemodAdd		13	39	0.9113	-5953.5284	21.3718
13	BldgType		14	43	0.9137*	-5967.8449*	20.8527*
* Optimal Value of Criterion							

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

Figure 1.29(R-Code)

```
## Data cleanup
# replace null values with null values in train dataset
train$PoolQC[is.na(train$PoolQC)] = "None"
```

```
train$MiscFeature[is.na(train$MiscFeature)] = "None"
train$Alley[is.na(train$Alley)] = "None"
train$Fence[is.na(train$Fence)] = "None"
train$FireplaceQu[is.na(train$FireplaceQu)] = "None"
train$GarageType[is.na(train$GarageType)] = "None"
train$GarageFinish[is.na(train$GarageFinish)] = "None"
train$GarageQual[is.na(train$GarageQual)] = "None"
train$GarageCond[is.na(train$GarageCond)] = "None"
train$BsmtExposure[is.na(train$BsmtExposure)] = "None"
train$BsmtCond[is.na(train$BsmtCond)] = "None"
train$BsmtQual[is.na(train$BsmtQual)] = "None"
train$BsmtFinType1[is.na(train$BsmtFinType1)] = "None"
train$BsmtFinType2[is.na(train$BsmtFinType2)] = "None"
train$MSZoning[is.na(train$MSZoning)] = "None"
train$MasVnrArea[is.na(train$MasVnrArea)] = 0
train$LotFrontage[is.na(train$LotFrontage)] = 0
train$MasVnrType[is.na(train$MasVnrType)] = "None"
train$Electrical[is.na(train$Electrical)] = "None"

# rename colmnns to make it match with ones in test dataset
colnames(train)[44] <- "FirstFlrSF"
colnames(train)[45] <- "SecondFlrSF"

train$logSalePrice = log(train$SalePrice)

# remove outliers
train=train[!train$Id %in% c(524,643,725,1299,1424),]

# replace null values with null values in test dataset
test = test[,-c(60)]
test$PoolQC[is.na(test$PoolQC)] = "None"
test$MiscFeature[is.na(test$MiscFeature)] = "None"
test$Alley[is.na(test$Alley)] = "None"
test$Fence[is.na(test$Fence)] = "None"
test$FireplaceQu[is.na(test$FireplaceQu)] = "None"
test$GarageType[is.na(test$GarageType)] = "None"
test$GarageFinish[is.na(test$GarageFinish)] = "None"
test$GarageQual[is.na(test$GarageQual)] = "None"
test$GarageCond[is.na(test$GarageCond)] = "None"
test$BsmtExposure[is.na(test$BsmtExposure)] = "None"
test$BsmtCond[is.na(test$BsmtCond)] = "None"
test$BsmtQual[is.na(test$BsmtQual)] = "None"
test$BsmtFinType1[is.na(test$BsmtFinType1)] = "None"
```

```
test$BsmtFinType2[is.na(test$BsmtFinType2)] = "None"
test$MSZoning[is.na(test$MSZoning)] = "None"
test$SaleType[is.na(test$SaleType)] = "None"
test$BsmtFinSF2[is.na(test$SaleType)] = 0
test$BsmtUnfSF[is.na(test$BsmtUnfSF)] = 0
test$GarageArea[is.na(test$GarageArea)] = 0
test$GarageCars[is.na(test$GarageCars)] = 0
test$BsmtFinSF1[is.na(test$BsmtFinSF1)] = 0
test$BsmtFinSF2[is.na(test$BsmtFinSF2)] = 0
test$MasVnrArea[is.na(test$MasVnrArea)] = 0
test$LotFrontage[is.na(test$LotFrontage)] = 0
test$TotalBsmtSF[is.na(test$TotalBsmtSF)] = 0
test$BsmtFullBath[is.na(test$BsmtFullBath)] = 0
test$BsmtHalfBath[is.na(test$BsmtHalfBath)] = 0
test$MasVnrType[is.na(test$MasVnrType)] = "None"
test$Electrical[is.na(test$Electrical)] = "None"
test$Exterior1st[is.na(test$Exterior1st)] = "None"
test$Exterior2nd[is.na(test$Exterior2nd)] = "None"
test$Functional[is.na(test$Functional)] = "None"
test$Utilities[is.na(test$Utilities)] = "None"

colnames(test)[44] <- "FirstFlrSF"
colnames(test)[45] <- "SecondFlrSF"
```

Figure 1.30(R-Code)
Kaggle : 0.14255

```
forward_train = train[,c("OverallQual","GrLivArea",
"Neighborhood","BsmtFinSF1", "YearBuilt", "OverallCond",
"GarageArea","GarageCars",
                             "BsmtUnfSF","BsmtFinSF2", "MSZoning",
"Fireplaces", "YearRemodAdd", "BldgType",  "logSalePrice")]

#create dummy variables
forward_train = dummy_cols(forward_train, select = c("Neighborhood",
"BldgType", "MSZoning"), remove_selected_columns = T)

#fit model
forward_fit = lm(logSalePrice~., data = forward_train)
summary(forward_fit)

foward_test = test
```

```
forward_test = dummy_cols(test, select = c("Neighborhood", "BldgType",
"MSZoning"), remove_selected_columns = T)
forward_test$CentralAir = ifelse(forward_test$CentralAir == "Y", 1,0)

# Make predictions
forward_pred = predict(forward_fit,newdata=forward_test)
forward_pred = exp(forward_pred)
forward_test$SalePrice = forward_pred

#Write prediction into its own file
forward_predictions = forward_test %>% dplyr::select(Id,SalePrice)
write_csv(forward_predictions, "forwardmodel_predictions.csv")
```

Figure 1.31(R-Code)
Kaggle Score: 0.14206

```
backward_train = train%>% select(OverallQual, OverallCond, YearBuilt,
YearRemodAdd, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, GrLivArea, FullBath,
HalfBath, BedroomAbvGr, TotRmsAbvGrd, Fireplaces, GarageCars, GarageArea,
WoodDeckSF, OpenPorchSF, EnclosedPorch, ScreenPorch, PoolArea, YrSold,
Neighborhood, MSZoning, LotShape, LotConfig, Condition1, BldgType,
BsmtFinType1, HeatingQC, CentralAir, Electrical, KitchenQual, GarageType,
GarageFinish, SaleType,logSalePrice)

# create dummy variables
backward_train$CentralAir = ifelse(backward_train$CentralAir == "Y", 1,0)
backcat_var = colnames(backward_train[, sapply(backward_train, class) %in%
c('character', 'factor')])
backward_train = dummy_cols(backward_train, select = backcat_var,
                            remove_selected_columns = T)

#fit model
backward_fit = lm(logSalePrice~., data = backward_train)
summary(backward_fit)

backward_test = dummy_cols(test, select = backcat_var,
remove_selected_columns = T)
backward_test$CentralAir = ifelse(backward_test$CentralAir == "Y", 1,0)

#create missing columns in test dataset
backward_test$KitchenQual_Ex[is.na(backward_test$KitchenQual_Ex)] = 0
backward_test$KitchenQual_Fa[is.na(backward_test$KitchenQual_Fa)] = 0
```

```
backward_test$KitchenQual_Gd[is.na(backward_test$KitchenQual_Gd)] = 0
backward_test$KitchenQual_TA[is.na(backward_test$KitchenQual_TA)] = 0
backward_test$Electrical_Mix = 0
backward_test$Electrical_None= 0

#make prediction
backward_pred = predict(backward_fit,newdata=backward_test)
backward_pred = exp(backward_pred)
backward_test$SalePrice = backward_pred

backward_predictions = backward_test %>% dplyr::select(Id,SalePrice)
write_csv(backward_predictions, "backwardmodel_predictions.csv")
```

Figure 1.32(R-Code)
Kaggle: 0.14272

```
stepwise_train = train[,c("OverallQual","GrLivArea",
"Neighborhood","BsmtFinSF1", "YearBuilt", "OverallCond", "GarageArea",
                             "BsmtUnfSF","BsmtFinSF2", "MSZoning",
"Fireplaces", "YearRemodAdd", "BldgType",  "logSalePrice")]
stepwise_train = dummy_cols(stepwise_train, select = c("Neighborhood",
"BldgType", "MSZoning"), remove_selected_columns = T)

#fit model
stepwise_fit = lm(logSalePrice~., data = stepwise_train)
summary(stepwise_fit)

#create dummy variables
stepwise_test = dummy_cols(test, select = c("Neighborhood", "BldgType",
"MSZoning"), remove_selected_columns = T)

# make predictions
stepwise_pred = predict(stepwise_fit,newdata=stepwise_test)
stepwise_pred = exp(stepwise_pred)
stepwise_test$SalePrice = stepwise_pred

# write predictions into own file
stepwise_predictions = stepwise_test %>% dplyr::select(Id,SalePrice)
write_csv(stepwise_predictions, "stepwisemodel_predictions.csv")
```

Figure 1.33(R Code)

Kaggle: 0.14137

```
# select interested columns
custom_train = train%>% select(OverallQual, OverallCond, YearBuilt,
YearRemodAdd, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, GrLivArea, FullBath,
HalfBath, BedroomAbvGr, TotRmsAbvGrd, Fireplaces, GarageCars, GarageArea,
WoodDeckSF, OpenPorchSF, EnclosedPorch, ScreenPorch, PoolArea, YrSold,
Neighborhood, MSZoning, LotShape, LotConfig, Condition1, BldgType,
HeatingQC, CentralAir, Electrical, KitchenQual, GarageType, GarageFinish,
logSalePrice)

# create dummy variables
custom_train = dummy_cols(custom_train, select = c("Neighborhood",
"BldgType", "MSZoning", "LotShape", "LotConfig", "HeatingQC",
"KitchenQual"), remove_selected_columns = T)
custom_train$CentralAir = ifelse(custom_train$CentralAir == "Y", 1,0)

#fit model
custom_fit = lm(logSalePrice~., data = custom_train)
summary(custom_fit)

# copy test dataset
foward_test = test
custom_test = dummy_cols(test, select = c("Neighborhood", "BldgType",
"MSZoning", "LotShape", "LotConfig", "HeatingQC", "KitchenQual"),
                         remove_selected_columns = T)

# create missing columns in test dataset
custom_test$CentralAir = ifelse(custom_test$CentralAir == "Y", 1,0)
#custom_test$Electrical_Mix = 0
#custom_test$Electrical_None= 0
custom_test$KitchenQual_Gd[is.na(custom_test$KitchenQual_Gd)] = 0
custom_test$KitchenQual_Ex[is.na(custom_test$KitchenQual_Ex)] = 0
custom_test$KitchenQual_Fa[is.na(custom_test$KitchenQual_Fa)] = 0
custom_test$KitchenQual_Gd[is.na(custom_test$KitchenQual_Gd)] = 0
custom_test$KitchenQual_TA[is.na(custom_test$KitchenQual_TA)] = 0

#predict
custom_pred = predict(custom_fit,newdata=custom_test)
custom_pred = exp(custom_pred)
custom_test$SalePrice = custom_pred

custom_predictions = custom_test %>% dplyr::select(Id,SalePrice)
write_csv(custom_predictions, "custommodel_predictions.csv")
```