PSTAT 126 - Regression Analysis Fall 2019

Prof. Bapat

Analysis of Home Sale Prices in Ames, Iowa

Nihaal Kumar (Section: Wed. 3 - 3:50PM Ke Wang)

Stephanie Zacharias (Section: Wed. 11:00 - 11:50AM Ke Wang)

Introduction

The Ames Housing dataset contains 2930 observations that describe the sale of homes in Ames, Iowa between 2006 and 2010. The dataset is rather large and contains 80 variables. In order to create a meaningful model, we decided to take a subset of 17 variables to consider. The variables we kept are sale price in dollars, overall quality, overall condition, year built, exterior condition, basement condition, heating system quality and condition, central air conditioning system (0: No, 1: Yes), above ground living area in square feet, number of bedrooms above ground, quality of the kitchen, total number of rooms above ground (not including bathrooms), car capacity of garage, total number of bathrooms, the year the property was sold, and the lot size in square feet. All of the variables that consider the condition or quality of an aspect contributing to the sale price are on a scale encoded from Excellent to Poor (5: Excellent, 4: Good, 3: Typical/Average, 2: Fair, 1: Poor). The goal of this project is to develop a regression model of the sale price of homes (SalePrice) on the remaining 18 variables in order to develop responses to our research questions.

Questions of Interest

We will consider the following research questions:

Question 1: Can we find evidence to prove that people in Ames, Iowa may prefer larger homes with an overall lower quality or homes that might be smaller but have a higher quality rating?

Question 2: Which increase of total bathrooms in the house (one to two, two to three,etc.) creates the most significant increase in the sale price of a house?

Question 3: Given a house with an average rating for all of the predictors in our final model, can we predict it's sale price?

Regression Method

To answer our research questions, we will develop an adequate linear regression model with our data set. To create this model, we will remove insignificant variables and use a stepwise regression to develop our first version of the model. To then make it better, we will analyze various diagnostic plots in addition to correlation values to make our model meet the four LINE conditions, and perform the necessary transformations to our variables if the LINE conditions are not met. Following that we will use F-tests, VIF(Variance Inflation Factors), and AIC values to finetune our model to what the regression deems to be the "best." Once the model is made to a level we're satisfied with, we will then proceed to answer our three research questions.

Regression Analysis, Results and Interpretation

Model Building Process

Preparation of the Data

We first begin by tailoring our data set for proper regression analysis methods. This process involved removing missing values and encoding some of the nominal variables to numeric factors. The variables which we had to refactor are listed below:

- ExterCond
- HeatingQC
- CentralAir
- BsmtCond
- KitchenQual

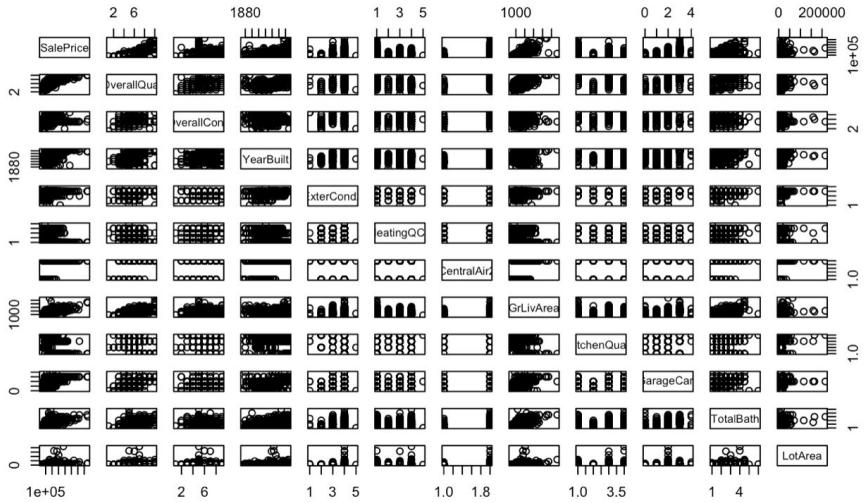
Once this was done, we created a "new" data set called newData including only the predictors we wanted as well as the predictors we encoded to work for our model.

Determining Ideal Model with Predictors using Stepwise Regression

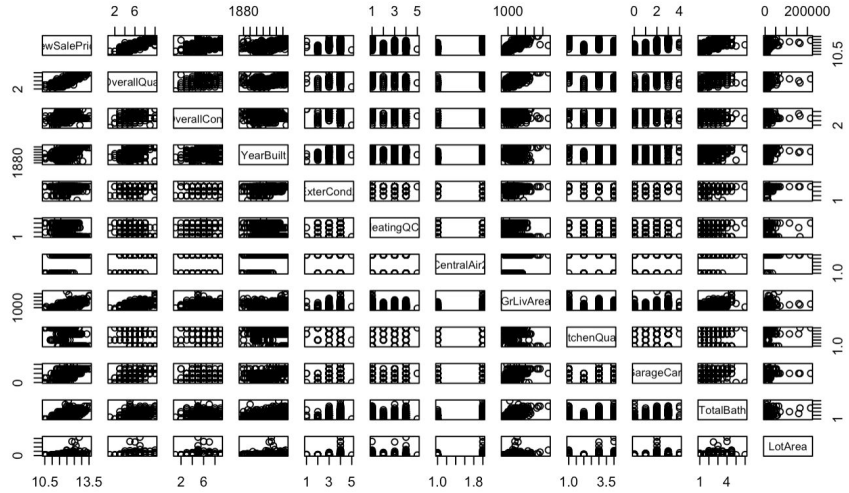
In order to determine which predictors were most impactful in our model to cut down the number of predictors, we performed a stepwise regression until we reached the model with the best predictors. From this, we got that the predictors to use to predict SalePrice were OverallQual, OverallCond, YearBuilt, ExterCond2, HeatingQC2, CentralAir, GrLivArea,KitchenQual2, GarageCars, TotalBath, and LotArea. From this process our model's AIC went from -5265.26 to -5267.6, which means our model was improved through the removal of the other predictors not included above. Normally we wouldn't perform the stepwise regression first, because in general we would first want to take a look at a scatterplot matrix. However, due to just how many predictors were in our model, it just wouldn't be practical to do a scatterplot matrix with that many predictors. As a result, we perform the stepwise regression earlier than normal to narrow down our set of predictors to have a better scatterplot matrix to move forward. However, as shown by the next section, even with our reduced model, we still run into issues with our data pre-transformation.

Removal of Multicollinearity of the Data

We continue our regression analysis by checking the correlations of the predictors in our improved model with our response variable, SalePrice. To achieve this, we plot the scatterplot matrix of SalePrice against all of the potential predictors.



While the scatterplot matrix could be analyzed as it is shown, we wanted to be able to better see the correlations between SalePrice and the predictors. Due to this, we chose to transform SalePrice earlier than usual as well in order to have a more "readable" matrix. Running the scatterplot matrix again, we get:



This still doesn't fix visibility well enough due to the number of predictors in the model, and so we continue with cleaning our data to reach a better model. In addition to the scatterplot matrix, we also computed the correlation matrix of the predictors, from which the correlations of TotRmAbvGrd and GrLivArea was the most notably high with a correlation value of 0.827. Many of the other correlations were also still relatively high, which leads us to wonder if multicollinearity plays a role between our predictors. To investigate, we check the VIF(Variance Inflation Factor) values of each of our predictors, which gives us:

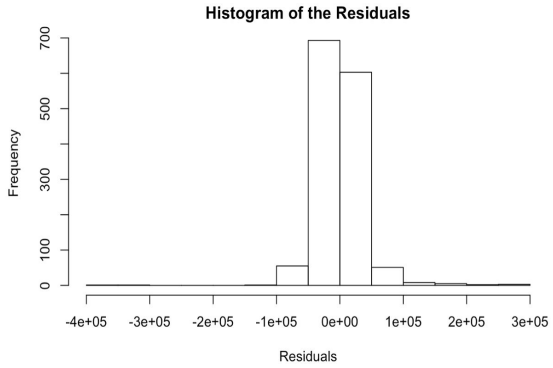
OverallQual	OverallCond	YearBuilt	ExterCond2	HeatingQC2	CentralAir2	GrLivArea	KitchenQual2	GarageCars	TotalBath	LotArea
2.902285	1.359522	3.041752	1.116554	1.379088	1.301463	2.390089	1.578568	1.870674	2.141125	1.102043

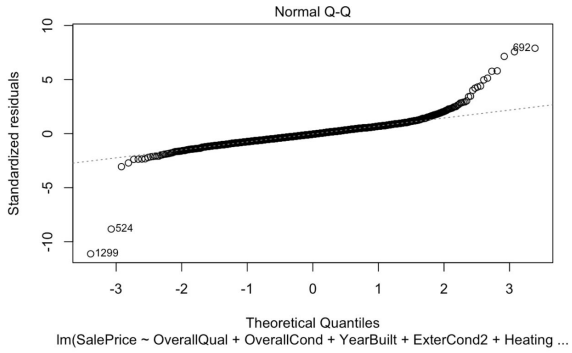
A VIF value is considered to imply non-collinearity if it is below 3 for its corresponding predictor variable. As is the case with our predictors, all of the predictors have a VIF value below 3, and so by this method we don't need to remove any predictors. One might notice that the VIF value for the YearBuilt predictor is just slightly above 3, but since it is so close to our

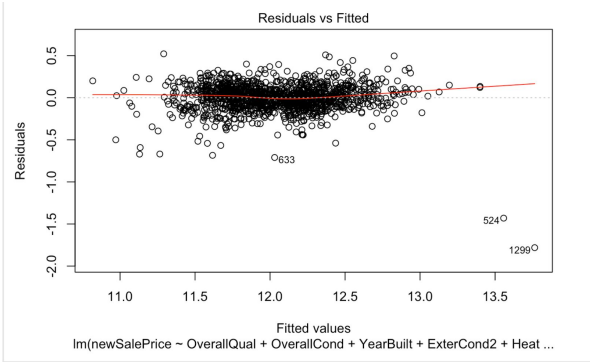
cutoff value we choose to keep it in the model for now. Should another analysis method later in our report flag this predictor again, we may then remove it. However, for now we proceed as is.

Residual Diagnostics and Data Transformations

We now fit a first order model on the remaining predictors of our model, from which we check the diagnostic plots to see if our updated model meets the "LINE" conditions:

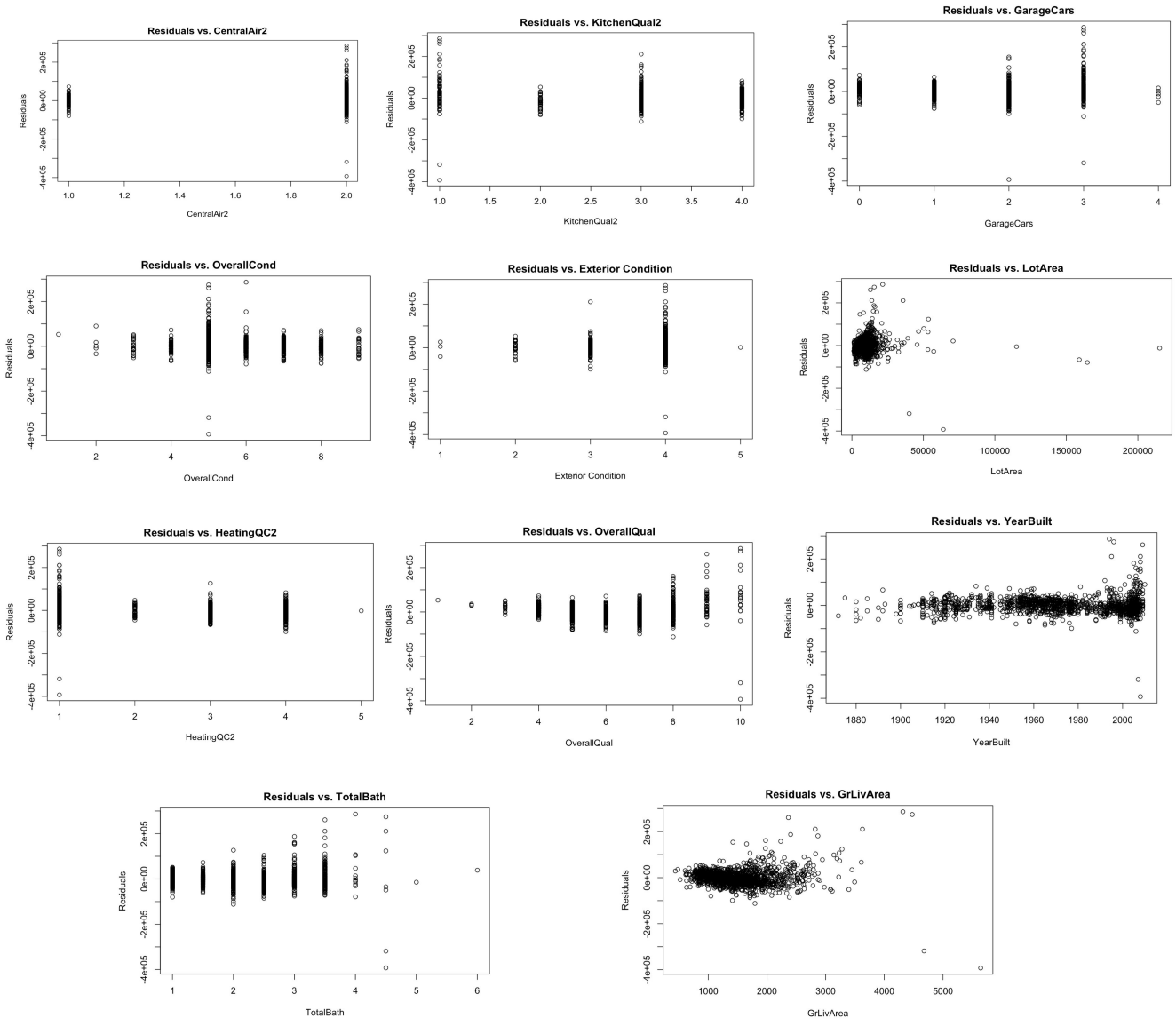






Observing the diagnostic plots above, the normal Q-Q plot shows a curvature, suggesting non-normality of the error terms. Furthermore, the histogram of the residuals indicate non-normality due to the curve not looking as even as we would hope for in a model that meets the LINE conditions. Lastly, the Residuals vs. Fitted Values plot shows a slight curvature as well, indicating a violation of the assumption of linearity.

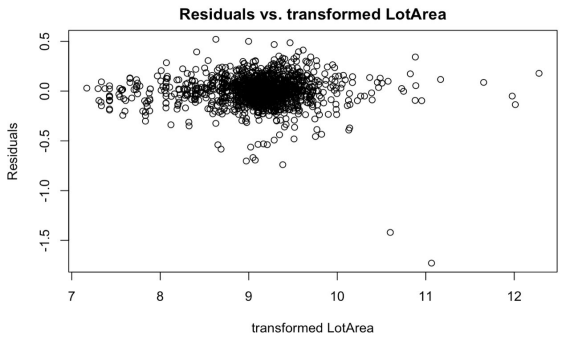
In working towards fixing these issues, we first look to fix the non-linearity issue by analyzing the residuals against each of the predictor variables in our model.

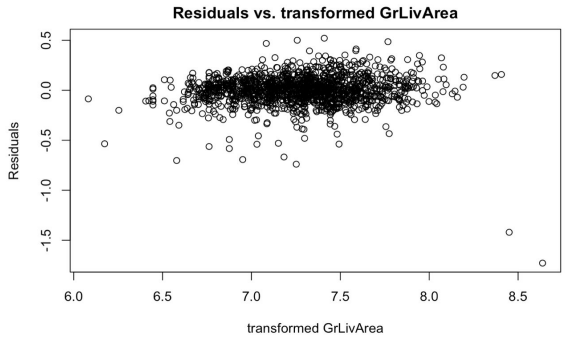


From the above plots, we see that some of the residual plots display characteristics of curvature indicating a violation of the assumption of linearity. Furthermore, some of the plots also have a "fanning" effect of their points, in which either the points fan out from left to right, or fan inwards going in the same direction. This is a sign of non-constant variance, which acts as

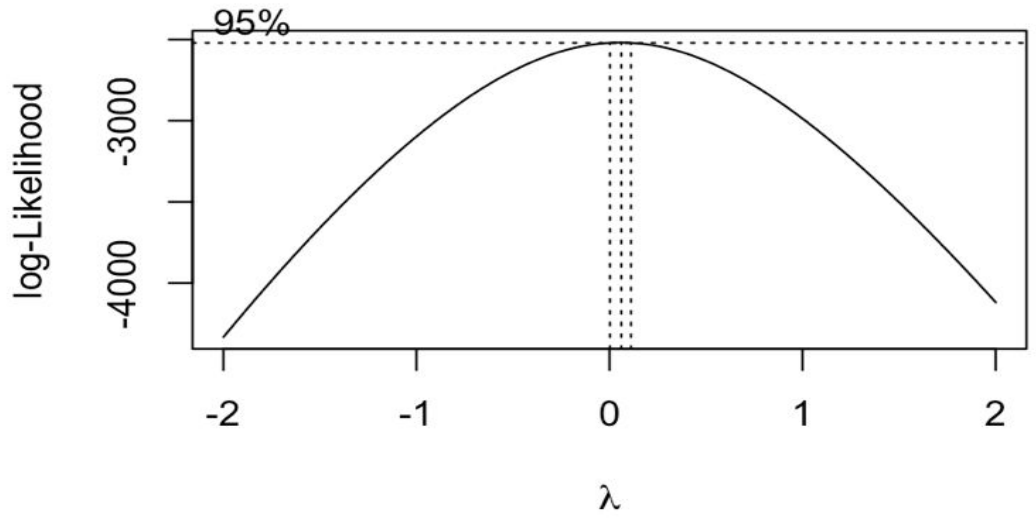
another obstacle as well. Finally, a few of the plots have points that cluster towards one side, indicating potential issue with the scale of the plot. We will address the issue of scale first.

Addressing the issue of scale can be done by taking the log of parameters in the respective plots. In this case of our data, we performed these log transformations on LotArea and GrLivArea, through which those respective residual plots came out looking much better and evenly spread out:





Now in order to address the two other issues of non constant variance and non-linearity as shown by some of the residual plots, we began by performing transformations on the predictors in question. First we log transformed the predictors and rechecked the residual plots to see if the issues we were encountering were fixed. However, the plots barely changed at all. After exhausting our transformation methods learned in this class, we fell back to our “last resort,” which was to use the Box-Cox method. Upon performing the Box-Cox, we saw that zero was within the 95 percent confidence interval, which meant that the optimal transformation for our model at this point would have been to log transform our response variable. However, we had already done that earlier in our preparation of the data, and so for now we’ve exhausted all of our options within the scope of this class. For the sake of clarity, we do not include the updated residual plot outputs below due to there not being much of a change to them from our transformations. However, the Box-Cox plot will be shown below, and the code for our attempted transformations can be referenced in the appendix.



Since we tried all the transformation methods in our repertoire yet were only able to fix a number of the residual plots, we wait to see if we can fix the LINE conditions not yet met by our model later in this report when we address outliers and influential points. For now, given the strong AIC of our model indicating that our data is still viable, we proceed with what we have.

Addition of Interaction Terms in the Model

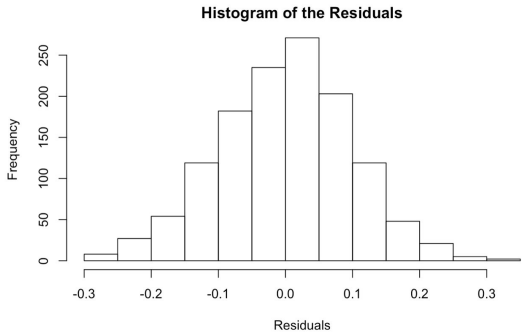
We now look to measure the interaction (if any) between our predictor variables. However, since our current model includes such a large number of predictors, we will only be looking to add the interaction terms of the predictors that will be important in answering our research questions. As such we check interaction between newGrLivArea(since we transformed this) and OverallQual. From this our model becomes:

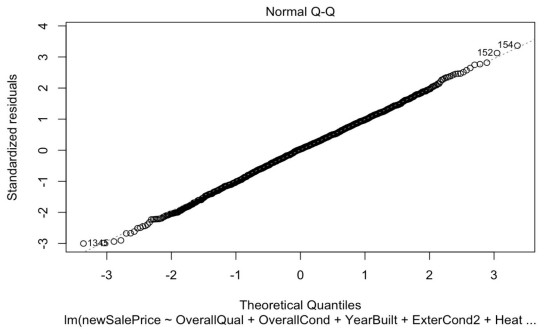
$$E(log(SalePrice_i)) = \beta_0 + \beta_1 OverallQual_i + \beta_2 OverallCond_i + \beta_3 YearBuilt_i + \beta_4 ExterCond_i$$
$$+ \beta_5 HeatingQC2_i + \beta_6 CentralAir2_i + \beta_7 log(GrLivArea_i) + \beta_8 KitchenQual2_i$$
$$+ \beta_8 GarageCars_i + \beta_9 TotalBath_i + \beta_{10} log(LotArea_i) + \beta_{17} OverallQual_i * log(GrLivArea_i)$$

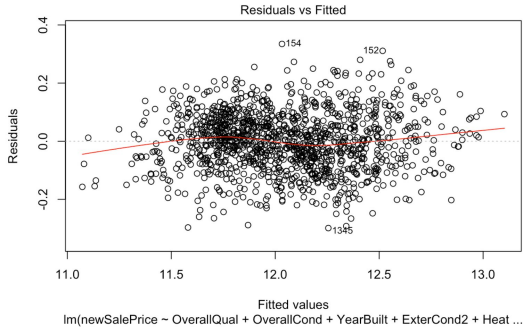
Determination of Influential Points

As mentioned earlier in our diagnostic plot analysis/transformations, since our model's violation of non-linearity and non constant variance was not able to be fixed using log-transformations or through the Box-Cox method, we now look to see if there are any influential points in our model to see if they are the cause of these issues. There were several possible ways by which we could do this, such as using DFFITs, leverages, or Cook's Distance values. However, because of just how many points we had in our data set, we chose to use Cook's Distance values. Under this criterion, a data point is deemed worthy of further investigation if it's Cook's Distance value is greater than 0.5. While we considered plotting the values, we instead chose to narrow down all of the points which had a Cook's Distance value greater than 0.5. Using this methodology, we find point 1299 and point 1264 to meet this condition.

In order to check if these points are actually influential, we delete them and then replot our model to recheck the diagnostic plots of what we hope to be our final "best model.":







Judging from these updated diagnostic plots, many of the issues in our model seem to be fixed by deleting the influential points. The assumptions of linearity, constant variance, and overall normal distribution of the histogram indicate that our model now adequately satisfies the LINE conditions. To confirm this idea, we perform the Shapiro-Wilks test, a test for normality of a model. The test is known to reject the null hypothesis of normality when the calculated p-value is less than 0.05. Upon calculation, we find the p-value to be greater than 0.05, which implies that our model is in fact normal now. Taking this one step further, we also calculate the R^2 value to be 0.9203, meaning that 92.3% of the variation in SalePrice is accounted for by the variation in the predictors in our final model. Now that we have our ideal model, we proceed to answer our research questions.

Research Questions

Can we find evidence to prove that people in Ames, Iowa may prefer larger homes with an overall lower quality or homes that might be smaller but have a higher quality rating?

While answering this question, we are effectively looking to see if OverallQual and GrLivArea interact. We perform a general linear F-test with null hypothesis

$$H_0 : \beta_{17} = 0$$

and alternative hypothesis

$$H_1 : \beta_{17} \neq 0$$

where β_{17} is the coefficient of the interaction term QverallQual*newGrLivArea. This test produces a p-value of $1.457e^{-09}$ (refer to appendix). Since the p-value is smaller than our significance level $\alpha = 0.05$, we reject the null hypothesis. Therefore, the data suggests that the effect of the overall quality of the home on the sale price depends on the square footage of the home.

Which increase of total bathrooms in the house (one to two, two to three,etc.) creates the most significant increase in the sale price of a house?

One of the most common characteristics of a house shown to potential homeowners is the number of bathrooms. As the number of bathrooms in a house increases, more often than not so

will the sale price. What interests us is which increment of total bathrooms in a house in Ames results in the most significant increase in the average sale price of those home. Within our data set the total number of bathrooms ranged from one to five. By performing a two sample t-test comparing the mean price of a house with each increment of number of bathrooms, we can compare the p-values of each test to answer this question. It should be noted that we did not perform the paired test for a house going from four to five bathrooms because the sample size of houses with five bathrooms was too small. Performing the tests for the rest of the increments, the p-value for the increase in sale price of a house with three vs. four bathrooms was not below our significance level of 0.05. As a result, we only compare the t-tests below:

```
        Welch Two Sample t-test

data:  OneBathPrice and TwoBathPrice
t = -14.468, df = 552.08, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -49139.27 -37391.17
sample estimates:
mean of x mean of y
 117121.1  160386.3
```

```
        Welch Two Sample t-test

data:  TwoBathPrice and ThreeBathPrice
t = -12.672, df = 259.15, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -77348.10 -56541.76
sample estimates:
mean of x mean of y
 160386.3  227331.3
```

The p-value of both tests comes out to be the same, indicating both of the increments in question are significant. However, looking at the confidence interval created, the confidence interval for

the increment from one to two bathrooms is narrower, indicating a higher accuracy. Taking that into account, we take the increase in sale price from going from a one to two bathroom house to be the most significant.

Given a house with an average rating for all of the predictors in our final model, can we predict it's sale price?

To predict the sale price of an average home we must first create a new dataset containing the average values of each predictor. We then use the dataset of averages to run our model on to create a 95% prediction interval. The interval that is returned after restoring the original units of the data is

```
      fit      lwr      upr
1 169029 138983.2 205570.2
```

Therefore, we predict that a home with completely average values would have sold for $169,029 and are 95% confident that the sale price lies between $138,983.20 and $205,570.20.

Conclusion

In conclusion, our regression analysis has shown that of the many variables that influence the sale price of a home in Ames, Iowa, the most important ones are overall quality, overall condition, year built, exterior condition, heating system quality, central air system, general living area, kitchen quality, garage space for cars, total bathrooms, lot area, and the interaction term between overall quality and general living area. Using this "ideal" model, we were able to successfully answer our research questions. We found that the overall quality of a home and the lot area size interacted with each other, meaning they both play a role in affecting the sale price of a home. We also found that the average increase in sales price of a house when adding additional bathrooms was most significant when going from one to two bathrooms. Last, we showed that a house with completely average values would sell for $169,029, and that we are 95% confident that given a house with all average values, it's sale price would be between $138,983.20 and $205,570.20.

It is important to restate that the data set we used only covered house prices in Ames from 2006 to 2010, and as such our results should only be applied towards houses sold within that time period. A factor such as inflation inherently affects these sale prices every new year, and so our model is not meant to predict/answer anything about home prices in Ames before 2006 or after 2010. Furthermore, it is no secret the economic crash in 2008 was a result of the housing crisis, and so it is not unlikely that the crash also factored into the prices of these homes. However, given the data and regression statistics we observed from our completed model, the model is very potent for further analysis of the data set as a whole. As mentioned in the introduction, since the original data set consisted of almost 80 predictor variables, we chose a subset of those predictors to begin our regression analysis. In the future, were we to return to this data set, we might consider using a completely different set of predictors to see what other potential models we could have made. Given a larger skill set we might have been able to perform an even more thorough analysis, but with the resources we have, we believe that our final model is in fact the most potent model for predicting sale prices of houses sold in Ames.

Appendix

Relevant code and outputs are shown below.

Code

Data Preparation

```
data <-read.table('train.csv',sep=',',header = TRUE)
names(data)
attach(data)

# encoding variables
data$ExterCond2 <- as.numeric(factor(data$ExterCond,
levels = c('Ex', 'Fa','Gd', 'TA','Po'),
labels = c(5,2,4,3,1) ,ordered = TRUE))
data$HeatingQC2 <- as.numeric(factor(data$HeatingQC,
levels = c('Ex', 'Fa','Gd', 'TA','Po'),
labels = c(5,2,4,3,1) ,ordered = TRUE))
data$CentralAir2 <- as.numeric(factor(data$CentralAir,
levels = c('N', 'Y'),
labels = c(0,1) ,ordered = TRUE))
data$BsmtCond2 <- as.numeric(factor(data$BsmtCond,
levels = c('Ex', 'Fa','Gd', 'TA','Po'),
labels = c(5,2,4,3,1) ,ordered = TRUE))
```

```
data$KitchenQual2 <- as.numeric(factor(data$KitchenQual,
levels = c('Ex', 'Fa','Gd', 'TA','Po'),
labels = c(5,2,4,3,1) ,ordered = TRUE))

#accumulating number of bathrooms
data$TotalBath <- data$FullBath + (data$HalfBath*0.5) + data$BsmtFullBath +
(data$BsmtHalfBath*0.5)

myPredictors <-
c('SalePrice','OverallQual','OverallCond','YearBuilt','ExterCond2',
'BsmtCond2','HeatingQC2',
'CentralAir2','GrLivArea','BedroomAbvGr','KitchenQual2',
'TotRmsAbvGrd','GarageCars',
'TotalBath',
'YrSold','LotArea')

# create a new table only containing the predictors we are interested in
newData <- data[,myPredictors]
attach(newData)
```

Stepwise Regression Using Akaike's Information Criterion

```
# limit the number of predictors using stepwise regression
newData <- na.omit(newData)
formula(step(lm(SalePrice ~. -SalePrice, data = newData)))
```

Removal of Multicollinearity from the Data

```
# check the scatterplot matrix of the regression model given by performing
stepwise regression
pairs(SalePrice ~ OverallQual + OverallCond + YearBuilt + ExterCond2 +
HeatingQC2 + CentralAir2 + GrLivArea + KitchenQual2 + GarageCars +
TotalBath + LotArea, data = newData)

# can't tell much from the original scatterplot matrix so let's transform the
response variable SalePrice using a log transformation)
# transform SalePrice of newData
newData$newSalePrice <- log(newData$SalePrice)
pairs(newSalePrice ~ OverallQual + OverallCond + YearBuilt + ExterCond2 +
HeatingQC2 + CentralAir2 + GrLivArea + KitchenQual2 + GarageCars + TotalBath +
LotArea, data = newData)

# determine correlation between predictors
cor(subset(newData, select = -c(newSalePrice, SalePrice)))

# determine Variance of Inflation Factors for predictors
full_model <- lm(newSalePrice ~ OverallQual + OverallCond + YearBuilt +
ExterCond2 + HeatingQC2 + CentralAir2 + GrLivArea + KitchenQual2 + GarageCars +
TotalBath + LotArea, data = newData)
#vif(full_model)
# all of the VIFs of the predictors are below at or below 3, this suggests that
there is no multicollinearity so we keep all of the predictors
```

Residual Diagnostics and Data Transformations

```
# plot the full model to evaluate the LINE conditions
plot(full_model)
hist(resid(full_model), main = 'Histogram of the Residuals', xlab =
'Residuals')
# the model is is experiencing non-normality and non-linearity
shapiro.test(resid(full_model))

plot(newData$OverallQual, resid(full_model), main = 'Residuals vs.
OverallQual', xlab = 'OverallQual', ylab = 'Residuals')
# suggests serial correlation; need a time series model - beyond scope of class
so we leave it as is

plot(newData$OverallCond, resid(full_model), main = 'Residuals vs.
OverallCond', xlab = 'OverallCond', ylab = 'Residuals')
# suggests serial correlation; need a time series model - beyond scope of class
so we leave it as is

plot(newData$YearBuilt, resid(full_model), main = 'Residuals vs. YearBuilt',
xlab = 'YearBuilt', ylab = 'Residuals')
# this looks good so leave it

plot(newData$ExterCond2, resid(full_model), main = 'Residuals vs. Exterior
Condition', xlab = 'Exterior Condition', ylab = 'Residuals')
# has slight fanning effect so may need to fix

plot(newData$HeatingQC2, resid(full_model), main = 'Residuals vs. HeatingQC2',
xlab = 'HeatingQC2', ylab = 'Residuals')
# suggests serial correlation; need a time series model - beyond scope of class
so we leave it as is

plot(newData$CentralAir2, resid(full_model), main = 'Residuals vs.
CentralAir2', xlab = 'CentralAir2', ylab = 'Residuals')
# suggests serial correlation; need a time series model - beyond scope of class
so we leave it as is

plot(newData$GrLivArea, resid(full_model), main = 'Residuals vs. GrLivArea',
xlab = 'GrLivArea', ylab = 'Residuals')
# might need transformation!

plot(newData$KitchenQual2, resid(full_model), main = 'Residuals vs.
KitchenQual2', xlab = 'KitchenQual2', ylab = 'Residuals')
# suggests serial correlation; need a time series model - beyond scope of class
so we leave it as is

plot(newData$GarageCars, resid(full_model), main = 'Residuals vs. GarageCars',
xlab = 'GarageCars', ylab = 'Residuals')
# suggests serial correlation; need a time series model - beyond scope of class
so we leave it as is

plot(newData$TotalBath, resid(full_model), main = 'Residuals vs. TotalBath',
xlab = 'TotalBath', ylab = 'Residuals')
```

```
# suggests serial correlation; need a time series model - beyond scope of class
so we leave it as is

plot(newData$LotArea, resid(full_model), main = 'Residuals vs. LotArea', xlab =
'LotArea', ylab = 'Residuals')
# might need transformation!

# transform GrLivArea of newData
newData$newGrLivArea <- log(newData$GrLivArea)

# transform LotArea of newData
newData$newLotArea <- log(newData$LotArea)

# transform ExterCond of newData
newData$newExterCond2 <- log(newData$ExterCond2)

# update model with transformed Predictors
full_model2 = update(full_model, ~. - LotArea + newLotArea - GrLivArea +
newGrLivArea)
formula(full_model2)
plot(full_model2)
hist(resid(full_model2), main = 'Histogram of the Residuals', xlab =
'Residuals')
shapiro.test(resid(full_model2))

# testing to see if we can fix
test_pred = update(full_model2, ~. - ExterCond2 + newData$newExterCond2)
plot(test_pred)
hist(resid(test_pred), main = 'Histogram of the Residuals', xlab = 'Residuals')

# plotting the new residual vs predictor plots to see if the transformations
were effective
plot(newData$newGrLivArea, resid(full_model2), main = 'Residuals vs.
log(GrLivArea)', xlab = 'log(GrLivArea)', ylab = 'Residuals')

plot(newData$newLotArea, resid(full_model2), main = 'Residuals vs.
log(LotArea)', xlab = 'log(LotArea)', ylab = 'Residuals')

plot(newData$newExterCond2, resid(test_pred), main = 'Residuals vs.
log(ExterCond2)', xlab = 'log(ExterCond2)', ylab = 'Residuals')
```

Addition of Interaction Term to the Model

```
# creating interaction terms of OverallQual with GrLivArea
interact = update(full_model2, ~. + OverallQual*newGrLivArea)
formula(interact)
plot(interact)
hist(resid(interact), main = 'Histogram of the Residuals', xlab = 'Residuals')
summary(interact)
```

Removal of Influential Points

```
# determination of influential points
hat = hatvalues(interact)
sum(hat)
df = dffits(interact)
rule = 2 * sqrt((13 + 1)/(2930 - 13 - 1))
test = (abs(df) > rule)
which(test == TRUE)
test2 = cooks.distance(interact)
which(test2 > 0.5)
# removing influential points
newData$cooksd <- cooks.distance(interact)

newData$cookyn <- ifelse(newData$cooksd < 4/2309, "keep","no")

no_influential <-subset(newData, cookyn == "keep")

final.model <- lm(newSalePrice ~ OverallQual + OverallCond + YearBuilt +
ExterCond2 + HeatingQC2 + CentralAir2 + KitchenQual2 + GarageCars + TotalBath +
newLotArea + newGrLivArea + OverallQual:newGrLivArea, data = no_influential)
plot(final.model)
hist(resid(final.model), main = 'Histogram of the Residuals', xlab =
'Residuals')
shapiro.test(resid(final.model))
# the final model now satisfies all LINE conditions and we can proceed with
answering our research questions

summary(final.model)
```

Research Questions

```
# Research Question 1
# lets do an F test to see if OverallQual and GrLivArea interact to influence
the sale price of a home

reduced = update(final.model, ~. - OverallQual*newGrLivArea + OverallQual +
newGrLivArea)
anova(final.model, reduced)
# here we can see the pvalue is small so we reject the null hypothesis (the
beta is = 0)

# Research Question 2

#Which number addition of total baths creates the most significant increase in
SalePrice?

OneBathPrice <- exp(no_influential$newSalePrice)[no_influential$TotalBath==1]
TwoBathPrice <- exp(no_influential$newSalePrice)[no_influential$TotalBath==2]
ThreeBathPrice <- exp(no_influential$newSalePrice)[no_influential$TotalBath==3]
FourBathPrice <- exp(no_influential$newSalePrice)[no_influential$TotalBath==4]
FiveBathPrice <- exp(no_influential$newSalePrice)[no_influential$TotalBath==5]
```

```
#The sample size for houses with 5 total baths wasn't big enough, so we only
test the increments below.
t.test(OneBathPrice,TwoBathPrice)
t.test(TwoBathPrice,ThreeBathPrice)
t.test(ThreeBathPrice,FourBathPrice)

# Research Question 3

# creating a dataframe with all average values
avg_data = data.frame(OverallQual = mean(no_influential$OverallQual),
OverallCond = mean(no_influential$OverallCond), YearBuilt =
mean(no_influential$YearBuilt), ExterCond2 = mean(no_influential$ExterCond2),
HeatingQC2 = mean(no_influential$HeatingQC2), CentralAir2 =
mean(no_influential$CentralAir2), KitchenQual2 =
mean(no_influential$KitchenQual2), GarageCars =
mean(no_influential$GarageCars), TotalBath = mean(no_influential$TotalBath),
newLotArea = mean(no_influential$newLotArea), newGrLivArea =
mean(no_influential$newGrLivArea), OverallQualnewGrLivArea =
mean(no_influential$OverallQual*no_influential$newGrLivArea))

# creating a prediction interval to find the sale price of a completely average
home
baby_predict = predict(final.model, avg_data, interval = 'prediction')
predict.price = exp(baby_predict) # since our model is using the log of
SalePrice we must restore the original units of SalePrice
predict.price
```

Relevant Code Outputs

Our Dataset

```
>head(newData)
```

	SalePrice <int>	OverallQual <int>	OverallCond <int>	YearBuilt <int>	ExterCond2 <dbl>	BsmtCond2 <dbl>	HeatingQC2 <dbl>	CentralAir2 <dbl>	GrLivArea <int>
1	208500	7	5	2003	4	4	1	2	1710
2	181500	6	8	1976	4	4	1	2	1262
3	223500	7	5	2001	4	4	1	2	1786
4	140000	7	5	1915	4	3	3	2	1717
5	250000	8	5	2000	4	4	1	2	2198
6	143000	5	5	1993	4	4	1	2	1362

6 rows | 1-10 of 16 columns

HeatingQC2 <dbl>	CentralAir2 <dbl>	GrLivArea <int>	BedroomAbvGr <int>	KitchenQual2 <dbl>	TotRmsAbvGrd <int>	GarageCars <int>	TotalBath <dbl>	YrSold <int>	LotArea <int>
1	2	1710	3	3	8	2	3.5	2008	8450
1	2	1262	3	4	6	2	2.5	2007	9600
1	2	1786	3	3	6	2	3.5	2008	11250
3	2	1717	3	3	7	3	2.0	2006	9550
1	2	2198	4	3	9	3	3.5	2008	14260
1	2	1362	1	4	5	2	2.5	2009	14115

6 rows | 8-17 of 16 columns

Stepwise Regression Using Akaike's Information Criterion

```
> formula(step(lm(SalePrice ~. -SalePrice, data = newData)))

Start:  AIC=29903.38
SalePrice ~ (OverallQual + OverallCond + YearBuilt + ExterCond2 +
    BsmtCond2 + HeatingQC2 + CentralAir2 + GrLivArea + BedroomAbvGr +
    KitchenQual2 + TotRmsAbvGrd + GarageCars + TotalBath + YrSold +
    LotArea) - SalePrice

               Df  Sum of Sq        RSS   AIC
- ExterCond2    1 1.0284e+08 1.8615e+12 29902
- YrSold        1 3.6399e+08 1.8618e+12 29902
- HeatingQC2    1 9.5134e+08 1.8624e+12 29902
- BsmtCond2     1 2.2645e+09 1.8637e+12 29903
<none>                      1.8614e+12 29903
- CentralAir2   1 4.6469e+09 1.8661e+12 29905
- TotRmsAbvGrd  1 6.3788e+09 1.8678e+12 29906
- TotalBath     1 2.4257e+10 1.8857e+12 29920
- OverallCond   1 3.2796e+10 1.8942e+12 29926
- BedroomAbvGr  1 3.4347e+10 1.8958e+12 29927
- YearBuilt     1 4.4275e+10 1.9057e+12 29935
- GarageCars    1 8.1369e+10 1.9428e+12 29962
- LotArea       1 8.3612e+10 1.9450e+12 29964
- KitchenQual2  1 1.4514e+11 2.0066e+12 30008
- GrLivArea     1 1.9237e+11 2.0538e+12 30041
- OverallQual   1 2.4008e+11 2.1015e+12 30074

Step:  AIC=29901.46
SalePrice ~ OverallQual + OverallCond + YearBuilt + BsmtCond2 +
    HeatingQC2 + CentralAir2 + GrLivArea + BedroomAbvGr + KitchenQual2 +
    TotRmsAbvGrd + GarageCars + TotalBath + YrSold + LotArea

               Df  Sum of Sq        RSS   AIC
- YrSold        1 3.5768e+08 1.8619e+12 29900
- HeatingQC2    1 9.8309e+08 1.8625e+12 29900
- BsmtCond2     1 2.1723e+09 1.8637e+12 29901
<none>                      1.8615e+12 29902
- CentralAir2   1 4.6115e+09 1.8661e+12 29903
- TotRmsAbvGrd  1 6.3327e+09 1.8679e+12 29904
- TotalBath     1 2.4568e+10 1.8861e+12 29918
- OverallCond   1 3.3878e+10 1.8954e+12 29925
- BedroomAbvGr  1 3.4290e+10 1.8958e+12 29925
- YearBuilt     1 4.4517e+10 1.9060e+12 29933
- GarageCars    1 8.1302e+10 1.9428e+12 29960
- LotArea       1 8.3550e+10 1.9451e+12 29962
- KitchenQual2  1 1.4538e+11 2.0069e+12 30006
- GrLivArea     1 1.9241e+11 2.0539e+12 30039
- OverallQual   1 2.4017e+11 2.1017e+12 30072

Step:  AIC=29899.74
SalePrice ~ OverallQual + OverallCond + YearBuilt + BsmtCond2 +
    HeatingQC2 + CentralAir2 + GrLivArea + BedroomAbvGr + KitchenQual2 +
```

```
    TotRmsAbvGrd + GarageCars + TotalBath + LotArea

               Df  Sum of Sq        RSS   AIC
- HeatingQC2    1 9.8463e+08 1.8629e+12 29898
- BsmtCond2     1 2.0644e+09 1.8639e+12 29899
<none>                       1.8619e+12 29900
- CentralAir2   1 4.6412e+09 1.8665e+12 29901
- TotRmsAbvGrd  1 6.3577e+09 1.8682e+12 29903
- TotalBath     1 2.4281e+10 1.8862e+12 29916
- OverallCond   1 3.3647e+10 1.8955e+12 29923
- BedroomAbvGr  1 3.4118e+10 1.8960e+12 29924
- YearBuilt     1 4.4648e+10 1.9065e+12 29932
- GarageCars    1 8.1615e+10 1.9435e+12 29959
- LotArea       1 8.3703e+10 1.9456e+12 29960
- KitchenQual2  1 1.4514e+11 2.0070e+12 30005
- GrLivArea     1 1.9263e+11 2.0545e+12 30038
- OverallQual   1 2.4051e+11 2.1024e+12 30071

Step:  AIC=29898.49
SalePrice ~ OverallQual + OverallCond + YearBuilt + BsmtCond2 +
    CentralAir2 + GrLivArea + BedroomAbvGr + KitchenQual2 + TotRmsAbvGrd +
    GarageCars + TotalBath + LotArea

               Df  Sum of Sq        RSS   AIC
- BsmtCond2     1 2.0608e+09 1.8649e+12 29898
<none>                       1.8629e+12 29898
- CentralAir2   1 4.7099e+09 1.8676e+12 29900
- TotRmsAbvGrd  1 6.5927e+09 1.8695e+12 29902
- TotalBath     1 2.4196e+10 1.8871e+12 29915
- OverallCond   1 3.4699e+10 1.8976e+12 29923
- BedroomAbvGr  1 3.5232e+10 1.8981e+12 29923
- YearBuilt     1 4.9241e+10 1.9121e+12 29934
- GarageCars    1 8.1353e+10 1.9442e+12 29957
- LotArea       1 8.2994e+10 1.9459e+12 29958
- KitchenQual2  1 1.5324e+11 2.0161e+12 30009
- GrLivArea     1 1.9371e+11 2.0566e+12 30037
- OverallQual   1 2.4578e+11 2.1086e+12 30073

Step:  AIC=29898.06
SalePrice ~ OverallQual + OverallCond + YearBuilt + CentralAir2 +
    GrLivArea + BedroomAbvGr + KitchenQual2 + TotRmsAbvGrd +
    GarageCars + TotalBath + LotArea

               Df  Sum of Sq        RSS   AIC
<none>                       1.8649e+12 29898
- CentralAir2   1 5.4670e+09 1.8704e+12 29900
- TotRmsAbvGrd  1 6.6982e+09 1.8716e+12 29901
- TotalBath     1 2.4222e+10 1.8891e+12 29914
- OverallCond   1 3.5405e+10 1.9003e+12 29923
- BedroomAbvGr  1 3.5559e+10 1.9005e+12 29923
- YearBuilt     1 5.1845e+10 1.9168e+12 29935
- GarageCars    1 8.1149e+10 1.9461e+12 29957
- LotArea       1 8.3712e+10 1.9486e+12 29958
- KitchenQual2  1 1.5215e+11 2.0171e+12 30008
```

```
- GrLivArea     1 1.9275e+11 2.0577e+12 30036
- OverallQual   1 2.4588e+11 2.1108e+12 30072
SalePrice ~ OverallQual + OverallCond + YearBuilt + CentralAir2 +
    GrLivArea + BedroomAbvGr + KitchenQual2 + TotRmsAbvGrd +
    GarageCars + TotalBath + LotArea
```

Our Correlation Matrix

```
> cor(subset(newData, select = -c(newSalePrice, SalePrice)))
```

	OverallQual	OverallCond	YearBuilt	ExterCond2	BsmtCond2	HeatingQC2	CentralAir2	GrLivArea	BedroomAbvGr	KitchenQual2	TotRmsAbvGrd	GarageCars
OverallQual	1.00000000	-0.122540745	0.57091235	0.13001768	0.068562565	-0.442980099	0.2295437399	0.59332677	0.09242371	-0.568627430	0.43473315	0.60883301
OverallCond	-0.12254074	1.000000000	-0.38940285	-0.22398561	-0.004224988	0.074753638	0.0897453595	-0.08850537	0.01453862	0.070803368	-0.06082124	-0.19537397
YearBuilt	0.57091235	-0.389402845	1.00000000	0.27559015	0.172593842	-0.424795290	0.3823495129	0.19127737	-0.08214935	-0.361191757	0.09046111	0.53777195
ExterCond2	0.13001768	-0.223985612	0.27559015	1.00000000	0.179974738	-0.057814691	0.1147627245	0.03963581	-0.02470092	-0.050263459	0.03154484	0.17944315
BsmtCond2	0.06856256	-0.004224988	0.17259384	0.17997474	1.000000000	-0.060275488	0.1830399022	-0.02537011	-0.04815600	-0.004948733	-0.03267162	0.06763541
HeatingQC2	-0.44298010	0.074753638	-0.42479529	-0.05781469	-0.060275488	1.000000000	-0.1736568955	-0.25209036	0.03768887	0.394633090	-0.17637035	-0.31703471
CentralAir2	0.22954374	0.089745359	0.38234951	0.11476272	0.183039902	-0.173656896	1.0000000000	0.06702698	-0.01408595	-0.051726993	0.01546875	0.21167586
GrLivArea	0.59332677	-0.088505372	0.19127737	0.03963581	-0.025370106	-0.252090361	0.0670269760	1.00000000	0.51740426	-0.389763000	0.82763133	0.46592041
BedroomAbvGr	0.09242371	0.014538618	-0.08214935	-0.02470092	-0.048156003	0.037688867	-0.0140859496	0.51740426	1.00000000	0.032936279	0.67030873	0.08263383
KitchenQual2	-0.56862743	0.070803368	-0.36119176	-0.05026346	-0.004948733	0.394633090	-0.0517269930	-0.38976300	0.03293628	1.000000000	-0.27698233	-0.37678504
TotRmsAbvGrd	0.43473315	-0.060821236	0.09046111	0.03154484	-0.032671623	-0.176370352	0.0154687522	0.82763133	0.67030873	-0.276982327	1.00000000	0.36328042
GarageCars	0.60883301	-0.195373966	0.53777195	0.17944315	0.067635412	-0.317034712	0.2116758551	0.46592041	0.08263383	-0.376785042	0.36328042	1.00000000
TotalBath	0.53583920	-0.187181417	0.52223944	0.09346407	0.063563740	-0.301207868	0.1944353432	0.59251770	0.23263261	-0.355661521	0.46189188	0.48335835
YrSold	-0.02628524	0.049214244	-0.01399356	-0.02227266	0.066976560	0.004242821	-0.0009742282	-0.03926643	-0.05108397	-0.006562478	-0.04716168	-0.03746290
LotArea	0.10429696	-0.006468598	0.01288947	0.01738908	0.016332896	0.002594360	0.0442395441	0.26258789	0.12111977	-0.058112666	0.19119625	0.15384278

	TotalBath	YrSold	LotArea
OverallQual	0.53583920	-0.0262852403	0.104296959
OverallCond	-0.18718142	0.0492142440	-0.006468598
YearBuilt	0.52223944	-0.0139935576	0.012889472
ExterCond2	0.09346407	-0.0222726580	0.017389080
BsmtCond2	0.06356374	0.0669765600	0.016332896
HeatingQC2	-0.30120787	0.0042428209	0.002594360
CentralAir2	0.19443534	-0.0009742282	0.044239544
GrLivArea	0.59251770	-0.0392664254	0.262587891
BedroomAbvGr	0.23263261	-0.0510839687	0.121119769
KitchenQual2	-0.35566152	-0.0065624779	-0.058112666
TotRmsAbvGrd	0.46189188	-0.0471616783	0.191196250
GarageCars	0.48335835	-0.0374629034	0.153842778
TotalBath	1.00000000	0.0204187466	0.205633758
YrSold	0.02041875	1.0000000000	-0.014505932
LotArea	0.20563376	-0.0145059317	1.000000000

Anova Table for Interaction Term

```
> summary(interact)
```

```
Call:
lm(formula = newSalePrice ~ OverallQual + OverallCond + YearBuilt +
    ExterCond2 + HeatingQC2 + CentralAir2 + KitchenQual2 + GarageCars +
    TotalBath + newLotArea + newGrLivArea + OverallQual:newGrLivArea,
    data = newData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43980 -0.07284  0.00508  0.08337  0.51668

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.3971466  0.5681342   4.219 2.61e-05 ***
OverallQual              0.0253086  0.0561809   0.450  0.65243
OverallCond              0.0498749  0.0040544  12.301  < 2e-16 ***
YearBuilt                0.0026669  0.0002202  12.112  < 2e-16 ***
ExterCond2               0.0197962  0.0098790   2.004  0.04527 *
HeatingQC2              -0.0105745  0.0033709  -3.137  0.00174 **
CentralAir2              0.1123354  0.0189535   5.927 3.88e-09 ***
KitchenQual2            -0.0415371  0.0059760  -6.951 5.54e-12 ***
GarageCars               0.0668772  0.0071801   9.314  < 2e-16 ***
TotalBath                0.0529077  0.0071197   7.431 1.86e-13 ***
newLotArea               0.1244730  0.0082290  15.126  < 2e-16 ***
newGrLivArea             0.2864320  0.0478395   5.987 2.70e-09 ***
OverallQual:newGrLivArea 0.0080217  0.0076401   1.050  0.29392
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1448 on 1410 degrees of freedom
Multiple R-squared:  0.866,     Adjusted R-squared:  0.8649
F-statistic: 759.5 on 12 and 1410 DF,  p-value: < 2.2e-16
```

SummaryOutput for Final Model

```
> summary(final.model)

Call:
lm(formula = newSalePrice ~ OverallQual + OverallCond + YearBuilt +
    ExterCond2 + HeatingQC2 + CentralAir2 + KitchenQual2 + GarageCars +
    TotalBath + newLotArea + newGrLivArea + OverallQual:newGrLivArea,
    data = no_influential)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29896 -0.06491  0.00335  0.06625  0.33478

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.6105995  0.4582271   7.879 6.97e-15 ***
OverallQual             -0.2214000  0.0495557  -4.468 8.60e-06 ***
OverallCond              0.0452249  0.0031881  14.185  < 2e-16 ***
YearBuilt                0.0027865  0.0001709  16.302  < 2e-16 ***
ExterCond2               0.0148835  0.0081014   1.837 0.066418 .
HeatingQC2              -0.0083624  0.0024585  -3.401 0.000691 ***
CentralAir2              0.1307754  0.0162903   8.028 2.23e-15 ***
KitchenQual2            -0.0420601  0.0045551  -9.234  < 2e-16 ***
GarageCars               0.0612891  0.0055779  10.988  < 2e-16 ***
TotalBath                0.0434442  0.0053406   8.135 9.67e-16 ***
newLotArea               0.1290550  0.0061470  20.995  < 2e-16 ***
newGrLivArea             0.0901189  0.0412053   2.187 0.028918 *
OverallQual:newGrLivArea 0.0412412  0.0067679   6.094 1.46e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09971 on 1281 degrees of freedom
Multiple R-squared:  0.9203,    Adjusted R-squared:  0.9195
F-statistic:  1232 on 12 and 1281 DF,  p-value: < 2.2e-16
```

Anova Table for Research Question 1

```
Analysis of Variance Table

Model 1: newSalePrice ~ OverallQual + OverallCond + YearBuilt + ExterCond2 +
    HeatingQC2 + CentralAir2 + KitchenQual2 + GarageCars + TotalBath +
    newLotArea + newGrLivArea + OverallQual:newGrLivArea
Model 2: newSalePrice ~ OverallCond + YearBuilt + ExterCond2 + HeatingQC2 +
    CentralAir2 + KitchenQual2 + GarageCars + TotalBath + newLotArea +
    OverallQual + newGrLivArea
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1281 12.736
2   1282 13.105 -1  -0.36918 37.133 1.457e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```