2017

# Assignment 1

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'll 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, conduct a data quality check, 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 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 600 $ft.^2$ or higher for the first level. With that in mind we also eliminated homes without a paved driveway or central air. With these transformations, the observations were reduced down to 1,519. 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 600 SqFt         1
11: No Central Air                     4
12: No Paved Driveway                 16
99: Eligible Sample                 1519
```

Data Quality Check:

We begin evaluating 20 variables that have the potential to be used in predicting sales prices. We dwindled down our variables to the following:

| | |
|---|---|
| SubClass | Identifies the type of dwelling involved in the sale. |
| Zoning | Identifies the general zoning classification of the sale. |
| LotArea | Lot size in square feet |
| Street | Type of road access to property |
| Utilities | Type of utilities available |
| BldgType | Type of dwelling |
| HouseStyle | Style of dwelling |
| OverallQual | Rates the overall material and finish of the house |
| OverallCond | Rates the overall condition of the house |
| YearBuilt | Original construction date |
| YearRemodel | Remodel date (same as construction date if no remodeling or additions) |

| | |
|---|---|
| ExterQual | Evaluates the quality of the material on the exterior |
| ExterCond | Evaluates the present condition of the material on the exterior |
| BsmtFinType1 | Rating of basement finished area |
| FirstFlrSF | First Floor square feet |
| GarageCars | Size of garage in car capacity |
| PavedDrive | Paved driveway |
| PoolArea | Pool area in square feet |
| CentralAir | Central air conditioning |

Continuous Variables:

When examining the continuous variables, we first looked at 'LotArea'. The lot area was pretty wide spread with a few outliers. One homes listed lot area was 215,245 sq. ft. This could be an error or this could be an anomaly.

We then examined 'FirstFlrSF'. Based on our assumption of a typical home, we decided to remove homes with the first floor being less than 600 ft.[2] With this we found our distribution to be fairly even with a few cases of very high square footage. No anomalies were found in this data.

Discrete Variables:

The majority of our sample size are part of SubClass 020, homes that are 1 level built after 1946. The distribution is a bit skewed, however, we've inferred that the majority of homes in Ames are single level. Zoning is heavily skewed to 'RL'. This means the majority of Ames is low density. This makes sense in a rural area. The main house style is a single family home, few townhomes and even fewer townhomes that are end units. Shown in Figure 2 below:

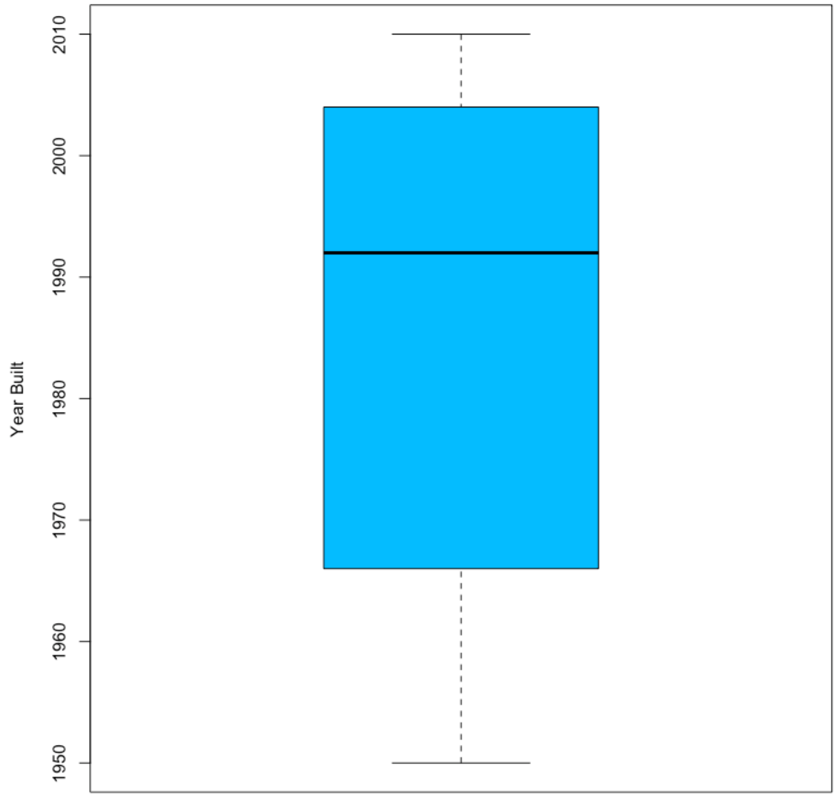
Figure 2:

| | Var1 | Freq |
|---|---|---|
| 1 | 20 | 889 |
| 2 | 60 | 520 |
| 3 | 80 | 110 |

| | Var1 | Freq |
|---|---|---|
| 1 | RH | 2 |
| 2 | RL | 1509 |
| 3 | RM | 8 |

Year built was fairly even distributed with no outliers identified as shown in Figure 3:

Figure 3:




Overall quality of the homes is evenly distributed statistically insignificant for any concerns. Figure 4 illustrates the spread:

Figure 4:

| | Var1 | Freq |
|---|---|---|
| 1 | 5 | 401 |
| 2 | 6 | 383 |
| 3 | 7 | 394 |
| 4 | 8 | 232 |
| 5 | 9 | 82 |
| 6 | 10 | 27 |

Exploratory Analysis:

To help predict the sales price for a home we should understand the response variable first. Figure 5 represents a summary for the 'SalesPrice' data in our sample population.

Figure 5:

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 62380 | 150600 | 185000 | 209600 | 242800 | 755000 |

As shown above, the mean pricing for a home in Ames according to the selected criteria is $209,600. We found no values that seemed out of the ordinary here. There is potential that the max value and min value can be outliers, but it's unable to be determined at this time.

Continuous Variables Analysis:

Beginning with the lot area, we decided to test the distribution for the data and also check for any correlation between sales price.

Figure 6:

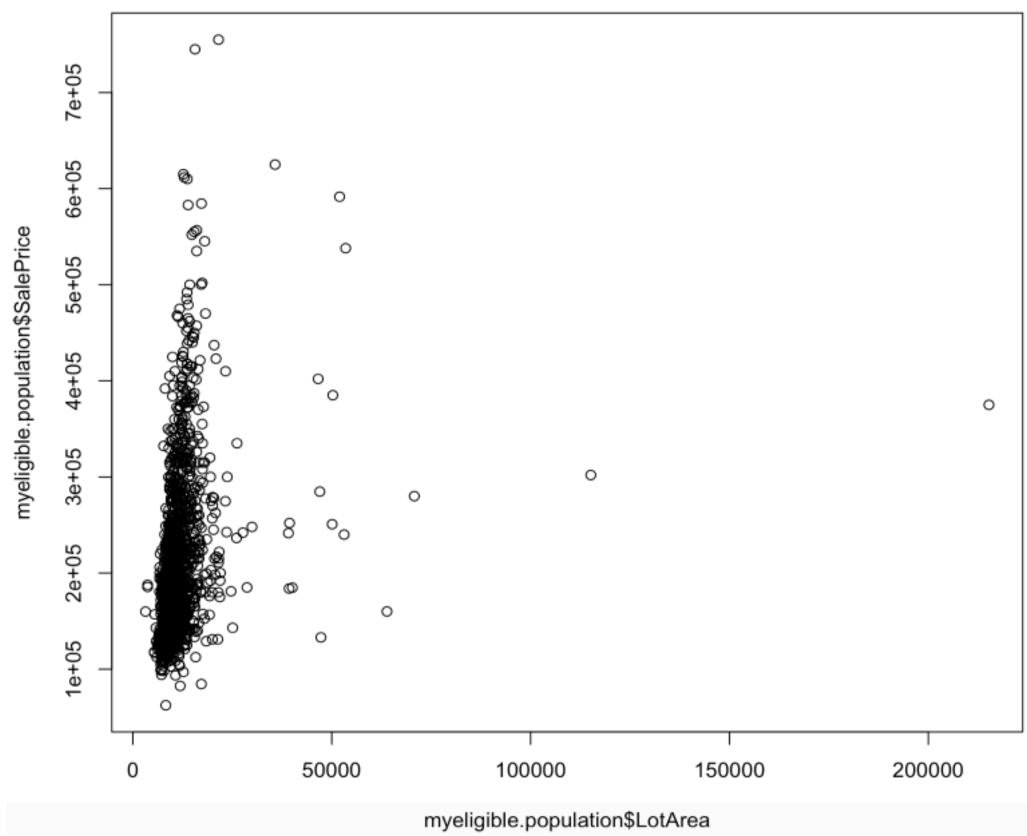
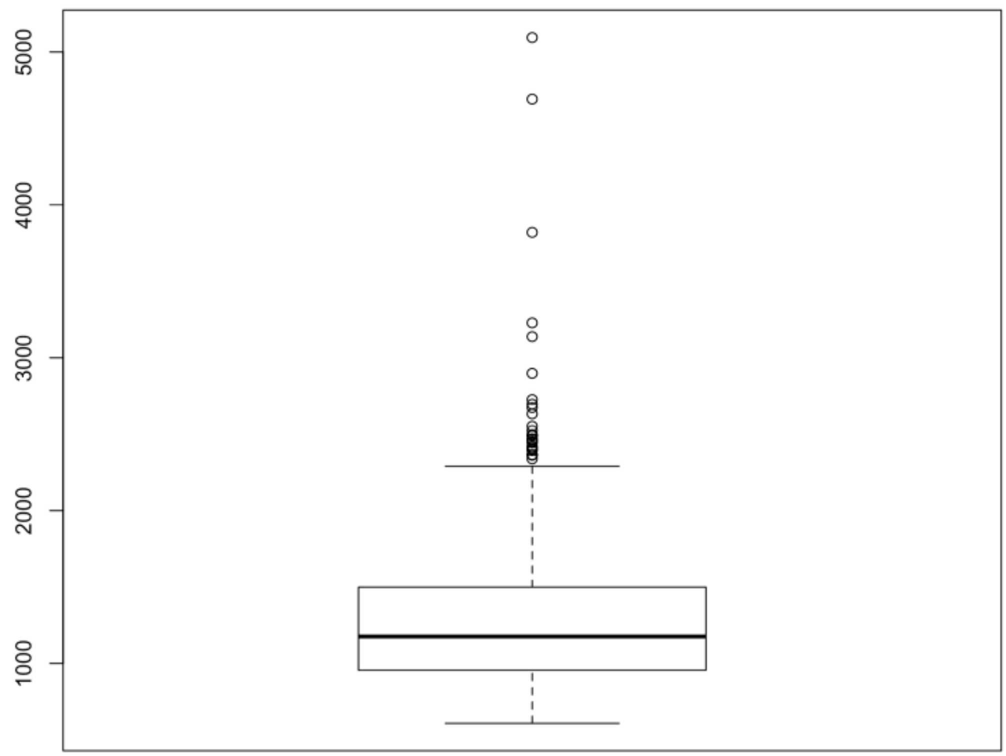


Figure 6 shows us that there is some correlation with the lot area and sales price. Some outliers are present. As shown with the home listed with a lot area of more than 200,000.

First floor square footage can be a reasonable predictor for sales price. When looking at the data, we found the following:

Figure 7:

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 608.0 | 955.5 | 1176.0 | 1263.0 | 1499.0 | 5095.0 |

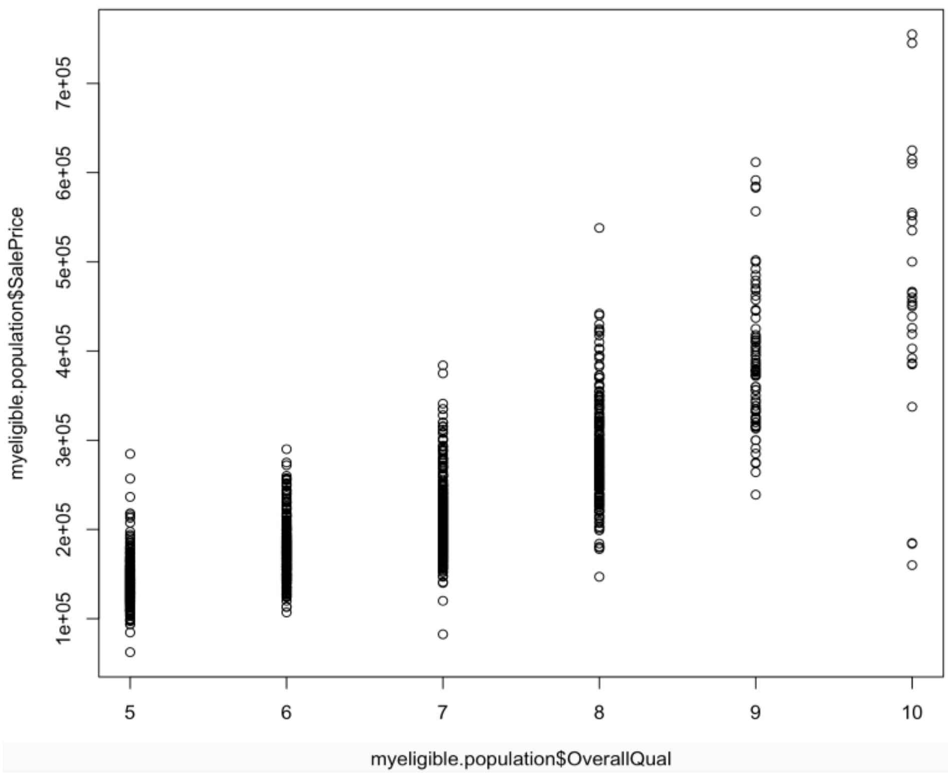



As shown in the figure above, the square footage for the first floor in Ames is about 1263 ft.$^{2}$. Further evaluation of the max value property is needed for it to be determined as an outlier.

Discrete Variables Analysis:

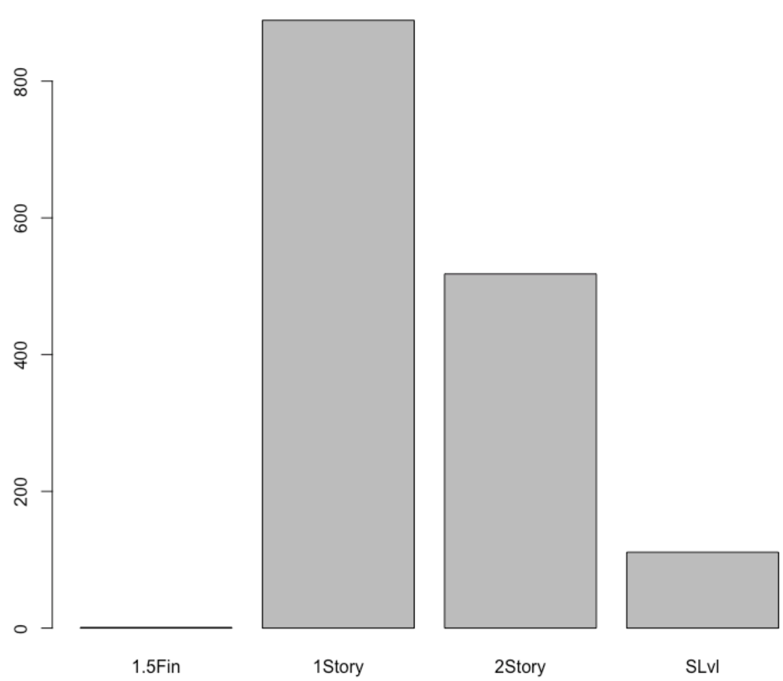
The figure below shows that population sales price in response to the overall quality of the home. Per our definition, homes ranked 5 or higher were selected. Based on this, we can reasonably infer that the quality of the home can tends to correlate well with the price of the home.

Figure 8:




The breakdown of our sample size in terms of housing style are shown below. As you can see the majority of the homes in Ames are one story single family homes.

Figure 9:



Of our sample size, when taking out several factors, only one townhome remained (end-unit) that met the criteria in our sample size. However, since there cannot be only one townhome, we've decided to keep it as part of the analysis. See breakdown below:

Figure 10:

| 1Fam | TwnhsE |
|---|---|
| 1518 | 1 |

Next we took a look at the 'Yearly Remodel' date. According to the data in our sample size, we can somewhat see a pattern of the price of the home increasing as the remodel date is relatively newer.

Figure 11:

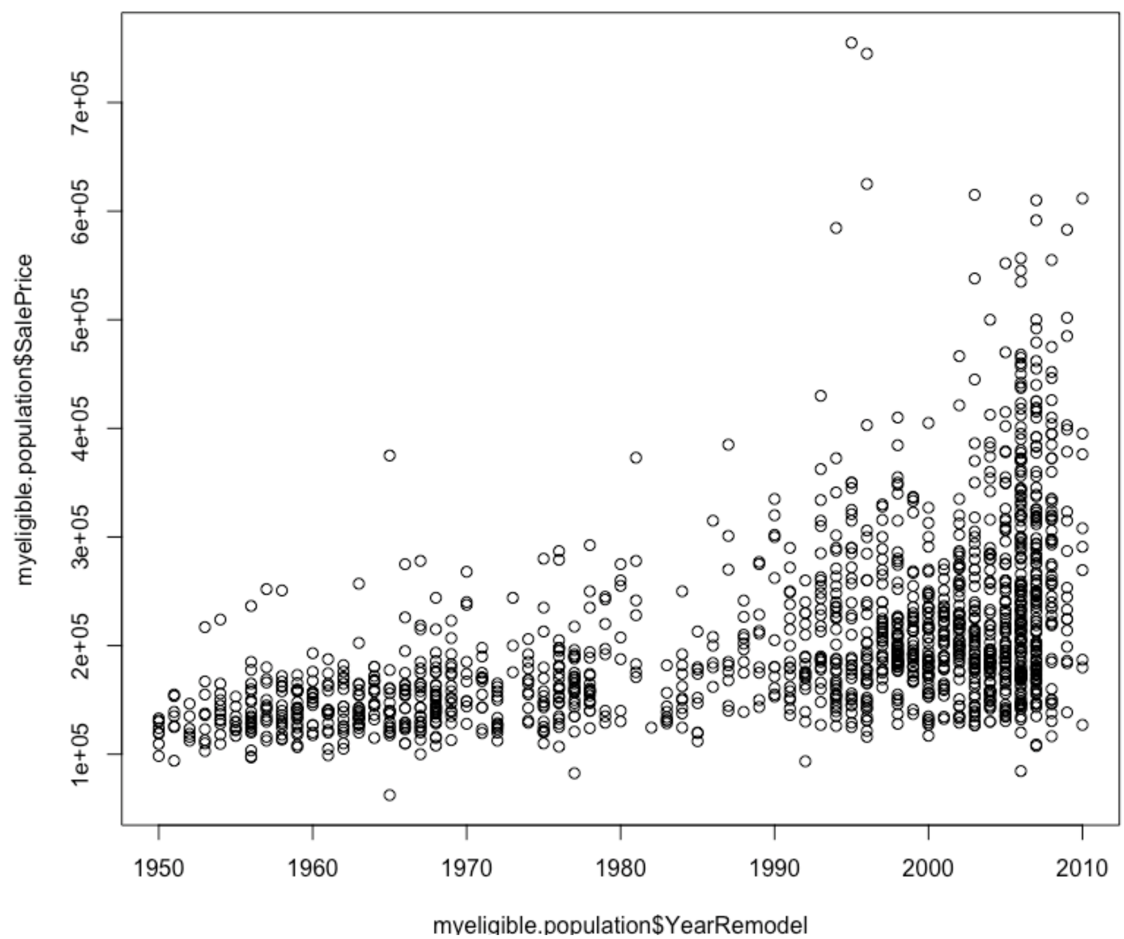
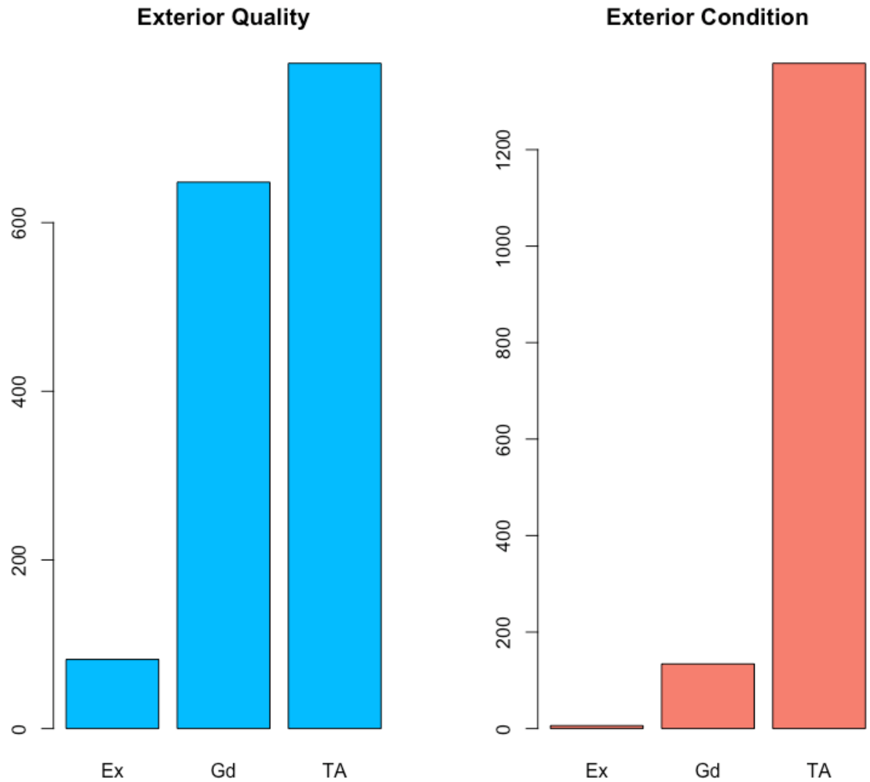


Figure 12:



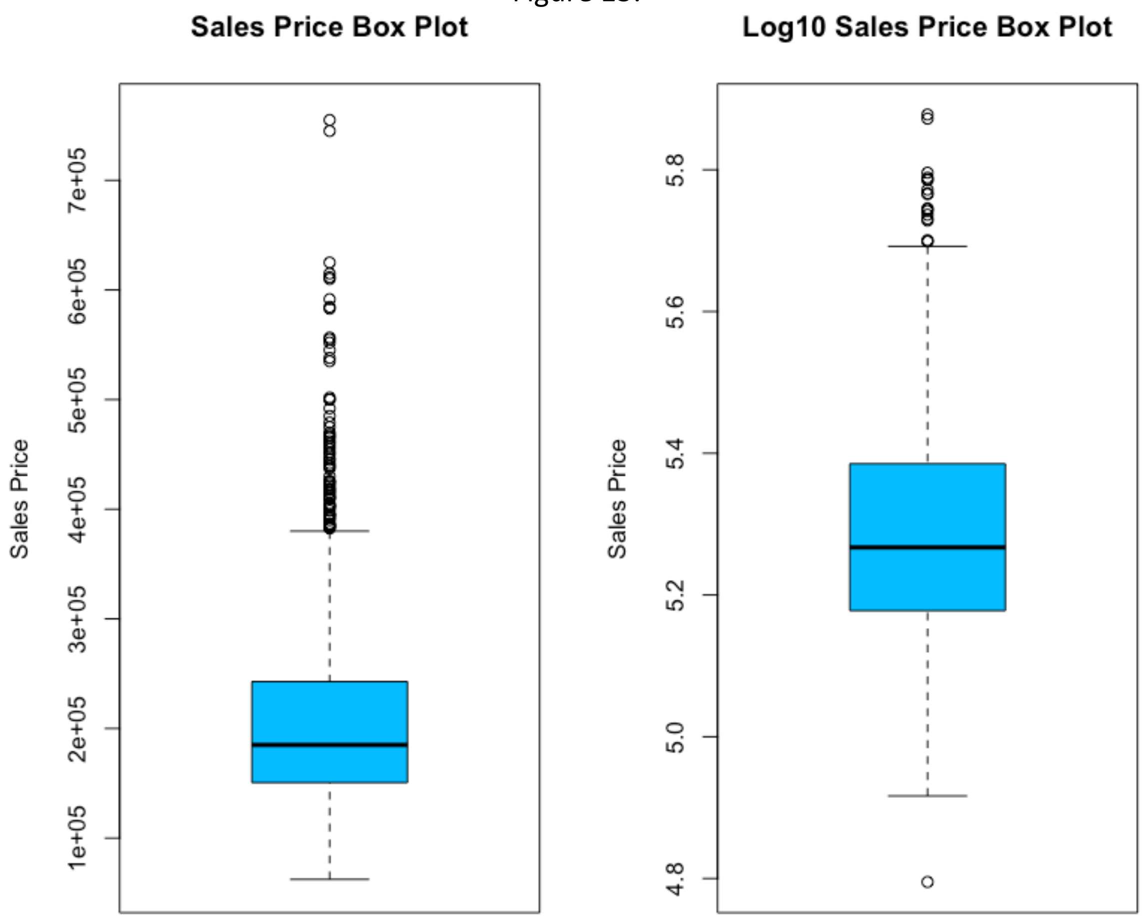

Above is the breakdown of the exterior quality and condition. Most of the homes in Ames are 'TA' or average. Especially when it comes to the exterior condition.

Regression Analysis:

After the exploratory analysis, we've found that three variables will be good predictors for our response variable, sales price. The three variables chosen were the lot area, first floor square footage, and the year the home was built.

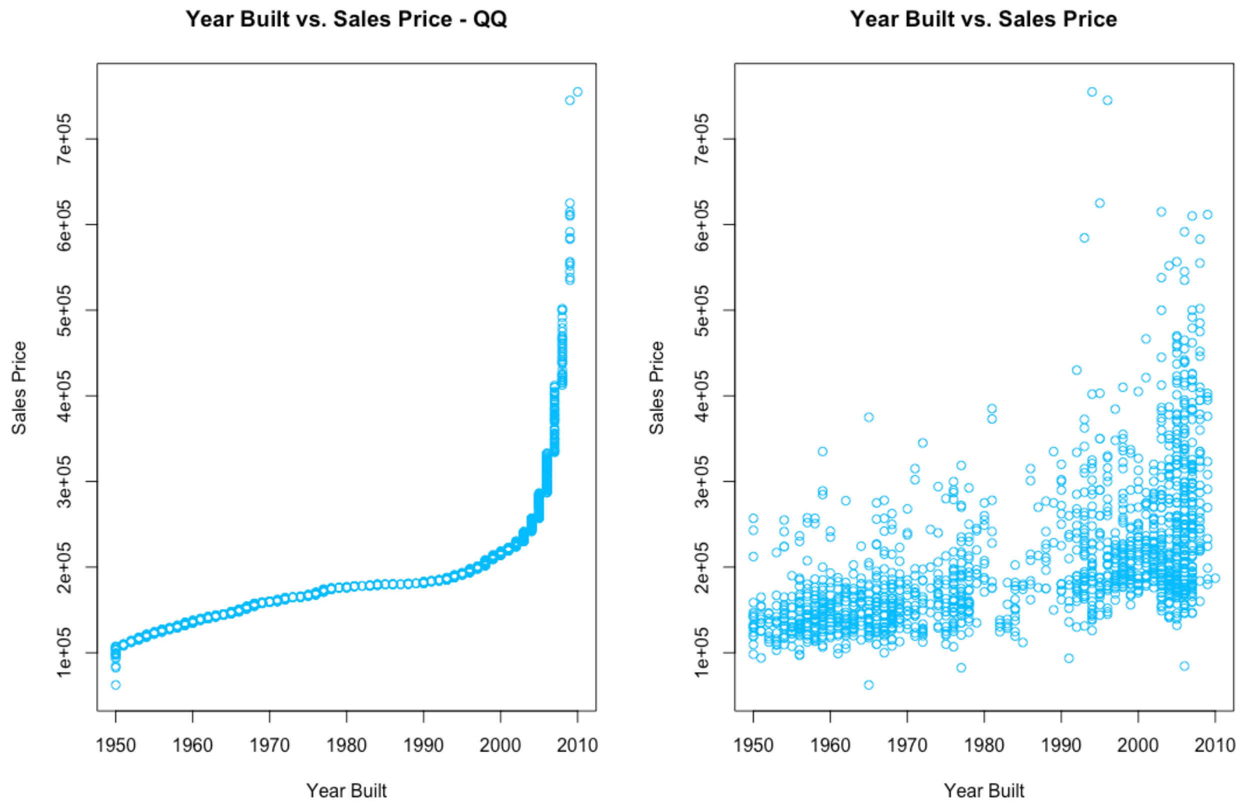
Before this we take a look at the sales price distribution as well as the log10 distribution for the sales price. After taking the log10, we see that we have a reasonable sample to conduct our analysis. There is some spread among the home sale prices, however, we believe this to be sufficient for our assessment.

Figure 13:



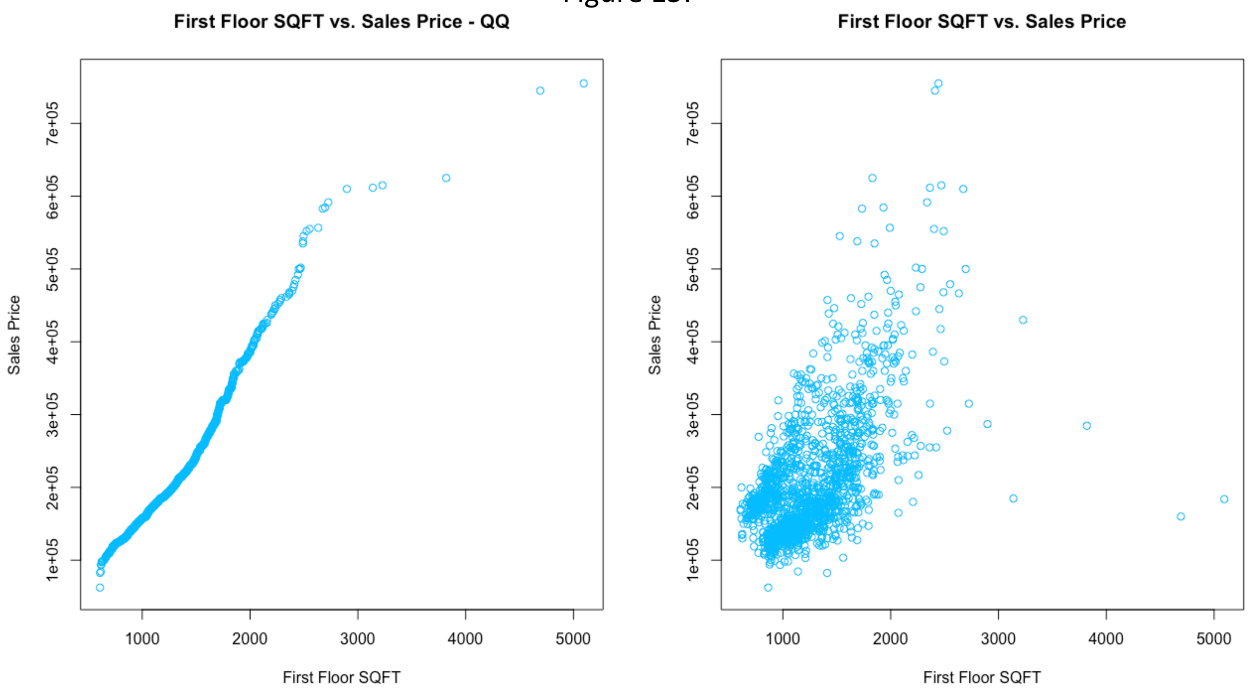
In the figure below we see the quantile-quantile plot as well as the scatter plot for the year the home was built versus the sales price. As shown, there is a certain point, as expected, where the cost drastically increases. This is due to the home being very new to market. Based on the visuals here we can reasonably infer a decent correlation between year built and the sales price. When calculating this in R, we received a value of 0.55 for the correlation coefficient.

Figure 14:



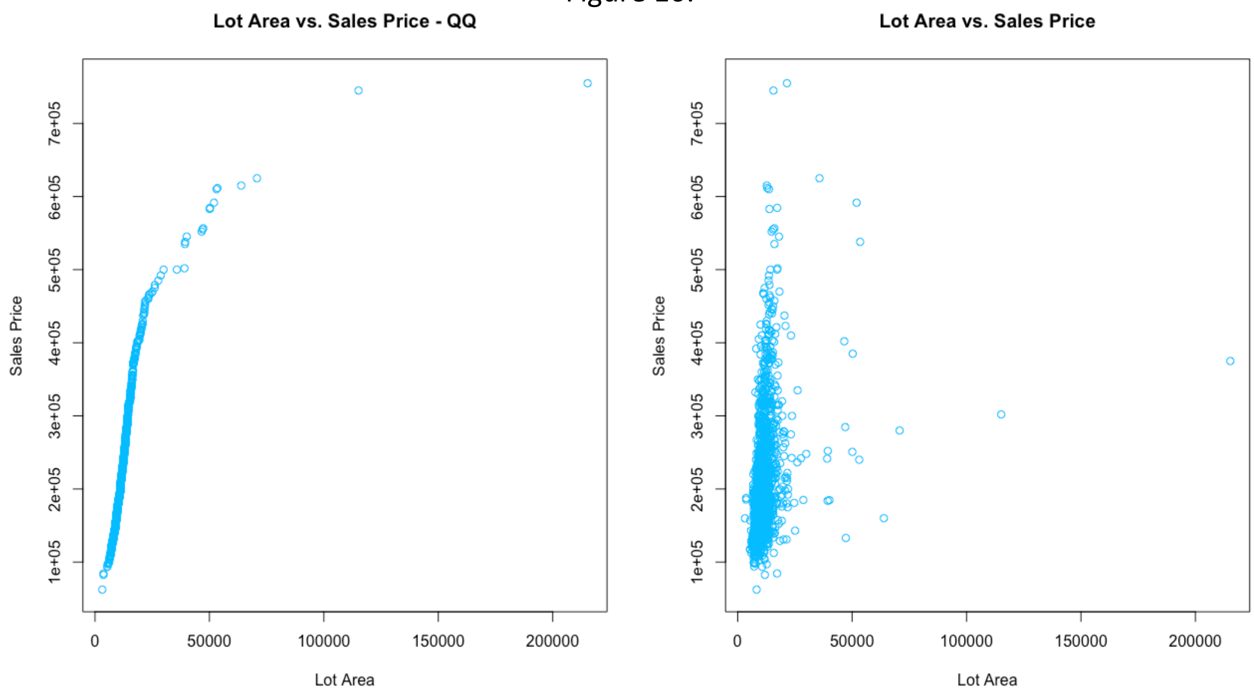
Next, we look at the relationship between the square footage on the first floor and the sales price for each home. We show both the QQ plot as well as the scatter plot for this relationship. When calculating the correlation coefficient, we received a value of 0.59. With this value, we can say there is relative correlation.

Figure 15:



Lot area was evaluated next. Based on our plots here in Figure 16, there is little variation of the lot area among homes in Ames, Iowa. The correlation coefficient here was 0.27, very little correlation.

Figure 16:




Conclusion:

Based on our current analysis, the results are inconclusive. There is somewhat of a correlation between the year built and the first floor square footage. Higher with the latter. However, based on our sample size, some optimization is required to better make an informed prediction model.

There are some corrections to be made with the sample size. After evaluation, we should remove the single townhome, remove the home with the 200,000 ft.2 lot size, as well as homes that cost below $100,000. This is to eliminate values that are less than what we would consider the typical home.

From our assessment above, we can infer that the first floor square footage and the year built are good predictors for the response variable, sales price. However, with some optimization and further assessment, we may have a more confirmed predictor variable.

Appendix:

R Script for Analysis

```r
# Zeeshan Latifi
# 9.23.2017
# 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$FirstFlrSF<600, '10: First Floor Under 600 SqFt',
  ifelse(amesiowa.df$CentralAir!='Y', '11: No Central Air',
  ifelse(amesiowa.df$PavedDrive!='Y', '12: No Paved Driveway',
  '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,11,1)


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

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

head(myeligible.population)

##############################################################################################
#Final Table

final.pop <- data.frame(myeligible.population$SubClass, myeligible.population$Zoning, myeligible.population$LotArea,
                        myeligible.population$Street, myeligible.population$Utilities, myeligible.population$BldgType,
                        myeligible.population$HouseStyle, myeligible.population$OverallQual, myeligible.population$OverallCond,
                        myeligible.population$YearBuilt, myeligible.population$YearRemodel,myeligible.population$ExterQual,
                        myeligible.population$ExterCond, myeligible.population$BsmtFinType1, myeligible.population$FirstFlrSF,
                        myeligible.population$GarageCars, myeligible.population$PavedDrive, myeligible.population$PoolArea,
                        myeligible.population$CentralAir, myeligible.population$SalePrice)

head(final.pop)

##############################################################################################
#Data Quality Check
as.data.frame(table(myeligible.population$SubClass))

as.data.frame(table(myeligible.population$Zoning))

summary(myeligible.population$LotArea)
myeligible.population[is.element(myeligible.population$LotArea, max(myeligible.population$LotArea)),]

as.data.frame(table(myeligible.population$Street))

as.data.frame(table(myeligible.population$Utilities))

as.data.frame(table(myeligible.population$BldgType))

as.data.frame(table(myeligible.population$HouseStyle))

as.data.frame(table(myeligible.population$OverallQual))

as.data.frame(table(myeligible.population$OverallCond))

as.data.frame(table(myeligible.population$YearBuilt))
summary(myeligible.population$YearBuilt)

as.data.frame(table(myeligible.population$YearRemodel))
summary(myeligible.population$YearRemodel)

as.data.frame(table(myeligible.population$ExterQual))

as.data.frame(table(myeligible.population$ExterCond))

as.data.frame(table(myeligible.population$BsmtFinType1))

as.data.frame(table(myeligible.population$OverallQual))

summary(myeligible.population$FirstFlrSF)
sd(myeligible.population$FirstFlrSF)

as.data.frame(table(myeligible.population$GarageCars))

as.data.frame(table(myeligible.population$PavedDrive))

summary(myeligible.population$PoolArea)
as.data.frame(table(myeligible.population$PoolArea))

as.data.frame(table(myeligible.population$CentralAir))

```

```r
summary(final.pop$myeligible.population.SalePrice)
sd(final.pop$myeligible.population.SalePrice)


##########################################################################################
#Exploratory Data Analysis
par(mfrow = c(1,1))

boxplot(myeligible.population$SalePrice)
qqplot(myeligible.population$YearBuilt, myeligible.population$SalePrice)
plot(myeligible.population$YearBuilt, myeligible.population$SalePrice)

boxplot(myeligible.population$FirstFlrSF)
qqplot(myeligible.population$FirstFlrSF, myeligible.population$SalePrice)
plot(myeligible.population$FirstFlrSF, myeligible.population$SalePrice)
summary(myeligible.population$FirstFlrSF)

plot(myeligible.population$OverallQual, myeligible.population$SalePrice)

plot(myeligible.population$OverallCond, myeligible.population$SalePrice)

plot(myeligible.population$LotArea, myeligible.population$SalePrice)
boxplot(myeligible.population$LotArea)


style_table <- table(myeligible.population$HouseStyle)
barplot(style_table)

bldg_table <- table(myeligible.population$BldgType)
barplot(bldg_table)

par(mfrow = c(1,2))
extqual_table <- table(myeligible.population$ExterQual)
barplot(extqual_table, col = 'deepskyblue', main = 'Exterior Quality')

extcond_table <- table(myeligible.population$ExterCond)
barplot(extcond_table,col = 'salmon', main = 'Exterior Condition')

par(mfrow = c(1,1))
plot(myeligible.population$YearRemodel, myeligible.population$SalePrice)
boxplot(myeligible.population$YearRemodel)


##########################################################################################
#Regression Analysis on 3 variables
par(mfrow = c(1,2))
boxplot(myeligible.population$SalePrice, ylim = c(60000,760000), col = 'deepskyblue', main = 'Sales Price Box Plot', ylab = 'Sales Price')
boxplot(log10(myeligible.population$SalePrice), col = 'deepskyblue', main = 'Log10 Sales Price Box Plot', ylab = 'Sales Price')




qqplot(myeligible.population$YearBuilt, myeligible.population$SalePrice, ylim = c(60000,760000), col = 'deepskyblue',
       main = 'Year Built vs. Sales Price - QQ', ylab = 'Sales Price', xlab = 'Year Built')
plot(myeligible.population$YearBuilt, myeligible.population$SalePrice, ylim = c(60000,760000),
     col = 'deepskyblue', main = 'Year Built vs. Sales Price', ylab = 'Sales Price', xlab = 'Year Built')
cor(myeligible.population$YearBuilt, myeligible.population$SalePrice)




qqplot(myeligible.population$FirstFlrSF, myeligible.population$SalePrice, ylim = c(60000,760000), col = 'deepskyblue',
       main = 'First Floor SQFT vs. Sales Price - QQ', ylab = 'Sales Price', xlab = 'First Floor SQFT')
plot(myeligible.population$FirstFlrSF, myeligible.population$SalePrice, ylim = c(60000,760000), col = 'deepskyblue',
     main = 'First Floor SQFT vs. Sales Price', ylab = 'Sales Price', xlab = 'First Floor SQFT')
cor(myeligible.population$FirstFlrSF, myeligible.population$SalePrice)




qqplot(myeligible.population$LotArea, myeligible.population$SalePrice, ylim = c(60000,760000), col = 'deepskyblue',
       main = 'Lot Area vs. Sales Price - QQ', ylab = 'Sales Price', xlab = 'Lot Area')
plot(myeligible.population$LotArea, myeligible.population$SalePrice, ylim = c(60000,760000), col = 'deepskyblue',
     main = 'Lot Area vs. Sales Price', ylab = 'Sales Price', xlab = 'Lot Area')
cor(myeligible.population$LotArea, myeligible.population$SalePrice)

boxplot(myeligible.population$YearBuilt, col = 'deepskyblue', ylab = 'Year Built')

```