ASSIGNMENT 3

Predict 410 Fall 2017

Zeeshan Latifi

Northwestern University

Introduction:

This is an exploratory data analysis of housing data for Ames, Iowa. In this analysis, we will be using determining factors that can help predict the sales prices for a typical home in Ames, Iowa. The data has been provided by DeCock (2011). We will be looking for several predictor variables that will help determine our response variable: Sales Price.

To do this, we'll work through many aspects of data analysis. Initially, we'll evaluate our data, define the sample population, perform analysis, and finally model our findings. From our model, we'll be able to assess whether our response variable can be predicted accurately using other predictor variables provided in the data set.

Sample Population:

There are 82 variables and 2,930 observations in the Ames, Iowa data set. We begin by conducting a waterfall in R to clean our data set. By evaluation of each variable, we've identified what constitutes a 'typical' home in Ames.

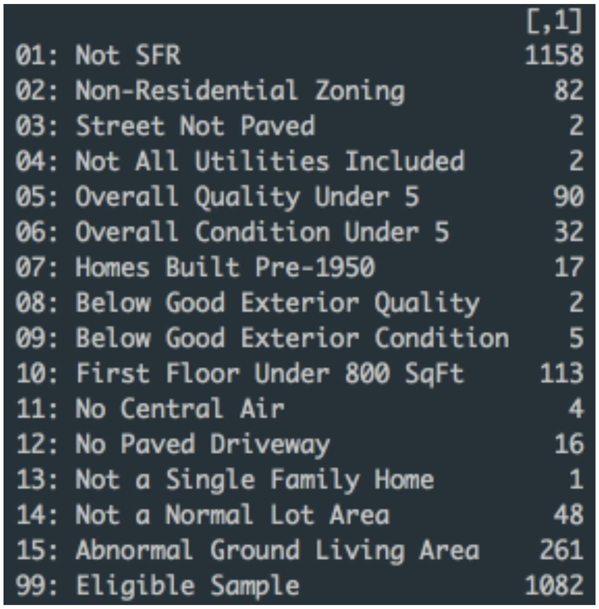
We began with filtering on the 'SubClass' field. This field identifies the class of the home. The decision was to keep only homes that are:

- 1-STORY 1946 & NEWER ALL STYLES
- 2-STORY 1946 & NEWER
- SPLIT OR MULTI-LEVEL

We then removed all non-residential zoning, keeping only residential high, medium, and low density. Next, removed all homes that were not on a paved street and did not have all public utilities included. A 'typical' home should have all standard utilities available.

To keep with our assumptions, the decision was made to only include homes that are in overall condition and quality of a 5 or higher. This means homes quality and condition ranked 'average' or higher. The same decision was made when filtering on the homes' exterior quality and condition. To eliminate homes that may skew our data set, only houses that were built in 1950 or later were included. Per our definition of the 'typical' home, we decided to only include homes that have square footage of 800 ft.^{2} or higher for the first level. With that in mind we also eliminated homes without a paved driveway or central air. Also, disregarded townhomes and homes with lot areas less than 5,000 ft^2 and above 20,000 ft^2. Additionally, we removed homes that had above 2,000 ft.^{2} of above ground living area, as this is atypical in Ames, Iowa. Finally, our sample will not include homes without a garage. With these transformations, the observations were reduced down to 1,082. Figure 1 displays the count for each of the reductions.

Figure 1:

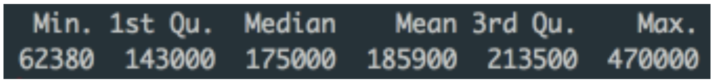


```
                                  [,1]
01: Not SFR                       1158
02: Non-Residential Zoning          82
03: Street Not Paved                 2
04: Not All Utilities Included       2
05: Overall Quality Under 5         90
06: Overall Condition Under 5       32
07: Homes Built Pre-1950            17
08: Below Good Exterior Quality      2
09: Below Good Exterior Condition    5
10: First Floor Under 800 SqFt     113
11: No Central Air                   4
12: No Paved Driveway               16
13: Not a Single Family Home         1
14: Not a Normal Lot Area           48
15: Abnormal Ground Living Area    261
99: Eligible Sample               1082
```

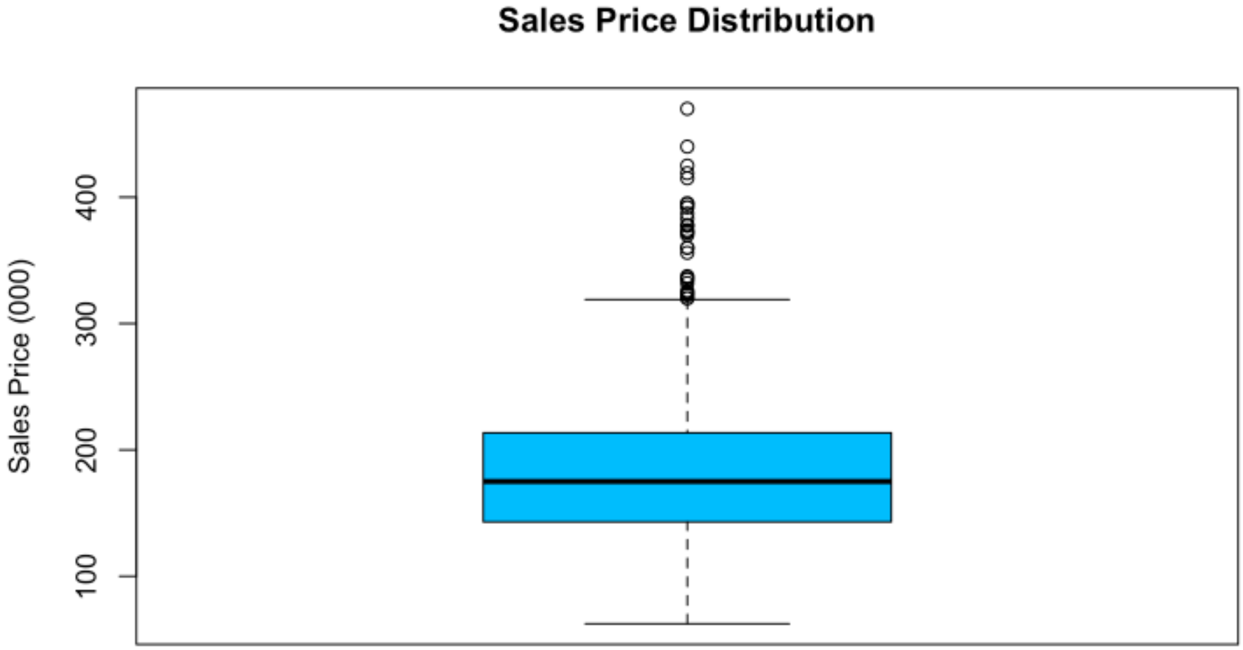
Exploratory Analysis:

To help predict the sale price for a home we should understand the response variable first. Figure 5 represents a summary for the 'SalePrice' data in our sample population as well the distribution of the sale prices within our sample for Ames.

Figure 2:



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  62380  143000  175000  185900  213500  470000
```

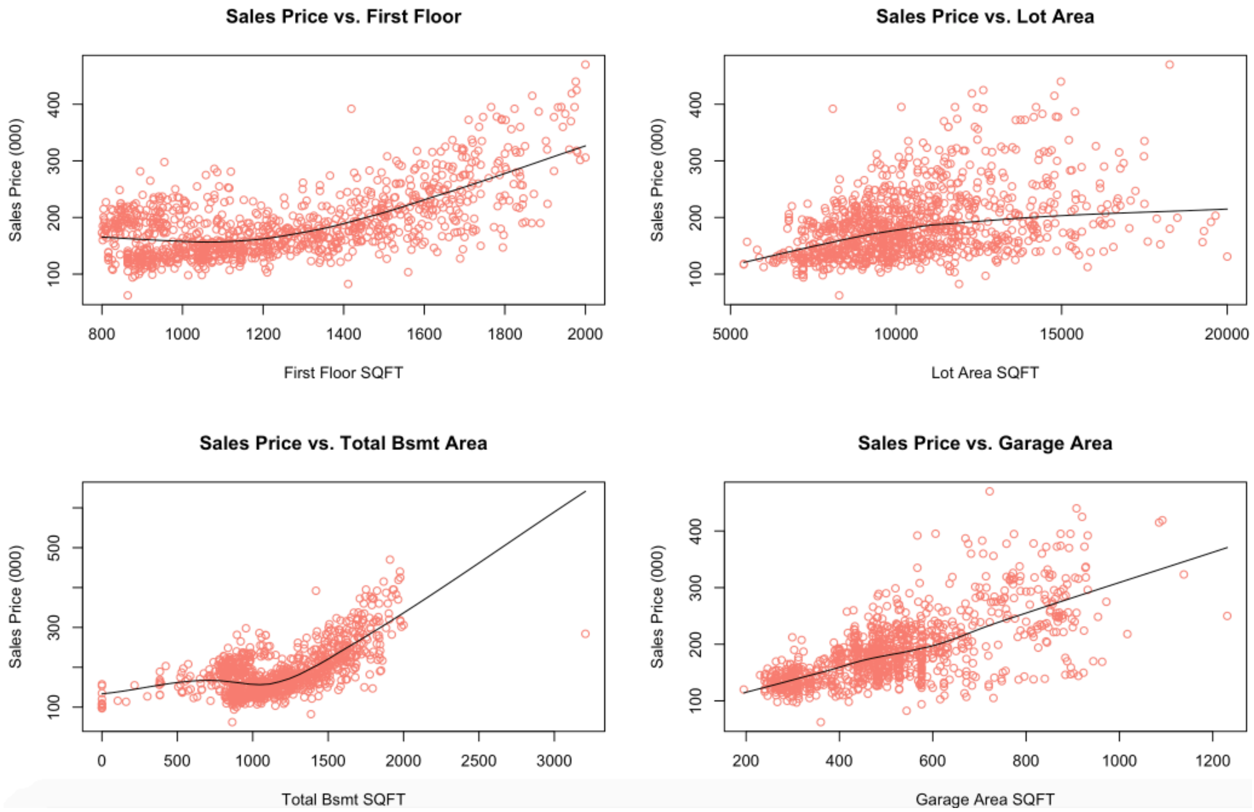


As shown above, the mean pricing for a home in Ames according to the selected criteria is $185,900. We found no values that seemed out of the ordinary here. There is potential that some max values can be outliers, but it's unable to be determined at this time; further analysis is required.

Within our analysis, we've decided to take a look at some continuous variables and their correlation with Sales Price along with the corresponding LOESS curve. In the figure below you will see:

- Sales Price vs. First Floor Area
- Sales Price vs. Lot Area
- Sales Price vs. Total Basement Area
- Sales Price vs. Garage Area

Figure 3:

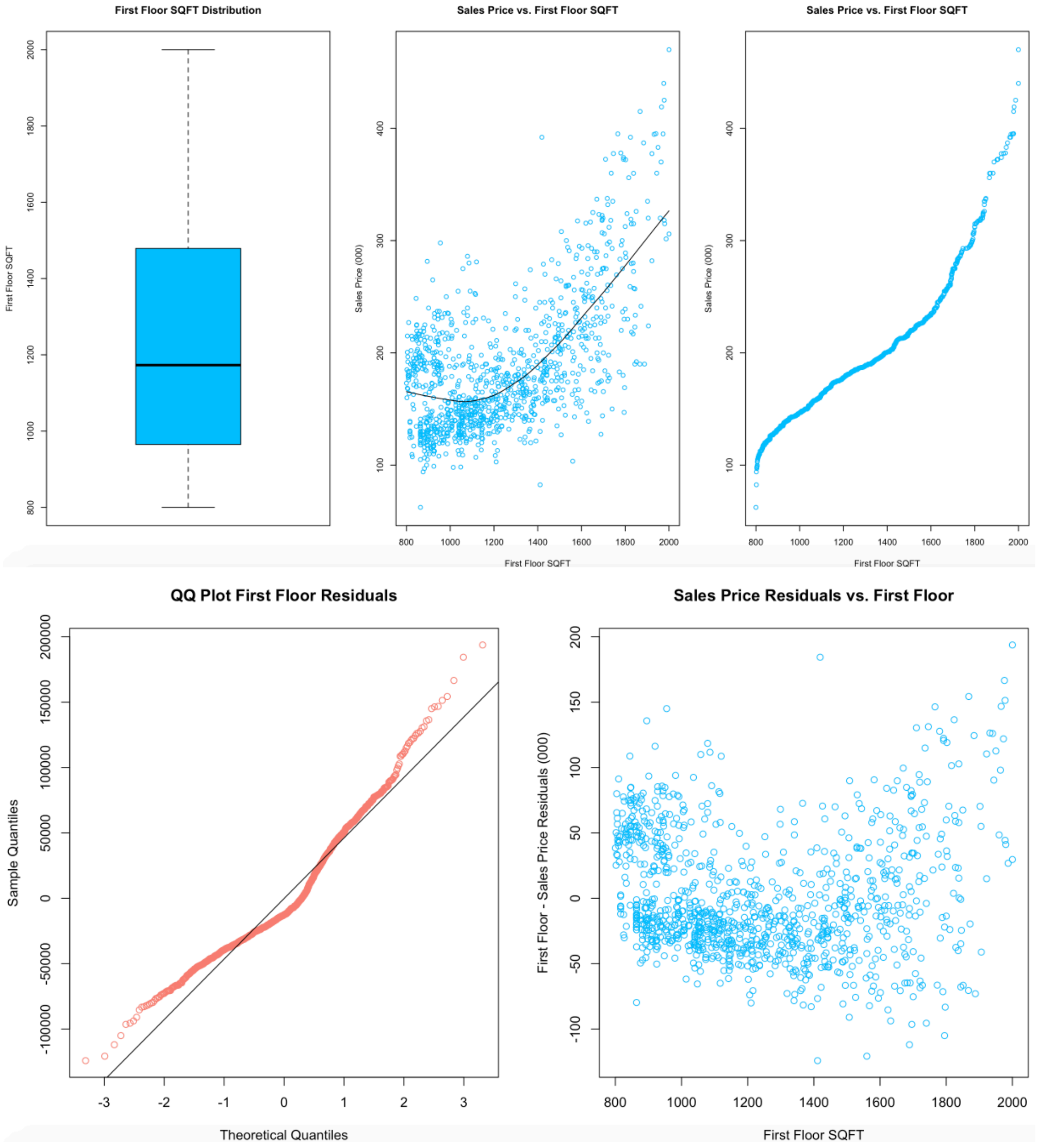


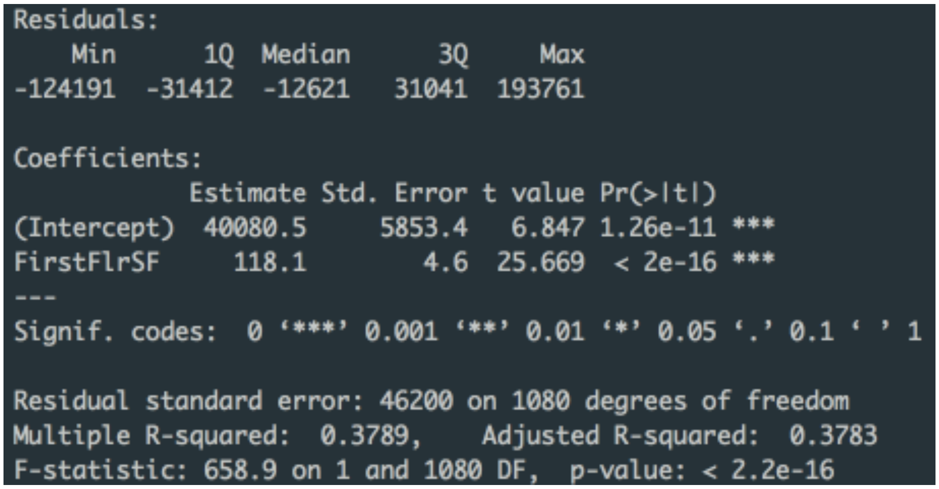
Based on the analysis above, the two best continuous variables to help predict ‘Sales Price’ are the ‘First Floor SQFT’ and ‘Lot Area SQFT’.

Predictor Variable Analysis – First Floor Square Footage:

For simple linear regression, only continuous variables can be used. First, we explore ‘First Floor SQFT’. Figure 4 shows us the distribution of the homes first floor square footage in the Ames, Iowa dataset after we’ve eliminated several factors. It also shows us the correlation with the sales price along with its LOESS line. To evaluate the goodness-of-fit, we’ve included the residual plot, as well as the QQ plot of residuals. As you can see in Figure 4 below, the residual plot shows ‘First Floor SQFT’ being an accurate predictor of ‘Sales Price’ but it may not be best predictor by itself.

Figure 4:


First Floor SQFT Distribution
First Floor SQFT
800
1000
1200
1400
1600
1800
2000
Sales Price vs. First Floor SQFT
Sales Price (000)
100
200
300
400
800
1000
1200
1400
1600
1800
2000
First Floor SQFT
Sales Price vs. First Floor SQFT
Sales Price (000)
100
200
300
400
800
1000
1200
1400
1600
1800
2000
First Floor SQFT
QQ Plot First Floor Residuals
Sample Quantiles
-100000
-50000
0
50000
100000
150000
200000
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Sales Price Residuals vs. First Floor
First Floor - Sales Price Residuals (000)
-100
-50
0
50
100
150
200
800
1000
1200
1400
1600
1800
2000
First Floor SQFT



```
Residuals:
    Min      1Q  Median      3Q     Max 
-124191  -31412  -12621   31041  193761 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  40080.5     5853.4   6.847 1.26e-11 ***
FirstFlrSF     118.1        4.6  25.669  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46200 on 1080 degrees of freedom
Multiple R-squared:  0.3789,	Adjusted R-squared:  0.3783 
F-statistic: 658.9 on 1 and 1080 DF,  p-value: < 2.2e-16
```

$$y = 40080.5 + 118.1x_1$$

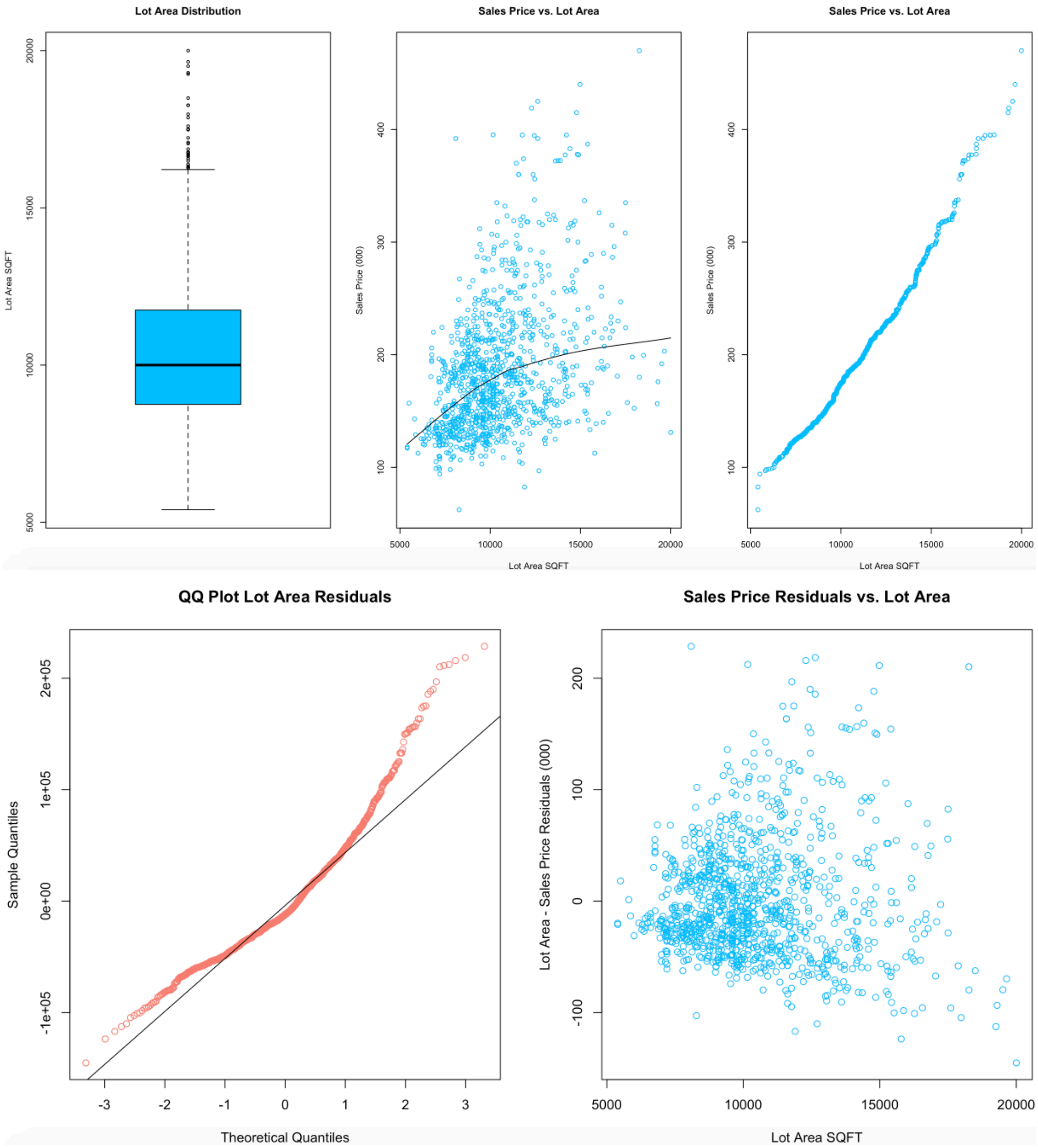
Per the summary statistics above, we have p-values that indicate that we can reject the null hypothesis. There seems to be a relationship between the first floor square footage and sales price. However, per the adjusted R^2 value of 0.3783, this may not be a good predictor of sales price.

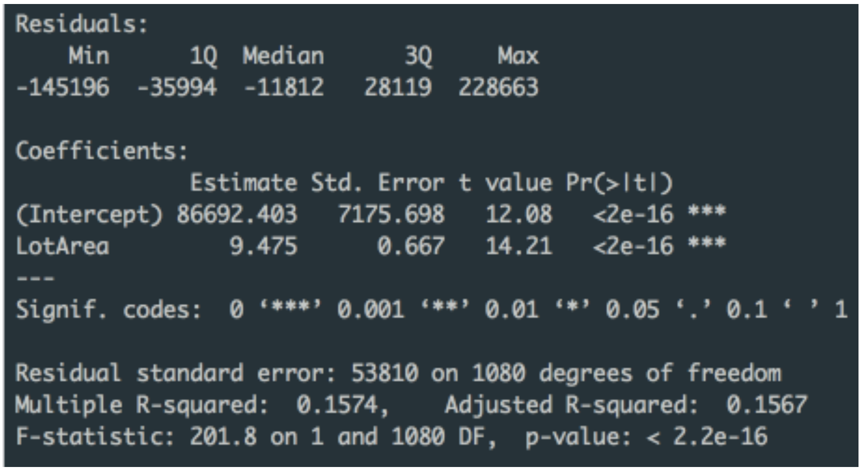
Predictor Variable Analysis – Lot Area Square Footage:

Next, we evaluate the ‘Lot Area’ in Figure 5. Similar to Figure 4, this shows us the distribution of the homes lot area square footage in the Ames, Iowa dataset after our elimination of several observations. It also shows us the correlation with the sale price along with its LOESS line. When evaluating the residuals, we’ve included the residual plot as well as the QQ plot of residuals.

The summary statistics in Figure 5 show that we have p-values that indicate a relationship between the lot area square footage and sales price. However, per the adjusted R^2 value of 0.1567, which is very low, this is not be a good predictor of sales price as a standalone variable. Per the residual plot, we may have some outliers as the visible shape seem to be a cone.

Figure 5:


Lot Area Distribution
Lot Area SQFT
5000
10000
15000
20000
Sales Price vs. Lot Area
Sales Price (000)
100
200
300
400
5000
10000
15000
20000
Lot Area SQFT
Sales Price vs. Lot Area
Sales Price (000)
100
200
300
400
5000
10000
15000
20000
Lot Area SQFT
QQ Plot Lot Area Residuals
Sample Quantiles
-1e+05
0e+00
1e+05
2e+05
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Sales Price Residuals vs. Lot Area
Lot Area - Sales Price Residuals (000)
-100
0
100
200
5000
10000
15000
20000
Lot Area SQFT



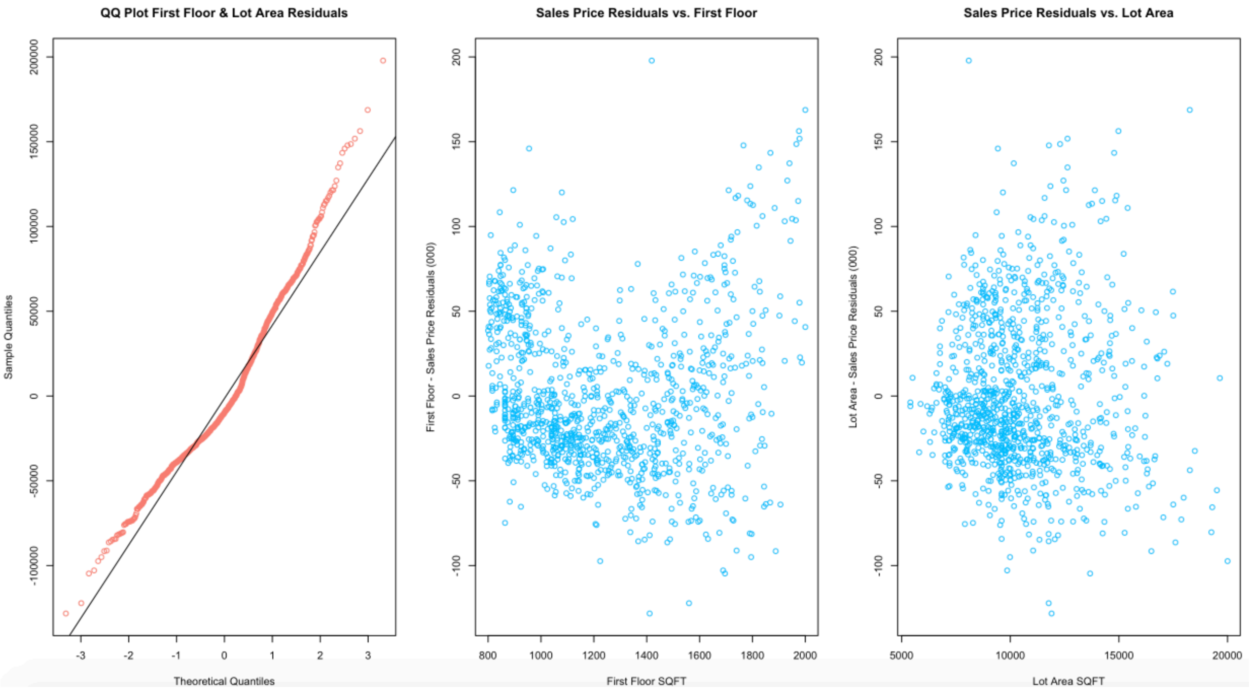
```
Residuals:
    Min      1Q  Median      3Q     Max
-145196  -35994  -11812   28119  228663

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 86692.403   7175.698   12.08   <2e-16 ***
LotArea         9.475      0.667   14.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53810 on 1080 degrees of freedom
Multiple R-squared:  0.1574,	Adjusted R-squared:  0.1567
F-statistic: 201.8 on 1 and 1080 DF,  p-value: < 2.2e-16
```

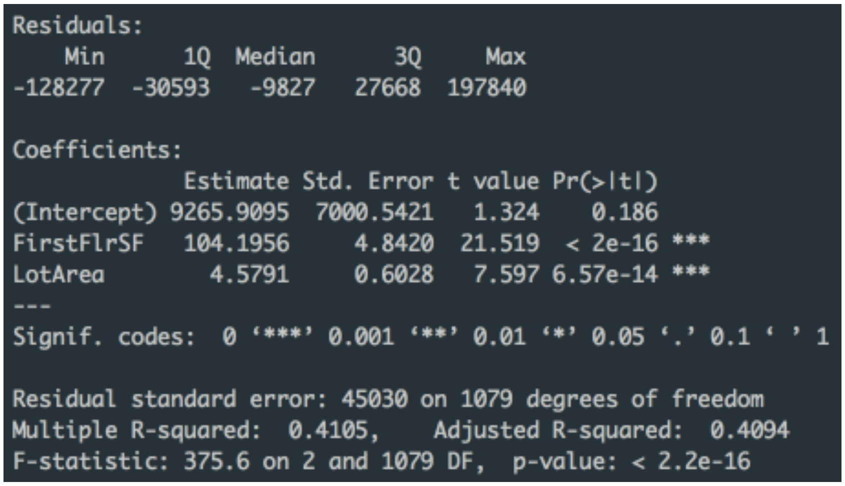
$$y = 86692.403 + 9.475x_1$$

Multiple Linear Regression:

After conducting our simple linear regression analysis, we’ll now look to combine our two variables in order to test whether we can fit a better model. In this model shown in Figure 6 below, ‘First Floor SQFT’ and ‘Lot Area’ are used in order to fit a better model. We’ve included the QQ-Plot for residuals as well as the combined residuals against both first floor square footage and lot area square footage.

Figure 6:





```
Residuals:
    Min      1Q  Median      3Q     Max 
-128277  -30593   -9827   27668  197840 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 9265.9095  7000.5421   1.324    0.186    
FirstFlrSF   104.1956     4.8420  21.519  < 2e-16 ***
LotArea        4.5791     0.6028   7.597 6.57e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45030 on 1079 degrees of freedom
Multiple R-squared:  0.4105,	Adjusted R-squared:  0.4094 
F-statistic: 375.6 on 2 and 1079 DF,  p-value: < 2.2e-16
```

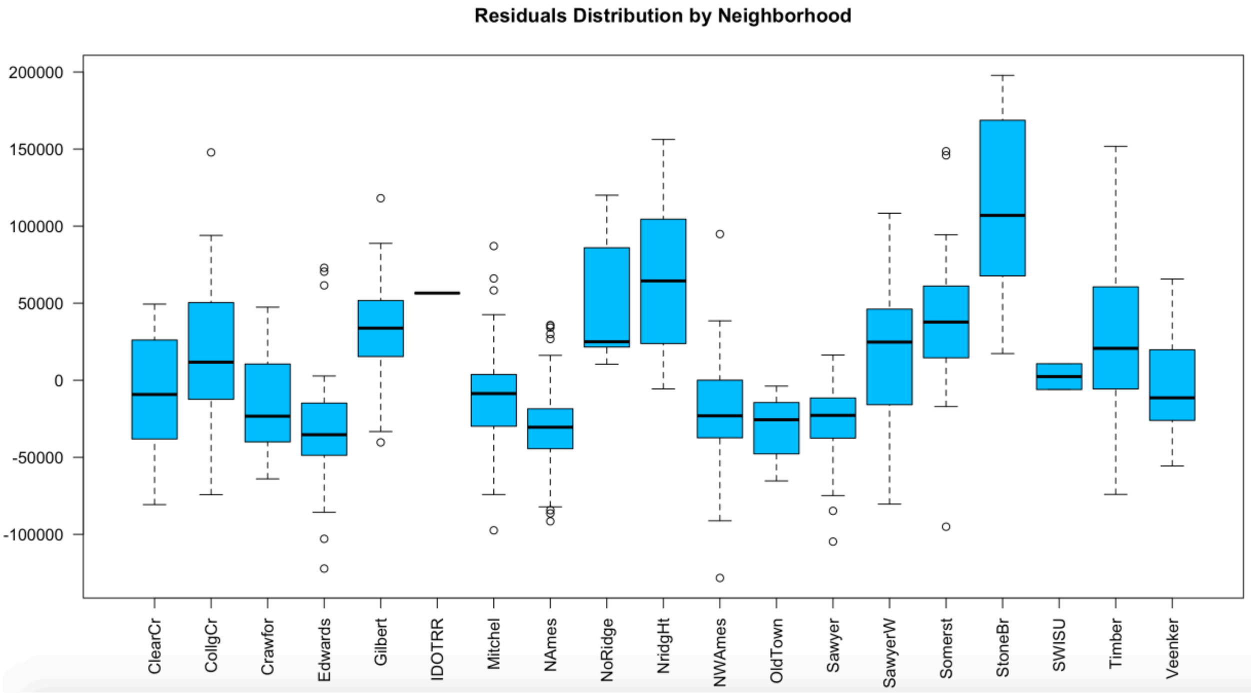
$$y = 9265.9095 + 104.1956x_1 + 4.5791x_2$$

Based on the adjusted R^2 value of 0.4094, we can say that the model that is combined using both first floor and lot area square footage is better than just the individual models by themselves. However, this regression model may not be the best predictor for sales price. In order to find the best model: adding additional variables will not always translate into a better model. The residual plot for ‘Lot Area’ is better distributed than the residual plot for ‘First Floor SQFT’, indicating a better goodness-of-fit for the ‘Lot Area’.

Neighborhood Assessment:

In order to better understand our sample set and how neighborhoods can affect the sample, we'll take a look at the residual distribution by neighborhood. Figure 7 below illustrates this:

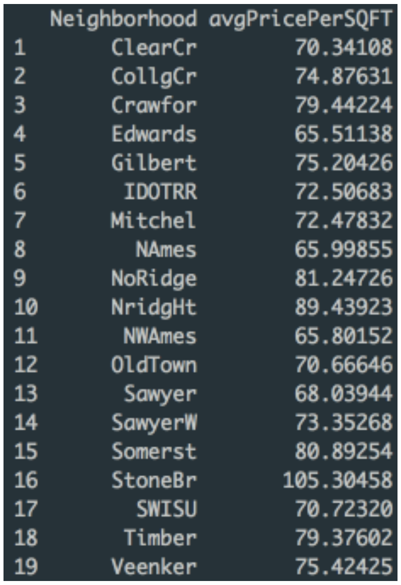
Figure 7:



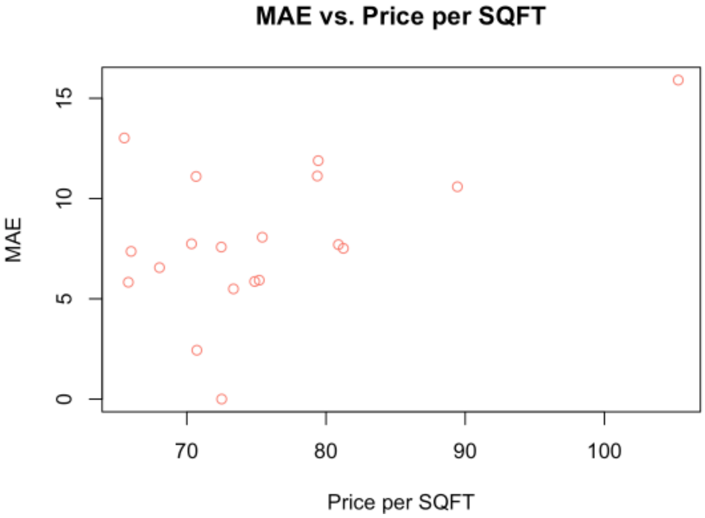
According to our data, 'StoneBr' seems to be a more expensive part of Ames. This means it can be over-predicted. The opposite can be said for 'NAmes' 'OldTown', 'Sawyer' and 'Edwards'. For places like 'IDOTRR' we don't enough information to accurately predict the sales price here. Only one home resides within our data set. 'StoneBr' will be most likely to cause the prediction of homes in Ames to show up higher than what may be the 'typical' home price. Some areas such as 'Timber' and 'SawyerW' have a broad range of prices that can help diversify our data set.

Diving a bit deeper, below is the data for price per square foot for Ames, by neighborhood. This was calculated by dividing the 'SalesPrice' from the sum of 'Above Ground SQFT' and 'Total Basement Area SQFT'. The plot in Figure 8 also displays the MAE (Mean Absolute Error) for each neighborhood within Ames versus the price per square footage. As shown, there is no relationship that can be inferred between these two values.

Figure 8:



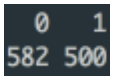
```
   Neighborhood avgPricePerSQFT
1       ClearCr        70.34108
2       CollgCr        74.87631
3       Crawfor        79.44224
4       Edwards        65.51138
5       Gilbert        75.20426
6        IDOTRR        72.50683
7       Mitchel        72.47832
8         NAmes        65.99855
9       NoRidge        81.24726
10      NridgHt        89.43923
11       NWAmes        65.80152
12      OldTown        70.66646
13       Sawyer        68.03944
14      SawyerW        73.35268
15      Somerst        80.89254
16      StoneBr       105.30458
17        SWISU        70.72320
18       Timber        79.37602
19      Veenker        75.42425
```



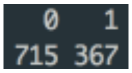
We'll now take a family of indicator variables to our multiple regression formula. We will break the price per square foot into 3 groups. Our baseline category being all homes that are under less than \$70 per ft.^{2}. The next two will be homes that are between \$70/ft.^{2} and \$80/ft.^{2} and homes above \$80/ft.^{2}. Homes will be marked as a binary value. If they fall under these categories they will be marked as 1 and if they do not, they will be marked as 0. As shown in the figure below, breaking out the homes into groups has a significant positive effect on the fit of our model. Adjusted R^2 is shown at 0.9718, which indicates a good model for predictability. Our MAE is also less than our original model (shown in Figure 6). The original model's MAE was calculated at 35,703.14 and our model when including the indicator variables calculated MAE at 23,970.27. As you can see the MAE is lower with our model in Figure 8, indicating a better fit when using indicator variables to group homes by price per square foot. Further assessment of the summary show that homes with less than \$70/ft.^{2} did not receive a p-value score (0.588) that would help us predict home prices in Ames. Further analysis of this variable is required.

Figure 9:

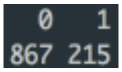
Homes under \$70/ft.^{2} Homes between \$70/ft.^{2} and \$80/ft.^{2} Homes over \$80/ft.^{2}



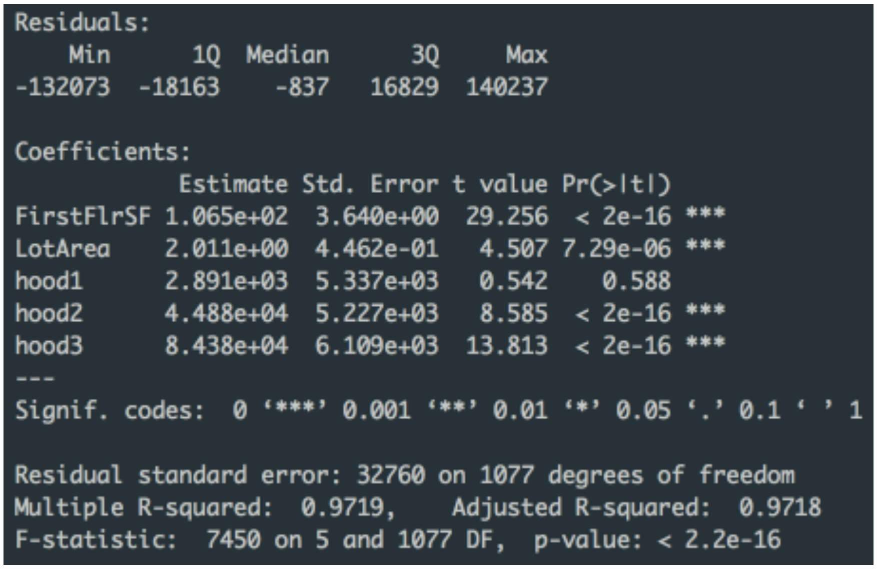
```
  0   1
582 500
```



```
  0   1
715 367
```



```
  0   1
867 215
```

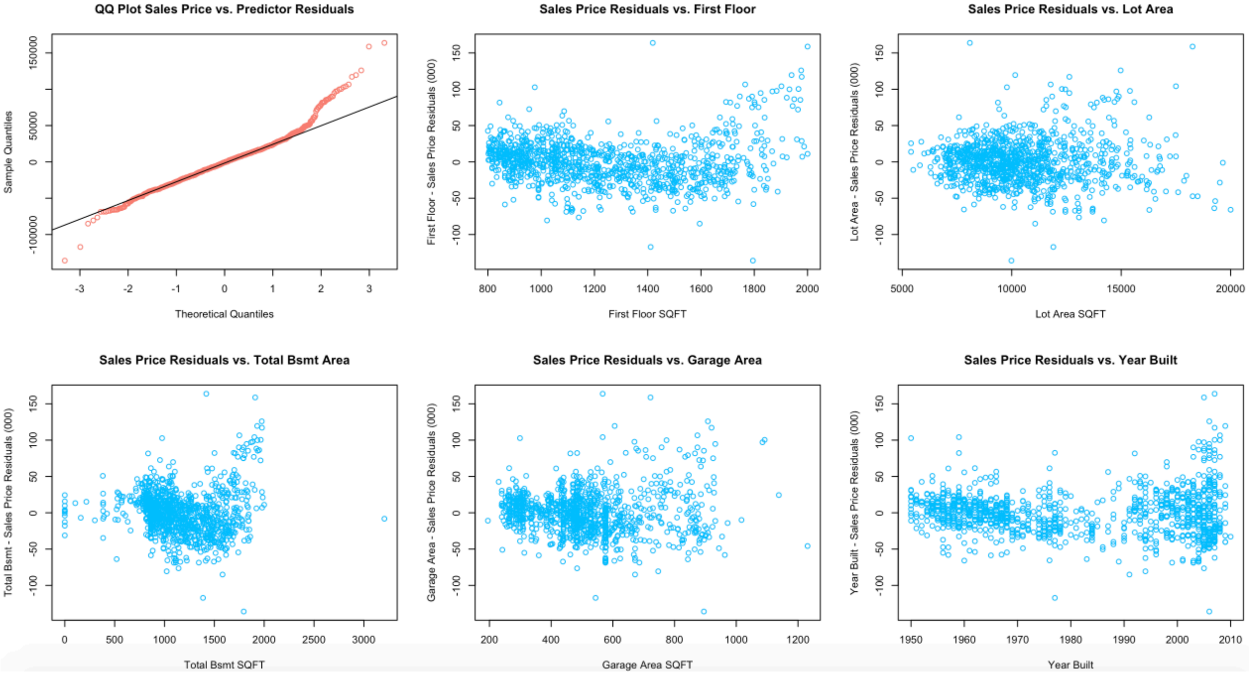


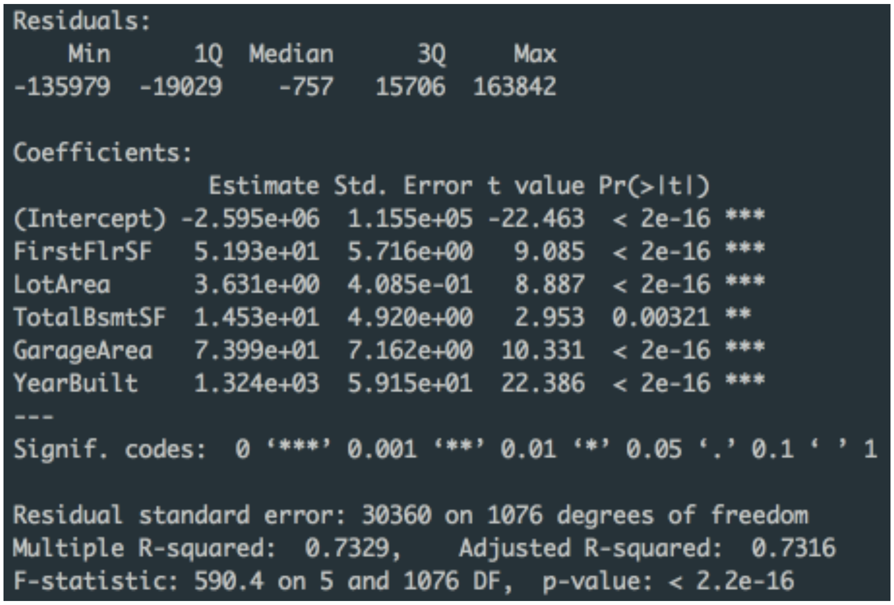
```
Residuals:
    Min      1Q  Median      3Q     Max 
-132073  -18163    -837   16829  140237 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
FirstFlrSF 1.065e+02  3.640e+00  29.256  < 2e-16 ***
LotArea    2.011e+00  4.462e-01   4.507 7.29e-06 ***
hood1      2.891e+03  5.337e+03   0.542    0.588    
hood2      4.488e+04  5.227e+03   8.585  < 2e-16 ***
hood3      8.438e+04  6.109e+03  13.813  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32760 on 1077 degrees of freedom
Multiple R-squared:  0.9719,    Adjusted R-squared:  0.9718 
F-statistic:  7450 on 5 and 1077 DF,  p-value: < 2.2e-16
```

Model Comparison SalePrice and Log(SalePrice):

We will now fit two models to predictor variables. We have chosen 5 variables (4 continuous and 1 discrete) that we believe are the best predictors for home prices in Ames. The continuous variables are (all in SQFT): ‘First Floor Area’, ‘Lot Area’, ‘Total Basement Area’, and ‘Garage Area’. The one discrete variable we will use for this analysis is ‘Year Built’. The comparison will be setting these predictor variables against ‘SalePrice’ and Log(SalePrice). Figure 10 below shows the summary and regression models for ‘SalePrice’. Figure 11 represents the log function. When interpreting the models one must take into account that the Log transformation of the response variable in Figure 11, is essentially normalizing the values for Sales Price. As you can see form our QQ-Plots for both Figures 10 and 11, 10 has more variation than does 11. Linear regressions may not be transparent in untransformed models. Therefore, generally, the log transformation of a response variable can improve the model by normalizing the variation. Which in this case, ‘Sale Price’ has some variation at the tail ends.

Figure 10:

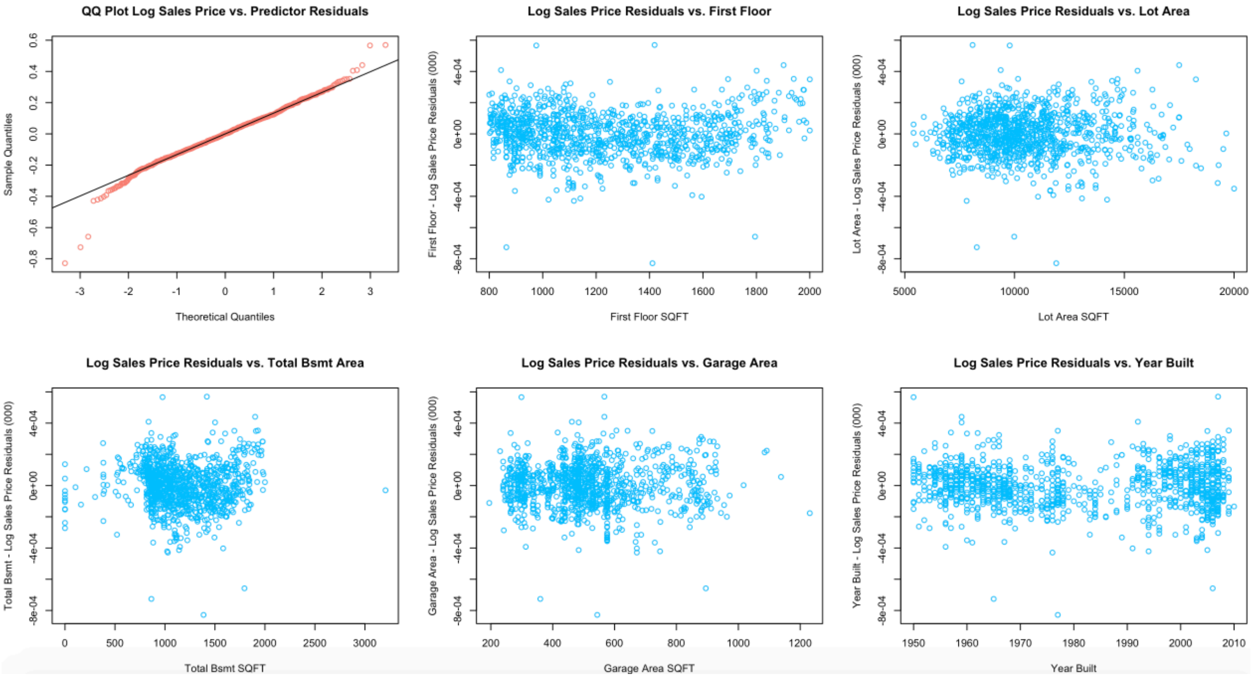


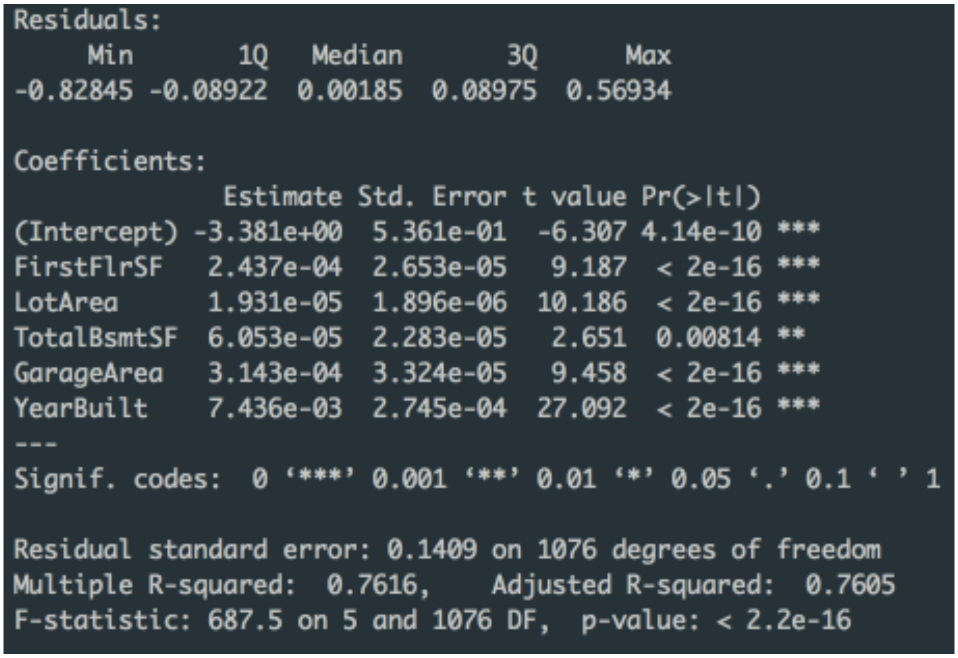


```
Residuals:
    Min      1Q  Median      3Q     Max
-135979  -19029    -757   15706  163842

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.595e+06  1.155e+05 -22.463  < 2e-16 ***
FirstFlrSF   5.193e+01  5.716e+00   9.085  < 2e-16 ***
LotArea      3.631e+00  4.085e-01   8.887  < 2e-16 ***
TotalBsmtSF  1.453e+01  4.920e+00   2.953  0.00321 **
GarageArea   7.399e+01  7.162e+00  10.331  < 2e-16 ***
YearBuilt    1.324e+03  5.915e+01  22.386  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30360 on 1076 degrees of freedom
Multiple R-squared:  0.7329,    Adjusted R-squared:  0.7316
F-statistic: 590.4 on 5 and 1076 DF,  p-value: < 2.2e-16
```

As shown above, the p-values all indicate that we can reject the null hypothesis, which proves good predictability for this model. Our adjusted R^2 value is also fairly high at 0.7316. We also have strong t-values for each variable.

Figure 11:




```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.82845 -0.08922  0.00185  0.08975  0.56934 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.381e+00  5.361e-01  -6.307 4.14e-10 ***
FirstFlrSF   2.437e-04  2.653e-05   9.187  < 2e-16 ***
LotArea      1.931e-05  1.896e-06  10.186  < 2e-16 ***
TotalBsmtSF  6.053e-05  2.283e-05   2.651  0.00814 ** 
GarageArea   3.143e-04  3.324e-05   9.458  < 2e-16 ***
YearBuilt    7.436e-03  2.745e-04  27.092  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1409 on 1076 degrees of freedom
Multiple R-squared:  0.7616,	Adjusted R-squared:  0.7605 
F-statistic: 687.5 on 5 and 1076 DF,  p-value: < 2.2e-16
```

Revealed in Figure 11, the p-values all indicate that we can reject the null hypothesis. The total basement square footage is a little larger than the other values, this is due to some homes not having a basement. However, still attesting good predictability for this model. Our adjusted R^2 value is also fairly high at 0.7605, larger than our model in Figure 10. We also have strong t-values for each variable. With the predictor variables fairly distributed as shown in our plots and based on our assessment, no further transformations are required.

Conclusion:

Established within the analyses above, the best model for predicting ‘Sale Price’ is the model that utilizes indicator variables, breaking out the price per square foot into three groups for Ames. There is a high correlation between the predictor variables and the response ‘Sale Price’. Adjusted R^2 is listed at 0.9718, which indicates a good model for predictability. When compared to our multiple regression analysis of lot area and first floor square footage, this is the best model compiled. The next best model compiled was using the Log(Sale Price) along with 4 continuous variables and one discrete variable. The adjusted R^2 value for this was 0.7605, not as high as our model with the indicator variables. However, if taken into account trying to predict the price of a typical home in Ames without accounting for neighborhood breakdown, it is a moderately accurate model for predictability.

Code Appendix:

```
# Zeeshan Latifi
# 10.7.2017
# ames_waterfall.R

# Read in csv file for Ames housing data;

# Note that back slash is an escape character in R so we use \\ when we want \;
path.name <- '/Users/Zeeshan/Desktop/PREDICT 410/Week 1/';
file.name <- paste(path.name,'ames_housing_data.csv',sep='');

# Read in the csv file into an R data frame;
amesiowa.df <- read.csv(file.name,header=TRUE,stringsAsFactors=FALSE);

# Single ifelse() statement
# ifelse(condition, value if condition is TRUE, value if the condition is FALSE)

# Nested ifelse() statement
# ifelse(condition1, value if condition1 is TRUE,
#           ifelse(condition2, value if condition2 is TRUE,
#           value if neither condition1 nor condition2 is TRUE
#           )
# )


# Create a waterfall of drop conditions;
# Work the data frame as a 'table' like you would in SAS or SQL;
amesiowa.df$dropConditions <- ifelse(amesiowa.df$SubClass!= 020 & amesiowa.df$SubClass != 060 &
amesiowa.df$SubClass != 080,'01: Not SFR',
                    ifelse(amesiowa.df$Zoning!='RH' & amesiowa.df$Zoning!='RL' & amesiowa.df$Zoning!='RM','02:
Non-Residential Zoning',
                    ifelse(amesiowa.df$Street!='Pave','03: Street Not Paved',
                    ifelse(amesiowa.df$Utilities!='AllPub', '04: Not All Utilities Included',
                    ifelse(amesiowa.df$OverallQual<5, '05: Overall Quality Under 5',
                    ifelse(amesiowa.df$OverallCond<5, '06: Overall Condition Under 5',
                    ifelse(amesiowa.df$YearBuilt<1950, '07: Homes Built Pre-1950',
                    ifelse(amesiowa.df$ExterQual!='TA' & amesiowa.df$ExterQual!='Gd'&
amesiowa.df$ExterQual!='Ex', '08: Below Good Exterior Quality',
                    ifelse(amesiowa.df$ExterCond!='TA' & amesiowa.df$ExterCond!='Gd'&
amesiowa.df$ExterCond!='Ex', '09: Below Good Exterior Condition',
                    ifelse(amesiowa.df$FirstFlrSF<800, '10: First Floor Under 800 SqFt',
                    ifelse(amesiowa.df$CentralAir!='Y', '11: No Central Air',
                    ifelse(amesiowa.df$PavedDrive!='Y', '12: No Paved Driveway',
                    ifelse(amesiowa.df$BldgType!='1Fam', '13: Not a Single Family Home',
                    ifelse(amesiowa.df$LotArea<5000 | amesiowa.df$LotArea>20000, '14: Not a Normal Lot Area',
                    ifelse(amesiowa.df$GrLivArea>2000, '15: Abnormal Ground Living Area',
                    ifelse(amesiowa.df$GarageFinish=='NA', '16: No Garage',
                    '99: Eligible Sample')
                    )))))))))))))));
```

```
table(amesiowa.df$dropConditions)

# Save the table
waterfalls <- table(amesiowa.df$dropConditions);

# Format the table as a column matrix for presentation;
as.matrix(waterfalls,15,1)


# Eliminate all observations that are not part of the eligible sample population;
myeligible.population <- subset(amesiowa.df,dropConditions=='99: Eligible Sample');

# Check that all remaining observations are eligible;
table(myeligible.population$dropConditions);

head(myeligible.population)



##########################################################################################
#Assignment 3
#part 2 initial EDA

par(mfrow = c(1,1))
boxplot(myeligible.population$SalePrice/1000, main = 'Sales Price Distribution', col = 'deepskyblue', ylab = 'Sales
Price (000)')
summary(myeligible.population$SalePrice)

par(mfrow = c(1,2))
scatter.smooth(myeligible.population$YearBuilt, myeligible.population$SalePrice/1000, col = 'deepskyblue',
            main = 'Sales Price vs. Year Built', xlab = 'Year Built', ylab = 'Sales Price (000)')

qqplot(myeligible.population$YearBuilt, myeligible.population$SalePrice/1000, col = 'deepskyblue',
      main = 'Sales Price vs. Year Built', xlab = 'Year Built', ylab = 'Sales Price (000)')

######****************************************************************
par(mfrow = c(1,3))
boxplot(myeligible.population$FirstFlrSF, main = 'First Floor SQFT Distribution', col = 'deepskyblue',
       ylab = 'First Floor SQFT')

#par(mfrow = c(1,2))
scatter.smooth(myeligible.population$FirstFlrSF, myeligible.population$SalePrice/1000, main = 'Sales Price vs. First
Floor SQFT',
            col = 'deepskyblue', ylab = 'Sales Price (000)', xlab = 'First Floor SQFT')

qqplot(myeligible.population$FirstFlrSF, myeligible.population$SalePrice/1000, main = 'Sales Price vs. First Floor
SQFT',
      col = 'deepskyblue', ylab = 'Sales Price (000)', xlab = 'First Floor SQFT')

summary(myeligible.population$FirstFlrSF)


model.first <- lm(SalePrice ~ FirstFlrSF, data=myeligible.population)
```

```
# Display model summary
summary(model.first)

# List out components of lm object
names(model.first)

model.first$coefficients

par(mfrow = c(1,2))
qqnorm(model.first$residuals, main = 'QQ Plot First Floor Residuals', col = 'salmon')
qqline(model.first$residuals)

# Make a scatterplot
plot(myeligible.population$FirstFlrSF,model.first$residuals/1000, main = 'Sales Price Residuals vs. First Floor',
    col = 'deepskyblue', xlab = 'First Floor SQFT', ylab = 'First Floor - Sales Price Residuals (000)')

######************************************************************
par(mfrow = c(1,3))
boxplot(myeligible.population$LotArea, main = 'Lot Area Distribution', col = 'deepskyblue',
      ylab = 'Lot Area SQFT')

#par(mfrow = c(1,2))
scatter.smooth(myeligible.population$LotArea, myeligible.population$SalePrice/1000, main = 'Sales Price vs. Lot
Area',
            col = 'deepskyblue', ylab = 'Sales Price (000)', xlab = 'Lot Area SQFT')

qqplot(myeligible.population$LotArea, myeligible.population$SalePrice/1000, main = 'Sales Price vs. Lot Area',
     col = 'deepskyblue', ylab = 'Sales Price (000)', xlab = 'Lot Area SQFT')

summary(myeligible.population$LotArea)

model.lot <- lm(SalePrice ~ LotArea, data=myeligible.population)

# Display model summary
summary(model.lot)

# List out components of lm object
names(model.lot)

model.lot$coefficients

par(mfrow = c(1,2))
qqnorm(model.lot$residuals, main = 'QQ Plot Lot Area Residuals', col = 'salmon')
qqline(model.lot$residuals)

# Make a scatterplot
plot(myeligible.population$LotArea,model.lot$residuals/1000, main = 'Sales Price Residuals vs. Lot Area',
    col = 'deepskyblue', xlab = 'Lot Area SQFT', ylab = 'Lot Area - Sales Price Residuals (000)')
```

```
###################################################################################################
#Assignment 3
#part 3 MLR
model.comb <- lm(SalePrice ~ FirstFlrSF + LotArea, data=myeligible.population)

# Display model summary
summary(model.comb)

# List out components of lm object
names(model.comb)

model.comb$coefficients

par(mfrow = c(1,3))
qqnorm(model.comb$residuals, main = 'QQ Plot First Floor & Lot Area Residuals', col = 'salmon')
qqline(model.comb$residuals)

# Make a scatterplot
plot(myeligible.population$FirstFlrSF,model.comb$residuals/1000, main = 'Sales Price Residuals vs. First Floor',
    col = 'deepskyblue', xlab = 'First Floor SQFT', ylab = 'First Floor - Sales Price Residuals (000)')

plot(myeligible.population$LotArea,model.comb$residuals/1000, main = 'Sales Price Residuals vs. Lot Area',
    col = 'deepskyblue', xlab = 'Lot Area SQFT', ylab = 'Lot Area - Sales Price Residuals (000)')


summary(model.comb)

mae.comb <- mean(abs(model.comb$residuals))
mae.comb




###################################################################################################
#Assignment 3
#part 4 Neighborhood Accuracy
par(mfrow = c(1,1))

boxplot(model.comb$residuals~myeligible.population$Neighborhood, main = 'Residuals Distribution by
Neighborhood',
      col = 'deepskyblue', las = 2)


# change the column names
avgSalePrice <- aggregate(myeligible.population$SalePrice,
by=list(Neighborhood=myeligible.population$Neighborhood), FUN=mean)
colnames(avgSalePrice) <- c('Neighborhood','AvgSalePrice')
avgSalePrice

myeligible.population$totalsqft <-myeligible.population$GrLivArea + myeligible.population$TotalBsmtSF

avgPricePerSQFT <- aggregate(myeligible.population$SalePrice/myeligible.population$totalsqft,
by=list(Neighborhood=myeligible.population$Neighborhood), FUN=mean)
```

```
colnames(avgPricePerSQFT) <- c('Neighborhood','avgPricePerSQFT')
avgPricePerSQFT

myeligible.population$avgPricePerSQFTCalc <- myeligible.population$SalePrice/myeligible.population$totalsqft

#plot(myeligible.population$Neighborhood, myeligible.population$SalePrice/myeligible.population$GrLivArea)
model.hood <- lm(myeligible.population$SalePrice/myeligible.population$totalsqft ~
myeligible.population$Neighborhood)

summary(model.hood)

mae.1 <- aggregate(abs(model.hood$residuals), by=list(Neighborhood=myeligible.population$Neighborhood),
FUN=mean)
mae.1

plot(avgPricePerSQFT$avgPricePerSQFT, mae.1$x, main = 'MAE vs. Price per SQFT', xlab = 'Price per SQFT',
    ylab = 'MAE', col = 'salmon')


table(myeligible.population$avgPricePerSQFTCalc)

myeligible.population$hood1 <- ifelse(myeligible.population$avgPricePerSQFTCalc<70,1,0);
myeligible.population$hood2 <- ifelse(myeligible.population$avgPricePerSQFTCalc>=70 &
myeligible.population$avgPricePerSQFTCalc<=80,1,0);
myeligible.population$hood3 <- ifelse(myeligible.population$avgPricePerSQFTCalc>80, 1,0);

table(myeligible.population$hood1)
table(myeligible.population$hood2)
table(myeligible.population$hood3)

model.IVhood <- lm(SalePrice ~ FirstFlrSF-1  + LotArea + hood1 + hood2 + hood3, data = myeligible.population)

# Display model summary
summary(model.IVhood)

# List out components of lm object
names(model.IVhood)

model.IVhood$coefficients

mae.IVhood <- mean(abs(model.IVhood$residuals))
mae.IVhood



########################################################################################
#Assignment 3
#part 5 Model Comparison
model.best1 <- lm(SalePrice ~ FirstFlrSF + LotArea + TotalBsmtSF + GarageArea + YearBuilt,
data=myeligible.population)

# Display model summary
summary(model.best1)
```

```
# List out components of lm object
names(model.best1)

model.best1$coefficients

par(mfrow = c(2,3))
qqnorm(model.best1$residuals, main = 'QQ Plot Sales Price vs. Predictor Residuals', col = 'salmon')
qqline(model.best1$residuals)

# Make a scatterplot
plot(myeligible.population$FirstFlrSF,model.best1$residuals/1000, main = 'Sales Price Residuals vs. First Floor',
    col = 'deepskyblue', xlab = 'First Floor SQFT', ylab = 'First Floor - Sales Price Residuals (000)')

plot(myeligible.population$LotArea,model.best1$residuals/1000, main = 'Sales Price Residuals vs. Lot Area',
    col = 'deepskyblue', xlab = 'Lot Area SQFT', ylab = 'Lot Area - Sales Price Residuals (000)')

plot(myeligible.population$TotalBsmtSF,model.best1$residuals/1000, main = 'Sales Price Residuals vs. Total Bsmt
Area',
    col = 'deepskyblue', xlab = 'Total Bsmt SQFT', ylab = 'Total Bsmt - Sales Price Residuals (000)')

plot(myeligible.population$GarageArea,model.best1$residuals/1000, main = 'Sales Price Residuals vs. Garage
Area',
    col = 'deepskyblue', xlab = 'Garage Area SQFT', ylab = 'Garage Area - Sales Price Residuals (000)')

plot(myeligible.population$YearBuilt,model.best1$residuals/1000, main = 'Sales Price Residuals vs. Year Built',
    col = 'deepskyblue', xlab = 'Year Built', ylab = 'Year Built - Sales Price Residuals (000)')

mae.best1 <- mean(abs(model.best1$residuals))
mae.best1

#********************************************************************************************
****************************
model.best2 <- lm(log(SalePrice) ~ FirstFlrSF + LotArea + TotalBsmtSF + GarageArea + YearBuilt,
data=myeligible.population)

# Display model summary
summary(model.best2)

# List out components of lm object
names(model.best2)

model.best2$coefficients

par(mfrow = c(2,3))
qqnorm(model.best2$residuals, main = 'QQ Plot Log Sales Price vs. Predictor Residuals', col = 'salmon')
qqline(model.best2$residuals)

# Make a scatterplot
plot(myeligible.population$FirstFlrSF,model.best2$residuals/1000, main = 'Log Sales Price Residuals vs. First Floor',
    col = 'deepskyblue', xlab = 'First Floor SQFT', ylab = 'First Floor - Log Sales Price Residuals (000)')

plot(myeligible.population$LotArea,model.best2$residuals/1000, main = 'Log Sales Price Residuals vs. Lot Area',
    col = 'deepskyblue', xlab = 'Lot Area SQFT', ylab = 'Lot Area - Log Sales Price Residuals (000)')

plot(myeligible.population$TotalBsmtSF,model.best2$residuals/1000, main = 'Log Sales Price Residuals vs. Total Bsmt Area',
    col = 'deepskyblue', xlab = 'Total Bsmt SQFT', ylab = 'Total Bsmt - Log Sales Price Residuals (000)')

plot(myeligible.population$GarageArea,model.best2$residuals/1000, main = 'Log Sales Price Residuals vs. Garage Area',
    col = 'deepskyblue', xlab = 'Garage Area SQFT', ylab = 'Garage Area - Log Sales Price Residuals (000)')

plot(myeligible.population$YearBuilt,model.best2$residuals/1000, main = 'Log Sales Price Residuals vs. Year Built',
    col = 'deepskyblue', xlab = 'Year Built', ylab = 'Year Built - Log Sales Price Residuals (000)')


mae.best2 <- mean(abs(model.best2$residuals))
mae.best2



###########################################################################################
#Assignment 3
#Additional Supplemental Material

par(mfrow = c(2,2))

# Make a scatterplot
scatter.smooth(myeligible.population$FirstFlrSF,myeligible.population$SalePrice/1000, main = 'Sales Price vs. First Floor',
    col = 'salmon', xlab = 'First Floor SQFT', ylab = 'Sales Price (000)')

scatter.smooth(myeligible.population$LotArea,myeligible.population$SalePrice/1000, main = 'Sales Price vs. Lot Area',
    col = 'salmon', xlab = 'Lot Area SQFT', ylab = 'Sales Price (000)')

scatter.smooth(myeligible.population$TotalBsmtSF,myeligible.population$SalePrice/1000, main = 'Sales Price vs. Total Bsmt Area',
    col = 'salmon', xlab = 'Total Bsmt SQFT', ylab = 'Sales Price (000)')

scatter.smooth(myeligible.population$GarageArea,myeligible.population$SalePrice/1000, main = 'Sales Price vs. Garage Area',
    col = 'salmon', xlab = 'Garage Area SQFT', ylab = 'Sales Price (000)')
```