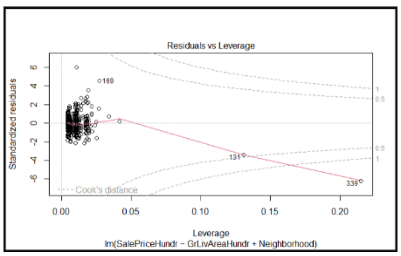
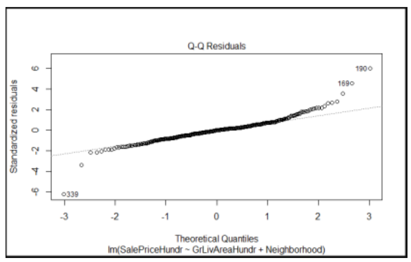
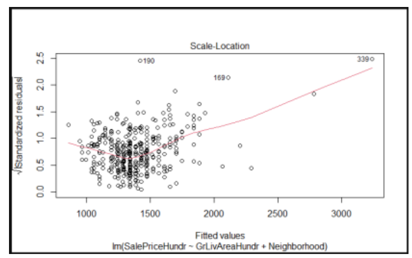
Analysis 1 Summary: Century 21 Ames is seeking to analyze the relationship of the square foot living area and the sale prices of houses in the North Ames, Edwards, and Brookside neighborhoods in Ames, Iowa to improve their customer service and competitive edge.

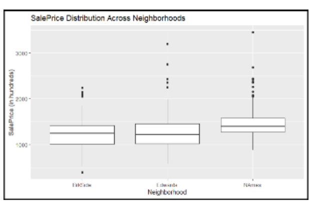
Problem: Using multiple linear regression analysis estimate if there is a difference in the price compared to the living space square footage in the North Ames, Edwards, and Brookside neighborhoods and how those neighborhoods statistically compare with each other.

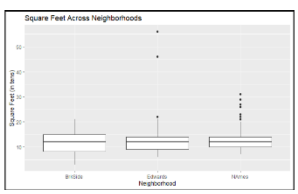
Assumptions Check: An initial visual check of ***normality*** using histograms and QQ-plots show some right skewedness, probably resulting from outliers with strong leverage (See graphics below). Because the sample size is greater than 30 the Central Limit Theorem (CLT) is expected to resolve inconsistencies in normality. Nonetheless, we will log transform some of the data to check for improved fit. A visual review of the residuals also shows ***some variance across the line*** despite a good variance inflation factor (VIF) for sale price and living space (1.007636, 1.003811). The variables appear to be ***independent*** and have good ***linearity***. The primary issue with this test is the ***outliers***. Because the outliers have so much leverage, we opted to review these extremes to check for human error and other anomalies (see leverage and boxplots below).

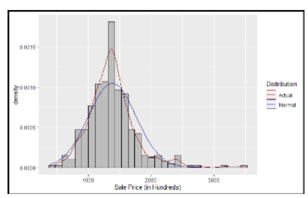


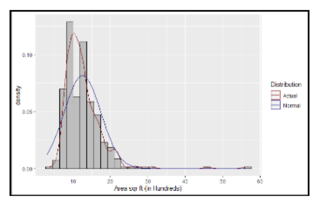








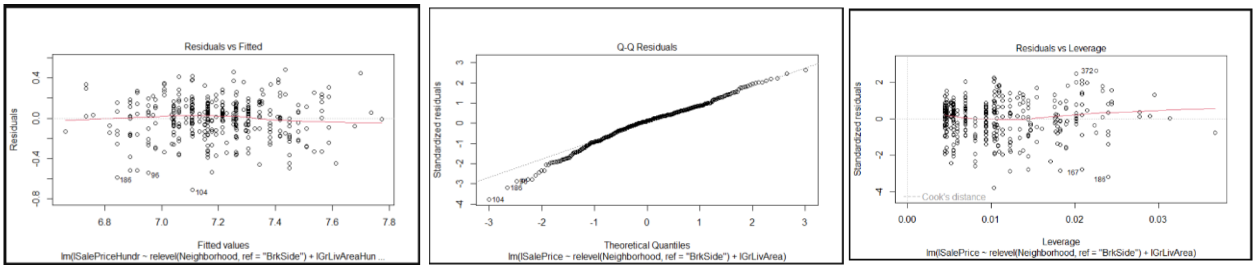


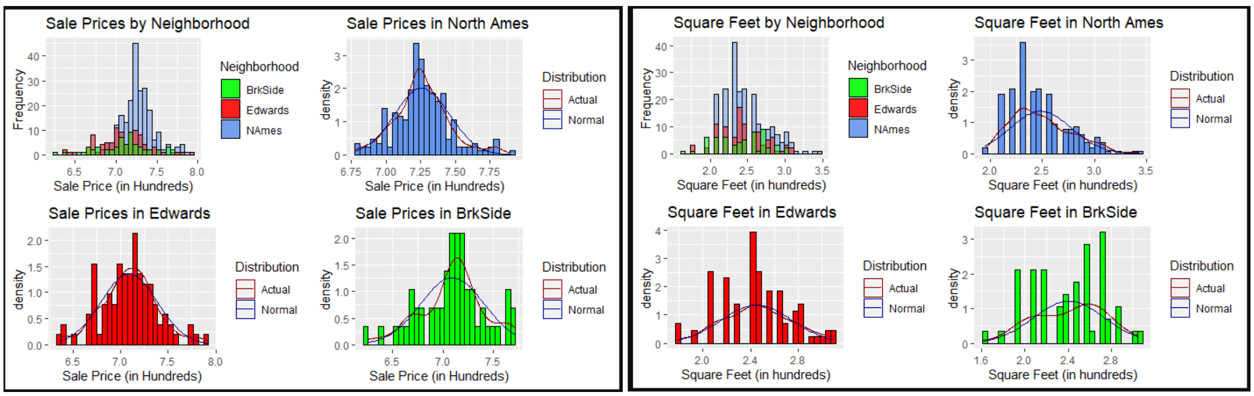


Influential Outliers: A review of the outliers showed that IDs 524 and 1299 were new construction and in an incomplete status. We use Zillow to review the properties in Edwards associated with these IDs and concluded that the square footage exceeded any properties in that neighborhood and the low prices were probably related to lot price rather than a newly constructed home. Additionally, ID 534 in Brookside was assessed to be missing a zero on both sales (39300) and square footage (324 sqft). Lastly, ID 725 in Edwards had an extraordinary sale price (320000) for the living space (1698) and condition (5). All these outliers were determined to be the result of human error or reporting requirements for planned construction and were removed from the model.

Comparing Competing models: All models in analysis 1 used one generalized formula (salePrice ~ GrLivArea + Neighborhood) adjusted for the logged variable. We tested a multiple linear regression (r^2 = 0.5081), log-linear (r^2 = 0.4785), linear-log (r^2 = 0.4897), log-log (r^2 = 0.4903), and a logged interaction (r^2 = 0.5094). The internal CV press for the log-log model returned a fairly good RMSE of 0.189923, suggesting strong performance, an r^2 of 0.49082, suggesting decent predictive performance, and a MAE of 0.145467 indicating only minor difference between the predicted and actual values.

Log-Log transformations: We opted to go with the Log-Log model with no interaction, primarily because it best met the assumptions. As you can see below, the log-log transformation and removed outliers improved ***normality***, ***linearity***, the ***residual distribution***, and ***variance*** across most neighborhoods, despite some right skewedness for living space in North Ames.





Parameters and Interpretation of the model:

```
-----------------------------Neighborhood no interaction---------------------

Call:
lm(formula = lSalePrice ~ relevel(Neighborhood, ref = "BrkSide") +
    lGrLivArea, data = AmesHousing_Data3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72335 -0.10484  0.02247  0.11660  0.48696

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   7.66029    0.24041  31.863  < 2e-16 ***
relevel(Neighborhood, ref = "BrkSide")Edwards -0.02755    0.03150  -0.874    0.382
relevel(Neighborhood, ref = "BrkSide")NAmes    0.12412    0.02809   4.419 1.3e-05 ***
lGrLivArea                                    0.57229    0.03388  16.891  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

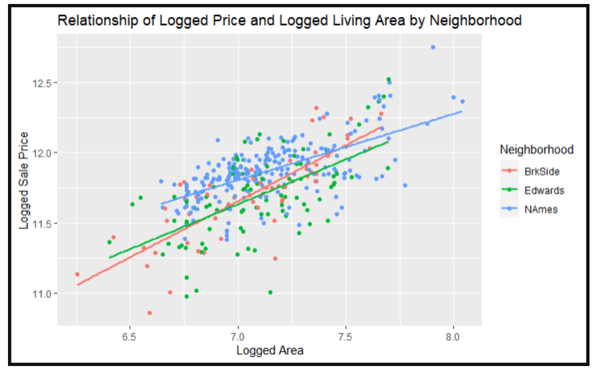
Residual standard error: 0.1886 on 375 degrees of freedom
Multiple R-squared:  0.4903,	Adjusted R-squared:  0.4863
F-statistic: 120.3 on 3 and 375 DF,  p-value: < 2.2e-16

                                                    2.5 %    97.5 %
(Intercept)                                   7.18756764 8.1330200
relevel(Neighborhood, ref = "BrkSide")Edwards -0.08949560 0.0344004
relevel(Neighborhood, ref = "BrkSide")NAmes    0.06889019 0.1793571
lGrLivArea                                    0.50566510 0.6389101
```

An analysis of the North Ames, Edwards, and Brookside neighborhoods shows relatively no statistically significant difference in the relationship of sales price and living area between Brookside (the reference neighborhood) and Edwards neighborhoods (p-value = 0.382). Although a purchase in the Edwards neighborhood is associated with a 0.02755 (or $97 non-logged) decrease in log-transformed price, this difference is not significant. We are 95 percent confident that the true logged sales price difference is found between (-0.0894956 and 0.0344004).

However, there is strong evidence (p-value < .0001) that the relationship between sales price and living area is different between Brookside and North Ames. For North Ames, there is evidence, holding all other variables constant, that purchasing in that neighborhood is associated with an estimated 0.12412 ($113 non-logged) increase in the log-transformed sale price for each unit of living space compared to Brookside. We are 95 percent confident that the true logged difference in the purchasing price between North Ames and Brookside is between (0.06889 and 0.179357).

Regarding the relationship between sales price and living area across all values, there is overwhelming evidence (p-value <0.0001) of a positive linear relationship. For each log-transformed unit of living space, the log-transformed sales price is increased on average by 0.57229 units. In real world terms, that is for every 100 square feet of living space there is an estimated increase of $171 in the base purchase price.

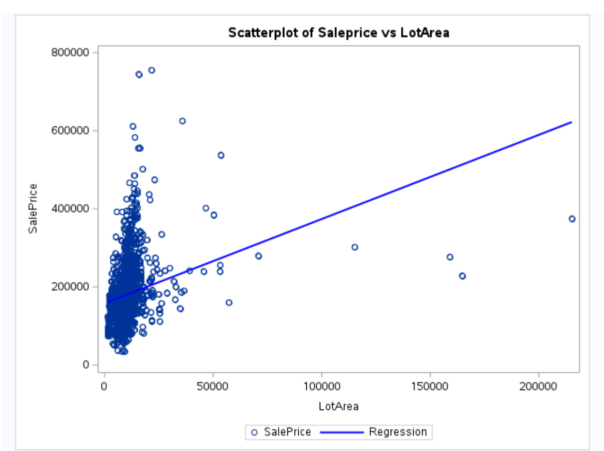


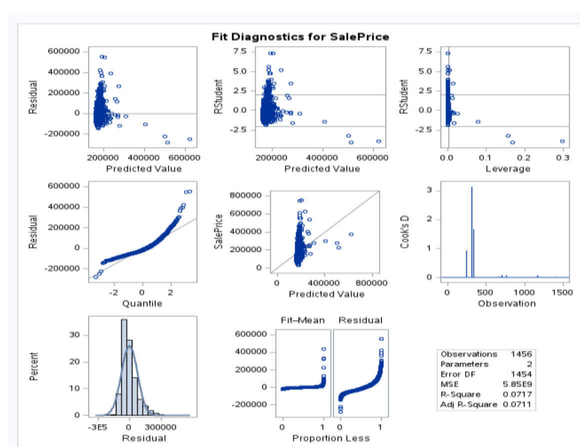
Conclusion: There is overwhelming evidence (p-value <0.0001) of a positive linear relationship between the sales price and the square footage of living space. However, this relationship does not mean there are significant differences between North Ames, Edwards, and Brookside neighborhoods in purchasing price and living space. A breakdown of the price to square footage between Edwards and Brookside shows no significant difference. Our best estimate was that square footage cost $97 dollars less in Edwards than Brookside, but this was not statistically significant. In North Ames, however, there was strong evidence that square footage cost $113 more per hundred square feet than the same living space in Brookside. This model explained only 49.03 percent of the variation and other variables should be considered to strengthen the model. (See Appendix 1)

Analysis 2 Summary: Expanding on the request from the customer, we will seek to build the most predictive model for the sale prices of homes in all of the neighborhoods in Ames Iowa.

The Problem: Build three models; a simple linear regression model, a multiple linear regression model comparing sales price, above ground living Area (GrLivArea), and number of full baths, and a custom multiple linear regression model that will show the relationship of the most predictive variables. These models will be compared to each other using metrics such as adjusted R squared, CV press and Kaggle score.

Model 1: Simple Regression – Examining the relationship between Salesprice and lot area.



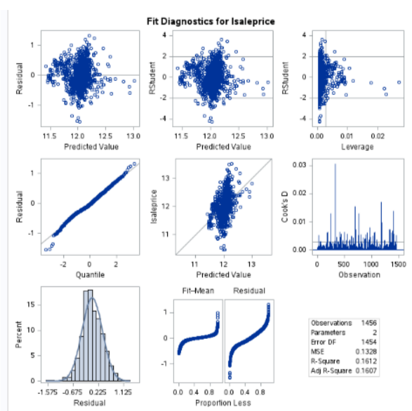


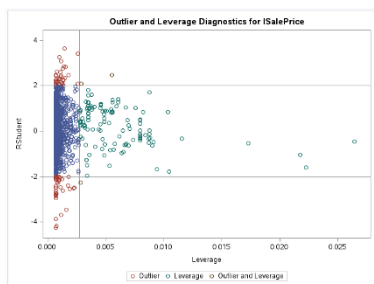
Assumptions Check

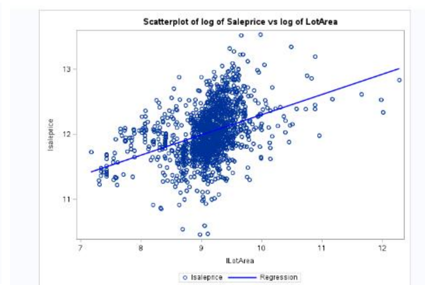
An initial visual review of the plots between saleprice and lot area shows some right skewedness in the histogram and q-q plot, suggesting the assumption for normal distribution is not met. Additionally, there appears to be clustering of the residuals and several outliers with heavy leverage, suggesting that there is not equal spread. In regard to the sample, we assume that the observations are independent and that the data represents the entire single family housing population in Ames.

Log-Log Transformation

To mitigate for the lot size outliers (cases with 2 to 5 acres) we will perform a log-log transformation on the sales price and lot area and examine the relationship between the two logged variables. (Note: Limited information on the cost associated with acreage precluded an analysis that would have allowed us to drop these most egregious outliers. Instead of ranged exclusion, we opted for log-log transformations.) By first plotting the log - log transformed data, there is a better fitting linear relationship between the logged saleprice (response variable) and the explanatory variable (Lot area). Furthermore, a first look at the graphics seem to meet the assumptions. There does not appear to be any potential influential point. Judging from the scatter plot, q-q plot and histogram of the residuals, there is no evidence that the residuals do not follow a normal distribution with constant variance. We continue to assume here that the observations are independent.







The REG Procedure
Model: MODEL1
Dependent Variable: lsaleprice

Number of Observations Read	1456
Number of Observations Used	1456

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	37.13003	37.13003	279.52	<.0001
Error	1454	193.14466	0.13284		
Corrected Total	1455	230.27468			

Root MSE	0.36447	R-Square	0.1612
Dependent Mean	12.02455	Adj R-Sq	0.1607
Coeff Var	3.03103		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	9.19260	0.16966	54.18	<.0001	8.85980	9.52540
lLotArea	1	0.31091	0.01860	16.72	<.0001	0.27443	0.34739

Pred log(Saleprice) = $\beta 0$ + $\beta 1$ log(LotArea)

Pred log(Saleprice) = 9.193 + 0.311log(LotArea)

We are 95% confident that for each doubling of the lot area, the median Saleprice rate will increase between approximately (2^0.27) 20.6% and (2^0.35) 27.5%. Our best estimate is an increase of (2^0.31) 24%.

Model 2: Multiple Regression with SalePrice~GrLivArea + Full Bath

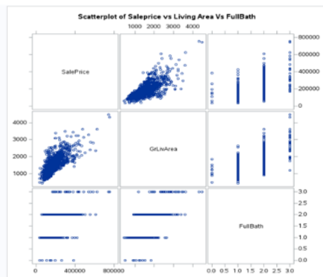
By plotting the data first, we can tell there is some evidence of a positive relationship between the explanatory variables and the response variable. Please see appendix 2 for additional visual plots.

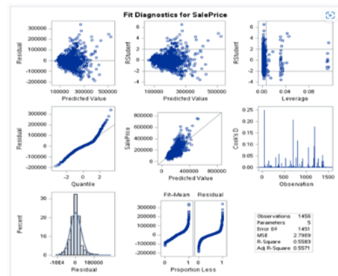
Assumptions check:

There does appear to be an influential point in the Rstudent plot. That observation will require further investigation. Judging from the scatter plot, q-q plot and histogram of the residuals, there is no evidence that the residuals do not follow a normal distribution but there is not enough evidence to indicate a constant variance. We assume here that the observations are independent.

Log – Log Transformation

We will do a log transformation of Saleprice and living area and plot those variables against each other to see if there is a linear relationship between them. Please see appendix 2 for plot.





Looking at the scatterplot matrix there is a positive linear relationship between the explanatory variables (Living area and full bath) and the response variable (Sales price). The log transformed variables show a more positive linear relationship between lsaleprice and lGrLivArea than the original data. We will build and fit a model with the log transformed data.

Assumptions Check

There do appear to be three influential points. That observation will require further investigation. Judging from the scatter plot, q-q plot and histogram of the residuals, there is no evidence that the residuals do not follow a normal distribution and a constant variance.
We assume here that the observations are independent.

We fit our model with ran on our explanatory (lsaleprice) and dependent (lgrlivArea and Full bath). This model has an Adjusted R square of 0.5637 which means this model estimates about 56.37% of the variation in salesprice is explained by the explanatory variables, a SBC of -3860.2586, and a CV press of 101.20 . Please see appendix for more details on the model.

The next step is to build a model with the log of sale price as the explanatory variable and fit the model.

$$pred\{lsaleprice\} = \beta\_0 + \beta\_1\ lgrlivArea + \beta\_2 FB0 + \beta\_3\ FB1 + \beta\_4\ FB2 \text{ Reference} = FB3$$
$$pred\{lsaleprice\} = 6.9797 + 0.723\ lgrlivArea - 0.2424FB0 - 0.2992FB1 - 0.142\ FB2$$
$$pred\{lSaleprice \mid lGrLivArea, FullBath=0\} = 6.9797 + 0.723\ lgrlivArea - 0.2424FB0 = 6.9209 + 0.723\ lgrlivArea$$
$$pred\{lSaleprice \mid lGrLivArea, FullBath=1\} = 6.9797 + 0.723\ lgrlivArea - 0.2992FB1 = 6.6805 + 0.723\ lgrlivArea$$
$$pred\{lSaleprice \mid lGrLivArea, FullBath=2\} = 6.9797 + 0.723\ lgrlivArea - 0.142\ FB2 = 6.8377 + 0.723 lgrlivArea$$
$$pred\{lSaleprice \mid lGrLivArea, FullBath=3\} = 6.9797 + 0.723\ lgrlivArea$$

We are 95% confident that for each doubling of the living area, the median Saleprice rate will increase between approximately 59% (2^0.67 = 1.59) and 71% (2^0.78=1.71). Our best estimate is an increase of 65 % (2^0.72 = 1.647) after holding all other variables (Full Bath) constant. To get the regression equation for a specific full bath number, we will need to adjust that full bath's coefficient with the intercept in our regression model.

Custom Model

For the custom model, we decided to go with an automatic selection technique and then finetune the model with manual selection techniques. We did this by running stepwise, forward and backward selection models on the explanatory variables in the housing dataset and our response variable (Saleprice). The stepwise method made the most sense with the highest R score and lowest CV press. (See the graphic for Adjusted R square, SBC, and CV press). So, for our custom model we chose the stepwise technique.

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	-2060.4605	176.3250
1	OverallQual		2	9	0.7033	-3370.7019	52.9725
2	lgrlivArea		3	10	0.7794	-3695.9381	39.5055
3	MSSubClass		4	24	0.8339	-3929.2796	30.7557
4	YearBuilt		5	25	0.8481	-4023.1361	28.1341
5	OverallCond		6	32	0.8735	-4185.9537	23.3779
6	BsmtFullBath		7	33	0.8852	-4288.7106	21.4082
7	lLotArea		8	34	0.8934	-4365.2298	19.8787
8	TotalBsmtSF		9	35	0.8994*	-4423.6168*	18.9170*

* Optimal Value of Criterion

Because our model was so large, we opted to reduce the number of variables manually by removing one variable at a time and retesting the model. We will fit our final model (lSalePrice ~lGrLivArea + OverallQual + MSSubClass + GarageCars + YearRemodelAdd + Neighborhood). Several of our variables were categorical and their coefficients require individual assessment, which we review in appendix 2.

The GLM Procedure

Dependent Variable: lsaleprice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	53	202.1354604	3.8138766	190.02	<.0001
Error	1402	28.1392236	0.0200708		
Corrected Total	1455	230.2746840			

R-Square	Coeff Var	Root MSE	lsaleprice Mean
0.877801	1.178184	0.141671	12.02455

Source	DF	Type I SS	Mean Square	F Value	Pr > F
lgrlivArea	1	124.3480171	124.3480171	6195.48	<.0001
OverallQual	9	52.6991006	5.8554556	291.74	<.0001
MSSubClass	14	13.8390804	0.9885057	49.25	<.0001
GarageCars	4	2.8719299	0.7179825	35.77	<.0001
YearRemodAdd	1	2.5344040	2.5344040	126.27	<.0001
Neighborhood	24	5.8429284	0.2434554	12.13	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
lgrlivArea	1	11.80342678	11.80342678	588.09	<.0001
OverallQual	9	7.45885323	0.82876147	41.29	<.0001
MSSubClass	14	3.60053039	0.25718074	12.81	<.0001
GarageCars	4	1.98659767	0.49664942	24.74	<.0001
YearRemodAdd	1	2.00373648	2.00373648	99.83	<.0001
Neighborhood	24	5.84292844	0.24345535	12.13	<.0001

Assumptions Check

Looking at our fit residuals, there does not appear to be any influential point on the residual plots based on our RStudent Plot. Judging from the scatter plot, q-q plot and there is no evidence that the residuals do not follow a normal distribution and a constant variance although our histogram looks slightly skewed to the left due to the presence of some outliers in the data. We will assume that all observations are independent of each other. (Please see appendix for fit residuals plot)

Comparing Competing Models

Predictive Models	Adjusted R2	CV PRESS	Kaggle Score
Simple Linear Regression	.16	103.5	.983
Multiple Linear Regression	.56	101.2	.864
Custom MLR Model	.899	31.53	.552
Other Models …	…	…	…

Conclusion

Comparing the three models, that is our simple Linear regression, our multiple linear regression (MLR), and the Custom MLR model, it is quite obvious that the custom MLR model with lGrLivArea + OverallQual + MSSubClass + GarageCars + YearRemodelAdd + Neighborhood as predictors for salesprice is the most useful model to predict the sale price of the homes in Iowa. It has the highest adjusted R2 of 0.899, and the lowest CV press of 31.53, our chosen metric for comparison. Approximately 89.9% of the variation in the salesprice is explained by explanatory variables in the model. Interpreting the slope for our regression equation will depend on what slope we want to interpret. For instance, interpreting the slope for living area will mean that we are 95% confident that for each doubling of the living area, the median Saleprice will increase between approximately 38% (2^0.46 = 1.38) and 45% (2^0.54). Our best estimate is an increase of 41 % (2^0.5 = 1.41) after holding all other variables constant. There is strong evidence (p-value < .0001) of a relationship with sales price.

Additionally, several categorical variables showed a statistically significant relationship with sales price (See appendix 2), including several neighborhood, Model type, and the number of cars to the garage, but varies based on the category. To get the regression equation for a specific categorical variables, we will need to adjust that variable's coefficient with the intercept in our regression model.

Project RShiny App: [Statistical Insights (weiprecht.github.io)]

Project RShiny Page: [SFDS Final Project (weiprecht.github.io)]

Interactive Graphic: [Ames Housing App (shinyapps.io)]

Project RShiny Github: [Ames Housing App (shinyapps.io)]

Appendix 1: Analysis 1 Code –

Initial analysis of original data

```
# Initial dataset for Analysis Question 1: check assumptions
```{r housing_regression}

# Fit a multiple linear regression model
lm_model <- lm(SalePrice ~ GrLivArea + Neighborhood, data = AmesHousing_Data2)

# Summarize the model
summary(lm_model)

# Calculate VIF for the variables in the model
vif_result <- vif(lm_model)
vif_result

# Calculate mean and standard deviation
mean_sale_price <- mean(AmesHousing_Data2$SalePriceHundr)
sd_sale_price <- sd(AmesHousing_Data2$SalePriceHundr)
mean_sqr_ft <- mean(AmesHousing_Data2$GrLivAreaHundr)
sd_sqr_ft <- sd(AmesHousing_Data2$GrLivAreaHundr)

# Check Assumptions Visually
plot(lm_model)
avPlots(lm(SalePrice ~ GrLivArea + Neighborhood, data = AmesHousing_Data2))

# For histograms
ggplot(AmesHousing_Data2, aes(x = SalePriceHundr)) +
  geom_histogram(aes(y = ..density..), color = "black", fill = "grey") +
  geom_density(aes(color = "Actual")) +
  stat_function(aes(color = "Normal"),
                fun = dnorm,
                args = list(mean = mean_sale_price, sd = sd_sale_price)) +
  xlab("Sale Price (in Hundreds)") +
  scale_colour_manual("Distribution", values = c("darkred", "darkblue"))

ggplot(AmesHousing_Data2, aes(x = GrLivAreaHundr)) +
  geom_histogram(aes(y = ..density..), color = "black", fill = "grey") +
  geom_density(aes(color = "Actual")) +
  stat_function(aes(color = "Normal"),
                fun = dnorm,
                args = list(mean = mean_sqr_ft, sd = sd_sqr_ft)) +
  xlab("Area sqr ft (in Hundreds)") +
  scale_colour_manual("Distribution", values = c("darkred", "darkblue"))

# Boxplot for SalePrice
ggplot(AmesHousing_Data2, aes(x = Neighborhood, y = SalePriceHundr)) +
  geom_boxplot() +
  labs(title = "SalePrice Distribution Across Neighborhoods",
       x = "Neighborhood",
       y = "SalePrice (in hundreds)")

# Boxplot for square feet
ggplot(AmesHousing_Data2, aes(x = Neighborhood, y = GrLivAreaHundr)) +
  geom_boxplot() +
  labs(title = "Square Feet Across Neighborhoods",
       x = "Neighborhood",
       y = "Square Feet (in hundreds)")

```
```

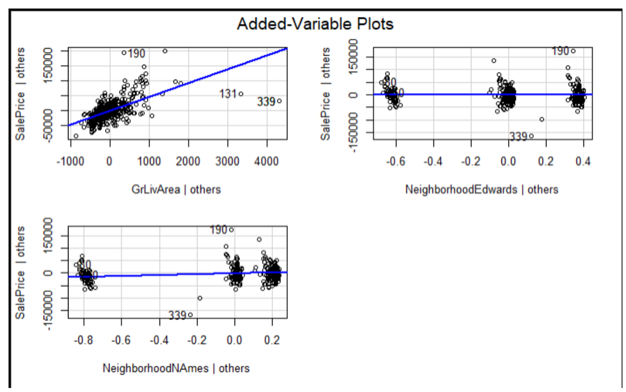
```
Call:
lm(formula = SalePrice ~ GrLivArea + Neighborhood, data = AmesHousing_Data2)

Residuals:
    Min      1Q  Median      3Q     Max
-165078  -16215     281   13578  175400

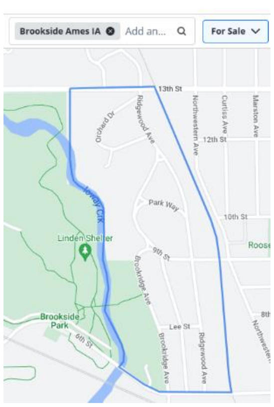
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          69781.538   5442.400  12.822  < 2e-16 ***
GrLivArea               45.760      3.149  14.533  < 2e-16 ***
NeighborhoodEdwards -2882.155   4930.632  -0.585 0.559204
NeighborhoodNAmes    16105.621   4395.352   3.664 0.000283 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

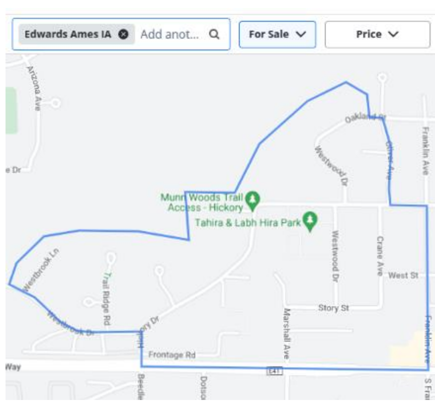
Residual standard error: 29760 on 379 degrees of freedom
Multiple R-squared:  0.3965,    Adjusted R-squared:  0.3917
F-statistic:    83 on 3 and 379 DF,  p-value: < 2.2e-16

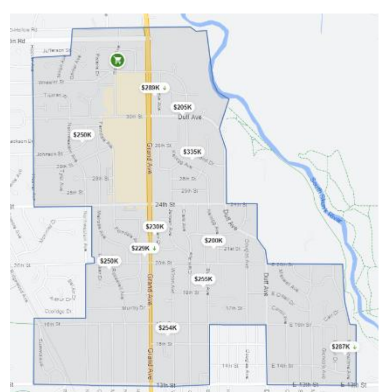
                GVIF Df GVIF^(1/(2*Df))
GrLivArea    1.008149  1        1.004066
Neighborhood 1.008149  2        1.002031
```



Zillow maps of Brookside, Edwards, and North Ames neighborhoods.







Internal Cross Validation:

````
```{r internal_cross_validation}
# Define the model formula
formula <- SalePrice ~ GrLivArea + Neighborhood

# Create a train control object for cross-validation
ctrl <- trainControl(method = "cv",  # Cross-validation method
                     number = 5,     # Number of folds (you can change this)
                     verboseIter = TRUE,  # Display iteration progress
                     summaryFunction = defaultSummary)  # Use default summary functio

# Train the model with cross-validation
lm_model <- train(formula,
                  data = AmesHousing_Data2,
                  method = "lm",  # Linear regression method
                  trControl = ctrl)

# View results
lm_model
```
````

```
383 samples
  2 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 307, 305, 307, 306, 307
Resampling results:

  RMSE     Rsquared   MAE
  29756.1  0.4114803  20897.76

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Interaction code:

````
```{r housing_regression_logged}
# ref is always the element that does not show up in the model.
cat("\n------------------------------Neighborhood no interaction--------------------\n")
pricefit = lm(lSalePrice ~ relevel(Neighborhood, ref = "BrkSide") + lGrLivArea, data = AmesHousing_Data3)
summary(pricefit)

confint(pricefit)

# Get coefficients and statistical summary in a tidy format
tidy_summary2 <- tidy(pricefit)
print(tidy_summary2)

plot(pricefit)

cat("\n------------------------Neighborhood interaction-------------------\n")

# No Interaction (Parallel Regression Lines)
pricefit2 = lm(lSalePrice ~ relevel(Neighborhood, ref = "BrkSide") * lGrLivArea, data = AmesHousing_Data3)
summary(pricefit2)

confint(pricefit2)

# Get coefficients and statistical summary in a tidy format
tidy_summary2 <- tidy(pricefit2)
print(tidy_summary2)

plot(pricefit2)
```
````

```
-------------------------Neighborhood interaction--------------------

Call:
lm(formula = lSalePrice ~ relevel(Neighborhood, ref = "BrkSide") *
    lGrLivArea, data = AmesHousing_Data3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72688 -0.10946  0.02184  0.10651  0.52051

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       6.06485    0.54988  11.029  < 2e-16 ***
relevel(Neighborhood, ref = "BrkSide")Edwards     1.09295    0.73317   1.491 0.136880
relevel(Neighborhood, ref = "BrkSide")NAmes       2.42788    0.63272   3.837 0.000146 ***
lGrLivArea                                        0.79836    0.07784  10.257  < 2e-16 ***
relevel(Neighborhood, ref = "BrkSide")Edwards:lGrLivArea -0.15917    0.10353  -1.537 0.125044
relevel(Neighborhood, ref = "BrkSide")NAmes:lGrLivArea   -0.32534    0.08933  -3.642 0.000309 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1855 on 373 degrees of freedom
Multiple R-squared:  0.5094,    Adjusted R-squared:  0.5029
F-statistic: 77.47 on 5 and 373 DF,  p-value: < 2.2e-16

                                                      2.5 %      97.5 %
(Intercept)                                       4.9836007  7.14609601
relevel(Neighborhood, ref = "BrkSide")Edwards    -0.3487146  2.53461692
relevel(Neighborhood, ref = "BrkSide")NAmes       1.1837337  3.67202489
lGrLivArea                                        0.6453021  0.95142031
relevel(Neighborhood, ref = "BrkSide")Edwards:lGrLivArea -0.3627466  0.04440939
relevel(Neighborhood, ref = "BrkSide")NAmes:lGrLivArea   -0.5009937 -0.14968157
```

Scatterplot code:

````
```{r}

library(ggplot2)
ggplot(AmesHousing_Data3, aes(x = lGrLivArea, y = lSalePrice, color = Neighborhood)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  labs(title = "Relationship of Logged Price and Logged Living Area by Neighborhood",
       x = "Logged Area", y = "Logged Sale Price", color = "Neighborhood")
```
````

Conversion code (transform coefficients back to original.)

````
```{r conversion}
# Coefficients from the regression model
intercept <- 7.66029
coeff_Edwards <- -0.02755
coeff_NAmes <- 0.12412
coeff_GrLivArea <- 0.57229

# Transforming coefficients back to original scale
original_intercept <- exp(intercept)
original_coeff_Edwards <- exp(coeff_Edwards)
original_coeff_NAmes <- exp(coeff_NAmes)
original_coeff_GrLivArea <- exp(coeff_GrLivArea)

# Interpretation in dollars and square feet
# Assuming GrLivArea is measured in square feet
# Assuming the SalePrice was measured in dollars
# Interpretation of coefficients in original scale
interpretation_intercept <- paste("Intercept (in hundreds of dollars):", round(original_intercept, 2))
interpretation_Edwards <- paste("Effect of Edwards (in hundreds of dollars):", round(original_coeff_Edwards, 2))
interpretation_NAmes <- paste("Effect of NAmes (in hundreds of dollars):", round(original_coeff_NAmes, 2))
interpretation_GrLivArea <- paste("Effect of GrLivArea (in hundreds of square feet):", round(original_coeff_GrLivArea, 2))

# Printing interpretations
print(interpretation_intercept)
print(interpretation_Edwards)
print(interpretation_NAmes)
print(interpretation_GrLivArea)
```
````

Appendix 2: Initial dataset for Analysis Question 2

```r
AmesHousing_Data$Neighborhood <- as.factor(AmesHousing_Data$Neighborhood)

# Create a new column that holds the'SalePrice', GrLivAre, and LotArea in hundreds
AmesHousing_Data$SalePriceHundr <- floor(AmesHousing_Data$SalePrice /100)
AmesHousing_Data$GrLivAreaHundr <- floor(AmesHousing_Data$GrLivArea /100)
AmesHousing_Data$LotAreaHundr <- floor(AmesHousing_Data$LotArea /100)

# Drop previously identified outliers rows 131, 136, and 339
rows_to_drop <- c( 131, 136, 190, 339)
AmesHousing_Data4 <- AmesHousing_Data[-rows_to_drop, ]

AmesHousing_Data4
summary(AmesHousing_Data4)


# Fit a multiple linear regression model
lm_model <- lm(SalePrice ~ LotArea, data = AmesHousing_Data4)

# Summarize the model
summary(lm_model)

# Calculate mean and standard deviation
mean_sale_price <- mean(AmesHousing_Data4$SalePriceHundr)
sd_sale_price <- sd(AmesHousing_Data4$SalePriceHundr)
lotmean_sqr_ft <- mean(AmesHousing_Data4$LotAreaHundr)
lotsd_sqr_ft <- sd(AmesHousing_Data4$LotAreaHundr)

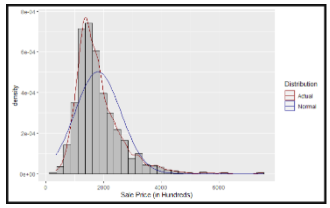
# Check Assumptions Visually
plot(lm_model)

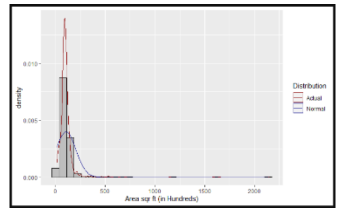
# For histograms
ggplot(AmesHousing_Data4, aes(x = SalePriceHundr)) +
  geom_histogram(aes(y = ..density..), color = "black", fill = "grey") +
  geom_density(aes(color = "Actual")) +
  stat_function(aes(color = "Normal"),
                fun = dnorm,
                args = list(mean = mean_sale_price, sd = sd_sale_price)) +
  xlab("Sale Price (in Hundreds)") +
  scale_colour_manual("Distribution", values = c("darkred", "darkblue"))

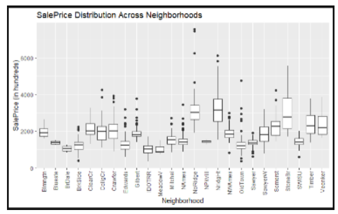
ggplot(AmesHousing_Data4, aes(x = LotAreaHundr)) +
  geom_histogram(aes(y = ..density..), color = "black", fill = "grey") +
  geom_density(aes(color = "Actual")) +
  stat_function(aes(color = "Normal"),
                fun = dnorm,
                args = list(mean = lotmean_sqr_ft, sd = lotsd_sqr_ft)) +
  xlab("Area sqr ft (in Hundreds)") +
  scale_colour_manual("Distribution", values = c("darkred", "darkblue"))

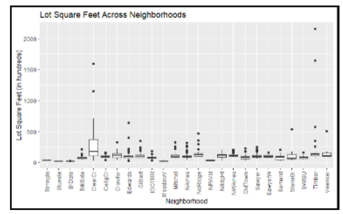
# Boxplot for SalePrice
ggplot(AmesHousing_Data4, aes(x = Neighborhood, y = SalePriceHundr)) +
  geom_boxplot() +
  labs(title = "SalePrice Distribution Across Neighborhoods",
       x = "Neighborhood",
       y = "SalePrice (in hundreds)") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

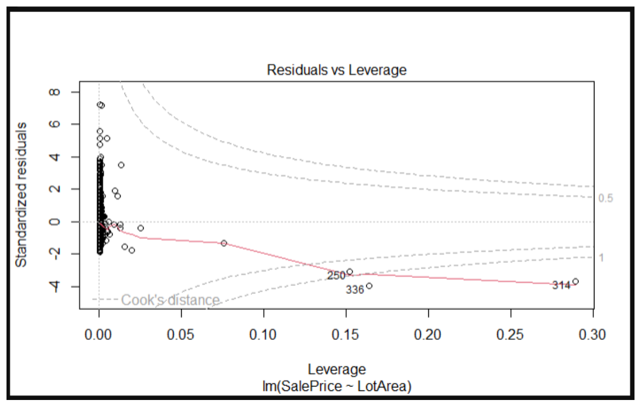
# Boxplot for square feet
ggplot(AmesHousing_Data4, aes(x = Neighborhood, y = LotAreaHundr)) +
  geom_boxplot() +
  labs(title = "Lot Square Feet Across Neighborhoods",
       x = "Neighborhood",
       y = "Lot Square Feet (in hundreds)") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```









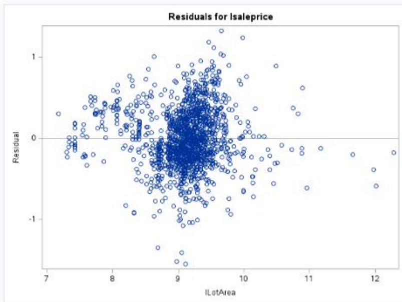


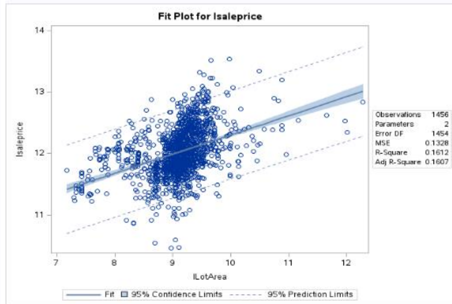
Additional Data for Model 1: This is the code for the sales price by the lot areas. The code includes a cv press details and then the simple linear model.

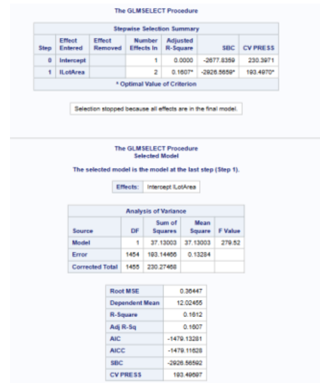
proc glmselect data = housing3;
/* where LotArea not in (63887,40094); */
model lSalePrice = llotArea / selection= Stepwise(stop = cv) cvmethod = random(10) CVDETAILS stats = adjrsq;
run;

/* Regression Code to get plots for SLR model */
proc reg data = housing3;

```
model lsaleprice = lLotarea;
run;
```







The GLMSELECT Procedure

Stepwise Selection Summary						
Step	Effect Entered	Effect Removed	Number Effects In	Adjusted R-Square	SBC	CV PRESS
0	Intercept		1	0.0000	-2677.8359	230.3971
1	lLotArea		2	0.1607*	-2926.5659*	193.4970*

* Optimal Value of Criterion

Selection stopped because all effects are in the final model.

The GLMSELECT Procedure
Selected Model

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

Effects: Intercept lLotArea

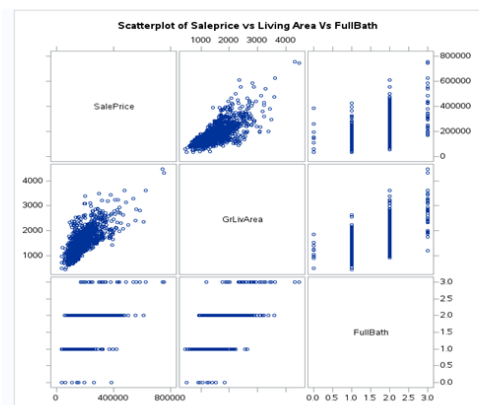
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	37.13003	37.13003	279.52
Error	1454	193.14466	0.13284	
Corrected Total	1455	230.27468		

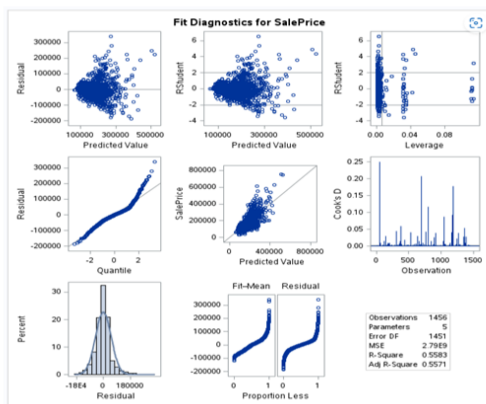
Root MSE	0.36447
Dependent Mean	12.02455
R-Square	0.1612
Adj R-Sq	0.1607
AIC	-1479.13281
AICC	-1479.11628
SBC	-2926.56592
CV PRESS	193.49697

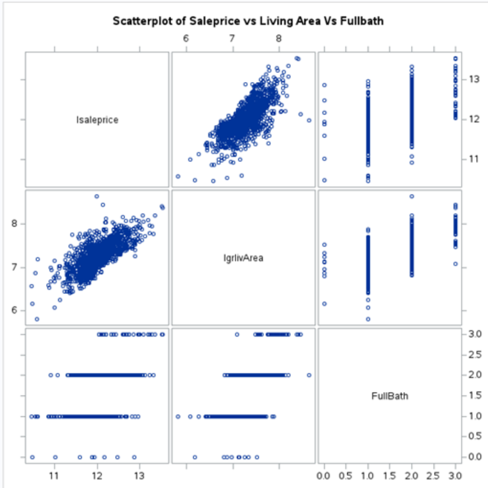
Code for Model 2: Includes CV Press and output.

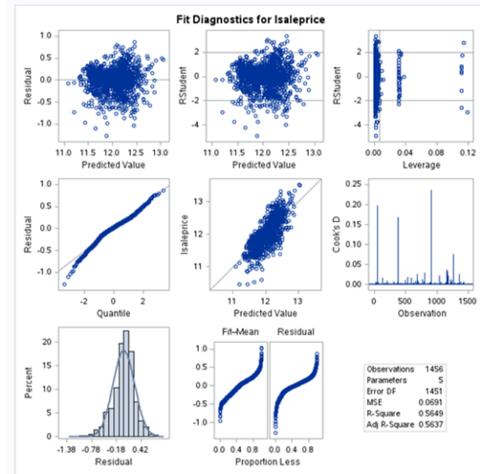
```
/* Backward Selection for 2nd Model */ proc glmselect data = housing3; Class Fullbath; model lSalePrice =
lGrLivArea FullBath / selection= backward(stop = cv) cvmethod = random(5) CVDETAILS stats = adjrsq; /* selection
= stepwise(stop = SL SLE = 0.05 SLS =0.05) STATS=adjrsq; */ run;

/*Proc glm model to fit 2nd Model*/
proc glm data = housing3 plots= all;
Class Fullbath;
model lSalePrice = lGrLivArea Fullbath / solution clparm;
run;
```









The GLMSELECT Procedure

Backward Selection Summary						
Step	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0		3	5	0.5637*	-3860.2586*	101.2437*

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

Stop Details				
Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Removal	FullBath	106.3716	>	101.2437

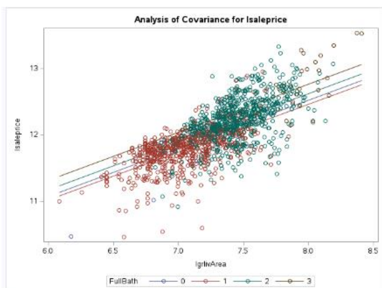
The GLMSELECT Procedure
Selected Model

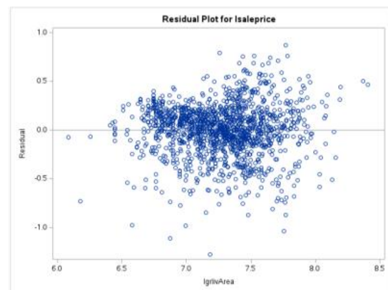
The selected model is the model at the last step (Step 0).

Effects:	Intercept lgrlivArea FullBath

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	4	130.07551	32.51888	470.91
Error	1451	100.19917	0.06906	
Corrected Total	1455	230.27468		

Root MSE	0.26278
Dependent Mean	12.02455
R-Square	0.5649
Adj R-Sq	0.5637
AIC	-2428.67580
AICC	-2428.61783
SBC	-3860.25856
CV PRESS	101.24369





Parameter estimates for model 2:

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	6.979739075	B	0.22497561	31.02	<.0001	6.538426854	7.421051297
lgrlivArea	0.723070081		0.02794304	25.88	<.0001	0.668257016	0.777883147
FullBath 0	-0.242434141	B	0.10202764	-2.38	0.0176	-0.442571579	-0.042296703
FullBath 1	-0.299228013	B	0.05304306	-5.64	<.0001	-0.403277290	-0.195178737
FullBath 2	-0.142160707	B	0.04898498	-2.90	0.0038	-0.238249661	-0.046071754
FullBath 3	0.000000000	B	.	.	.	.	.

Code for Model 3:
Stepwise Selection For custom model
proc glmselect data = housing3;

Class MsSubclass Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle OverallQual OverallCond RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model lSalePrice = MSSubClass LotFrontagen lLotArea Utilities Lotconfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical fstFlrSF ScndFlrSF

LowQualFinSF lGrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch SsnPorch3 ScreenPorch PoolArea PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition / selection=stepwise (stop = cv) cvmethod = random(10) CVDETAILS stats = adjrsq;

/* selection=backward (stop = cv) cvmethod = random(10) CVDETAILS stats = adjrsq; */ /* selection = backward(stop = SL SLS = .01) stats = adjrsq; */

run;

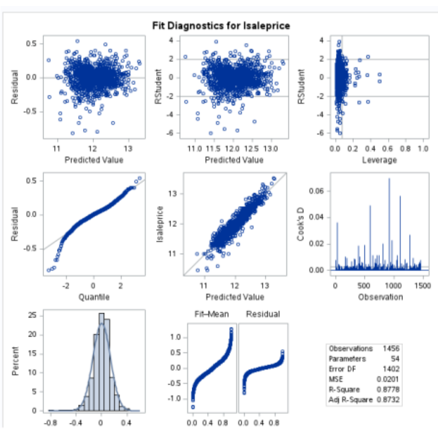
Building and fitting custom model with proc glm

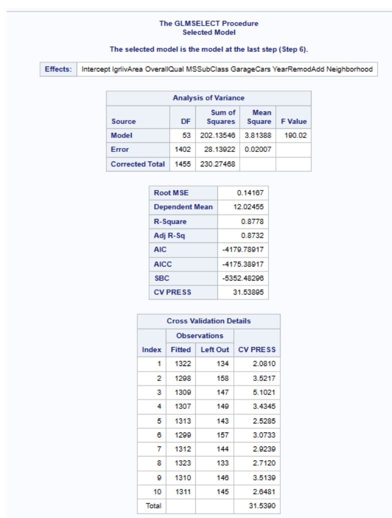
proc glm data=housing3 alpha=0.05 plots=all; class Neighborhood OverallQual MsSubclass GarageCars ;
model lSalePrice = lGrLivArea OverallQual MsSubclass GarageCars YearRemodAdd Neighborhood / solution clparm;
run;

Proc Glmselect on custom model to get the CV press score

proc glmselect data=housing3;
class Neighborhood OverallQual MsSubclass GarageCars ;
model lSalePrice = lGrLivArea OverallQual MsSubclass GarageCars YearRemodAdd Neighborhood /
selection=stepwise (stop = cv) cvmethod = random(10) CVDETAILS stats = adjrsq;
run;

Plots and Parameter Estimates for the custom model 3:





The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 6).

Effects:	Intercept lgrlivArea OverallQual MSSubClass GarageCars YearRemodAdd Neighborhood

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	53	202.13546	3.81388	190.02
Error	1402	28.13922	0.02007	
Corrected Total	1455	230.27468		

Root MSE	0.14167
Dependent Mean	12.02455
R-Square	0.8778
Adj R-Sq	0.8732
AIC	-4179.78917
AICC	-4175.38917
SBC	-5352.48296
CV PRESS	31.53895

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	1322	134	2.0810
2	1298	158	3.5217
3	1309	147	5.1021
4	1307	149	3.4345
5	1313	143	2.5285
6	1299	157	3.0733
7	1312	144	2.9239
8	1323	133	2.7120
9	1310	146	3.5139
10	1311	145	2.6481
Total			31.5390

Final Custom Model Coefficients:

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	4.280743248	B	0.52245785	8.19	<.0001	3.255859905	5.305626591
lgrlivArea	0.496418591		0.02047039	24.25	<.0001	0.456262705	0.536574476
OverallQual 1	-0.916865062	B	0.15065716	-6.09	<.0001	-1.212402816	-0.621327308
OverallQual 2	-0.965093114	B	0.09622608	-10.03	<.0001	-1.153855727	-0.776330501
OverallQual 3	-0.809112472	B	0.05374896	-15.05	<.0001	-0.914549516	-0.703675428
OverallQual 4	-0.640796589	B	0.04512535	-14.20	<.0001	-0.729317063	-0.552276114
OverallQual 5	-0.562656995	B	0.04264635	-13.19	<.0001	-0.646314521	-0.478999469
OverallQual 6	-0.500773143	B	0.04183898	-11.97	<.0001	-0.582846883	-0.418699403
OverallQual 7	-0.425310777	B	0.04032303	-10.55	<.0001	-0.504410742	-0.346210812
OverallQual 8	-0.328527673	B	0.03884786	-8.46	<.0001	-0.404733866	-0.252321480
OverallQual 9	-0.145029325	B	0.04238874	-3.42	0.0006	-0.228181507	-0.061877144
OverallQual 10	0.000000000	B	.	.	.	.	.
MSSubClass 20	0.124482872	B	0.03019460	4.12	<.0001	0.065251413	0.183714331
MSSubClass 30	0.017673844	B	0.03303650	0.53	0.5927	-0.047132449	0.082480136
MSSubClass 40	0.043169378	B	0.07655780	0.56	0.5729	-0.107010808	0.193349564
MSSubClass 45	0.053570117	B	0.05050201	1.06	0.2890	-0.045497523	0.152637758
MSSubClass 50	-0.001424324	B	0.02926683	-0.05	0.9612	-0.058835818	0.055987170
MSSubClass 60	0.059897913	B	0.03136839	1.91	0.0564	-0.001636129	0.121431955
MSSubClass 70	-0.045014524	B	0.03298705	-1.36	0.1726	-0.109723815	0.019694766
MSSubClass 75	-0.040022595	B	0.04510565	-0.89	0.3751	-0.128504432	0.048459242
MSSubClass 80	0.114089289	B	0.03520347	3.24	0.0012	0.045032148	0.183146430
MSSubClass 85	0.180177468	B	0.04432706	4.06	<.0001	0.093222966	0.267131970
MSSubClass 90	-0.035581181	B	0.03436749	-1.04	0.3007	-0.102998416	0.031836055
MSSubClass 120	0.074112962	B	0.03658759	2.03	0.0430	0.002340642	0.145885281
MSSubClass 160	-0.123874888	B	0.03968541	-3.11	0.0019	-0.202116410	-0.045633367
MSSubClass 180	0.112673690	B	0.06346229	1.78	0.0760	-0.011817579	0.237164960
MSSubClass 190	0.000000000	B	.	.	.	.	.
GarageCars 0	-0.336307896	B	0.06687641	-5.03	<.0001	-0.467496504	-0.205119289
GarageCars 1	-0.229013060	B	0.06586194	-3.48	0.0005	-0.358211628	-0.099814492
GarageCars 2	-0.182450281	B	0.06522846	-2.80	0.0052	-0.310406177	-0.054494385
GarageCars 3	-0.112052432	B	0.06677217	-1.68	0.0935	-0.243036554	0.018931690
GarageCars 4	0.000000000	B	.	.	.	.	.
YearRemodAdd	0.002476537		0.00024786	9.99	<.0001	0.001990320	0.002962753
Neighborhood Blmngtn	-0.193593512	B	0.05764461	-3.36	0.0008	-0.306672486	-0.080514539
Neighborhood Blueste	-0.140708997	B	0.11214701	-1.25	0.2098	-0.360703022	0.079285028
Neighborhood BrDale	-0.240913052	B	0.06222150	-3.87	0.0001	-0.362970332	-0.118855772
Neighborhood BrkSide	-0.172058189	B	0.04985307	-3.45	0.0006	-0.269852837	-0.074263541
Neighborhood ClearCr	-0.021640270	B	0.05120345	-0.42	0.6726	-0.122083892	0.078803352
Neighborhood CollgCr	-0.145262395	B	0.04497194	-3.23	0.0013	-0.233481944	-0.057042846
Neighborhood Crawfor	-0.013758384	B	0.04917063	-0.28	0.7797	-0.110214315	0.082697547
Neighborhood Edwards	-0.234898557	B	0.04671790	-5.03	<.0001	-0.326543080	-0.143254035
Neighborhood Gilbert	-0.180440449	B	0.04678802	-3.86	0.0001	-0.272222525	-0.088658372
Neighborhood IDOTRR	-0.345221880	B	0.05194359	-6.65	<.0001	-0.447117420	-0.243326341
Neighborhood MeadowV	-0.260048959	B	0.06314308	-4.12	<.0001	-0.383914045	-0.136183873
Neighborhood Mitchel	-0.167261571	B	0.04810205	-3.48	0.0005	-0.261621321	-0.072901821
Neighborhood NAmes	-0.186076366	B	0.04482673	-4.15	<.0001	-0.274011049	-0.098141683
Neighborhood NPkVill	-0.122759836	B	0.06599543	-1.86	0.0631	-0.252220273	0.006700601
Neighborhood NWAmes	-0.181263362	B	0.04656076	-3.89	0.0001	-0.272599624	-0.089927100
Neighborhood NoRidge	-0.031855927	B	0.04978914	-0.64	0.5224	-0.129525164	0.065813310
Neighborhood NridgHt	-0.055224300	B	0.04718702	-1.17	0.2421	-0.147789077	0.037340477
Neighborhood OldTown	-0.309081793	B	0.04758637	-6.50	<.0001	-0.402429961	-0.215733625
Neighborhood SWISU	-0.209232442	B	0.05467654	-3.83	0.0001	-0.316489095	-0.101975788
Neighborhood Sawyer	-0.205185796	B	0.04702151	-4.36	<.0001	-0.297425901	-0.112945690