Here is my Github link: https://github.com/Cherrycoder22/Stats-RE-Project/tree/main

Introduction

For analysis one, a real estate company has contracted us to research the estimate sale price of the three neighborhoods: NAmes, Edwards, and BrkSide, and to see if there's a relationship between the sale price and square footage of the living area (Grlivingarea). The purpose of focusing on the three neighborhoods is to see if the sale price in relation to square footage can depend on where the house is located. The method of solving this problem is by building and fitting a model, provide an estimate with confidence intervals, along with a formal conclusion of our findings. At the end of this paper, we will be able to communicate the relationships between sale price, square footage, and neighborhoods. This will be effective for the real estate company to make informed decisions in the houses they sell.

For analysis two, we are focusing on getting the best Kaggle score by building a model that will predict sales prices in all neighborhoods in Ames, Iowa. First, building a linear regression model then a multiple linear regression model. With a table showing the adjusted R squared, CV press, and our final Kaggle score. We will be able to create the most accurate predictive model using tools such as R and SAS studio.

Data Description

The dataset comes from a Kaggle competition and is a large dataset with 2919 observations. However, for the purpose of this analysis, we will only focus on the sales price, square footage of the living area, total basement square footage, fireplaces, wood deck square footage, garage area, lot footage, overall quality and condition, and sale condition.

Analysis Question 1:

The purpose of our model will be to determine if the sale price has a relationship to the square footage of the living area and if that relationship can depend on the NAmes, Edwards, and BrkSide neighborhoods.

First, I filtered the dataset to focus on the three neighborhoods since the real estate company only sells in those three. As well as scaling the Grlivingarea in increments of 100 sq ft as requested by the real estate company. Our model was built with the sales price as the dependent variable and Grlivingarea and neighborhoods as the explanatory variables. Before starting the analysis of the estimates and predictions, we fitted the model into scatterplots, q-q plots, histogram, and a line chart. The graphs are referenced in the appendix, graphs 1,2, and 3.

Assumptions:

The original dataset has a lot of clustering which is not entirely a bad thing for our model. The purpose of reviewing the assumptions of linearity, normality, constant variance, and independence is to see if any of them are violated and to see if there's a need for any transformations. The histogram and line chart, there is sufficient evidence of linearity as well as normality. There is proof of constant variance by the residual plots and scatterplots. With the clustering in the scatterplots, there seems to be evidence of dependence but for the purpose of

this analysis, we will assume independence. There are a few influential points in the fit plot, one point that is more than 500 sq feet but only at about 150,000 sale prices. The other similar point is at around 145,000 sale price and around 450 sq feet. These are on the low end of sale price but high point for sq footage which is extreme compared to the majority of the observations. On the other extreme end, there are two points that are less than 300 sq feet but more than 350,000 sale prices.

After analyzing the original data, we wanted to research how a log transformation on Grlivingarea to compare which model may perform better before proceeding with testing the model. The log transformed graphs that are referenced are in the appendix (graphs 4, 5, and 6)

Assumptions:

With the log transformation, there doesn't seem to be a large difference in the plots. The assumptions are almost identical to the original data. The histogram and line chart, there is sufficient evidence of linearity as well as normality. There is proof of constant variance by the residual plots and scatterplots. With the clustering in the scatterplots, there seems to be evidence of dependence but for the purpose of this analysis, we will assume independence. The influtional points are the same as in the original data set. Since the log transformation doesn't change much about the model, we will proceed using the original data set to test our model.

Comparing Competing Models

The referenced numbers are in the parameter 1 table that's in the appendix.

Adj R^2

When performing three model selection (stepwise, forward, and backward), the result was consistent with all three. Pictured below is the relevant output. The adjusted R2 is only at 0.4474 which means the model is only accounting for 44.74% of the variance. The variance is at a moderate level which was most likely caused by the Grlivingarea rather than neighborhoods due to its high significance compared to the other variables. This suggests that there is a lot more room for improvement in our model and would need more than the neighborhoods and Grlivingarea as predictors.

AIC

The AIC is used to help compare models and to see which one would be best for the final predication model. In all the models, the AIC was at 8249.72388 and indicates that any model of choice is at equal accuracy to each other. It can also signify overfitting, but in this case that would be unlikely due to the filtering of the dataset. The most plausible cause of the AIC remaining the same would be the model is too simple as signified by the r squared. Since the AIC doesn't provide much information on which model to choose. We decided to proceed with no selection.

Parameters

Estimates

In this section, we will be referencing parameters 1 from the appendix. The intercept is in the Names neighborhood with an estimate of 74,676.40 when all other variables are constant and is highly significant indicated by the p-value <.0001. This can signal that this particular neighborhood has a relationship with the sale price. Also, as the intercept for the model, it creates a baseline for the predictions. With square footage of the ground living area the estimate is 543.15863 when neighborhoods are at zero and is also highly significant indicated by the p-value <.0001. With this significance, we can estimate that there is a relationship between sale price and square footage of the living area. In the neighborhood of Brkside, where Grlivearea and the two other neighborhoods are constant. The estimate is -54,704.88774 which indicates a negative relationship with the response variable. It is statistically significant since the p-value is <.0001. The Edwards neighborhood's estimate when all other variables are constant is 13676.70234 compared to the intercept and is least statistically significant with the p-value .1336 which is higher than our significance level of 0.005. This estimate shows that the Edwards neighborhood does not have a significant correlation with the sale price and square footage of the living area. When Grlivingarea is considered with the neighborhoods, the variance between the estimates decreases significantly. With Brkside, the estimate becomes 328.46670 and is also statistically significant (p-value = 0.0026). The Edwards neighborhood's estimate when Grlivingarea is considered is 63.61391 with a statistically significant p-value of 0.0001.

Interpretation

The estimates helped us interpret that Grlivingarea is the most significant predictor in this model, and we can estimate that the sale price can increase by 543.16 for every estimated increase in Grlivingarea. For the neighborhoods, only Names and Brkside had a significant correlation to the sales price. In Names or the intercept, for every additional house in this neighborhood, the estimated sales price increases by 74,676.40. In Brkside, for every additional house in this neighborhood, the estimated sales price decreases by 54,704.8. However, when Grlivingarea is a considered variable, sales price is estimated to increase by 328.47. The only significant estimate for Edwards was an estimated decrease of 245.66 with every estimated additional house in this neighborhood. This model is very simple and would need to add additional predictors for a more accurate prediction of the sale price, but there is sufficient evidence that Grlivingarea has a strong relationship with sales prices and certain neighborhoods can affect the price as well.

Confidence Intervals

The confidence intervals as pictured in the appendix (confidence interval 1) have some large differences particularly in the neighborhoods. However, zero is not in any of the intervals meaning that all variables are statistically significant in this model.

Conclusion

With square footage of the living area, and the neighborhoods Brkside and Names having significant p-values, there is significant evidence that there is a relationship between sale price and square footage of the living area along with certain neighborhoods can affect the sale price. The strongest predictor of sale price is the square footage of the living area suggesting that a larger living area can be associated with a higher sale price. The second highest predictor is the Names neighborhood, with those properties having a higher average sales price than those in Brkside which has an average sales price. However, this model only explains the 44% variance in the sales price which indicates there are other predictors besides neighborhood and square footage of the living area that can help determine sales price and improve the model's prediction power.

R Shiny: Price v. Living Area Chart

Below is my link to the RShiny app that can allow you to manipulate the price s the living area. It provides a deeper look into the potential sales prices if someone wants a specific living area.

http://cherokee32carr.shinyapps.io/RERShiny

Analysis Question 2

The purpose of this analysis is to build a predictive model that will predict sales prices in all neighborhoods in Ames, Iowa using a simple linear regression model, and two multiple linear regression model. Comparing the competing models to see which one provides the most accurate depiction of future sales prices in Ames, Iowa. Also, we will provide a table that will present the adjusted R squared, CV press, and our final Kaggle score. The first model that we built was a simple regression model with sales prices as the dependent variable and square footage of the living area. Before comparing the different models, we first checked assumptions in the graphs and plots. The residual plots, histograms, and scatter plots that are referenced are graphs 7,8, and 9 in the appendix.

The first multiple regression I performed had sales prices as the dependent variable, and square footage of the living room and full baths as the explanatory variables. The graphs and plots are pictured in the appendix (graph 10,11).

The second multiple regression model has the sales prices as the dependent variable and the explanatory variables are square footage of the living room, the garage area, number of fireplaces, total basement square footage, wood deck square footage, overall condition, lot frontage, overall quality, and sale condition. The graphs and plots are pictured in the appendix (graph 12).

Assumptions:

In the simple linear regression model, we are only using the Grlvingarea in relation to sales prices, the dataset still has a lot of clustering. The histogram and line chart, there is sufficient evidence of linearity as well as normality. There is proof of constant variance by the residual plots and scatterplots. With the clustering in the scatterplots, there seems to be evidence of dependence but for the purpose of this analysis, we will assume independence. There are a few influential points in the fit plot, two points are more than 4500 Grlivingarea but less than 200,000 sale prices. On the other end, there's two points that are about 4100 Grlivingarea but more than 700,000 sales prices. These outliers don't seem to affect the model in a drastic way, so we will keep them for our test.

The first multiple linear regression model, only focused on Grlvingarea and full bath in relation to sales prices, the dataset still has a lot of clustering. The histogram and line chart, there is sufficient evidence of linearity as well as normality. There is proof of constant variance by the residual plots and scatterplots. With the clustering in the scatterplots, there seems to be evidence of dependence but for the purpose of this analysis, we will assume independence. The influential points are the same as in the previous model. For full baths, there are a few in homes that have 3 full bathrooms.

In the second multiple linear regression model, we selected square footage of the living room, the garage area, number of fireplaces, total basement square footage, wood deck square footage, overall condition, lot frontage, overall quality, and sale condition in relation to sales prices helps offer a fuller picture. The histogram and line chart, there is sufficient evidence of linearity as well as normality. There is proof of constant variance by the residual plots and scatterplots. With the clustering in the scatterplots, there seems to be evidence of dependence but for the purpose of this analysis, we will assume independence. For these variables, it would be best to see how a log transformation would change the plots.

Comparing Competing Models

To see which multiple linear regression model will perform better, we performed a stepwise selection focusing on the AIC and CV press for both models. First comparing the adjusted R squared to see which model explains the variance better. The referenced numbers are in parameters 2, 3, and 4.

Adj R^2

The simple regression's model adjusted R2 is 0.5018 which means the model is only accounting for 50.18% of the variance. Since the variance is at a moderate level and only half of the model is accurate. It suggests that we can add more predictors to our model, so we can get a better prediction of the sales price.

In the first multiple regression model, the adjusted r squared is at 0.5231 or the variance of the model is at 52.31%. This about the same level of variance with only Grliving in the model which means full baths doesn't help build the model and is not a strong predictor of sales price. However, in the second multiple regression model, the adjusted R squared is 0.7604. The

variance of the model is much larger at 76.04% and this is after a log transformation of the garage area and total basement square footage.

Internal CV Press

The cv press for the simple linear regression model is extremely high at 4.615811E12 which means the simple regression model is most likely inaccurate. To reduce it, we can remove the outliers, add predictors, and also perform a transformation on them. We will research this portion further in the multiple regression models.

The cv press for the first multiple regression model is extremely high at 4.430873E12 which means this model is still too simple and potentially underfitted. In the second multiple linear regression model, the cv is 2.068395E12 which is still high but somewhat lower than the first model.

AIC

The simple linear regression model's AIC was at 33392 and indicates that any model of choice is at equal accuracy to each other. It can also signify underfitting which would align with the model since there's only one explanatory variable. In the first multiple linear regression model, the AIC is 33330 while in the second model the AIC is 29857. Since the second model's AIC is lower than the other models, we decided to proceed with the second multiple regression model and a stepwise selection.

Table

Predictive Models	**Adjusted R squared**	**CV Press**	**Kaggle Score**	**AIC**
Simple Linear	0.5018	4.615811E12		**33392**
MLS 1	0.5231	4.430873E12		**33330**
MLS2	0.7604	2.068395E12		**29857**

Assumptions with Log transformation:

After deciding to focus on the second multiple regression model, we log transformed the garage area and total basement area. The histogram and line chart, there is sufficient evidence of linearity as well as normality. There is proof of constant variance by the residual plots and scatterplots. With the clustering in the scatterplots, there seems to be evidence of dependence but for the purpose of this analysis, we will assume independence. For these variables, it would be best to see how a log transformation would change the plots. As referenced in graph 13.

Parameters

Estimates

For Grlivingarea, the estimate is 44.0483 and is statistically significant with the p-value >.0001. The log transformed garage area's estimate is 28853.9641 and the log transformed total basement square footage area's estimate is 42061.6458. For fireplaces, the estimate is 7497.8872. With the overall condition, the estimate is 4966.2297 and overall quality's estimate is 26197.4885. These are all statistically significant since all p-values are <.0001.

Interpretation

The Grlivingarea estimate remained the strongest throughout the analysis, with every estimated additional square footage, there's an estimated 44.05 increase in sales prices. For every estimated increase in garage area, there's an estimated 28,853.96 increase in sales price. For every estimated increase in total basement square footage area, there's an estimated increase of 42,061.65 increase in sales price. For every estimated additional fireplace, there's an estimated increase of 7,497.89 increase in sales prices. For every estimated increase in overall condition, there's an estimated increase of 4,966.23 in sales price. For every estimated increase in overall quality, there's an estimated increase of 26,197.49 in sales price.

Confidence Intervals

The confidence intervals as pictured in the appendix (confidence interval 2) have much smaller margins. The lowest is the Grlivingare and total basement area. There is one variable that is not statistically significant which is the lot frontage since zero is the interval.

Conclusion: In our simple regression model, it reenforced that Grliving area is a significant predictor of sales price. The key to getting a better model was shown in our second multiple regression model where we added more predictors and performed a log transformation. In that model, all variables were statistically significant. The main issues in our models were the cv press and adjusted r squared since they indicated more predictors may improve the model's accuracy.

Appendix

SAS CODE

```
FILENAME REFFILE '/home/u61516948/test.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=RETest; GETNAMES=YES; RUN;

FILENAME REFFILE '/home/u61516948/train.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV OUT=RETrain; GETNAMES=YES; RUN;

/*Join train and test sets */

data RE;

set RETrain RETest;

run;

/*Filter Neighborhoods*/

Data RE_filtered;

Set RE; If Neighborhood in ('NAmes', 'Edwards', 'BrkSide');

 Run;

/*Scale sq footage8/

 Data RE_scaled;

Set RE_filtered;

 GrLivAreaScaled = GrLivArea/10;

Run;

proc print data= RE_scaledE; run;
```

```
/*Analysis 1*/

data RE_filtered;

set RE;

/* Keep only the specified neighborhoods */

 if Neighborhood in ('NAmes', 'Edwards', 'BrkSide');

 run;

data RE1;

data RETrain;

/*Creat new column with filtered neighborhood names */

 data RE2;

set RE_scaled;

if Neighborhood = 'NAmes' then Neighborhood1= 'NAmes';

else if Neighborhood = 'Edwards' then Neighborhood2 = 'Edwards';

else if Neighborhood = 'BrkSide' then Neighborhood3 = 'BrkSide';

else Neighborhood_num = .; / Missing value for other categories */

run;

/*First model*/

 proc glmselect data=RE_scaled;

class Neighborhood;

model SalePrice = GrLivAreaScaled / selection=none /* Stepwise selection based on AIC /
cvmethod=split(10) / 10-fold cross-validation / stats=all; / Detailed statistics for each split
*/ run;

/*Regression and plots */
```

```
proc glm data= RE2;
 class Neighborhood;
model SalePrice = GrLivAreaScaled|Neighborhood/ solution;
run;
proc glm data=RE2 plots=all;
class Neighborhood;
 model SalePrice = GrLivAreaScaled;
run;
/*Analysis 2*/
 /*Simple LR*/
proc glm data= RE; class Neighborhood;
 model SalePrice = GrLivArea|Neighborhood/ solution;
run;
proc glm data=RE plots=all;
   model SalePrice = GrLivArea;
Run;
/*MLS1*/
proc glmselect data= RE_scaled;
 class Neighborhood; model SalePrice = GrLivAreaScaled FullBath / selection=stepwise
(stop= CV) /* Stepwise selection based on AIC / cvmethod=split(10) / 10-fold cross-
validation / stats=AIC; / Detailed statistics for each split */
run;
proc glm data=RE_scaled plots=all;
```

```
model SalePrice = GrLivAreaScaled FullBath;

 run;


/*MLS 2 and log transform*/

data RE_log;

set RE;

logGarageArea = log(GarageArea);

logTotalBsmtSF = log(TotalBsmtSF);

run;

 proc glmselect data= RE_log;


class Neighborhood;

 model SalePrice = GrLivArea logGarageArea logTotalBsmtSF Fireplaces OverallCond
OverallQual / selection=stepwise (stop= CV) /* Stepwise selection based on AIC /
cvmethod=split(10) / 10-fold cross-validation / stats=AIC; / Detailed statistics for each split
*/ run;

proc glm data=RE_log plots=all;

 model SalePrice = GrLivArea logGarageArea logTotalBsmtSF Fireplaces OverallCond
OverallQual;

run;
```

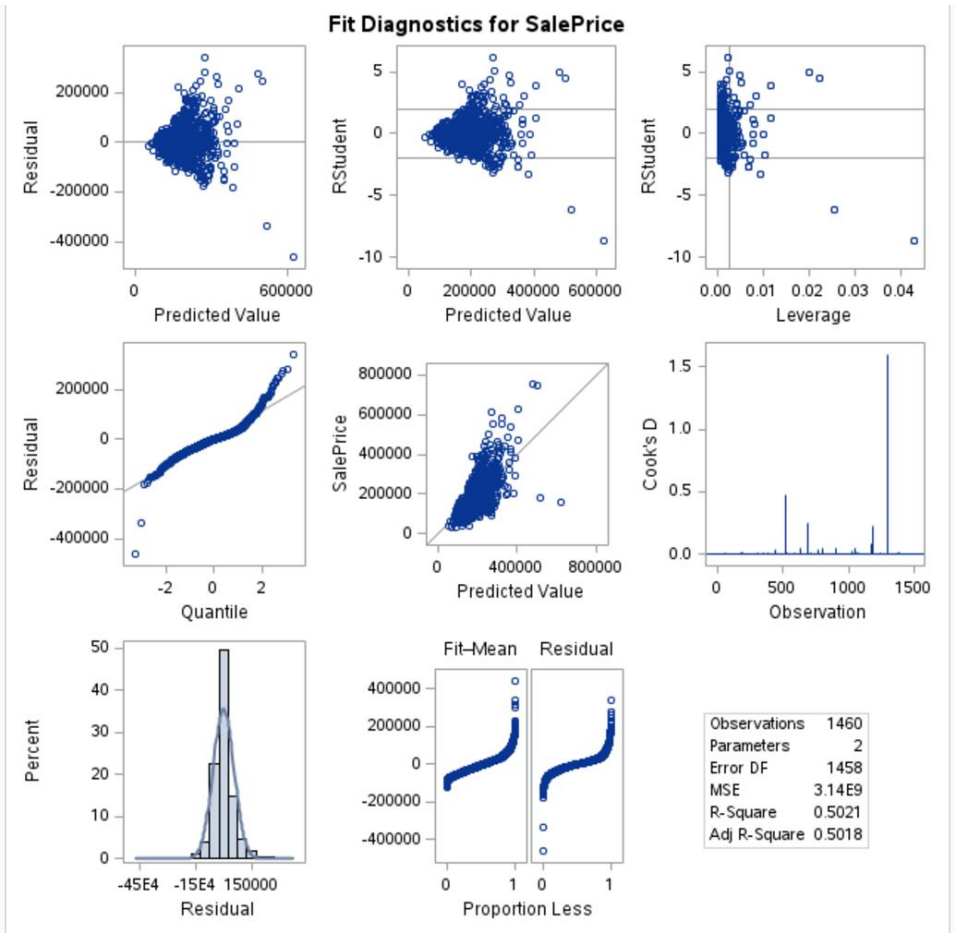
R CODE

```
RE <-read.csv("/Users/cherokeecarr/Documents/house-prices-advanced-regression-
techniques/train.csv")

print(RE)
```

```
na.omit(RE)

library(dplyr)          # Load dplyr package


RE1<- RE %>% filter(Neighborhood %in% c("Edwards", "BrkSide","NAmes"))


RE1$Neighborhood <- factor(RE1$Neighborhood)


RE1$Neighborhood <- relevel(RE1$Neighborhood, ref = "NAmes")


RE1$GrLivArea <- log(RE1$GrLivArea)


fit <- lm(SalePrice ~ GrLivArea + Neighborhood, data = RE1 )

summary(fit)

aic_value <- AIC(fit)


# Print the AIC value

print(aic_value)

plot(fit)


fit <- lm(SalePrice ~ GrLivArea, data = RE1 )

summary(fit)

RE$GrLivArea <- log(RE$GrLivArea)
```

```
RE$LotFrontage <- log(RE$LotFrontage)


# Fit the model again

fit <- lm(SalePrice ~ GrLivArea+GarageArea+
LotFrontage+TotalBsmtSF+Fireplaces+OverallCond+OverallQual, data = RE_clean)

summary(fit)


confint(fit)
```

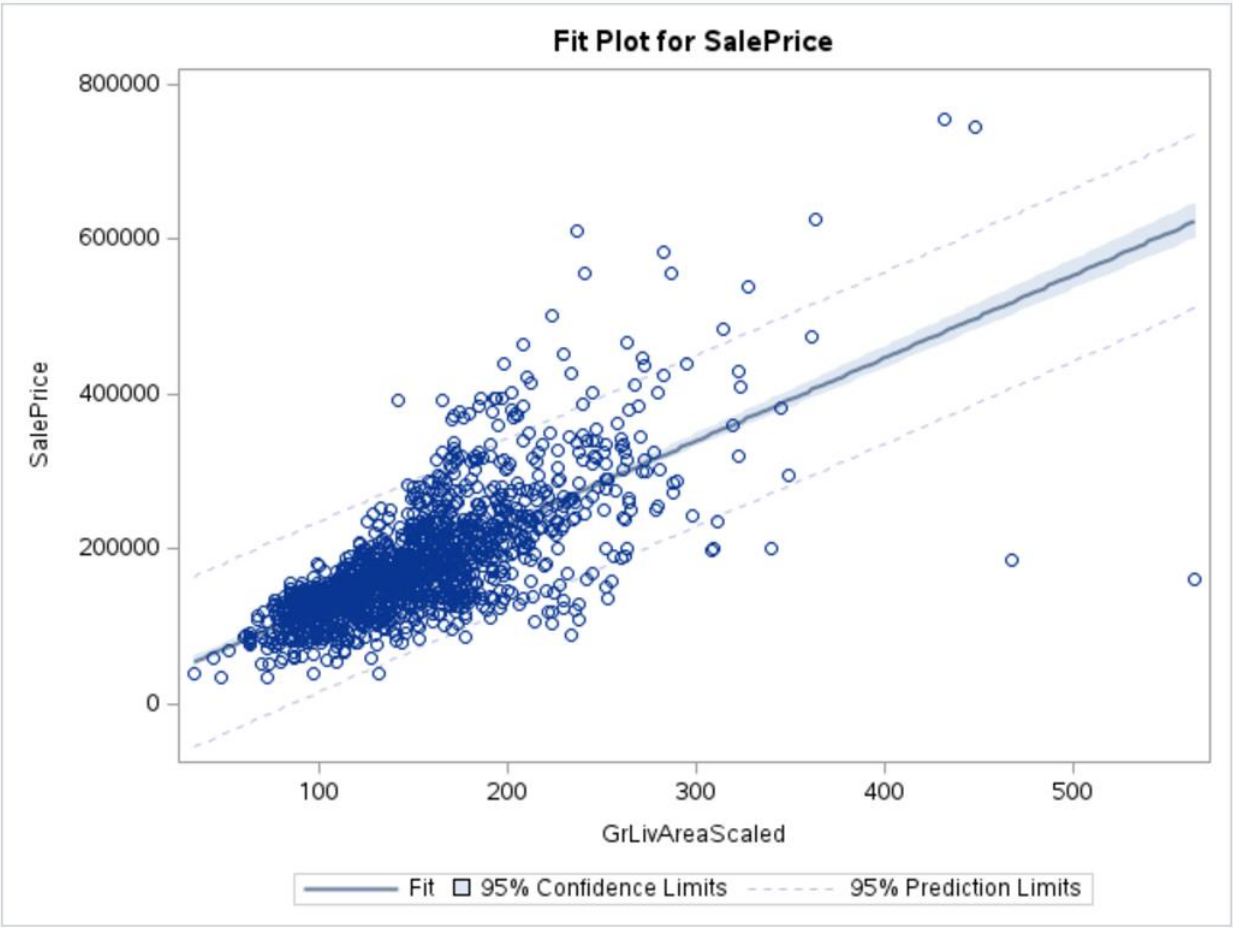
Graph 1


Fit Diagnostics for SalePrice
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
SalePrice
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 1460
Parameters 2
Error DF 1458
MSE 3.14E9
R-Square 0.5021
Adj R-Square 0.5018

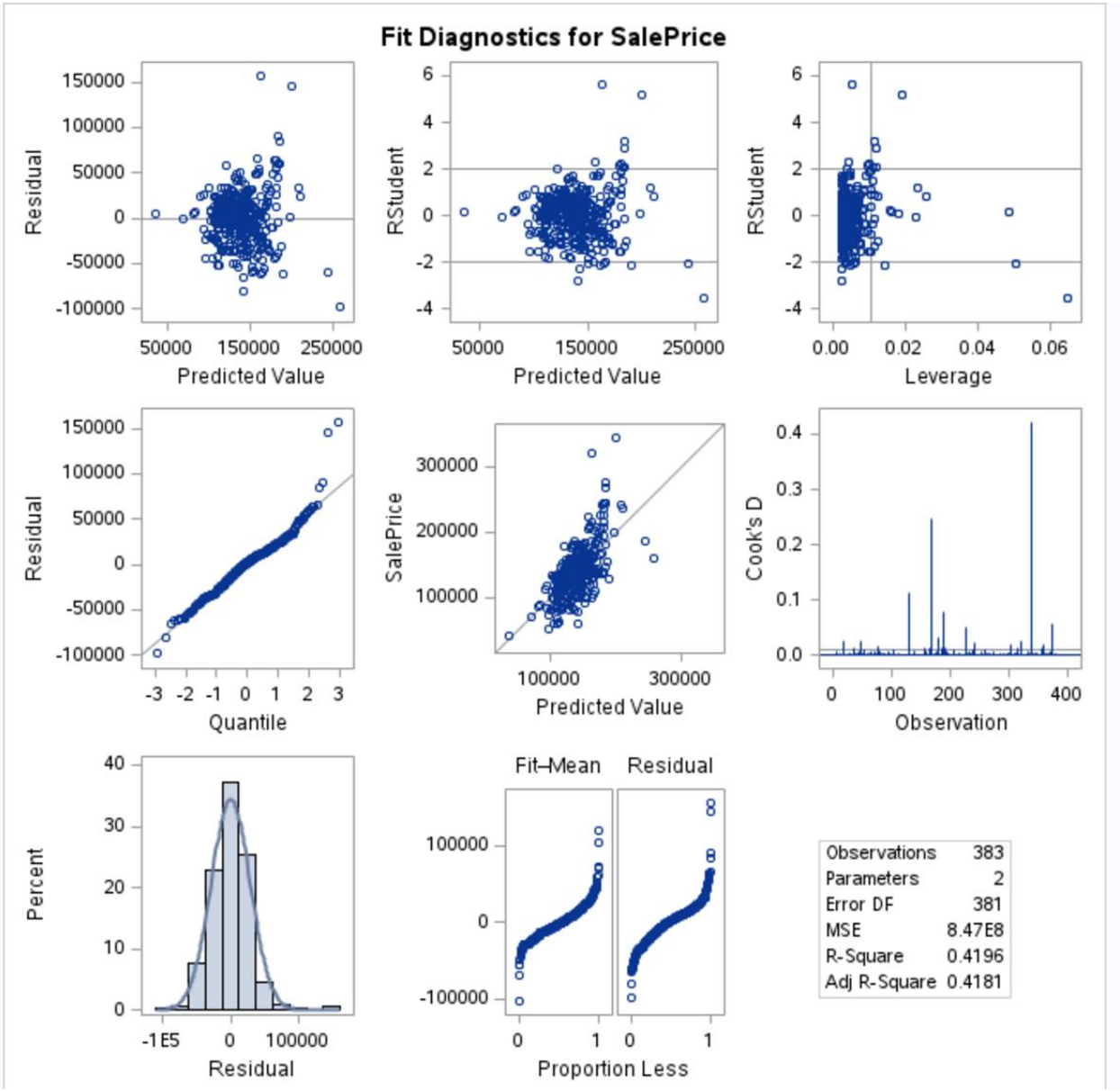
Graph 2


Residual Plot for SalePrice
Residual
GrLivAreaScaled

Graph 3



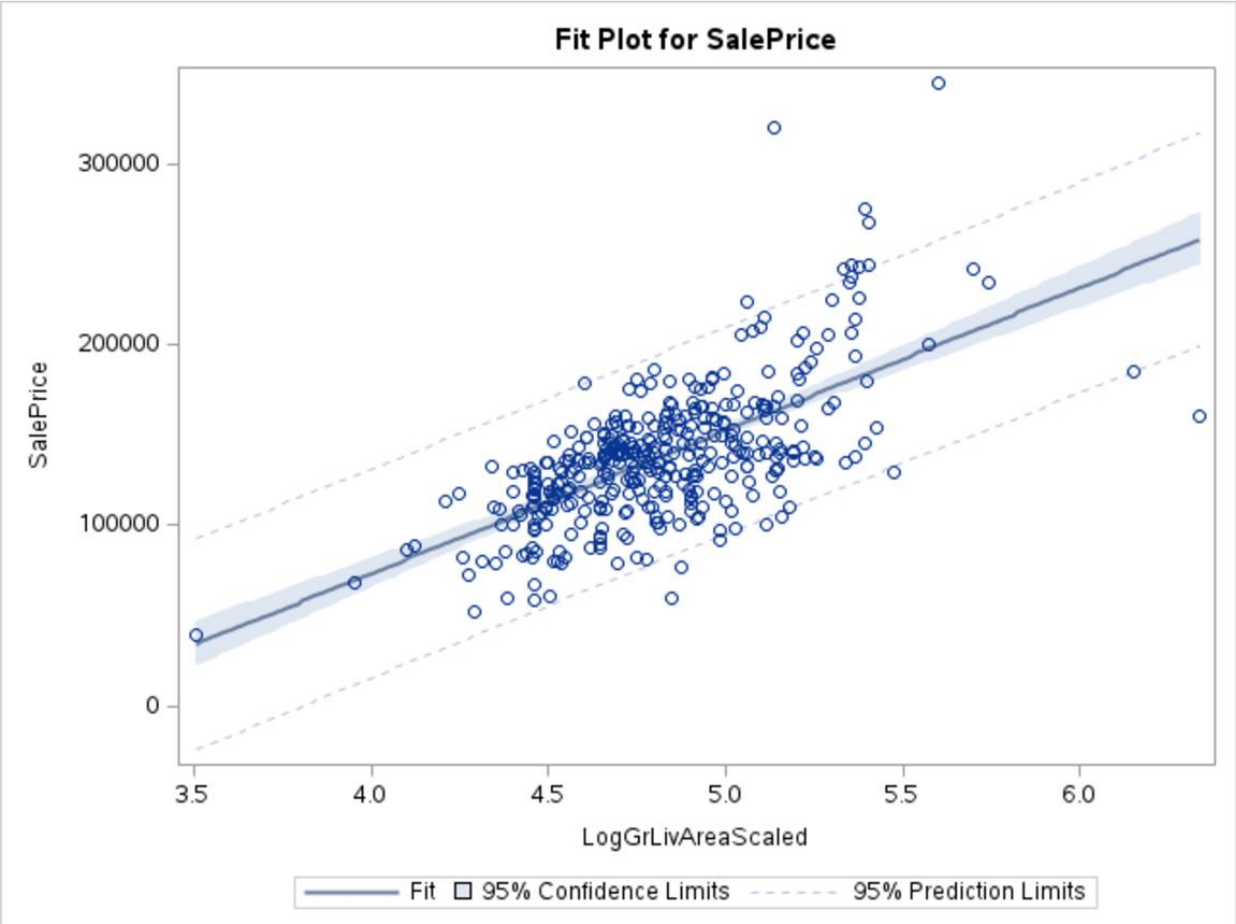
Graph 4


Fit Diagnostics for SalePrice
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
SalePrice
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 383
Parameters 2
Error DF 381
MSE 8.47E8
R-Square 0.4196
Adj R-Square 0.4181

Graph 5


Residual Plot for SalePrice
Residual
150000
100000
50000
0
-50000
-100000
3.5
4.0
4.5
5.0
5.5
6.0
LogGrLivAreaScaled

Graph 6



Parameters 1

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	5	2.48809E11	49761808941	61.04
Error	377	3.073431E11	815233644	
Corrected Total	382	5.561521E11		

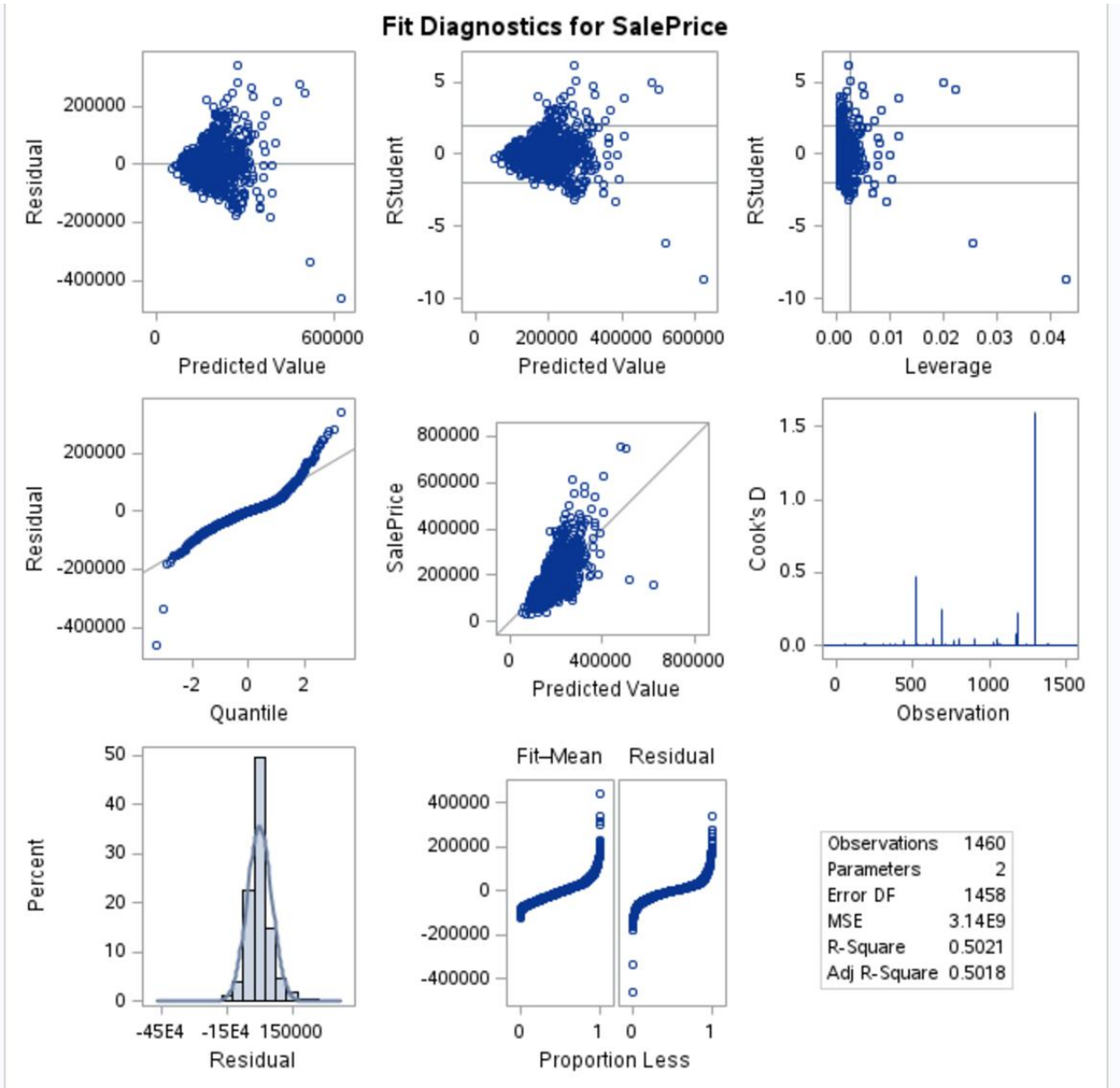
Root MSE	28552
Dependent Mean	138063
R-Square	0.4474
Adj R-Sq	0.4400
AIC	8249.72388
AICC	8250.02255
SBC	7888.41209

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	74676	6337.893993	11.78
Neighborhood BrkSide	1	-54705	13882	-3.94
Neighborhood Edwards	1	13677	9097.574652	1.50
Neighborhood NAmes	0	0	.	.
GrLivArea*Neighborho BrkSide	1	871.625326	97.819580	8.91
GrLivArea*Neighborho Edwards	1	297.503024	43.796861	6.79
GrLivArea*Neighborho NAmes	1	543.158627	46.136363	11.77

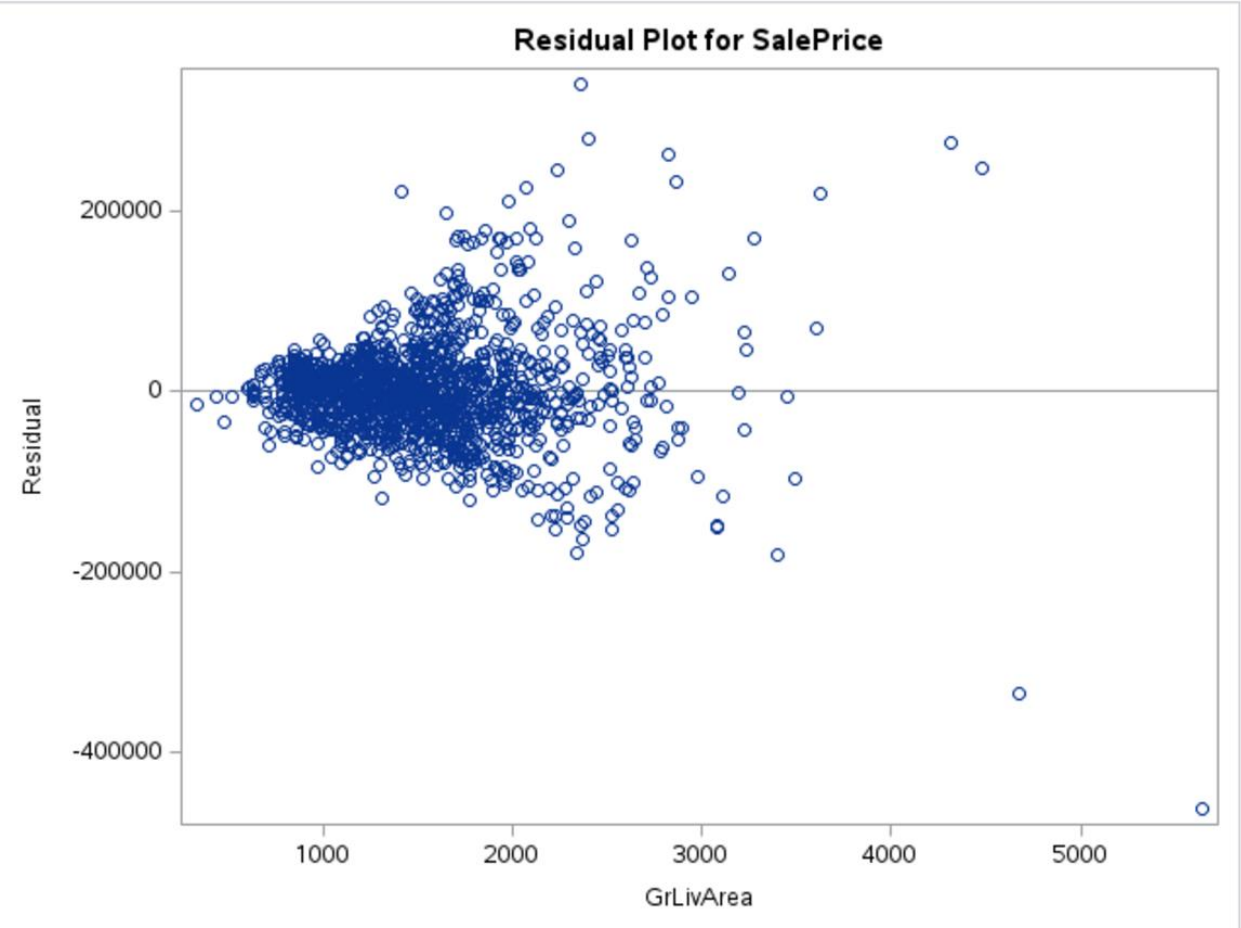
Confidence Intervals 1

```
> confint(fit)
                              2.5 %       97.5 %
(Intercept)             76885.46901  94888.84915
GrLivArea                  39.56876     51.95137
NeighborhoodBrkSide -24747.95040  -7463.29116
NeighborhoodEdwards -26022.65245 -11952.89930
>
```

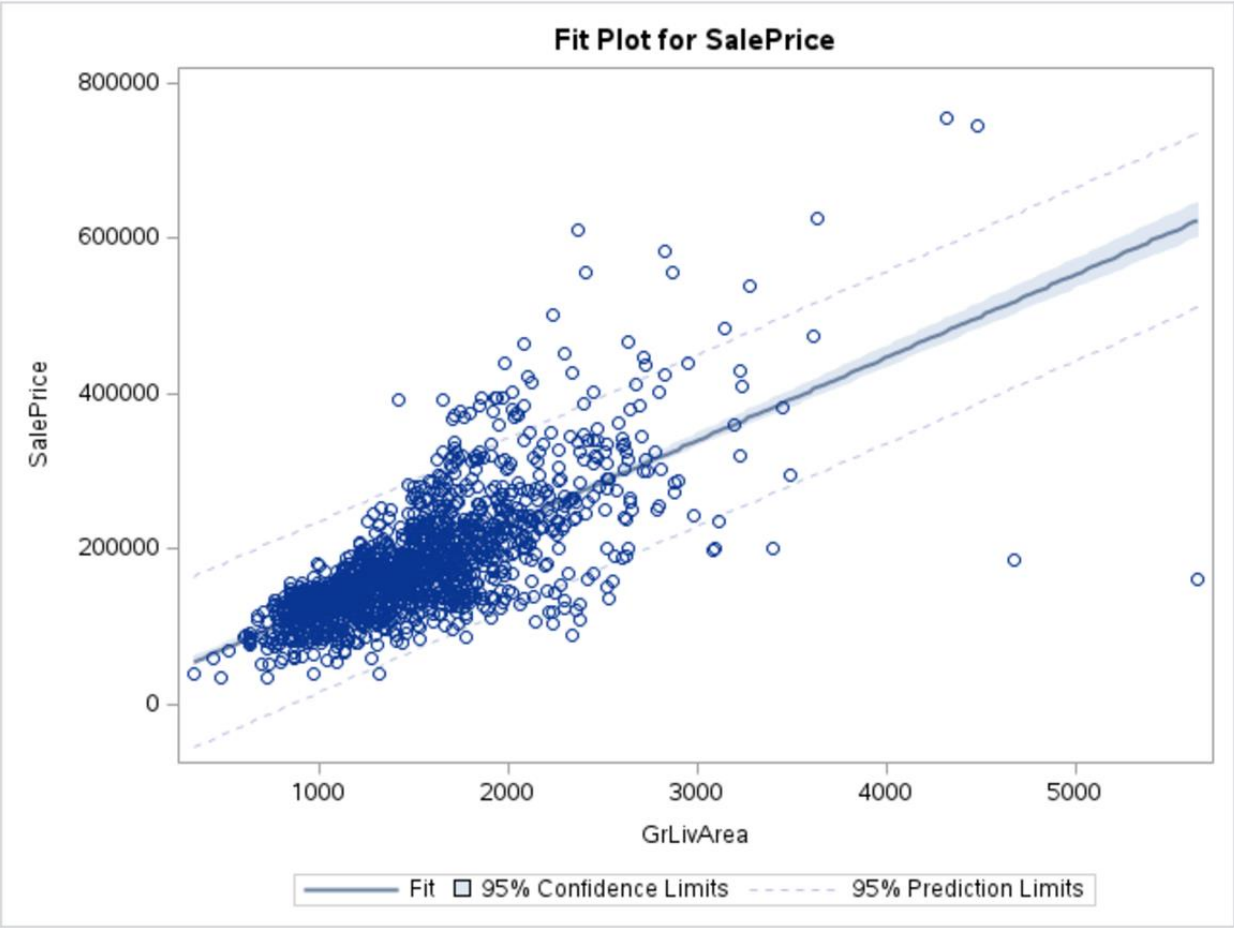
Graph 7



Graph 8



Graph 9



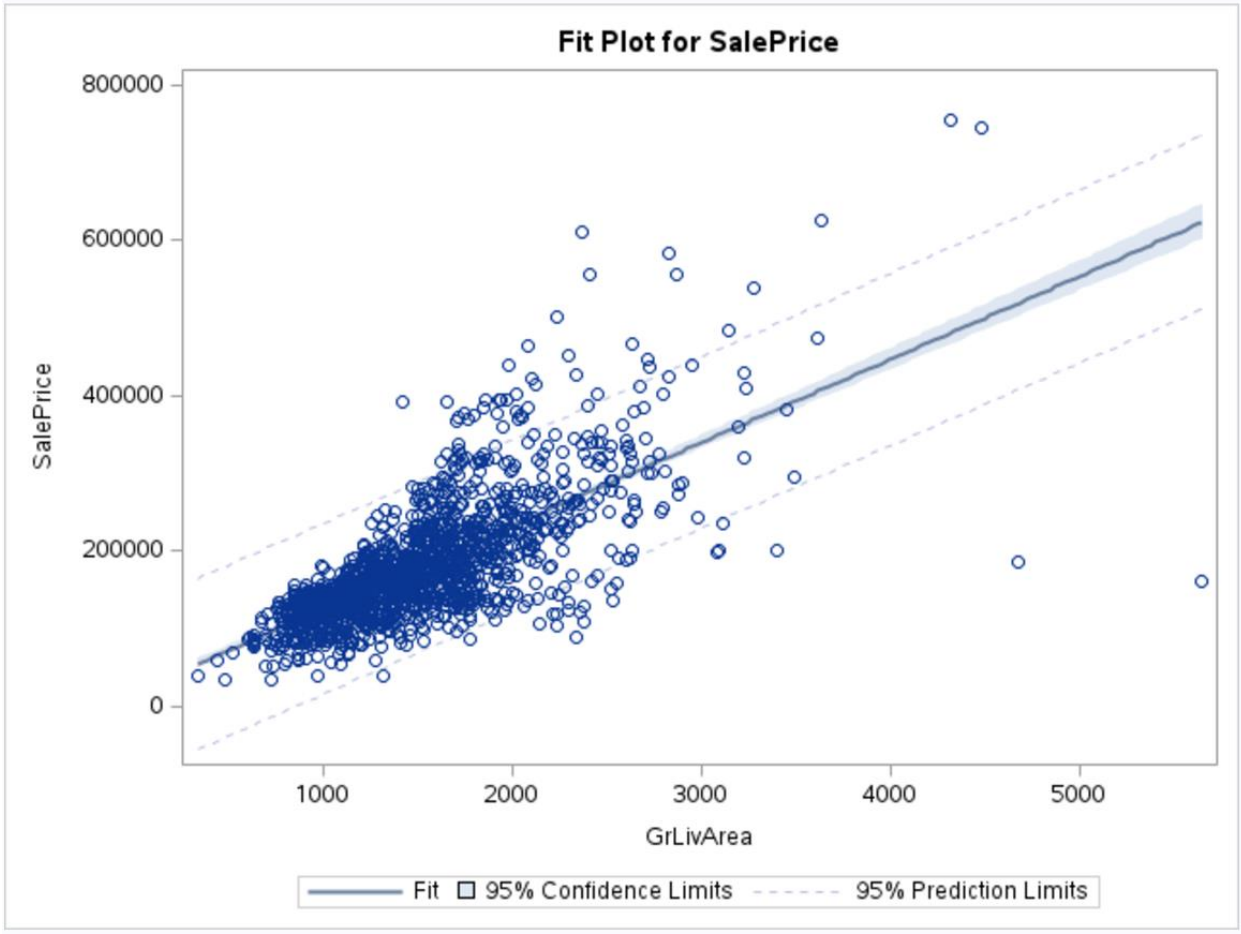
Model 1

Root MSE	56073
Dependent Mean	180921
R-Square	0.5021
Adj R-Sq	0.5018
AIC	33392
AICC	33392
SBC	31941
CV PRESS	4.615811E12

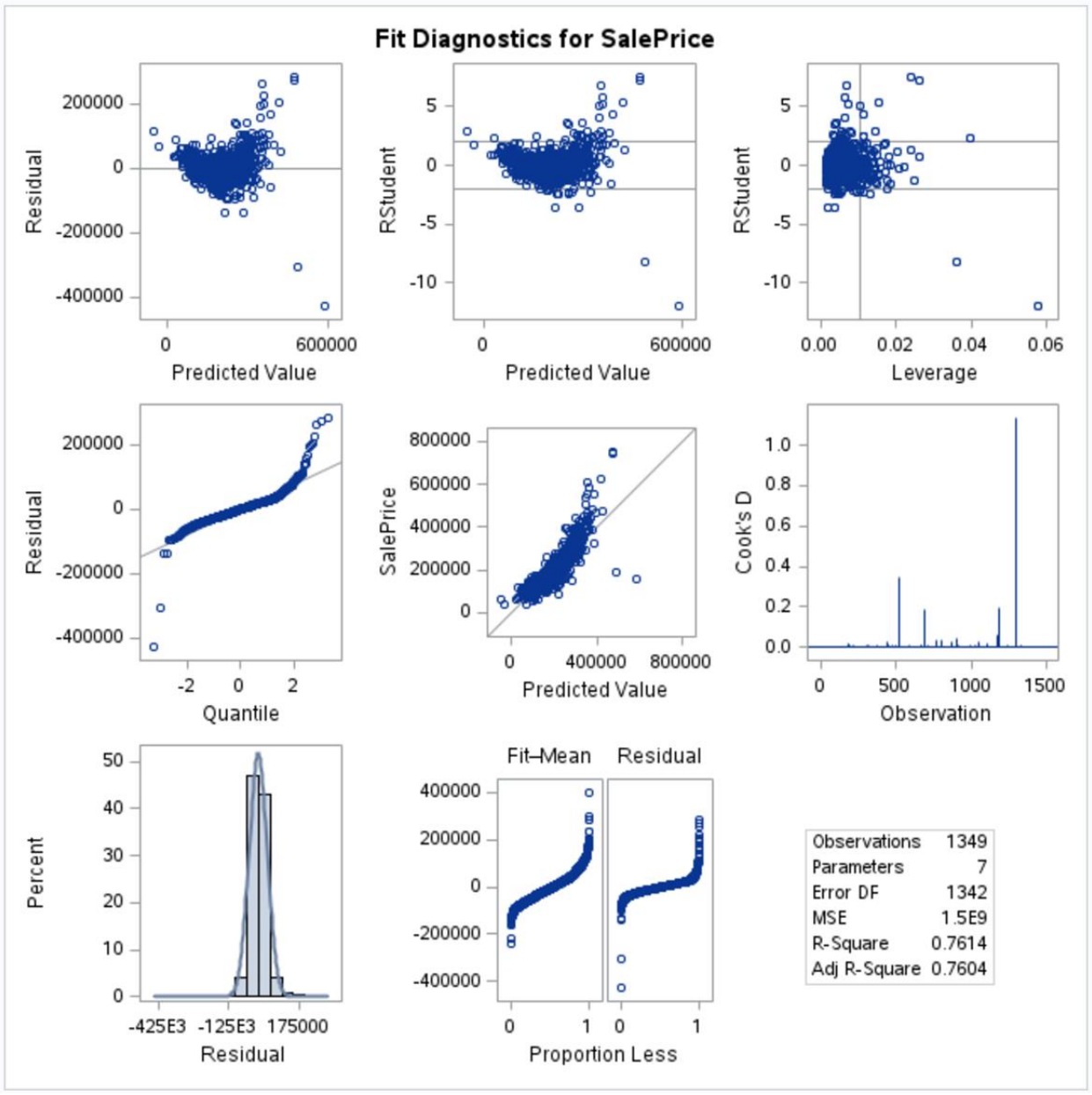
Graph 10


Residual Plots for SalePrice
Residual
100000
0
-100000
100
200
300
400
500
GrLivAreaScaled
0.0
0.5
1.0
1.5
2.0
2.5
3.0
FullBath

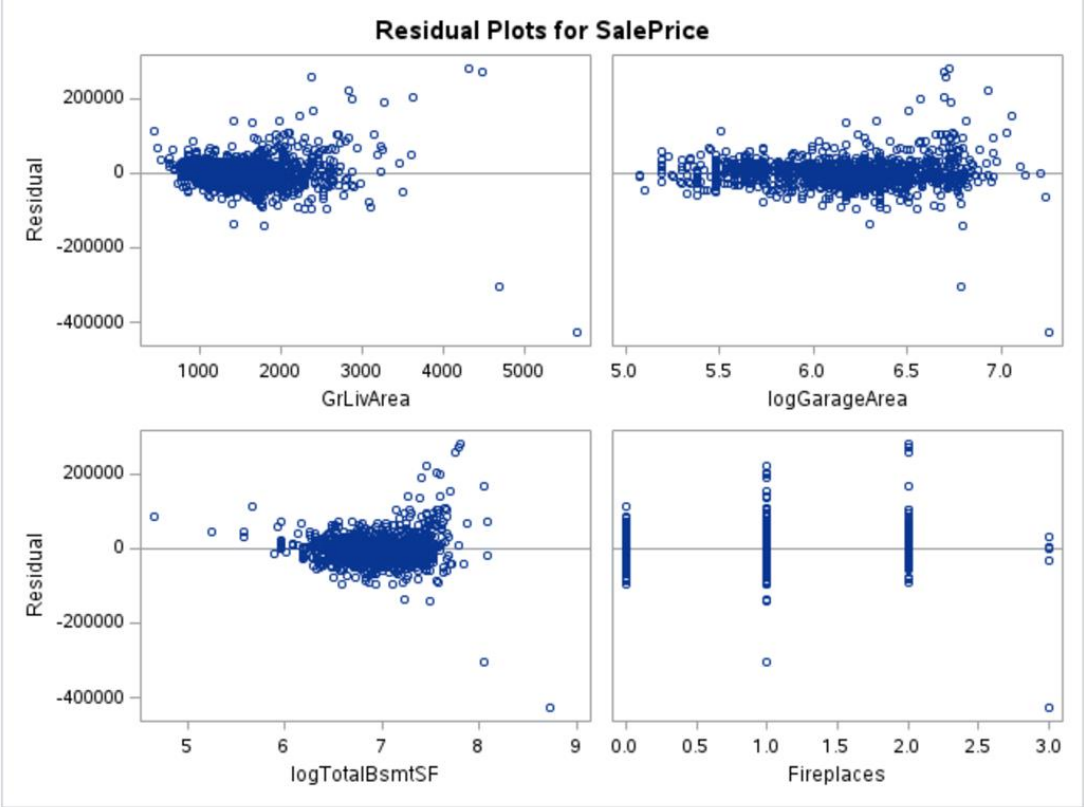
Graph 11

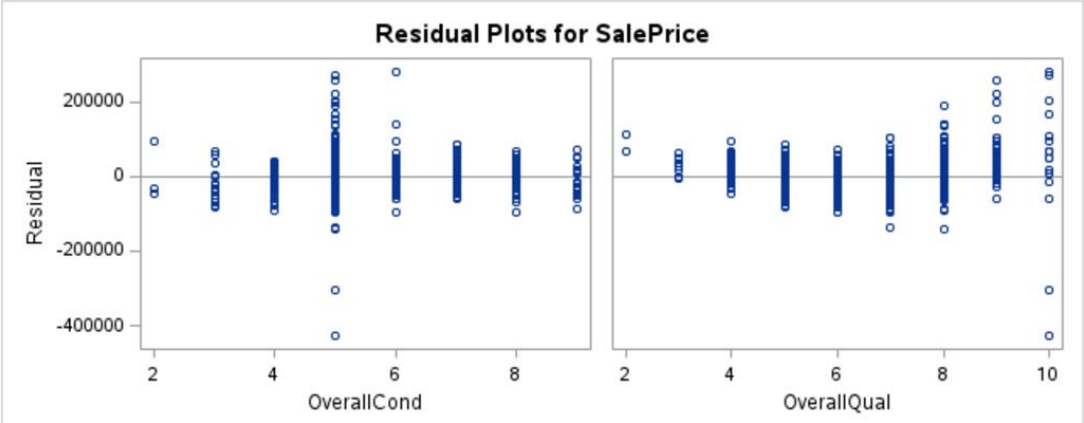

Fit Plot for SalePrice
800000
600000
400000
200000
0
SalePrice
1000
2000
3000
4000
5000
GrLivArea
Fit
95% Confidence Limits
95% Prediction Limits

Graph 12


Fit Diagnostics for SalePrice
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
SalePrice
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 1349
Parameters 7
Error DF 1342
MSE 1.5E9
R-Square 0.7614
Adj R-Square 0.7604

Graph 13





Parameter 2

Root MSE	56073
Dependent Mean	180921
R-Square	0.5021
Adj R-Sq	0.5018
AIC	33392
AICC	33392
SBC	31941
CV PRESS	4.615811E12

Parameter 3

Root MSE	54860
Dependent Mean	180921
R-Square	0.5238
Adj R-Sq	0.5231
AIC	33330
AICC	33330
SBC	31883
CV PRESS	4.430873E12

Parameter 4

Root MSE	38682
Dependent Mean	187119
R-Square	0.7614
Adj R-Sq	0.7604
AIC	29857
AICC	29857
SBC	28543
CV PRESS	2.068395E12

Confidence Interval 2

```
> confint(fit)
                    2.5 %       97.5 %
(Intercept) -539564.20976 -427343.23247
GrLivArea     50340.62483   66991.39211
GarageArea       49.91087      74.20074
TotalBsmtSF      26.21369      37.85622
Fireplaces     6106.99976   13421.01765
OverallCond    1482.89057    5256.18311
OverallQual   22430.93542   26698.49019
```