Economics 430: Homework-2

Anshika Sharma (UID: 305488635)

Contents

Question 1

The dataset train.csv contains 79 explanatory variables. The data description and csv filecan be downloaded directly from kaggle. Your task, as suggested on the kaggle website, is to build a model to predict final home prices. Note, this is part of a kaggle competition which you might consider participating in later on. Before you start the parts below, identify any 10 variables of your choice and write a brief paragraph of why you selected them. These are the predictors you will use for solving the problem.

The variables I chose include the following: Numerical variables: LotArea, GarageArea, TotalBsmtSF (Total Basement in Squarefoot), Lot Frontage, YearBuilt, YearRemodAdd (Year in which Remodelling of the house was done), GrLivArea (Gross Living Area), Overall Quality

Categorical Variables: Building Type (Blgd Type), House Style

After reading a few articles, I concluded that these are some of the important variables that influence the sale price of a house. I feel that the year in which the house was built and the year of remodeled have a direct influence on the price. Latest construction and remodeling should command a higher sale price.Higher Gross Living Area should have a positive impact on Saleprice. GarageArea and TotalBsmtSF too will impact the sale price to an extent.

(a) Provide a descriptive analysis of your variables. This should include histograms and fitted distributions, quantile plots, correlation plot, boxplots, scatterplots, and statistical summaries (e.g., the five-number summary). All figures must include comments.

Since there are NA values, I first omit the NA. The procedure for the same is as below:

```r
#reading the data into R
library(readr)
train.kaggle <- read.csv("C:\\Users\\anshi\\Desktop\\UCLA MAE\\ECON 430\\R\\Homework 2-20201020\\train.
```

Now I create my own data frame keeping variables that I selected above.

```r
#Creating data frame, Train
Train = train.kaggle[c(4,5,16,17,18,20,21,39,47,63,81)]

#Converting to factor variable
```

```r
Train[,"BldgType"] <- as.factor(Train[,"BldgType"])
Train[,"HouseStyle"] <- as.factor(Train[,"HouseStyle"])

#Since there are NA values, I first omit the NA.
#The procedure for the same is as below:
Train = na.omit(Train)
#Re-check that there are no NAs
anyNA(Train)
```

```
## [1] FALSE
```

```r
#This output shows us that there are no NAs in our data frame, Train.
```

1. *Description of LotArea*

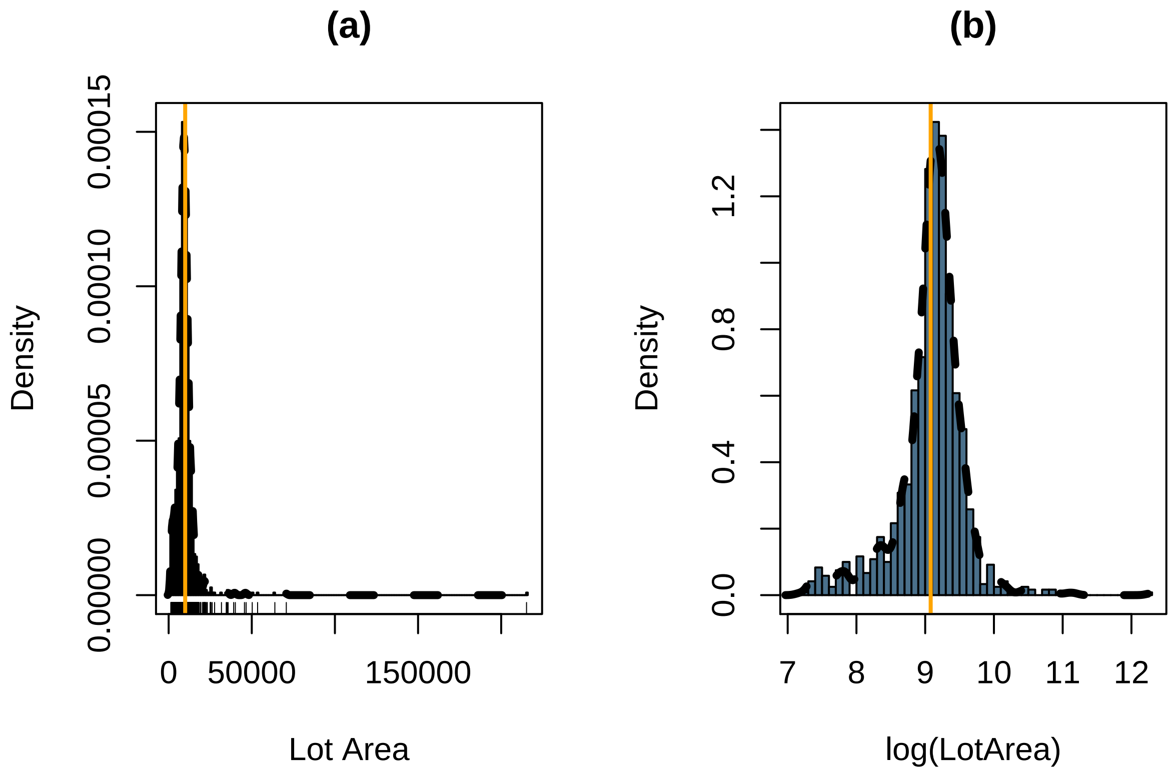
```r
library(pastecs)
stat.desc(Train$LotArea)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range 
## 1.201000e+03 0.000000e+00 0.000000e+00 1.300000e+03 2.152450e+05 2.139450e+05 
##          sum       median         mean      SE.mean CI.mean.0.95          var 
## 1.195199e+07 9.262000e+03 9.951699e+03 2.286611e+02 4.486201e+02 6.279539e+07 
##      std.dev     coef.var 
## 7.924354e+03 7.962816e-01
```

The statistical summaries are as follows: the mean is 9951.69 while the median is 9262.00, showing it is a right skewed distribution. The maximum value is 215245 and the minimum value is 1300, making it a large range.

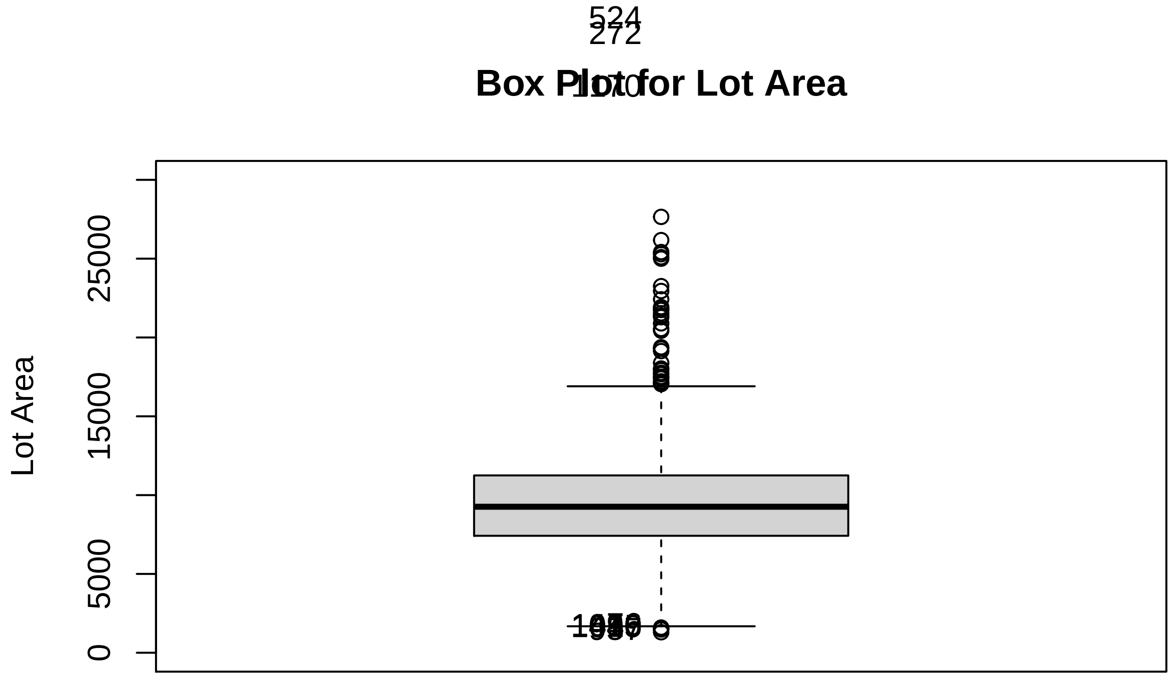
```r
par(mfrow = c(1,2))

hist(Train$LotArea, breaks = "FD", col="skyblue4", main = "(a)",
     freq = FALSE, xlab = "Lot Area", ylab = "Density")
lines(density(Train$LotArea, na.rm = TRUE),lwd=4,col='black',lty=2)
rug(Train$LotArea)
abline(v = mean(Train$LotArea), col = "orange", lwd =2, lty = 1 )
box()
hist(log(Train$LotArea), breaks = "FD", col="skyblue4",
     main = "(b)", freq = FALSE, xlab = "log(LotArea)",
     ylab = "Density" )
lines(density(log(Train$LotArea), na.rm = TRUE),lwd=4,col='black',lty=2)
rug(Train$LotArea)
abline(v = mean(log(Train$LotArea)), col = "orange", lwd =2, lty = 1)
box()
```



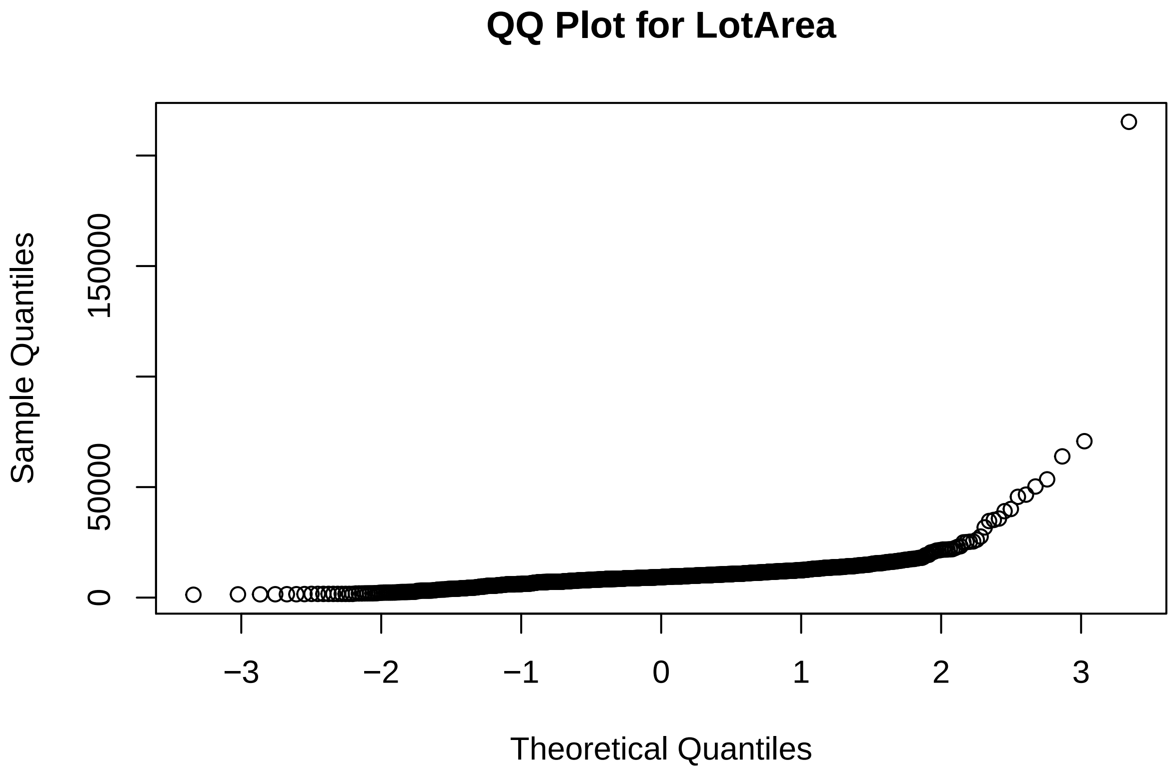
I plotted two histograms. Graph (a) plots LotArea on the X axis and Graph (b) plots log(Lotarea) on the x axis. This is based on my intuition that we will require a log transformation. For the Histogram in part (a), it seems to be positively skewed which means that there a few values that are very large and many small values. The same idea is suggested by the rugplot also. So, in such a case, taking a log transformation spreads out the small values and compresses the large ones. Hence we get a more symmetric distribution and a better visual representation as can be seen in the graph (b). After transforming the distribution, the graph suggests a symmetric distribution, approximating to the Normal distribution as suggested by the fitted density curve.

```r
Boxplot(~LotArea, data=Train, ylab = "Lot Area",
          main = "Box Plot for Lot Area", ylim = c(0, 30000))
```



```
##  [1] "76"   "490"  "615"  "957"  "1039" "1040" "1450" "314"  "452"  "1299"
## [11] "770"  "54"   "662"  "849"  "524"  "272"  "1170"
```

```
qqnorm(Train$LotArea, main = "QQ Plot for LotArea")
```



The qqplot and boxplot suggests similar results as that of the histogram. The quantile plot shows that its distribution is not normal and it is right skewed with some extreme large values.The boxplot of LotArea shows that the median is around 10000 square feet. But there are some extreme large houses up to 20000 square feet. These can be outliers impacting our result.

2. *Description of GarageArea*

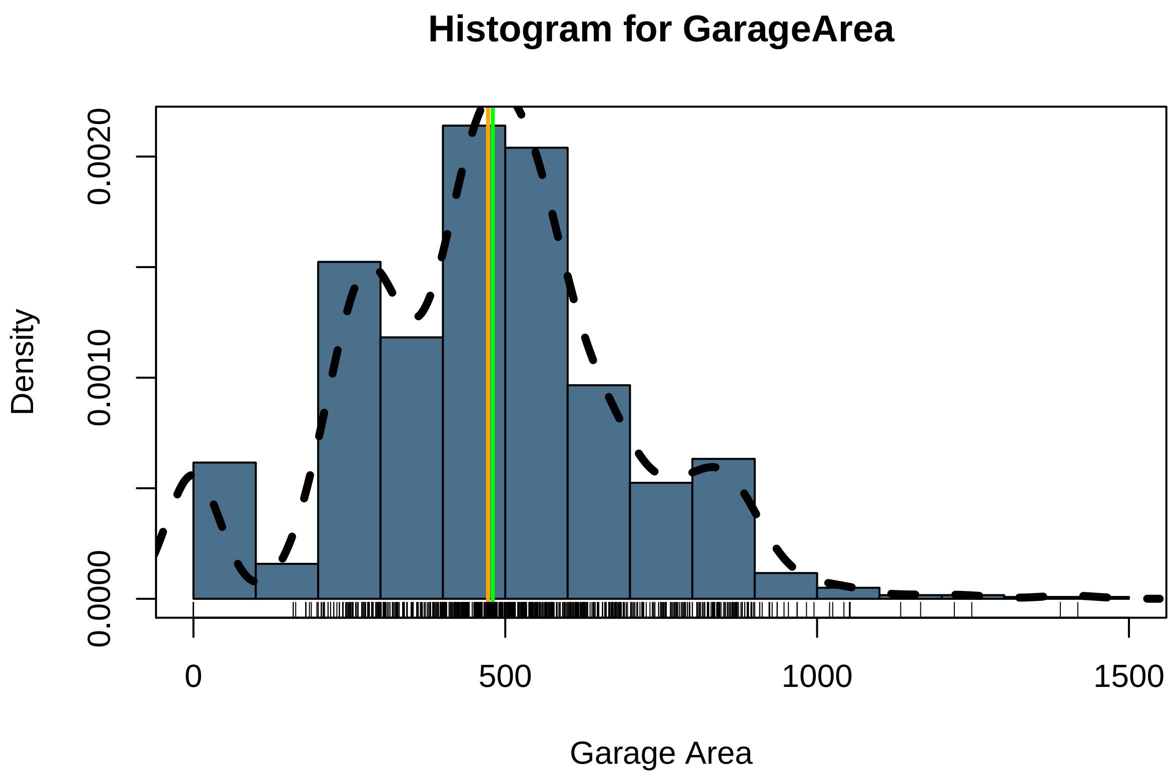
```
stat.desc(Train$GarageArea)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
## 1.201000e+03 7.400000e+01 0.000000e+00 0.000000e+00 1.418000e+03 1.418000e+03
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 5.674060e+05 4.800000e+02 4.724446e+02 6.382062e+00 1.252124e+01 4.891759e+04
##      std.dev     coef.var
## 2.211732e+02 4.681463e-01
```

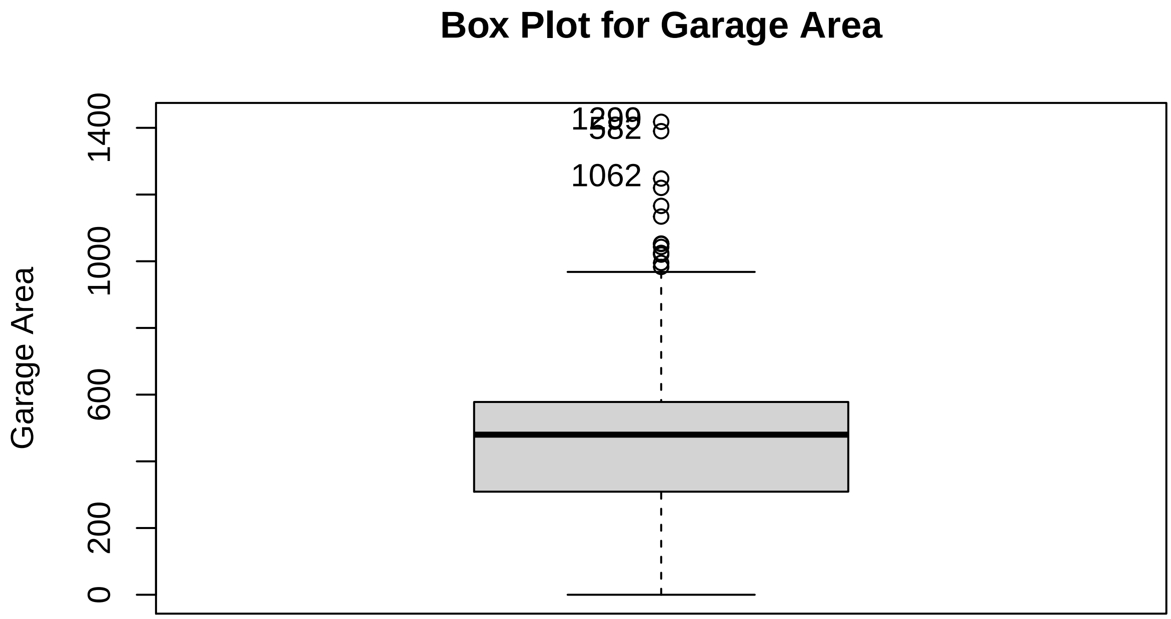
The statistical summaries are as follows: the mean is 472.44 while the median is 480. The maximum value is 1418 and the minimum value is 0, making it a large range.

```
#Histogram
k = 1 + log2(length(Train$GarageArea))
hist(Train$GarageArea, breaks = k, col="skyblue4",
     main = "Histogram for GarageArea",
     freq = FALSE, xlab = "Garage Area", ylab = "Density")
lines(density(Train$GarageArea, na.rm = TRUE),lwd=4,col='black',lty=2)
```

```r
abline(v = mean(Train$GarageArea), col = "orange", lwd =2, lty = 1 )
abline(v = median(Train$GarageArea), col = "green", lwd =2, lty = 1 )
rug(Train$GarageArea)
box()
```

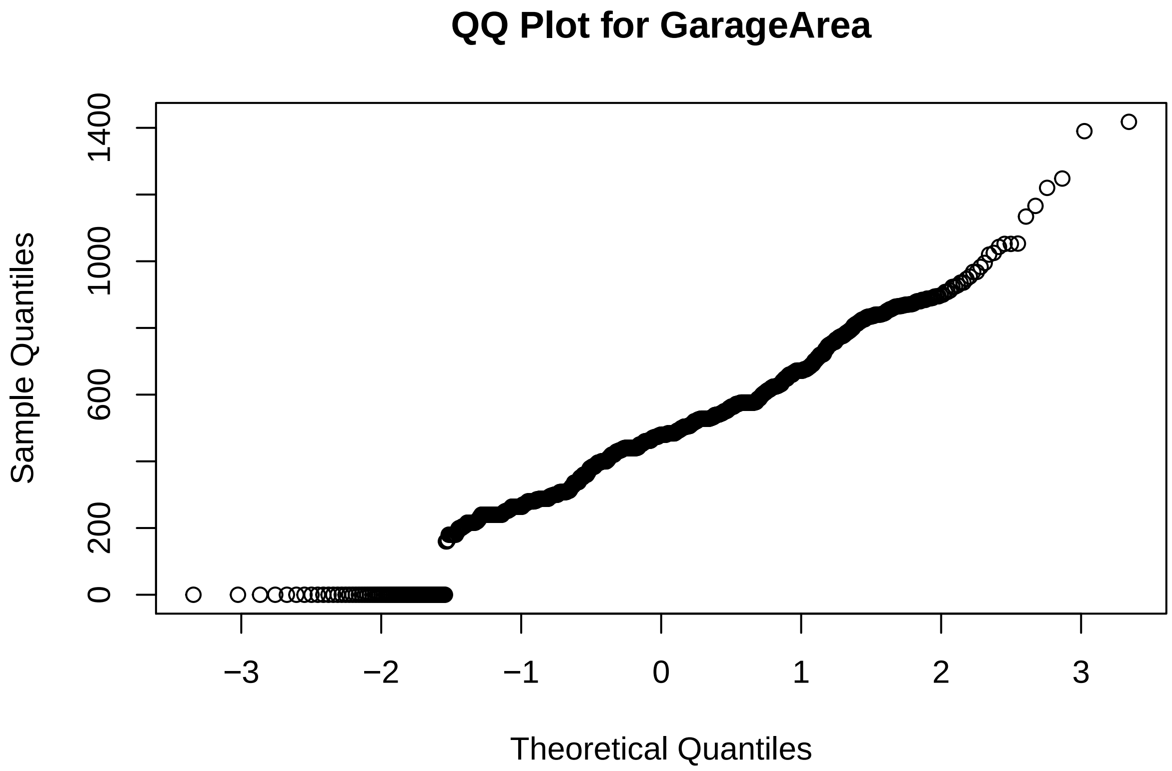


```r
#Boxplot
Boxplot(~GarageArea, data=Train,id=list(n=3), ylab = "Garage Area",
main = "Box Plot for Garage Area")
```



```
## [1] "1299" "582"  "1062"
```

```
#QQ-plot
qqnorm(Train$GarageArea, main = "QQ Plot for GarageArea")
```



I plotted a histogram for Garage Area with near optimal number of bins based on the number of observations (given by $k = 1 + log_2(n)$) The Histogram and density curve suggest that the data is multimodal. There is an indication of outlier in the data beyond approx 1000 sqft area. A similar idea is suggested both by the boxplot and the QQplot. The rugplot suggests that most of the data is concentrated between 250 and 950 sqft area. The orange line indicates the mean garage area which is roughly 470 sqft. The green line indicates the median of the garage area which is roughly 480 sqft.The mean and median match each other. The Boxplot helps us visualize median as well. The IQR as seen in the boxplot is roughly 241.5

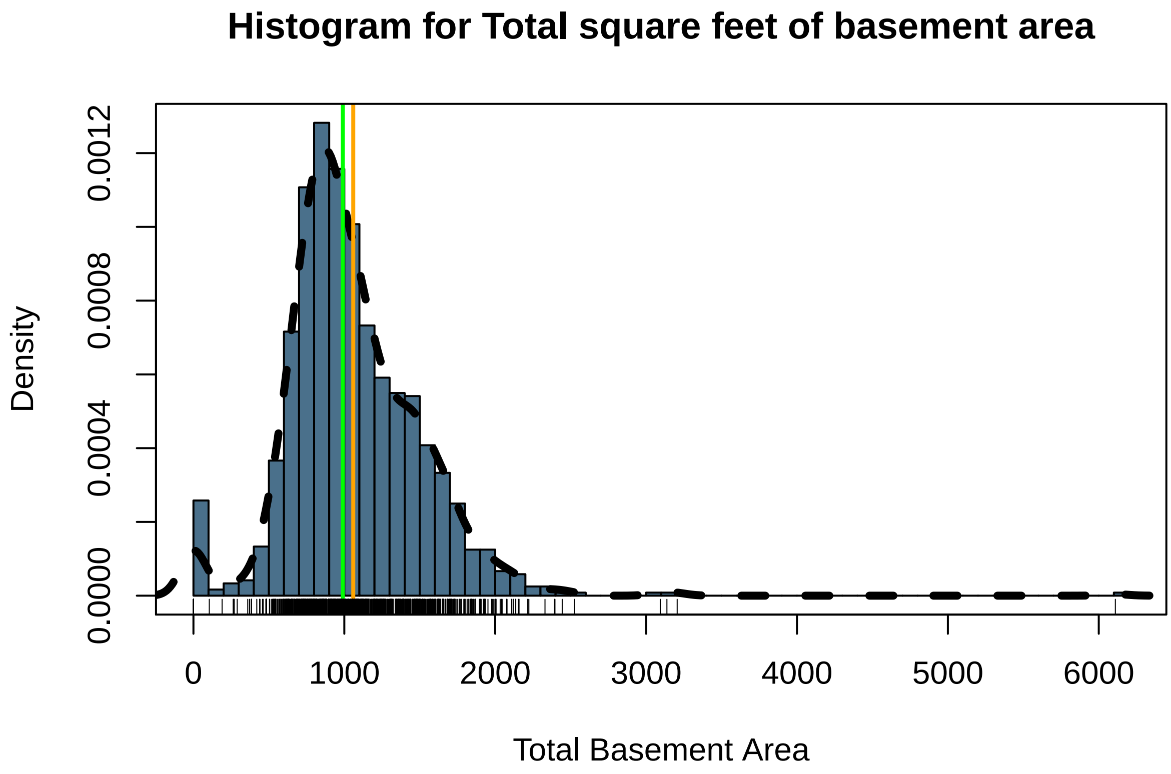
3. *Description of Total BsmtSF*

```
stat.desc(Train$TotalBsmtSF)
```

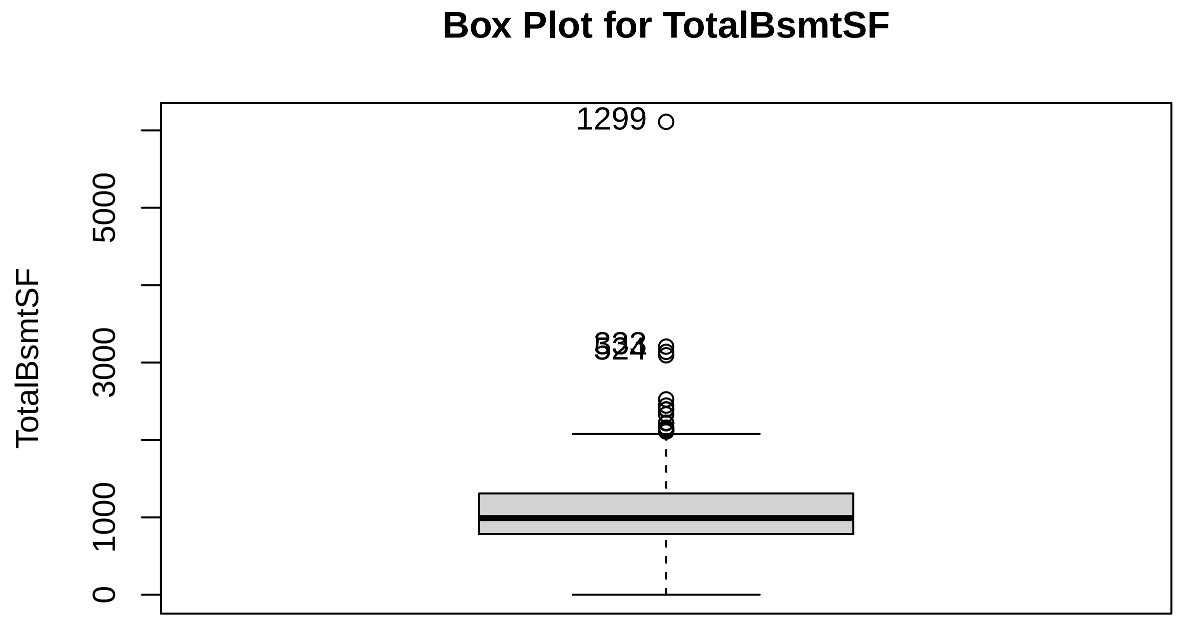
```
##      nbr.val     nbr.null       nbr.na          min          max        range 
## 1.201000e+03 3.100000e+01 0.000000e+00 0.000000e+00 6.110000e+03 6.110000e+03 
##          sum       median         mean      SE.mean CI.mean.0.95          var 
## 1.272321e+06 9.900000e+02 1.059385e+03 1.293612e+01 2.537993e+01 2.009793e+05 
##      std.dev     coef.var 
## 4.483071e+02 4.231769e-01
```

The statistical summaries are as follows: the mean is 1059.385 while the median is 990. The maximum value is 6110 and the minimum value is 0, making it a large range.

```
hist(Train$TotalBsmtSF, breaks = "FD", col="skyblue4",
     main="Histogram for Total square feet of basement area",
     freq = FALSE,
     xlab = "Total Basement Area",
     ylab ="Density")
lines(density(Train$TotalBsmtSF),lwd=4,col='black',lty=2)
abline(v = mean(Train$TotalBsmtSF), col = "orange", lwd =2, lty = 1 )
abline(v = median(Train$TotalBsmtSF), col = "green", lwd =2, lty = 1 )
rug(Train$TotalBsmtSF)
box()
```

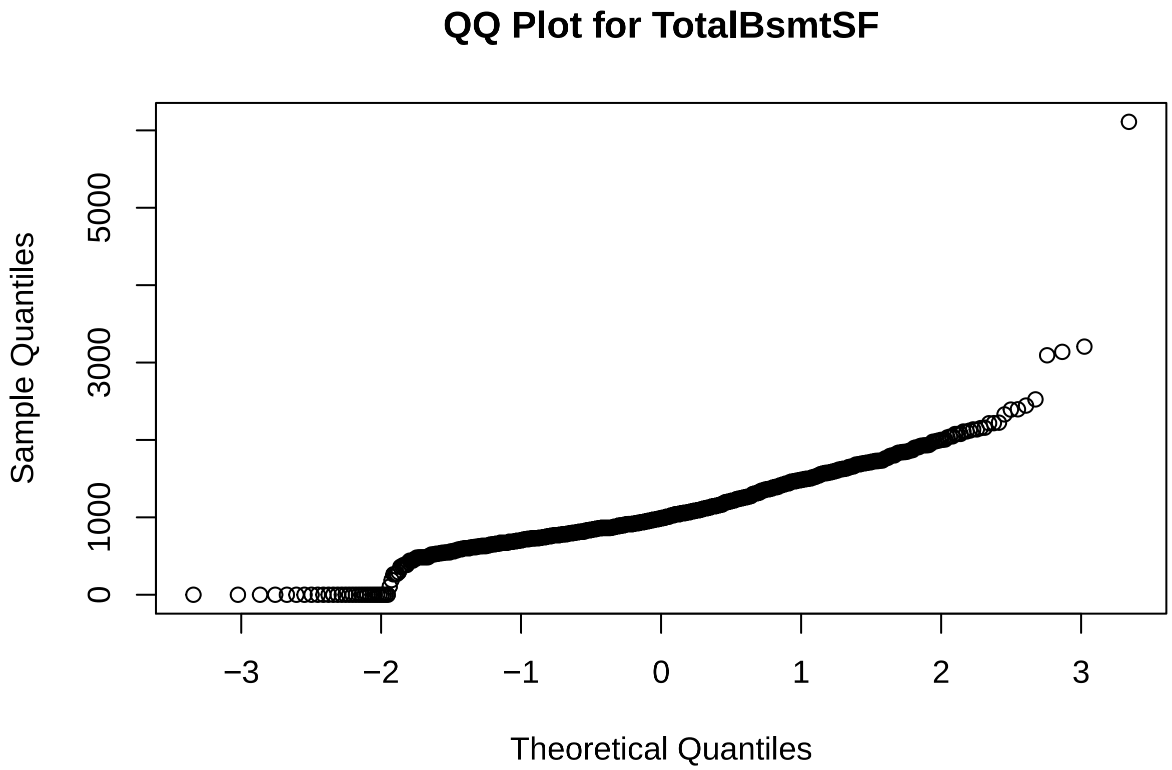


```
#Boxplot
Boxplot(~TotalBsmtSF, data=Train,id=list(n=3), ylab = "TotalBsmtSF",
main = "Box Plot for TotalBsmtSF")
```



```
## [1] "1299" "333"  "524"
```

```
#QQ-plot
qqnorm(Train$TotalBsmtSF, main = "QQ Plot for TotalBsmtSF")
```



I plotted the Boxplot, qqplot and histogram.They suggests outliers in the data especially beyond 2000 sqft. The median as can be seen from the boxplot is approximately is 990 sqft. The IQR is approx 500 sqft.

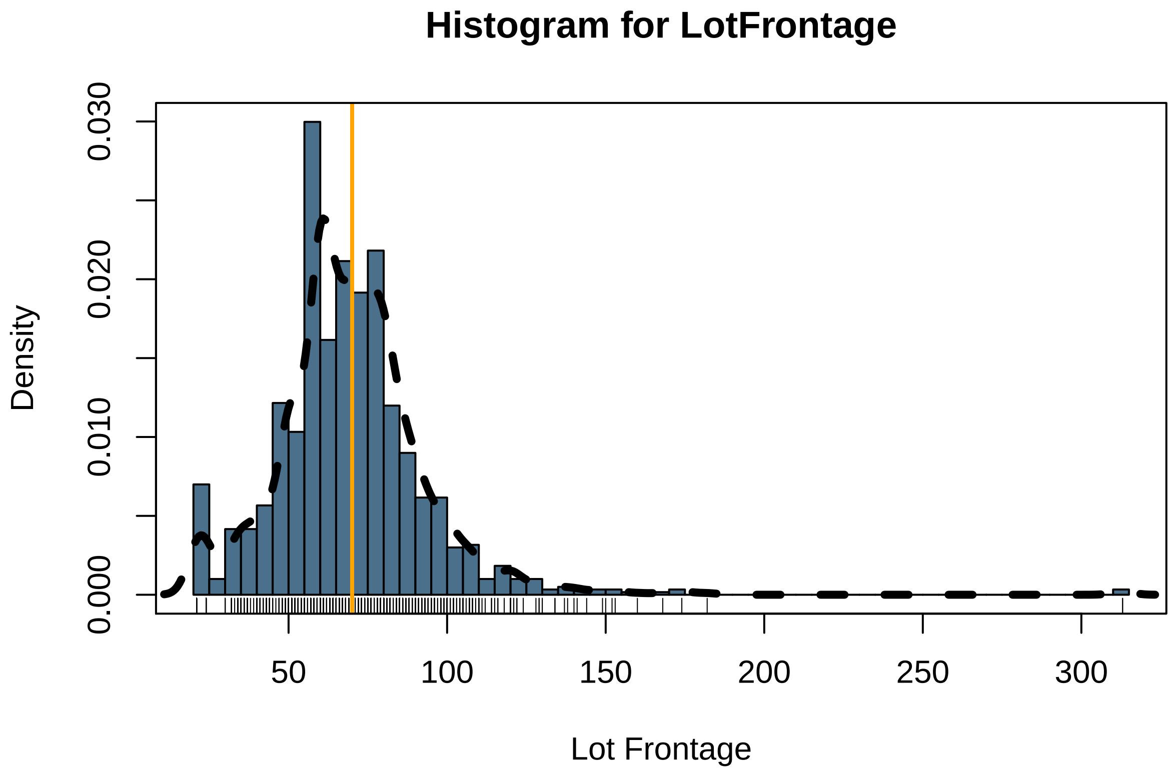
4. *Description of Lot Frontage*

```
stat.desc(Train$LotFrontage)
```

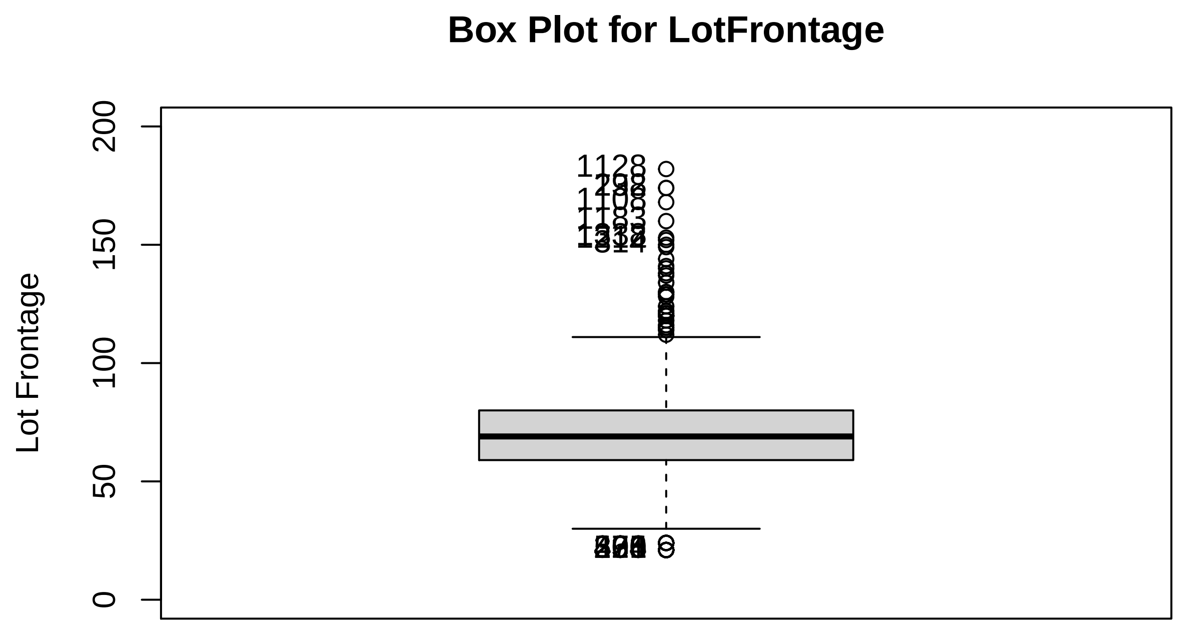
```
##      nbr.val     nbr.null       nbr.na          min          max        range
## 1.201000e+03 0.000000e+00 0.000000e+00 2.100000e+01 3.130000e+02 2.920000e+02
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 8.413000e+04 6.900000e+01 7.004996e+01 7.007485e-01 1.374828e+00 5.897492e+02
##      std.dev     coef.var
## 2.428475e+01 3.466776e-01
```

The statistical summaries are as follows: the mean is 70.049 while the median is 69. The maximum value is 313 and the minimum value is 21.

```
#histogram
hist(Train$LotFrontage, breaks = "FD", col="skyblue4",
     main = "Histogram for LotFrontage",
     freq = FALSE, xlab = "Lot Frontage", ylab = "Density")
lines(density(Train$LotFrontage),lwd=4,col='black',lty=2)
rug(Train$LotFrontage)
abline(v = mean(Train$LotFrontage), col = "orange", lwd =2, lty = 1 )
box()
```

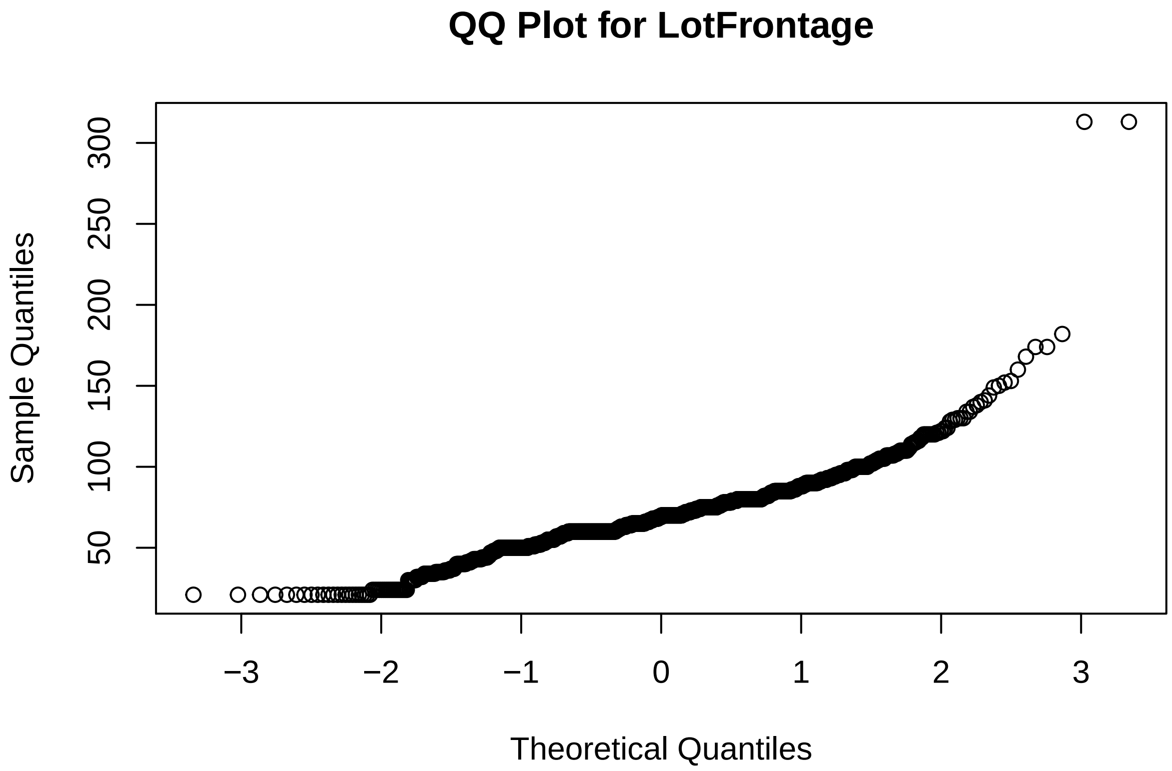


```
#boxplot
Boxplot(~LotFrontage, data=Train, ylab = "Lot Frontage",
main = "Box Plot for LotFrontage", ylim = c(0,200))
```



```
##  [1] "76"   "226"  "228"  "233"  "236"  "364"  "431"  "435"  "490"  "501"
## [11] "935"  "1299" "1128" "198"  "232"  "1108" "1183" "1338" "1212" "314"
```

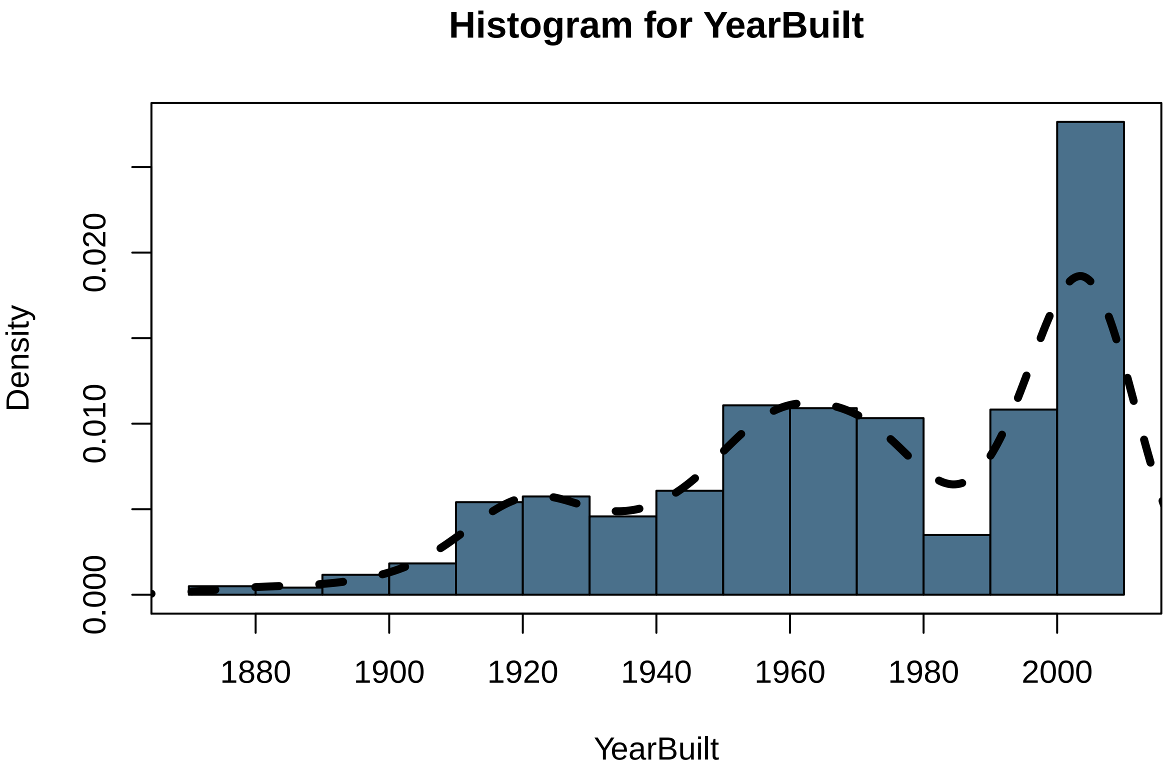
```
#QQPlot
qqnorm(Train$LotFrontage, main = "QQ Plot for LotFrontage")
```



Both Histograms and Boxplot suggest that the distribution is positively skewed which means that there are a few values are very large and many small values. It is visible form the graphs that there are outliers in the data which renders the graphs non-symmetric. The quantile plot shows that its distribution is not normal and it is right skewed.

5. *Description of Year Built*

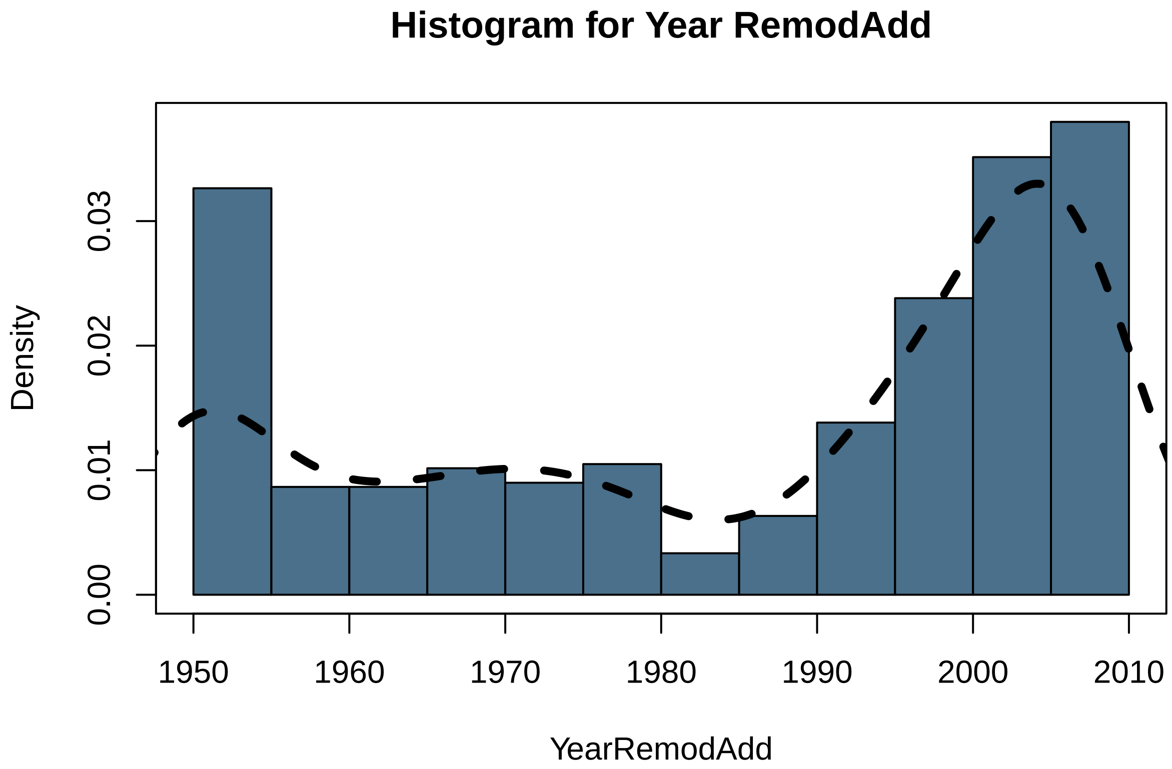
```
hist(Train$YearBuilt, breaks = "FD", col="skyblue4",
     main = "Histogram for YearBuilt",
     freq = FALSE, xlab = "YearBuilt", ylab = "Density")
lines(density(Train$YearBuilt),lwd=4,col='black',lty=2)
box()
```



Histogram and Boxplot suggests that data is negatively skewed. The old houses constructed in the 1880s are few in number as compared to houses constructed in the 1990s houses constructed in the 1880s are very few and 1960s saw an increase in the construction. In the data, the maximum number of houses were built in the year 2000

6. *Description of Year RemodAdd*

```
hist(Train$YearRemodAdd, breaks = "FD", col="skyblue4",
     main = "Histogram for Year RemodAdd",
     freq = FALSE, xlab = "YearRemodAdd", ylab = "Density")
lines(density(Train$YearRemodAdd),lwd=4,col='black',lty=2)
box()
```



The histogram showed that the remodel date of the houses are either in 1950s or after 21st century. That might also be because YearRemodAdd can be same as YearBuilt if no remodeling or additions are done. There are more houses built recently, so the YearRemodAdd dates also tend to be more recent.

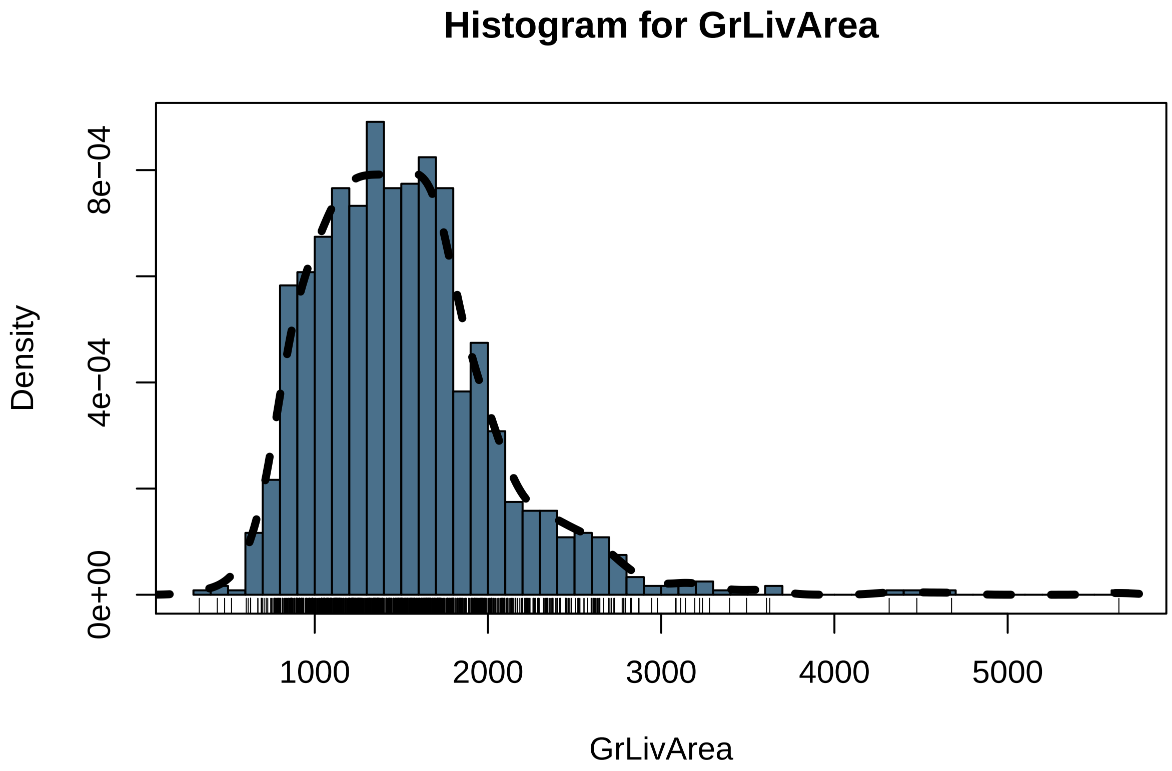
7. *Description of GrLiv Area*

```
stat.desc(Train$GrLivArea)
```

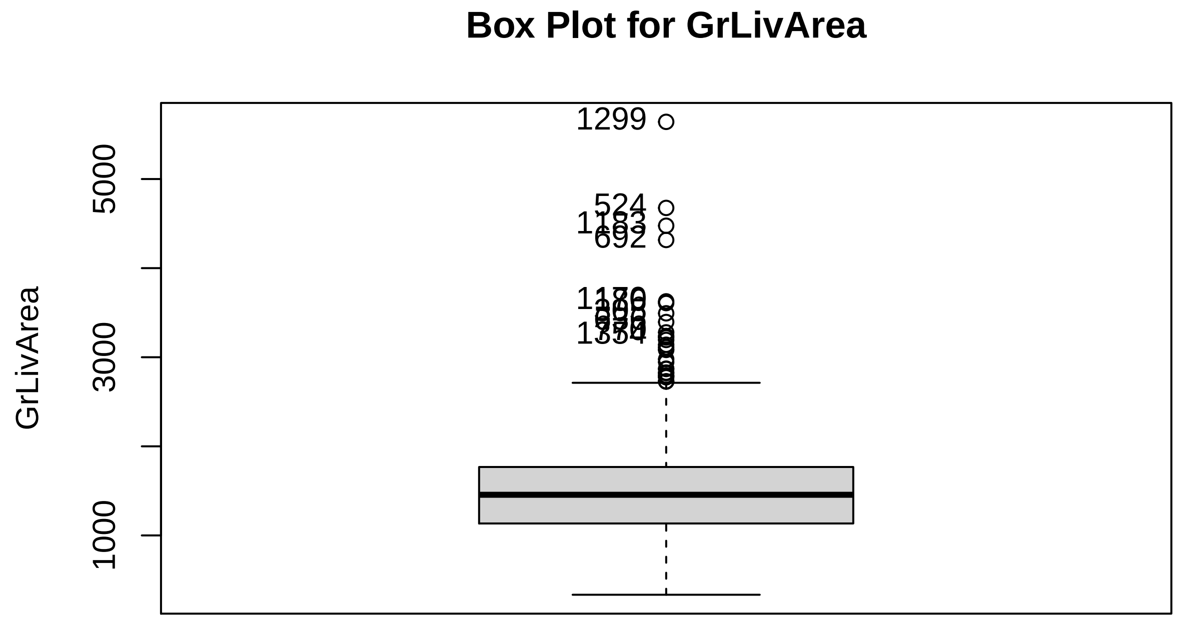
```
##      nbr.val     nbr.null       nbr.na          min          max        range 
## 1.201000e+03 0.000000e+00 0.000000e+00 3.340000e+02 5.642000e+03 5.308000e+03 
##          sum       median         mean      SE.mean CI.mean.0.95          var 
## 1.814870e+06 1.456000e+03 1.511132e+03 1.529134e+01 3.000073e+01 2.808239e+05 
##      std.dev     coef.var 
## 5.299282e+02 3.506828e-01
```

The statistical summaries are as follows: the mean is 1511.132 while the median is 1456. The maximum value is 5642 and the minimum value is 334.

```
hist(Train$GrLivArea, breaks = "FD", col="skyblue4",
     main = "Histogram for GrLivArea",
     freq = FALSE, xlab = "GrLivArea", ylab = "Density")
lines(density(Train$GrLivArea, na.rm = TRUE),lwd=4,col='black',lty=2)
rug(Train$GrLivArea)
box()
```

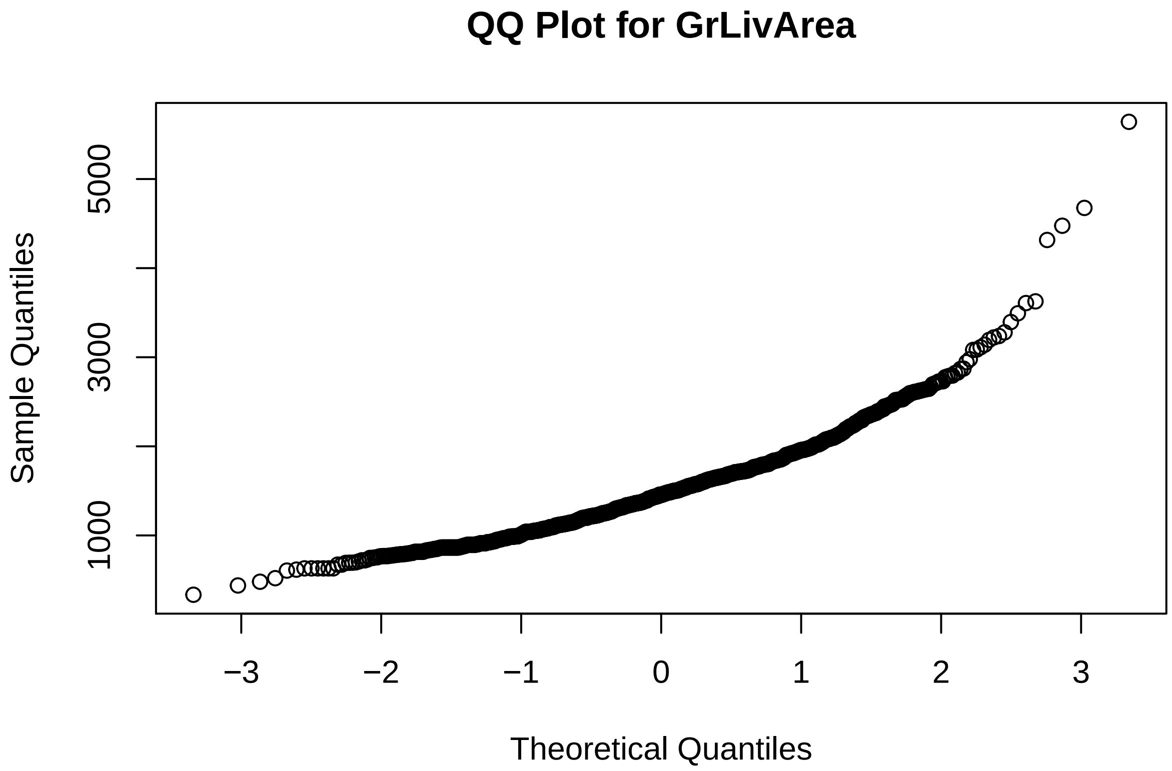


```
#boxplot
Boxplot(~GrLivArea, data=Train, ylab = "GrLivArea",
main = "Box Plot for GrLivArea")
```



```
## [1] "1299" "524"  "1183" "692"  "1170" "186"  "305"  "636"  "770"  "1354"
```

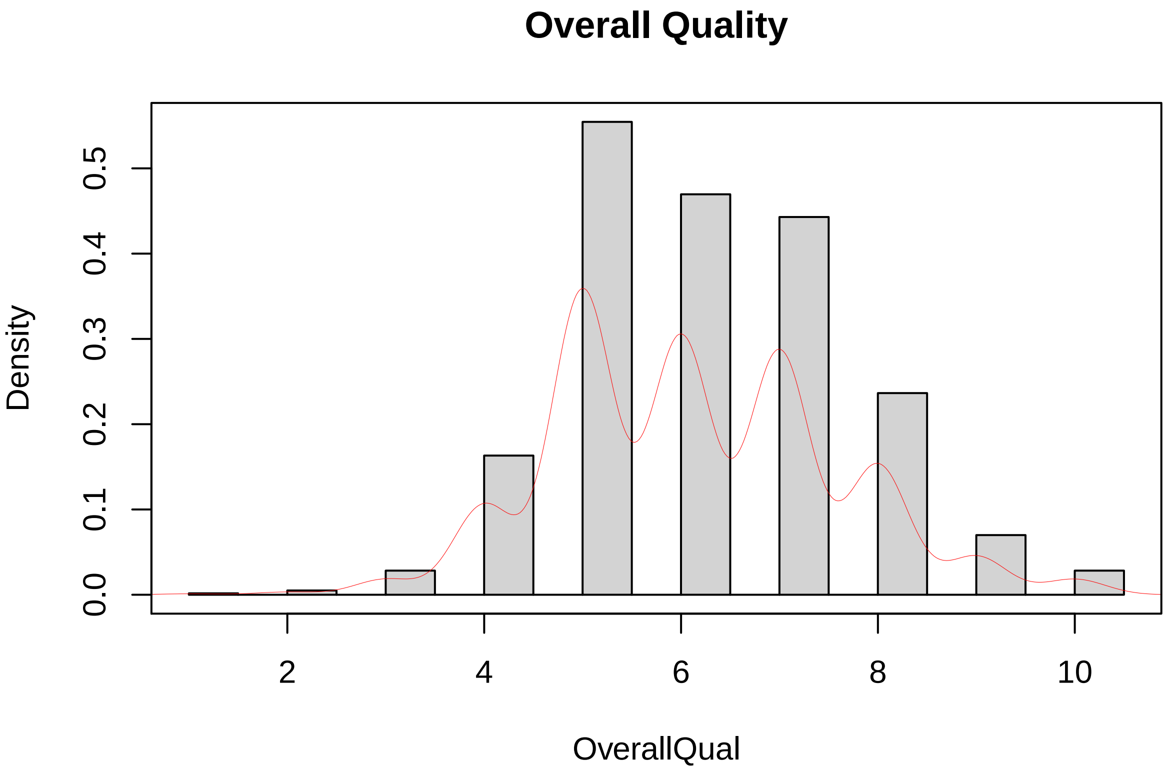
```
#QQPlot
qqnorm(Train$GrLivArea, main = "QQ Plot for GrLivArea")
```



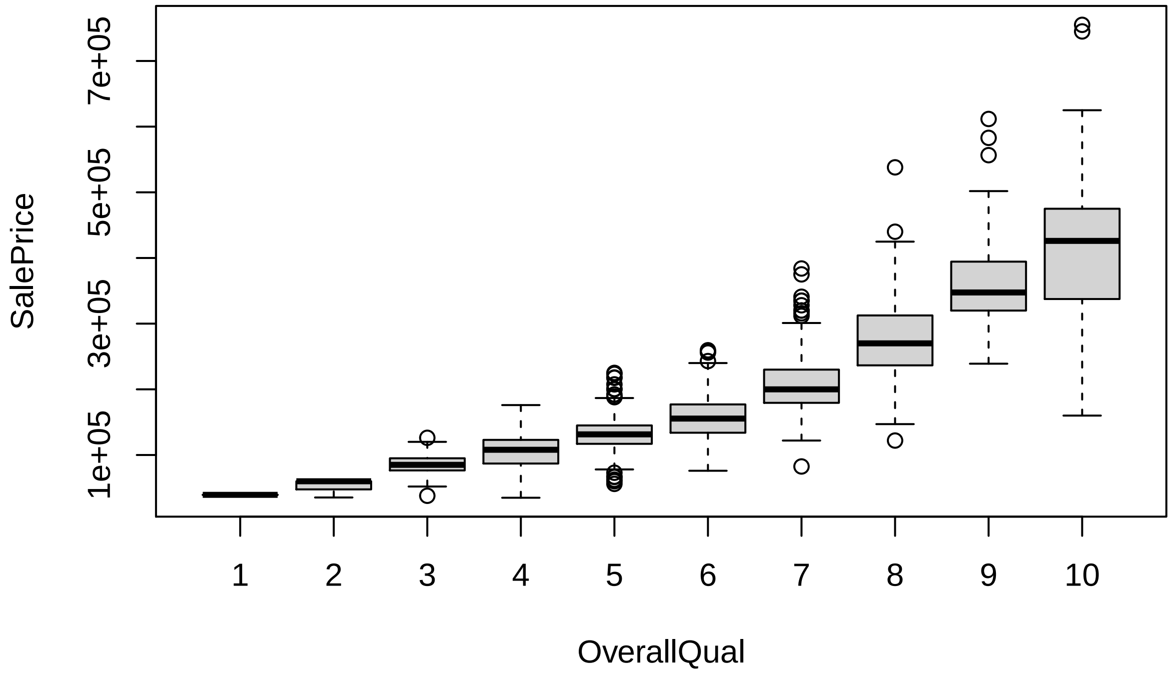
The histogram, QQplot and Boxplot seems to be positively skewed which means that there are a few values are very large and many small values. The same idea is suggested by the rugplot also. So, in such a case, I feel that taking a log transformation spreads out the small values and compresses the large one.However formal transformation test is required. As is seen visually, the data has outliers beyond 3000.

8. *Description of Overall Quality*

```r
#Histogram
truehist(Train$OverallQual,col='lightgrey',xlab="OverallQual", ylab="Density",
         main = "Overall Quality")
lines(density(Train$OverallQual),lwd=0.1,col='red')
box()
```



```
#Boxplot
boxplot(Train$SalePrice~Train$OverallQual, xlab = "OverallQual", ylab = "SalePrice")
```



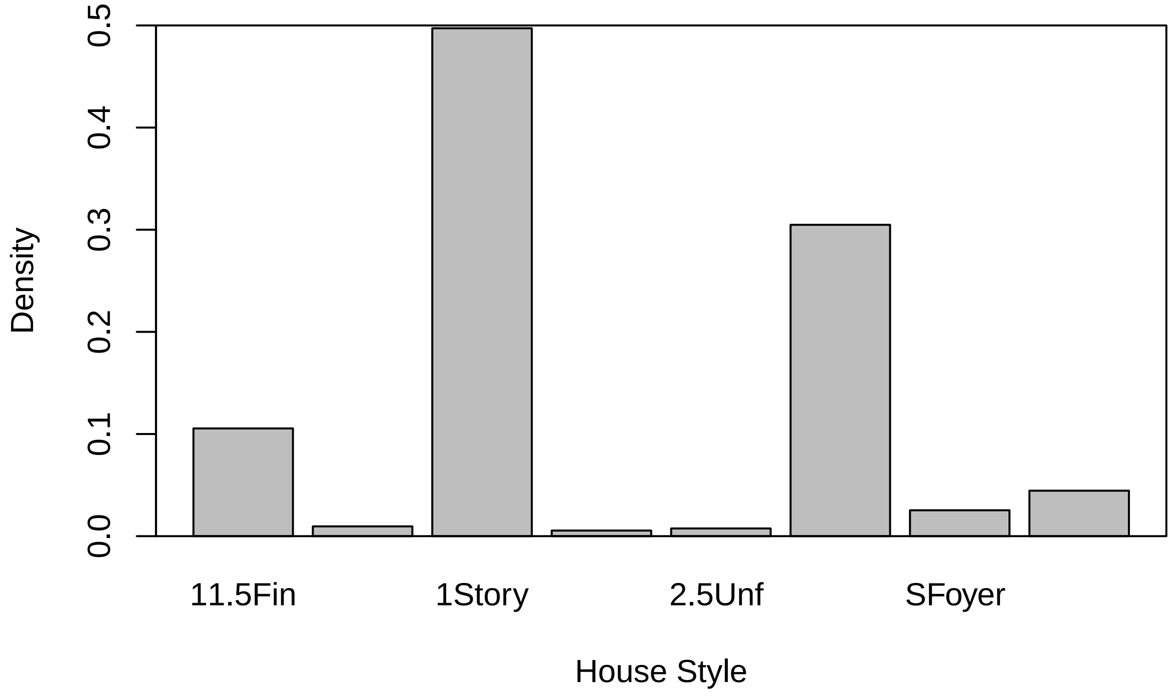
```r
#Statistical Description
stat.desc(Train$OverallQual)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range 
## 1.201000e+03 0.000000e+00 0.000000e+00 1.000000e+00 1.000000e+01 9.000000e+00 
##          sum       median         mean      SE.mean CI.mean.0.95          var 
## 7.353000e+03 6.000000e+00 6.122398e+00 4.095217e-02 8.034581e-02 2.014173e+00 
##      std.dev     coef.var 
## 1.419216e+00 2.318071e-01
```

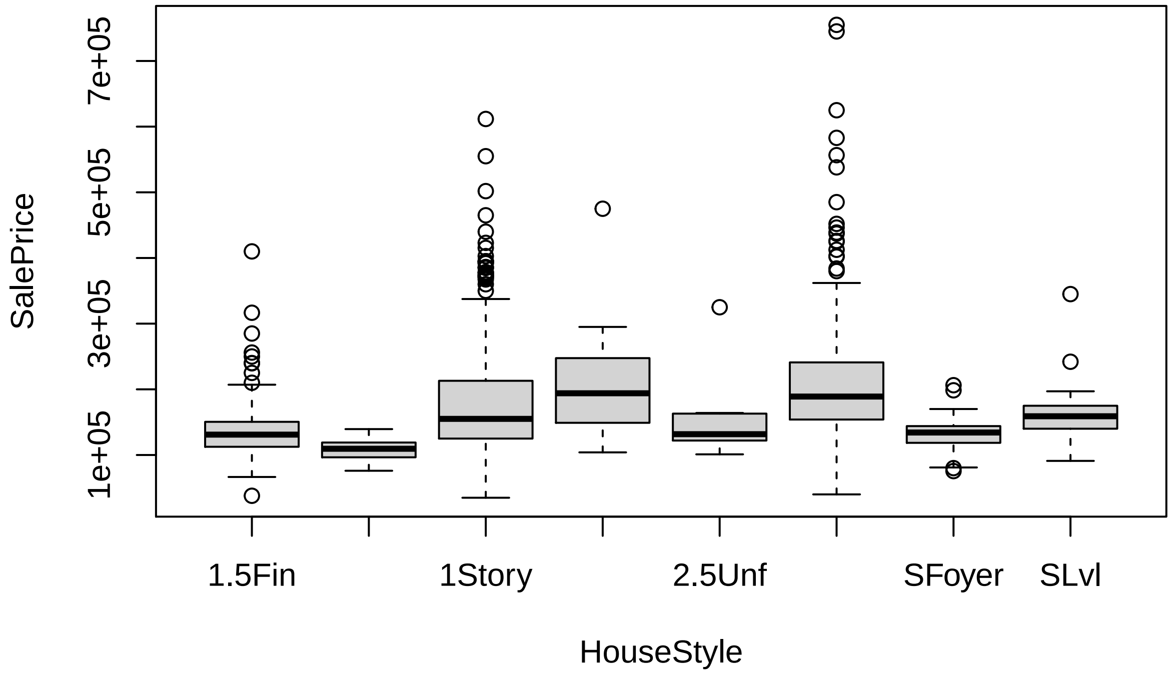
The histogram shows that most houses are rated around 6 points, which is above average. Houses rating 5 takes up the most proportion. The boxplot shows that Rates the overall material and finish of the house has a significant influence on price. The mean rating is 6.12, which is above average. The median is 6, maximum rating is 10 and minimum rating is 1.

9. *Description of House Style*

```r
#Barplot
barplot(c(154,14,726,8,11,445,37,65)/1460,
        names.arg = c("11.5Fin","1.5Unf ","1Story"," 2.5Fin ","2.5Unf ","2Story",
                      "SFoyer "," SLvl "),xlab = "House Style",ylab = "Density",
                      ylim=c(0,0.5))
box()
```



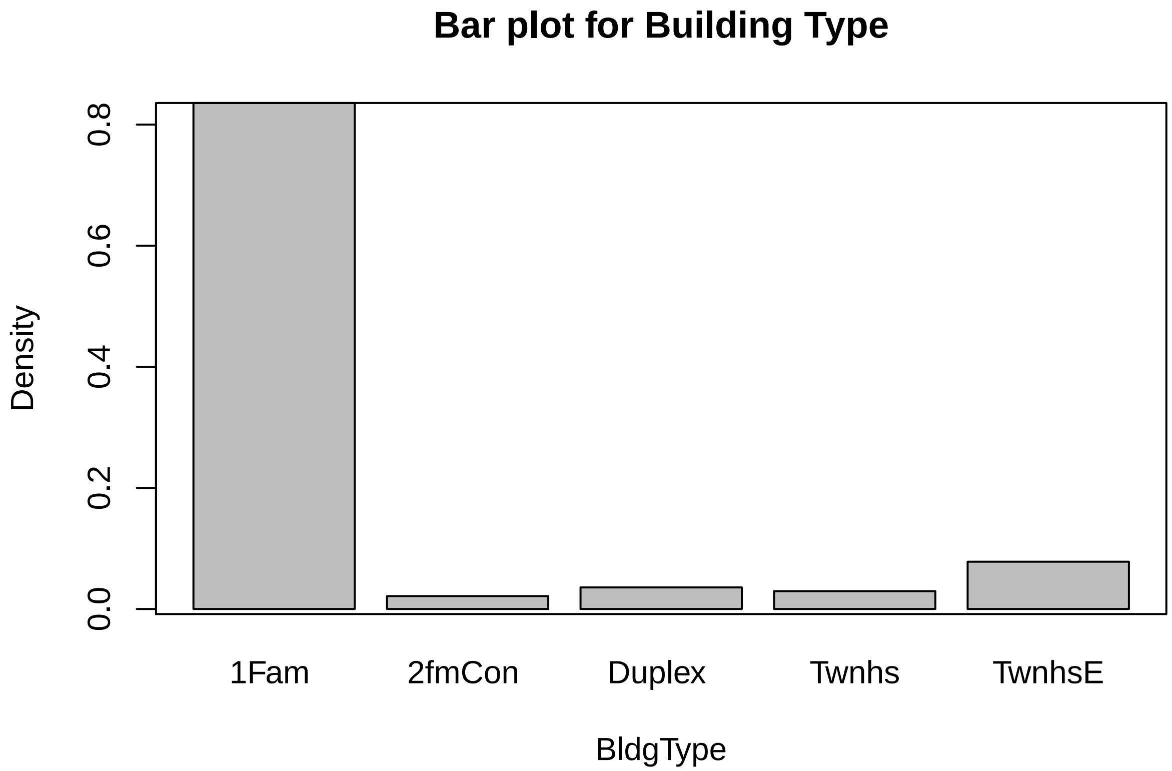
```
#Boxplot
boxplot(Train$SalePrice~Train$HouseStyle, xlab = "HouseStyle", ylab = "SalePrice")
```



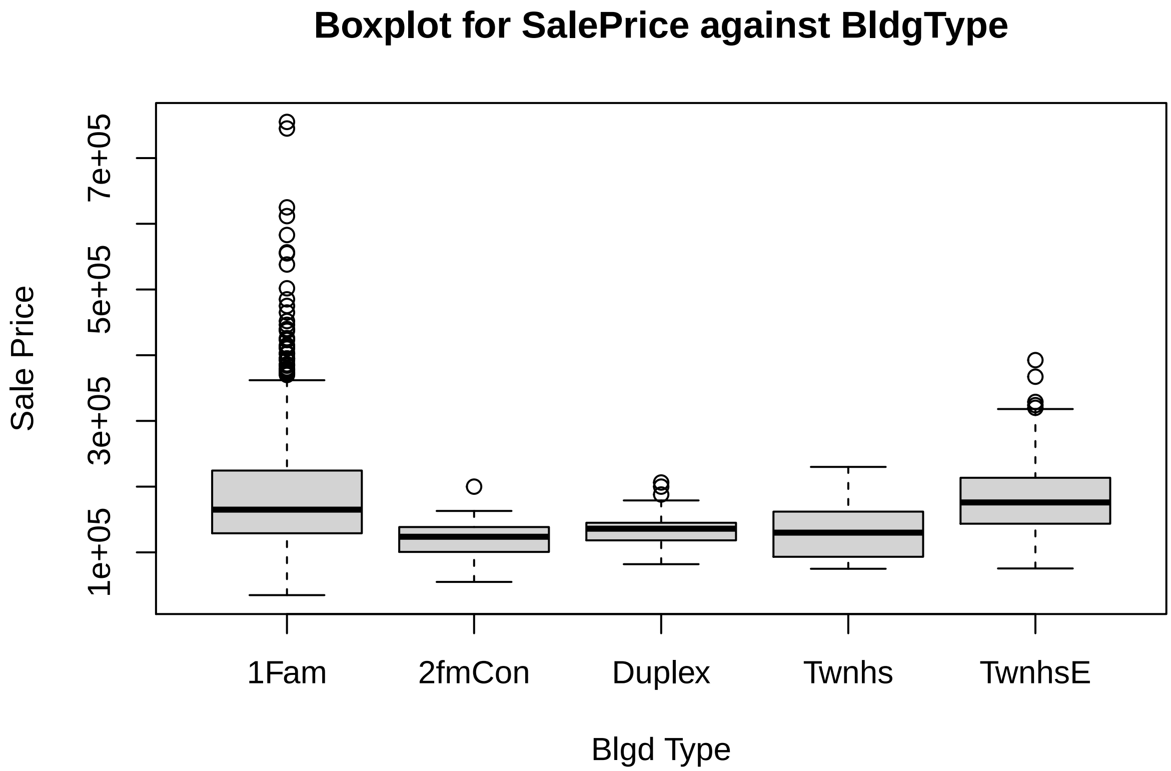
The distribution shows that most of the houses are one story and two and one-half story: 2nd level unfinished.The box plot shows that houses with Two story or Two and one-half storey: 2ndlevel finished tend to have higher price.

9. *Description of Bldg Type*

```
#Barplot
barplot(c(1220,31,52,43,114)/1460,names.arg = c("1Fam ","2fmCon ","Duplex ",
                                               " Twnhs ","TwnhsE"),
                                         xlab = "BldgType",ylab = "Density",
                                        main = "Bar plot for Building Type")
box()
```



```r
boxplot(Train$SalePrice~Train$BldgType,
main = "Boxplot for SalePrice against BldgType", xlab = "Blgd Type",
ylab ="Sale Price")
```



The distribution showed that most of the houses are Single-family Detached dwelling type. Only small proportion are Two-family Conversion, Duplex, TownhouseEnd Unit and Townhouse Inside Unit. The boxplot shows that Townhouse Inside Unit houses have higher prices. The prices for Single-family Detached houses are also higher, but with biggerspread.

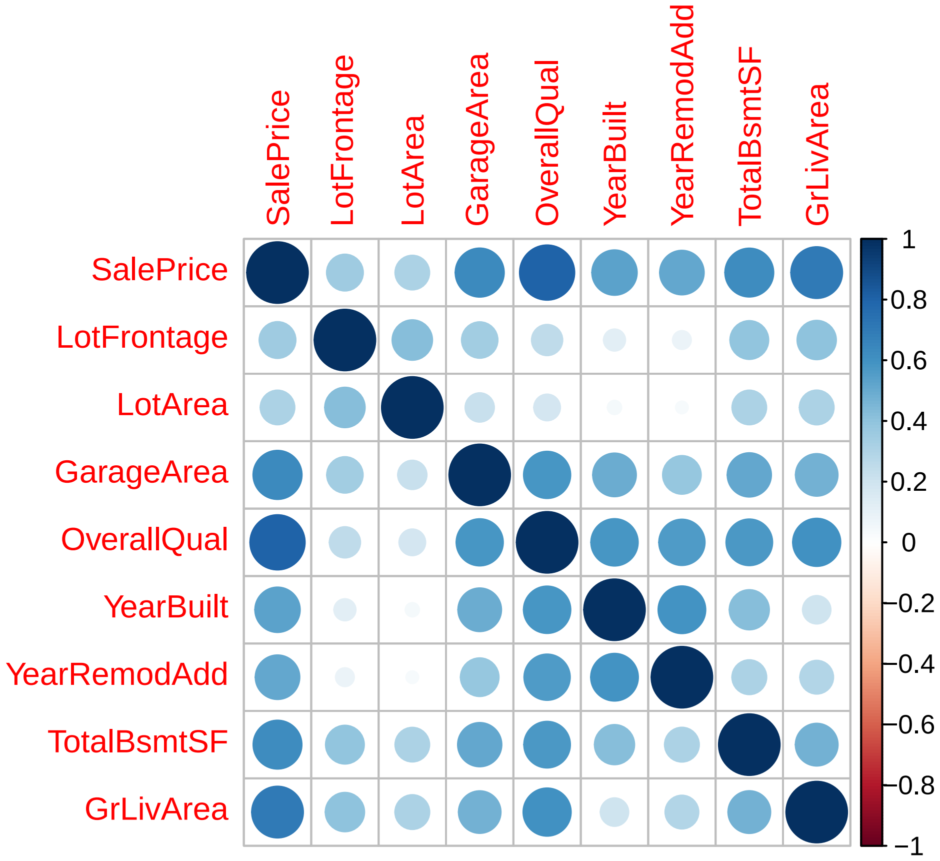
To get an overall picture of the variables, we plot the correlation matrix

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.0.3
```

```
## corrplot 0.84 loaded
```

```
corrplot(cor(Train[c('SalePrice','LotFrontage','LotArea','GarageArea','OverallQual',
                     'YearBuilt','YearRemodAdd','TotalBsmtSF','GrLivArea')]))
```



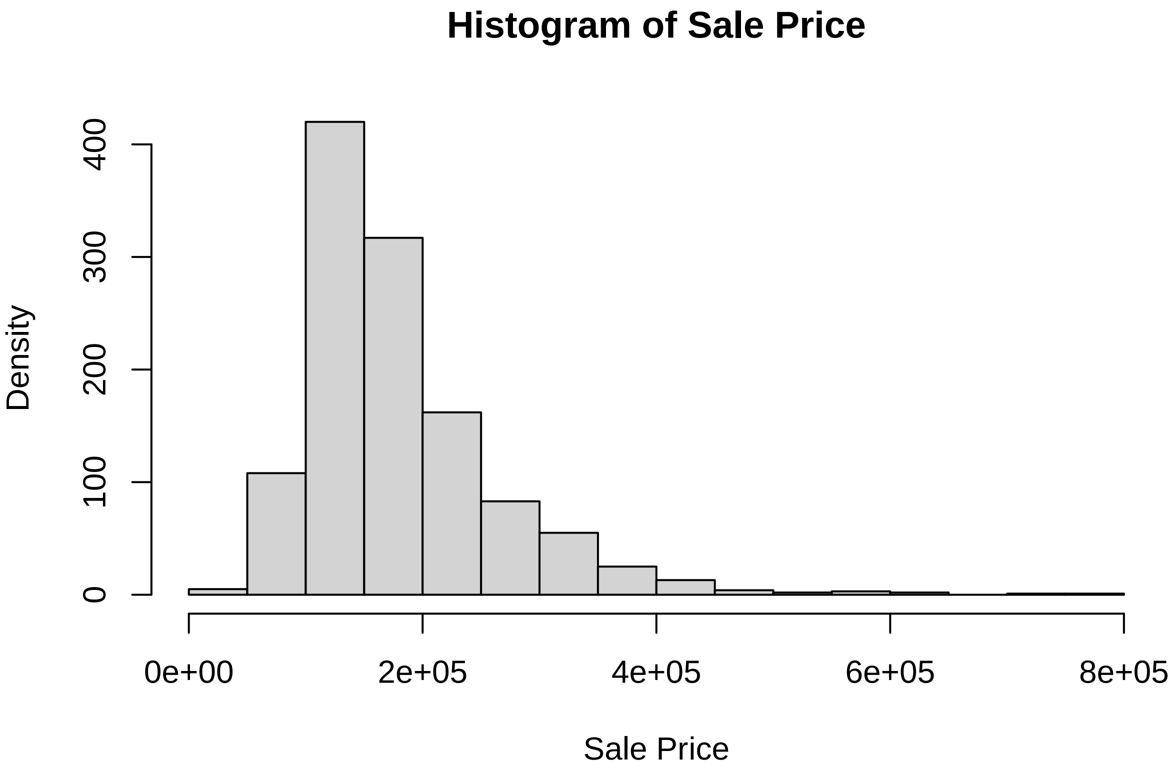
GrLivArea, GrarageArea,TotalBsmtSF have a strong positive correlation with sale price. OverallQual has a stronger correlation to the sale price. It can be seen that YearRemodAdd has a correlation with YearBuilt as per the plot atleast. Overallqual and Yearbuilt also have a strong correlation.

(b) For each variable (except indicator ones), test if a transformation to linearity is appropriate, and if so, apply the respective transformation, and comment on the transformed predictor(s).

We carry out the transformation of the numerical variables. This is based on scatter plots, box-plots of power transformations, histograms and box-cox transformation. I decided to perform the box-cox transformation on some of my numeric variables individually because after plotting my numeric variables it suggested that certain variables did not require a box-cox transformation. However for a sanity check, I performed the box-cox transformation test on numeric variable.

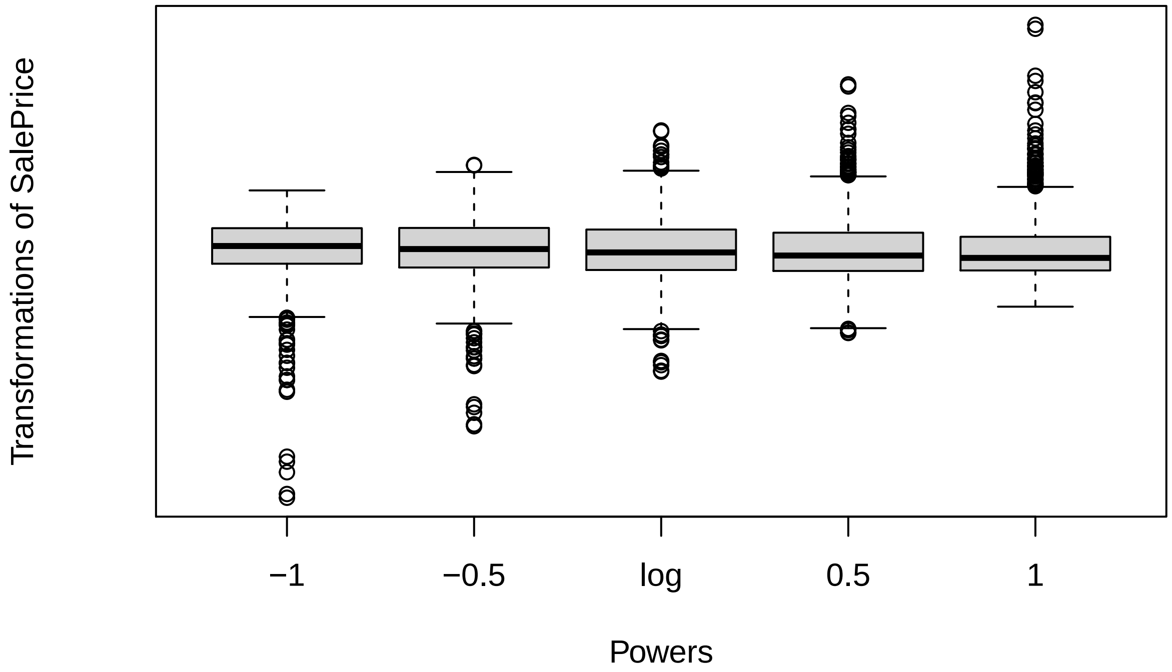
1. $Transformation\ of\ Sale\ Price:$

```r
hist(Train$SalePrice, main = "Histogram of Sale Price", xlab = "Sale Price",
     ylab = "Density")
```



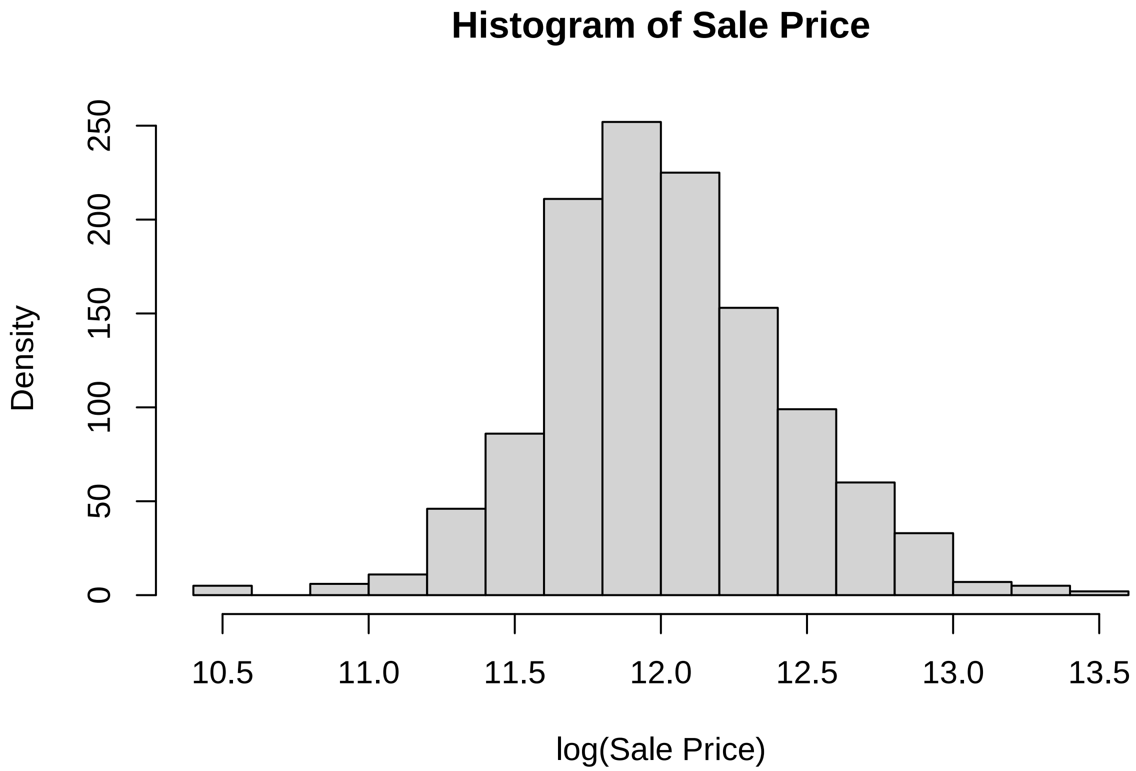
The graph tells us that we require transformation of the y variable. My intuition tells me that we should do a log transformation since the distribution looks right skewed. Now we plot a boxplot.

```
symbox(~SalePrice, data = Train)
```



The Boxplot of Power Transformation tells us that we should do a log transformation. Now we will plot the histogram of $log(SalePrice)$.

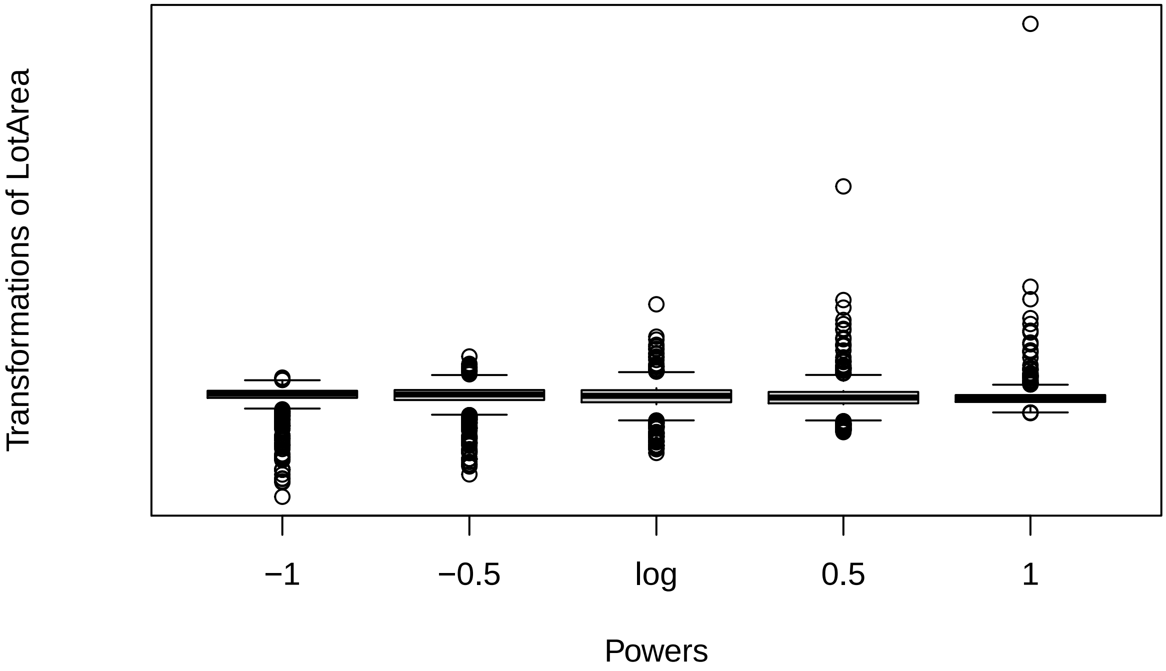
```
hist(log(Train$SalePrice), main = "Histogram of Sale Price",
     xlab = "log(Sale Price)", ylab = "Density")
```



After log transformation, the graph appears to be following a symmetric distribution, So we go ahead with carrying out log transformation for Sale Price.

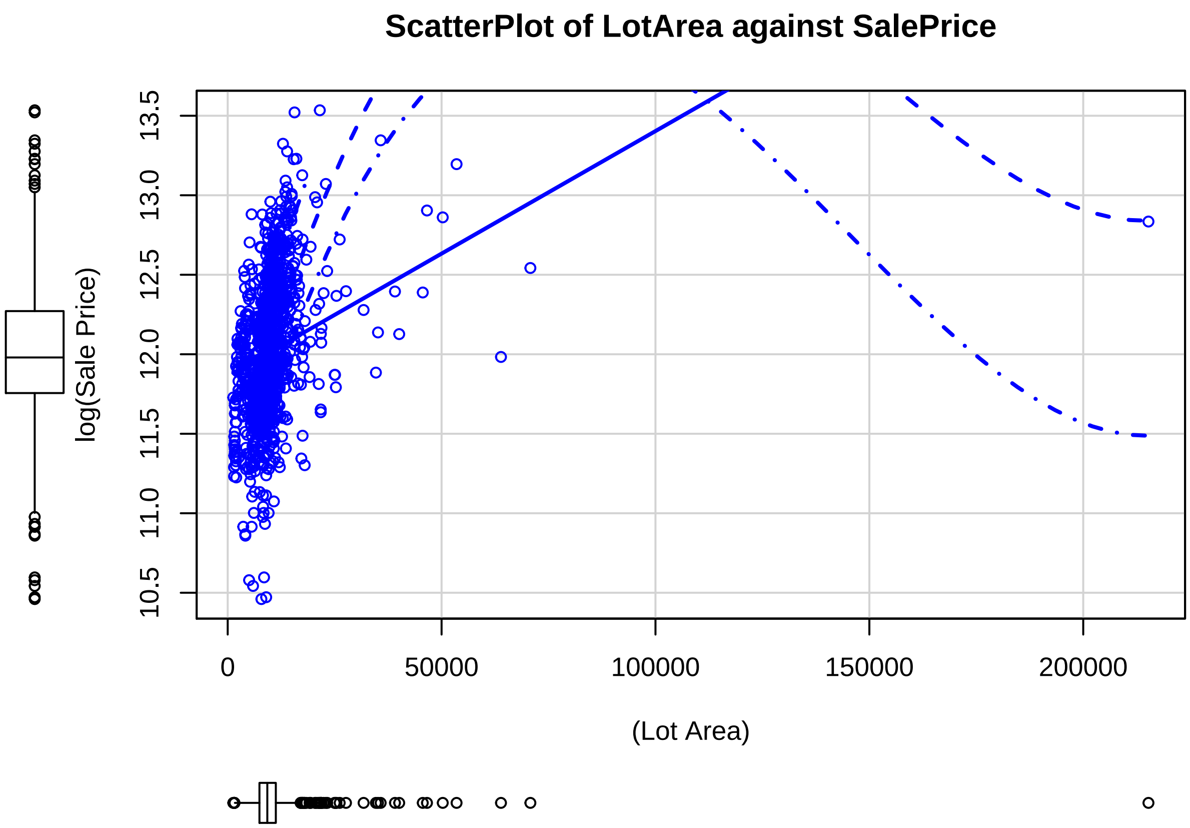
2. *Transformation of Lot Area* :

```
symbox(~LotArea, data = Train)
```

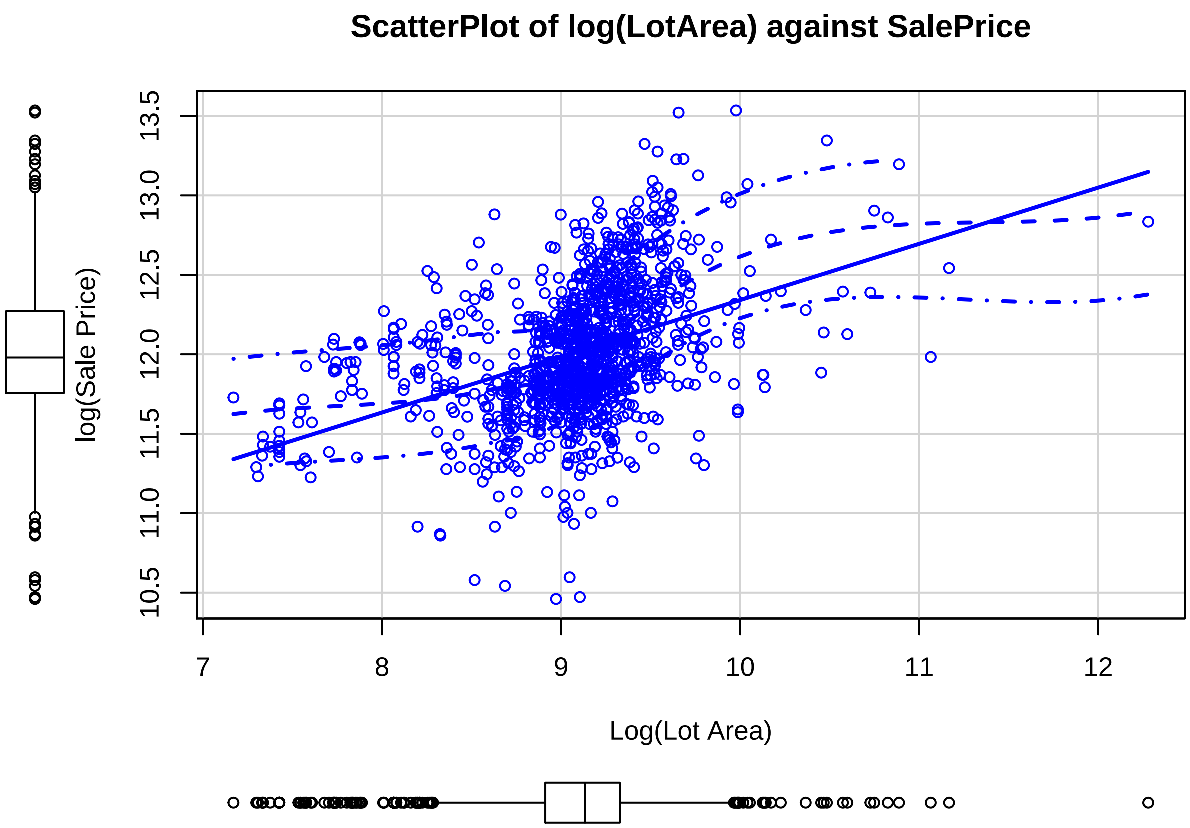


The boxplot suggests that transformation to Log should be done. The distribution seems to be symmetric. This also suggests linearity when we make scatter plot of transformed variable against sale price

```
par(mfrow = c(1,2))
scatterplot(log(SalePrice)~LotArea, data = Train,
            xlab = "(Lot Area)", ylab = "log(Sale Price)",
            main = "ScatterPlot of LotArea against SalePrice")
```



```
scatterplot(log(SalePrice)~log(LotArea), data = Train, xlab = "Log(Lot Area)",
            ylab = "log(Sale Price)",
            main = "ScatterPlot of log(LotArea) against SalePrice")
```



Linearity is shown by transformed graph.

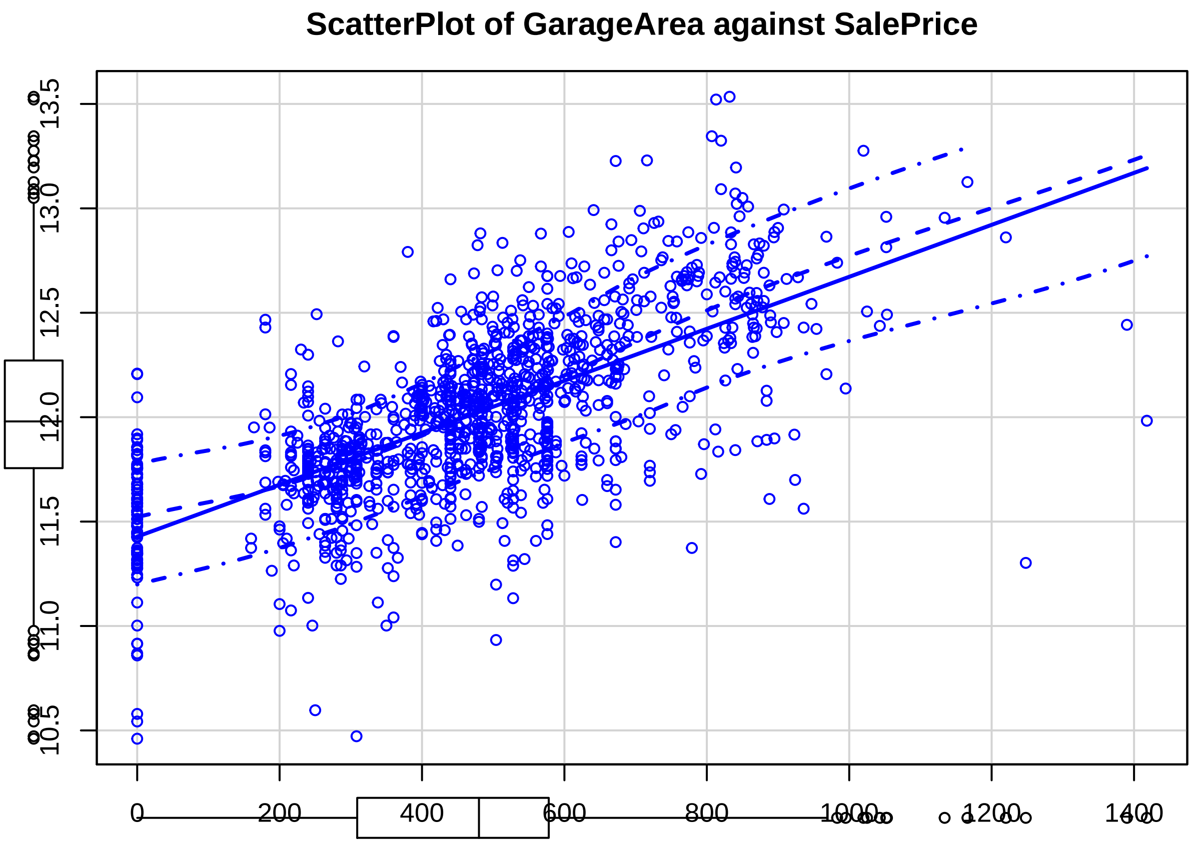
3. *Transformation of Garage Area* :

```
#Note that we add `1` to Garage Area because the minimum area is 0
p2 <- powerTransform((Train$GarageArea + 1 ) ~ 1, data = Train, family = "bcPower")
summary(p2)
```

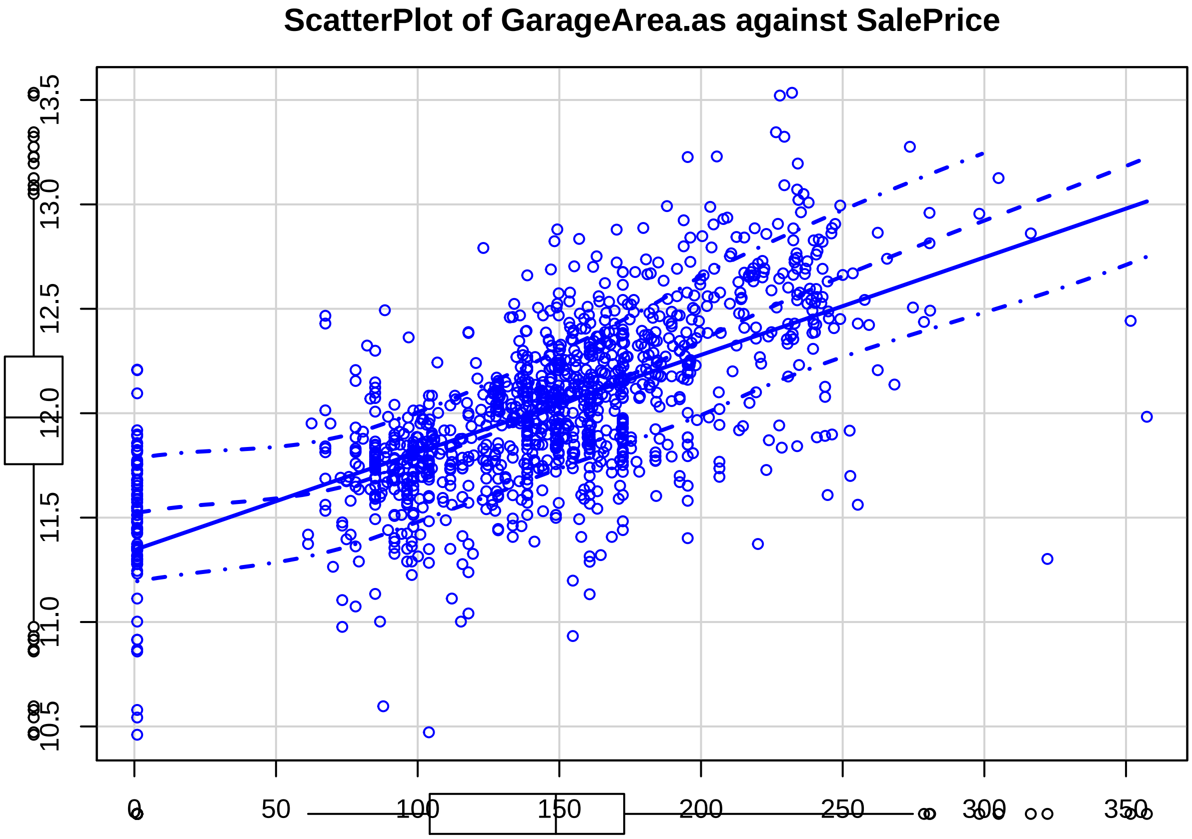
```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.7952         0.8       0.7471       0.8434
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df       pval
## LR test, lambda = (0) 1970.816  1 < 2.22e-16
##
## Likelihood ratio test that no transformation is needed
##                             LRT df       pval
## LR test, lambda = (1) 59.31765  1 1.3434e-14
```

The Box-cox suggests that we do a transformation of Garage Area. We make a scatter plot to see if we want to transform the variable or not.

```
GarageArea.New <- (Train$GarageArea + 1)^(0.81)
par(mfrow = c(1,3), mar =  c(1,1,3,2))
scatterplot(log(SalePrice)~GarageArea, data = Train, xlab = "Garage Area",
            ylab = "log(Sale Price)",
            main = "ScatterPlot of GarageArea against SalePrice")
```



```
scatterplot(log(SalePrice)~GarageArea.New, data = Train, xlab = "Garage Area",
            ylab = "log(Sale Price)",
            main = "ScatterPlot of GarageArea.as against SalePrice")
```



We decide that we aren't going to do transformation for Garage Area since the scatter plot before transformation is linear.

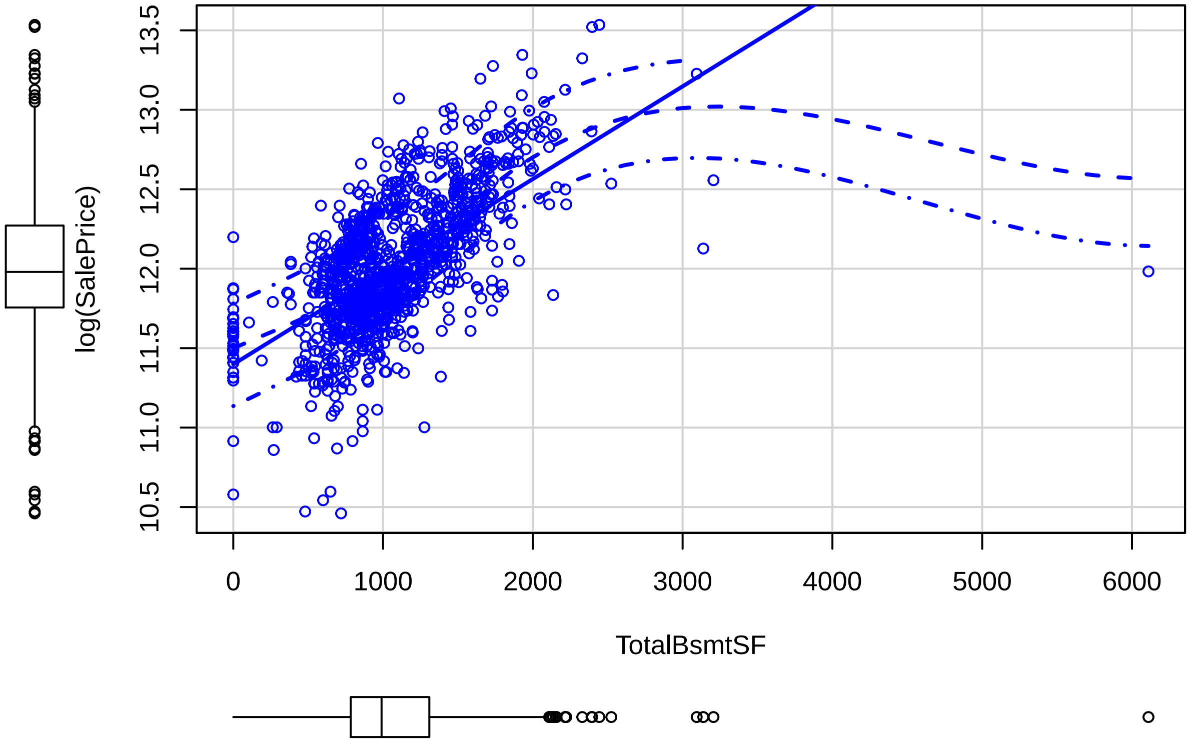
4. *Transformation of Total BasementSF* : Carrying out the Box-cox transformation:

```
p3 <- powerTransform((Train$TotalBsmtSF +1) ~ 1, data = Train, family="bcPower")
summary(p3)
```

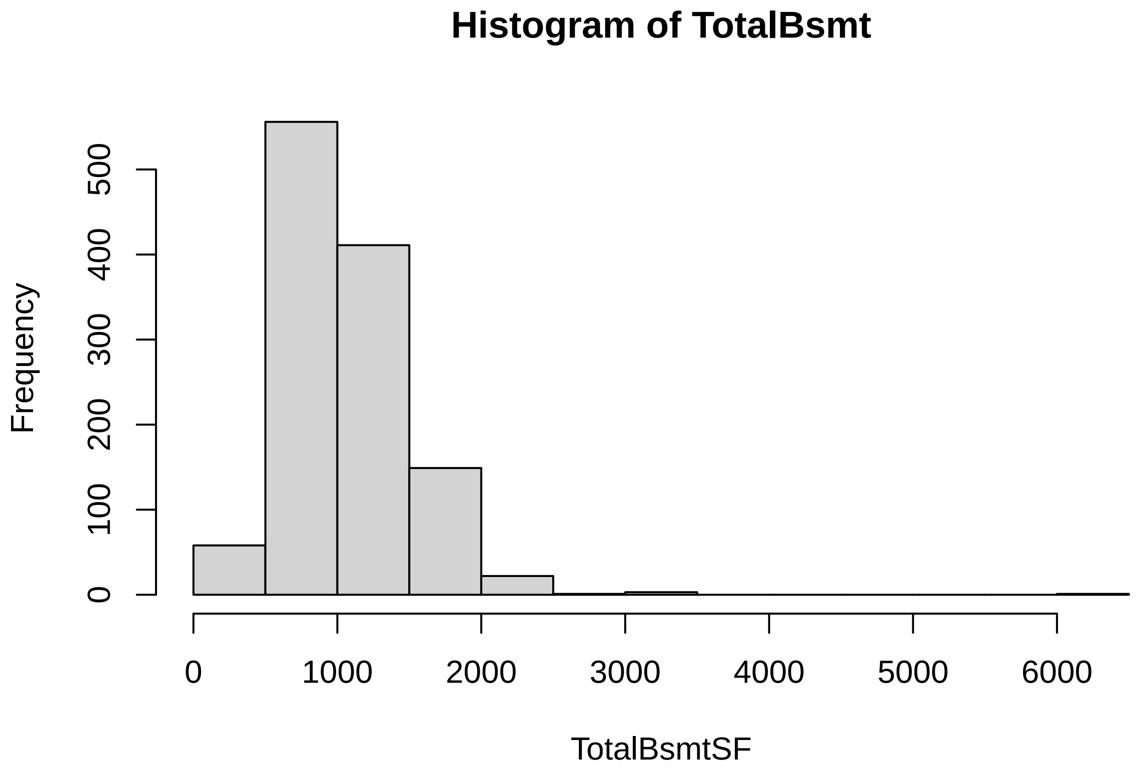
```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.7278        0.73        0.681       0.7745
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                               LRT df       pval
## LR test, lambda = (0) 1997.986  1 < 2.22e-16
##
## Likelihood ratio test that no transformation is needed
##                               LRT df       pval
## LR test, lambda = (1) 106.1039  1 < 2.22e-16
```

The Box-cox suggests us that we use do a transformation on Total BasementSF We can take our decision by making the scatter plot and histograms:

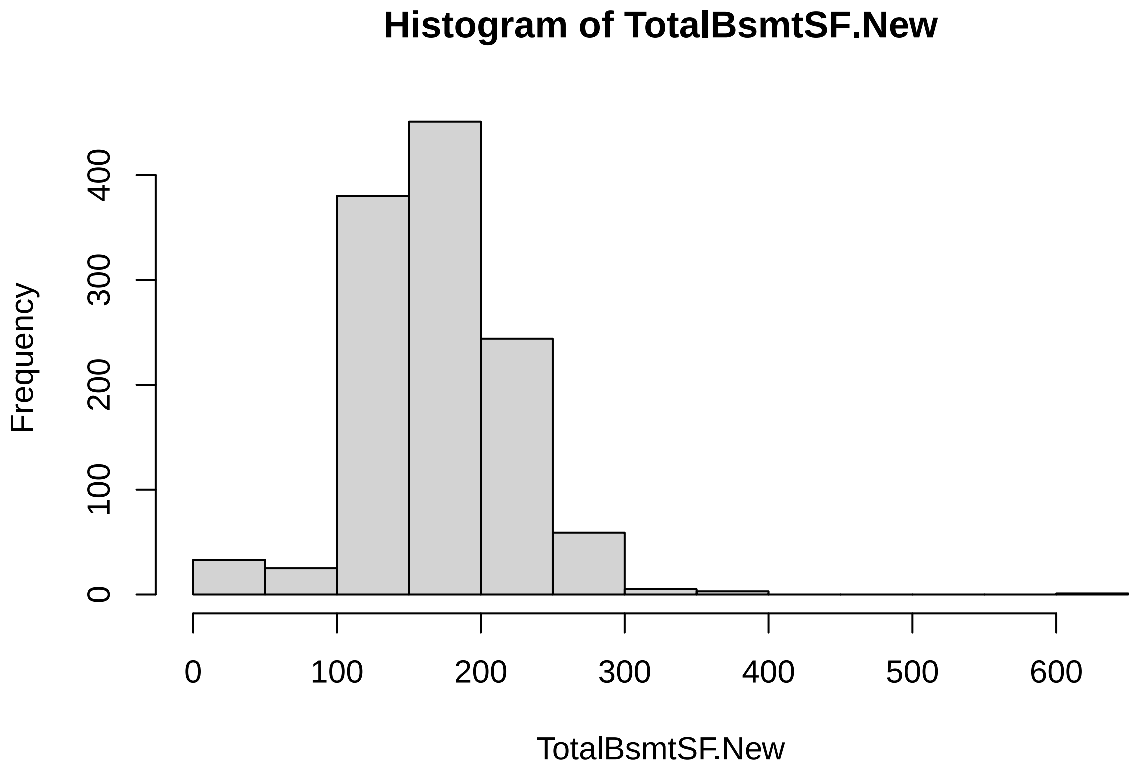
```r
TotalBsmtSF.New <- ((Train$TotalBsmtSF) + 1)^(0.74)
scatterplot(log(SalePrice)~TotalBsmtSF, data = Train)
```



```r
hist(Train$TotalBsmtSF, xlab = "TotalBsmtSF", main = "Histogram of TotalBsmt")
```



```
hist(TotalBsmtSF.New)
```



There is no difference in the distribution of Total BsmtSF and the transformed variable's distribution. Also the scatter plot for the transformed variable is not very different. So, I won't do the transformation for this.

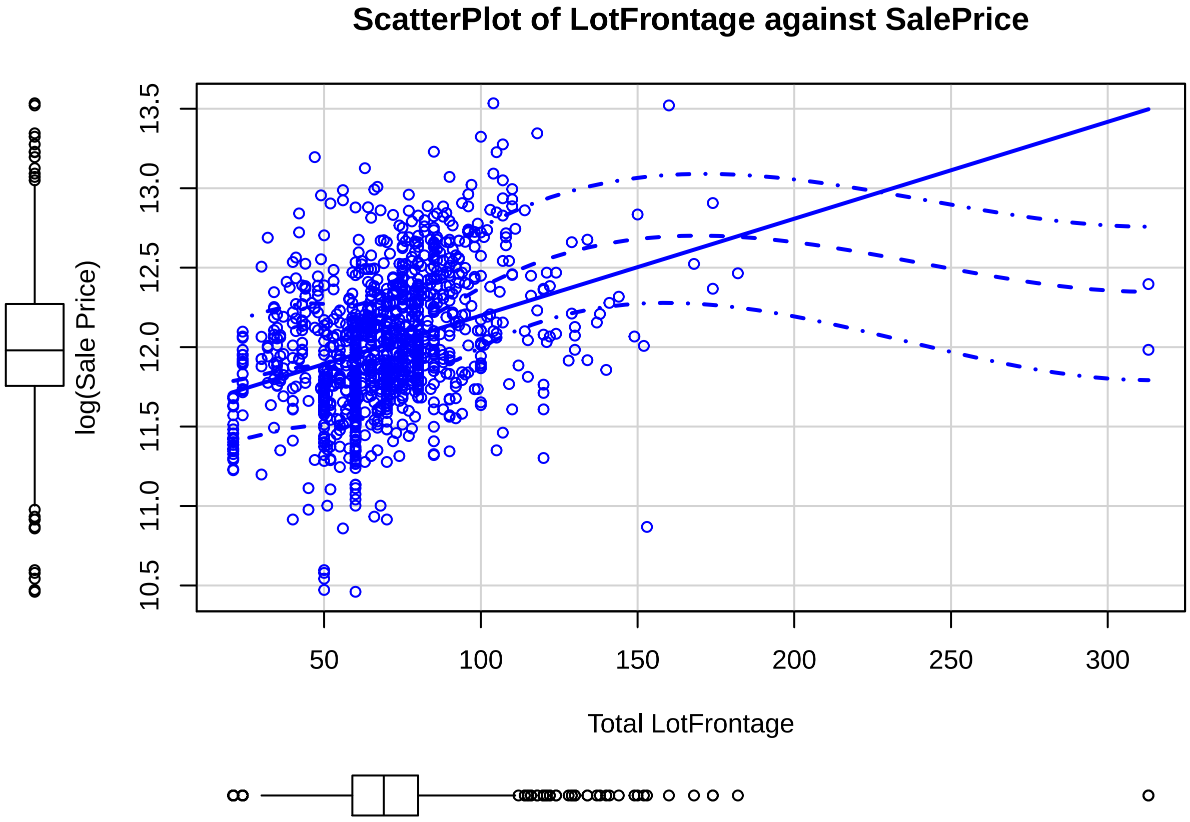
5. *Transformation of Lot Frontage* :

```
p3 <- powerTransform(Train$LotFrontage ~ 1, data = Train, family = "bcPower")
summary(p3)
```

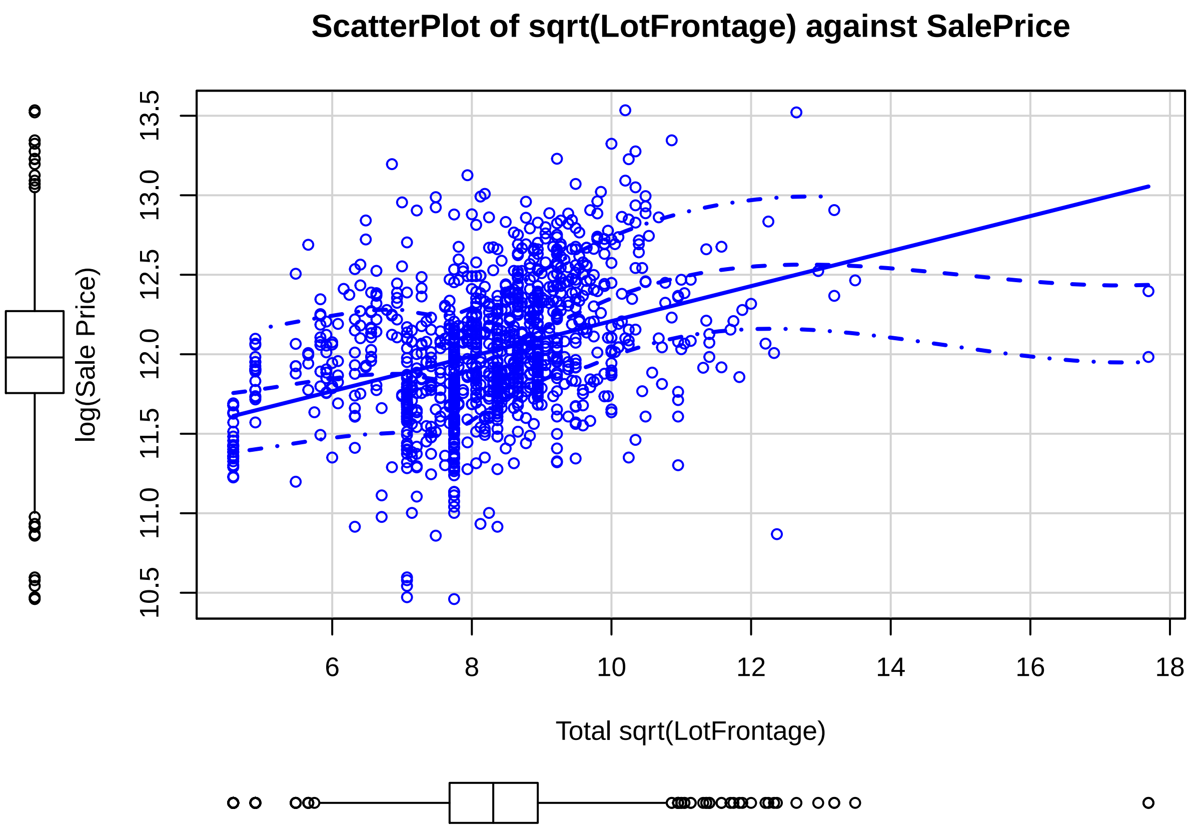
```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.4008        0.33       0.3056       0.4959
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 65.75027  1 5.5511e-16
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 163.2759  1 < 2.22e-16
```

Box-cox suggests us that we do a transformation. We do square root transformation. Let us see and compare the scatter plots to make decision.

```
par(mfrow = c(1,2))
scatterplot(log(SalePrice)~LotFrontage, data = Train, xlab = "Total LotFrontage",
            ylab = "log(Sale Price)",
            main = "ScatterPlot of LotFrontage against SalePrice")
```



```
scatterplot(log(SalePrice)~sqrt(LotFrontage), data = Train, xlab =
              "Total sqrt(LotFrontage)", ylab = "log(Sale Price)",
            main = "ScatterPlot of sqrt(LotFrontage) against SalePrice")
```



The values are concentrated towards the left in the non transformed graph. But after carrying out transformation, there is a better visual representation of the values and graph looks linear. So we go ahead with the transformation.

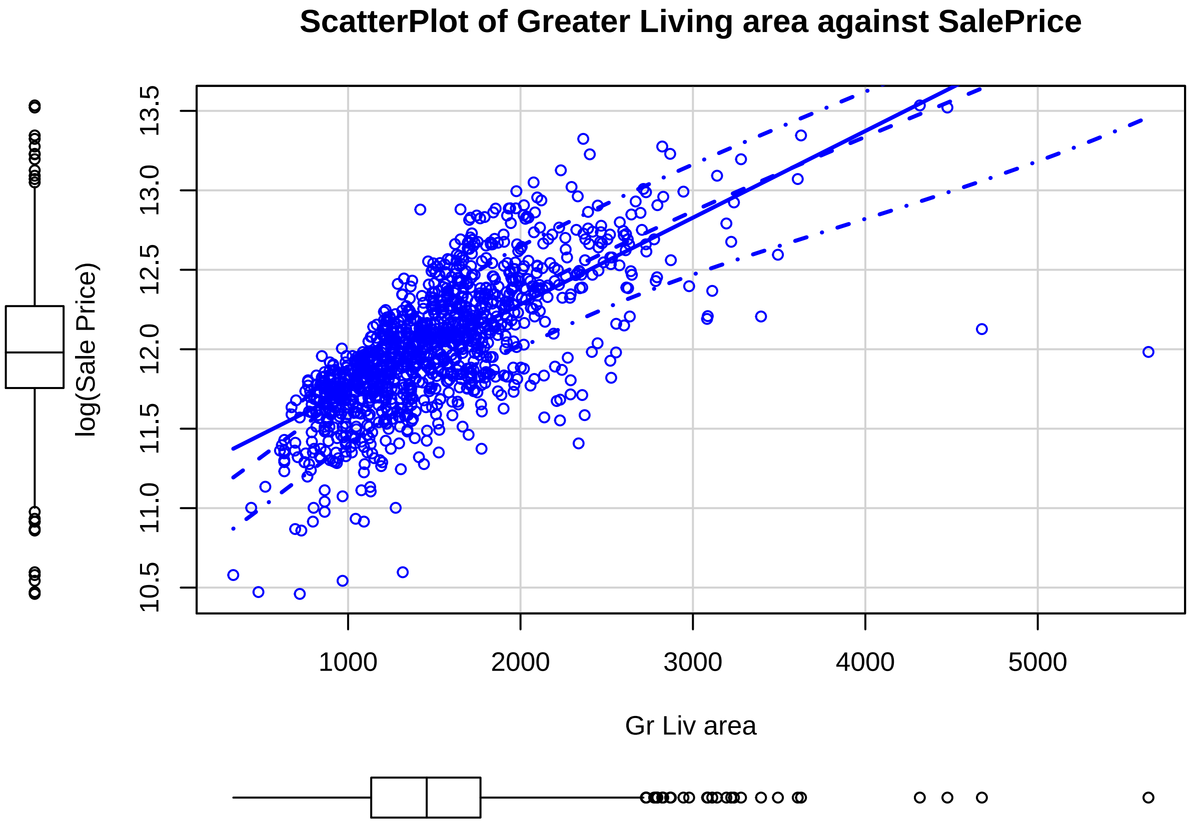
6. *Transformation of Gr Living Area* :

```
p4 <- powerTransform(GrLivArea ~ 1, data = Train, family = "bcPower")
summary(p4)
```

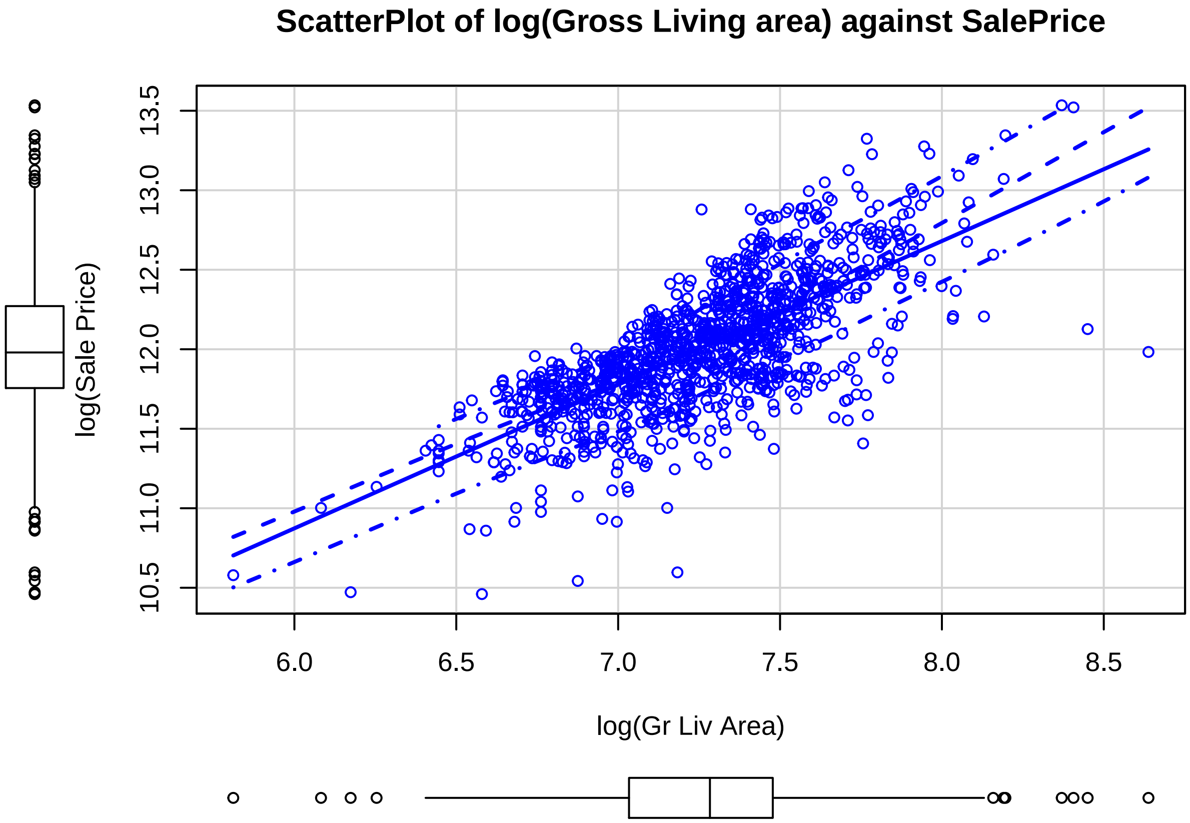
```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.0082           0      -0.1186        0.135
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                                LRT df    pval
## LR test, lambda = (0) 0.01609894  1 0.89903
##
## Likelihood ratio test that no transformation is needed
##                              LRT df       pval
## LR test, lambda = (1) 243.6167  1 < 2.22e-16
```

The box-cox suggests that we do log transformation for Gross Living Area Plotting the scatterplots to see visually if log transformation as suggested by Box cox improves linearity

```
scatterplot(log(SalePrice)~GrLivArea, data = Train, xlab = "Gr Liv area",
            ylab = "log(Sale Price)",
            main = "ScatterPlot of Greater Living area against SalePrice")
```



```
scatterplot(log(SalePrice)~log(GrLivArea), data = Train, xlab = "log(Gr Liv Area)",
            ylab = "log(Sale Price)",
            main = "ScatterPlot of log(Gross Living area) against SalePrice")
```



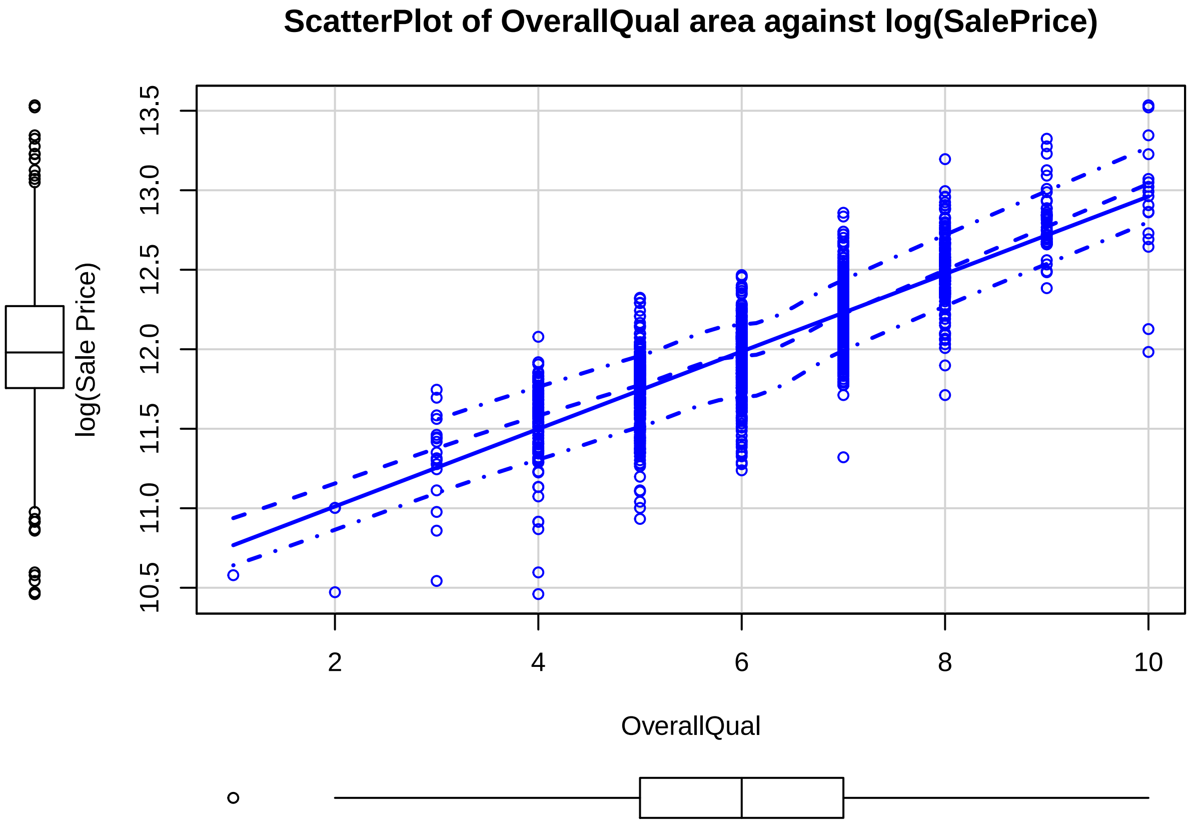
The observations are more evenly spread and the graph looks more linear, so we to the log transformation.

7. *Transformation of Overall Quality* :

```
p5<-powerTransform(OverallQual ~ 1, data=Train, family="bcPower")
summary(p5)
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.6885        0.69        0.511        0.866
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 67.53411  1 2.2204e-16
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 11.21731  1 0.00081038
```

```
scatterplot(log(SalePrice)~OverallQual, data = Train, xlab = "OverallQual",
            ylab = "log(Sale Price)",
            main = "ScatterPlot of OverallQual area against log(SalePrice)")
```



The Box cox suggests us that we do transformation on Overall Quality to the power 0.69. But since the ratings are already linear, we do not do any transformations.

As for Year Built and YearRemodadd, although numerical, it does not make sense to do a transformation, since that would only change the intercept. The remaining variables, are categorical and hence no transformation will be done.

(c) Estimate a multiple linear regression model that includes all the main effects only (i.e.,no interactions nor higher order terms). We will use this model as a baseline. Comment on the statistical and economic significance of your estimates. Also, make sure to provide an interpretation of your estimates. Note: You can use any combination of transformed and untransformed variables from the model in part (b).

```
Model1.1 <- lm(log(SalePrice) ~ log(LotArea) + GarageArea + sqrt(LotFrontage)+
                 OverallQual +BldgType +HouseStyle, data = Train)

summary(Model1.1)
```

```
##
## Call:
## lm(formula = log(SalePrice) ~ log(LotArea) + GarageArea + sqrt(LotFrontage) +
##     log(GrLivArea) + TotalBsmtSF + YearBuilt + YearRemodAdd +
##     OverallQual + BldgType + HouseStyle, data = Train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67953 -0.07474  0.01207  0.09104  0.49771
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.614e+00  6.306e-01  -2.559 0.010633 *
## log(LotArea)        1.163e-01  1.613e-02   7.213 9.73e-13 ***
## GarageArea          1.786e-04  2.894e-05   6.171 9.33e-10 ***
## sqrt(LotFrontage) -1.210e-02  5.196e-03  -2.329 0.020052 *
## log(GrLivArea)      4.852e-01  3.012e-02  16.106  < 2e-16 ***
## TotalBsmtSF         4.597e-05  1.763e-05   2.608 0.009228 **
## YearBuilt           2.029e-03  2.375e-04   8.544  < 2e-16 ***
## YearRemodAdd        2.216e-03  2.963e-04   7.478 1.46e-13 ***
## OverallQual         9.548e-02  5.923e-03  16.121  < 2e-16 ***
## BldgType2fmCon     -5.440e-02  3.194e-02  -1.703 0.088781 .
## BldgTypeDuplex     -1.342e-01  2.644e-02  -5.076 4.48e-07 ***
## BldgTypeTwnhs      -5.215e-02  3.523e-02  -1.480 0.139018
## BldgTypeTwnhsE      3.647e-04  2.323e-02   0.016 0.987473
## HouseStyle1.5Unf    9.146e-02  4.918e-02   1.860 0.063199 .
## HouseStyle1Story    7.054e-02  2.002e-02   3.523 0.000442 ***
## HouseStyle2.5Fin   -1.710e-02  6.131e-02  -0.279 0.780316
## HouseStyle2.5Unf   -9.714e-02  5.411e-02  -1.795 0.072847 .
## HouseStyle2Story   -9.369e-04  1.843e-02  -0.051 0.959475
## HouseStyleSFoyer    1.543e-01  3.751e-02   4.113 4.18e-05 ***
## HouseStyleSLvl      7.798e-02  2.956e-02   2.638 0.008449 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1616 on 1181 degrees of freedom
## Multiple R-squared:  0.8515, Adjusted R-squared:  0.8491
## F-statistic: 356.3 on 19 and 1181 DF,  p-value: < 2.2e-16
```

```
AIC(Model1.1)
```

```
## [1] -947.3456
```

```
BIC(Model1.1)
```

```
## [1] -840.4365
```

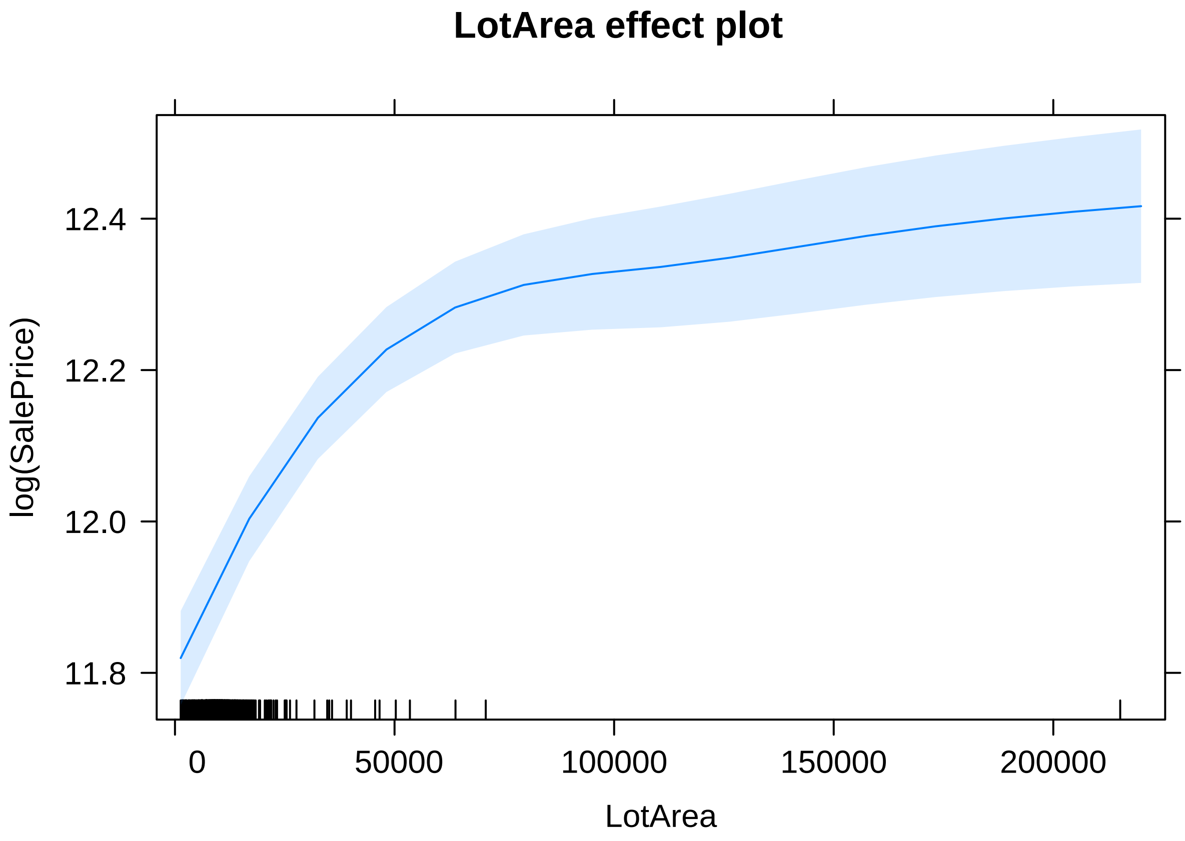
My R^2 is 0.8515 which means that 85.15% of variation in log(SalePrice) is explained by explanatory variables. Economically, the variables :log(LotArea),GarageArea,log(GrLivArea),TotalBsmtSF,YearBuilt, YearRemodAdd and OverallQual have a positive coefficient. This makes intutive sense. With increase in area, Sale price tends to increase. A better quality house tends to have a higher Sale price. As far as year variables are considered, with new houses or recently renovated houses, the Price would be higher.

We can look at the marginal effects plots to understand the relationships visually.

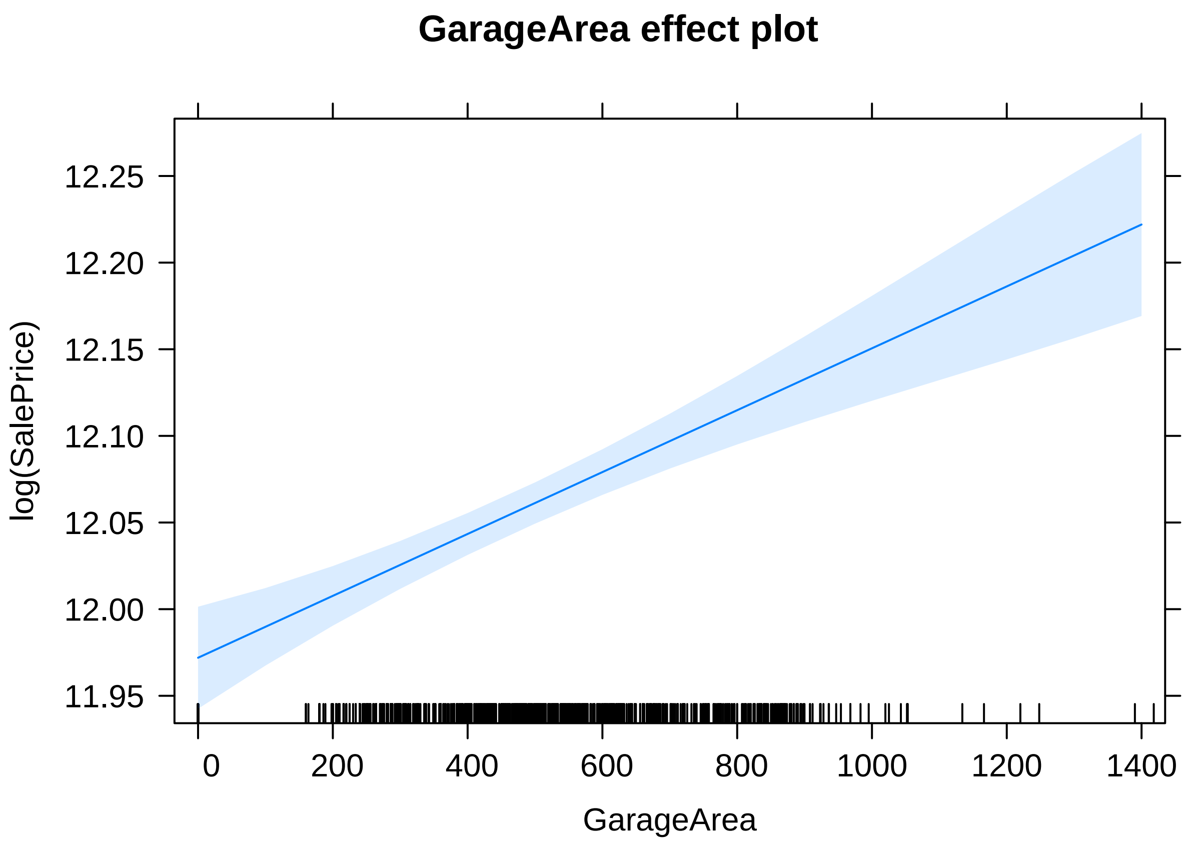
```
library(effects)
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
## Use the command
##     lattice::trellis.par.set(effectsTheme())
##   to customize lattice options for effects plots.
## See ?effectsTheme for details.
```

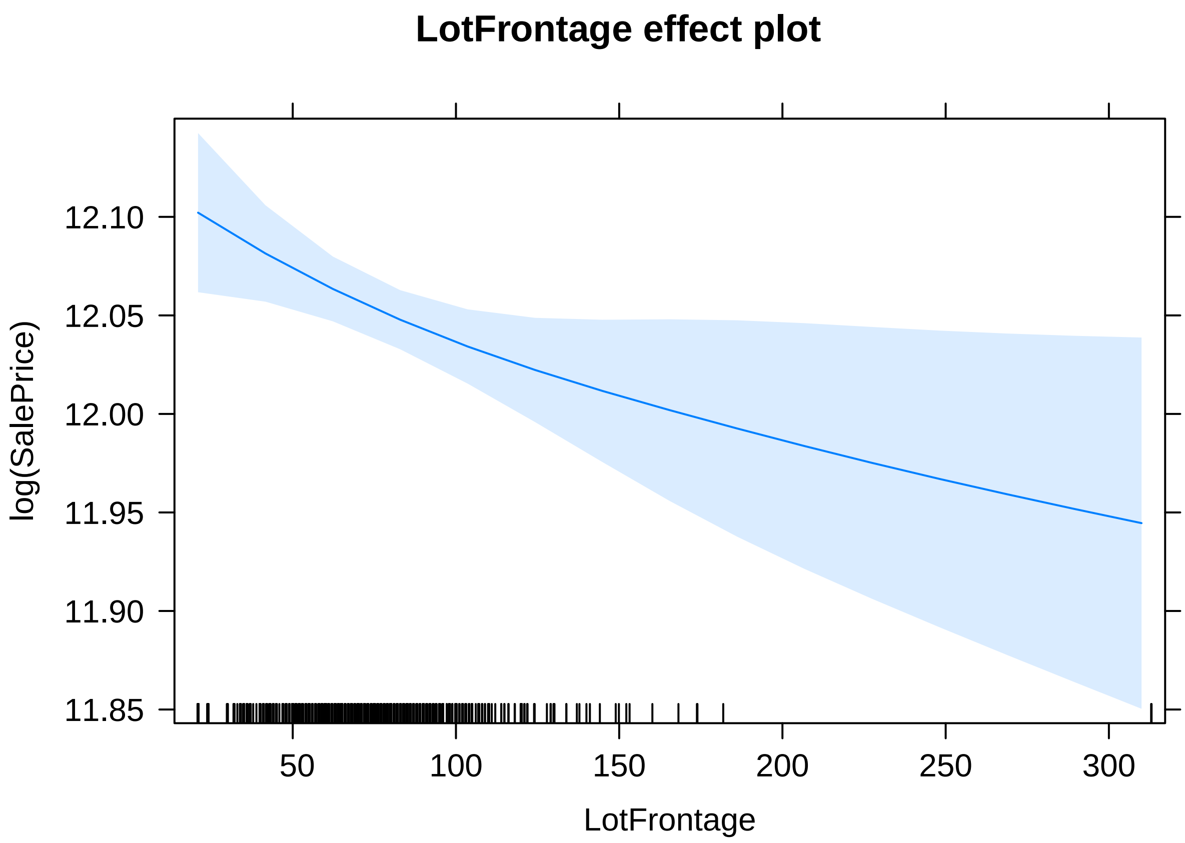
```
#plot(allEffects(mod = Model1.1))
plot(effect(mod=Model1.1,"log(LotArea)"))
```



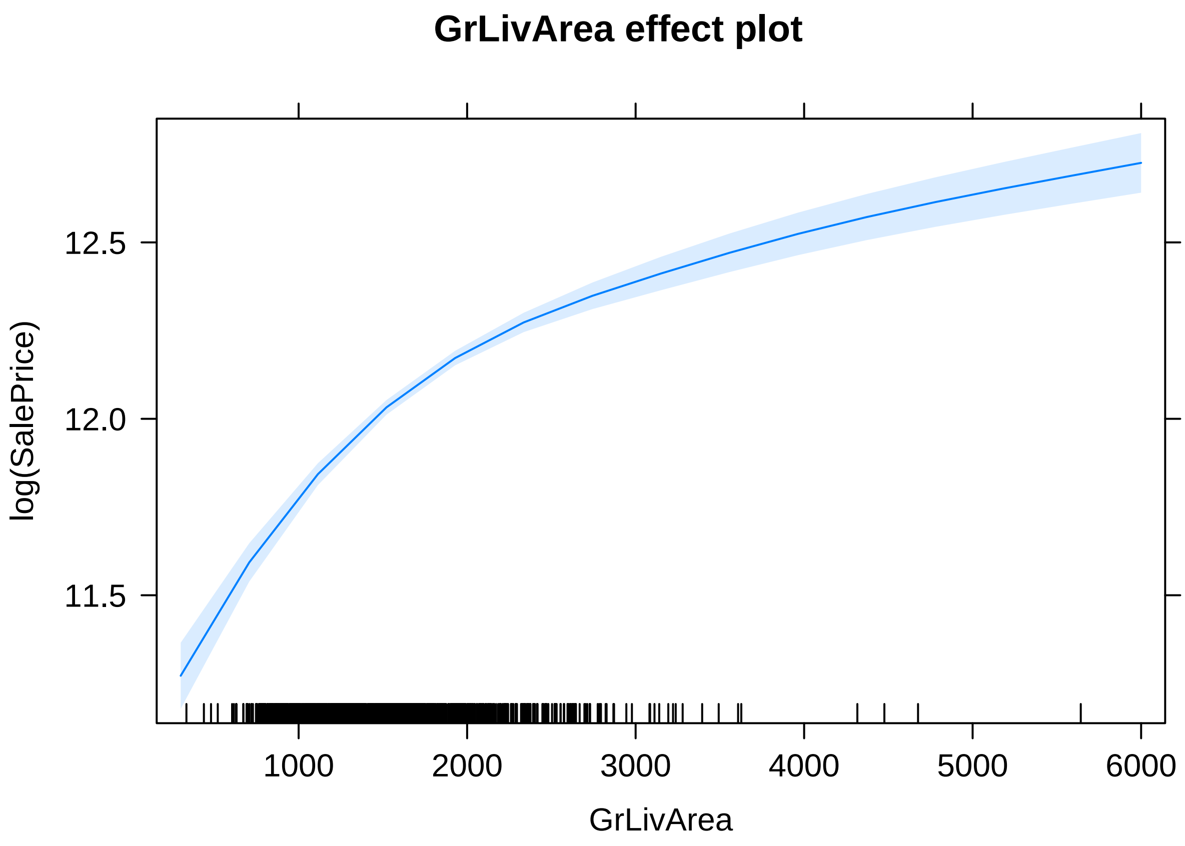
```
plot(effect(mod=Model1.1,"GarageArea"))
```



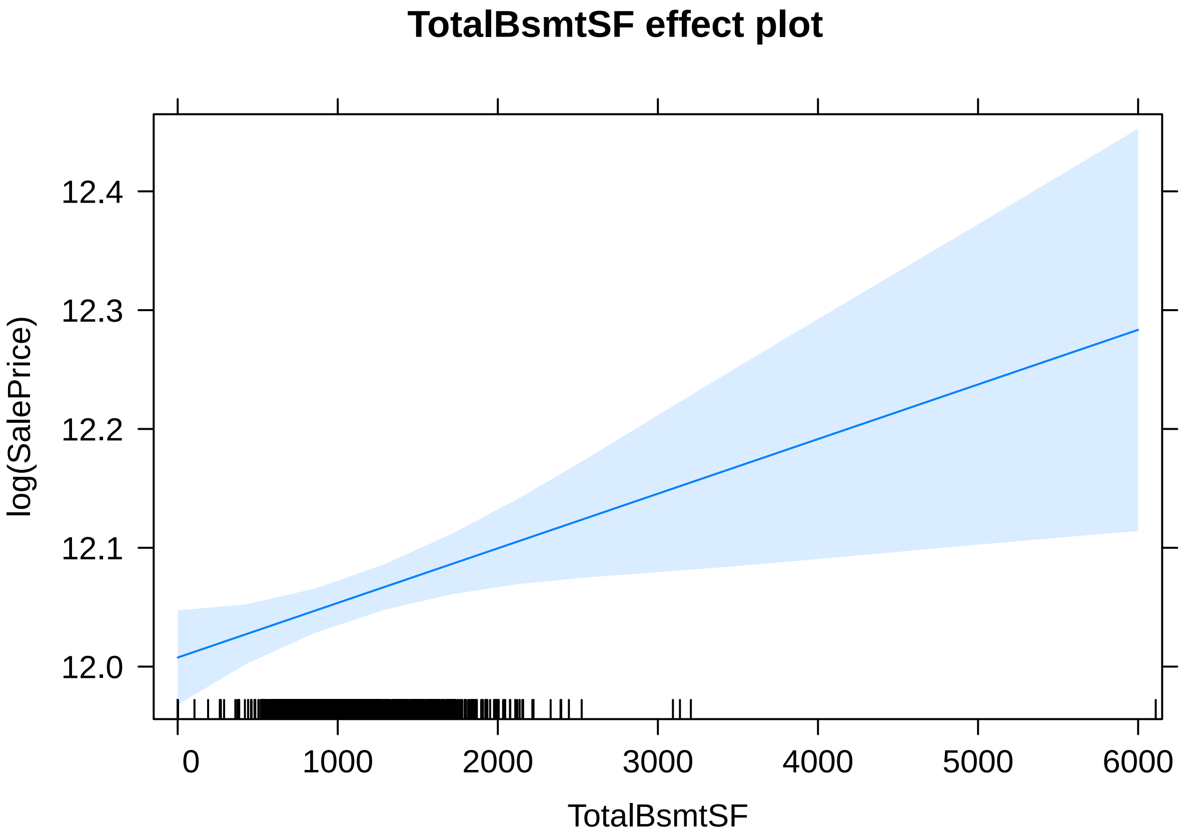
```
plot(effect(mod=Model1.1,"sqrt(LotFrontage)"))
```



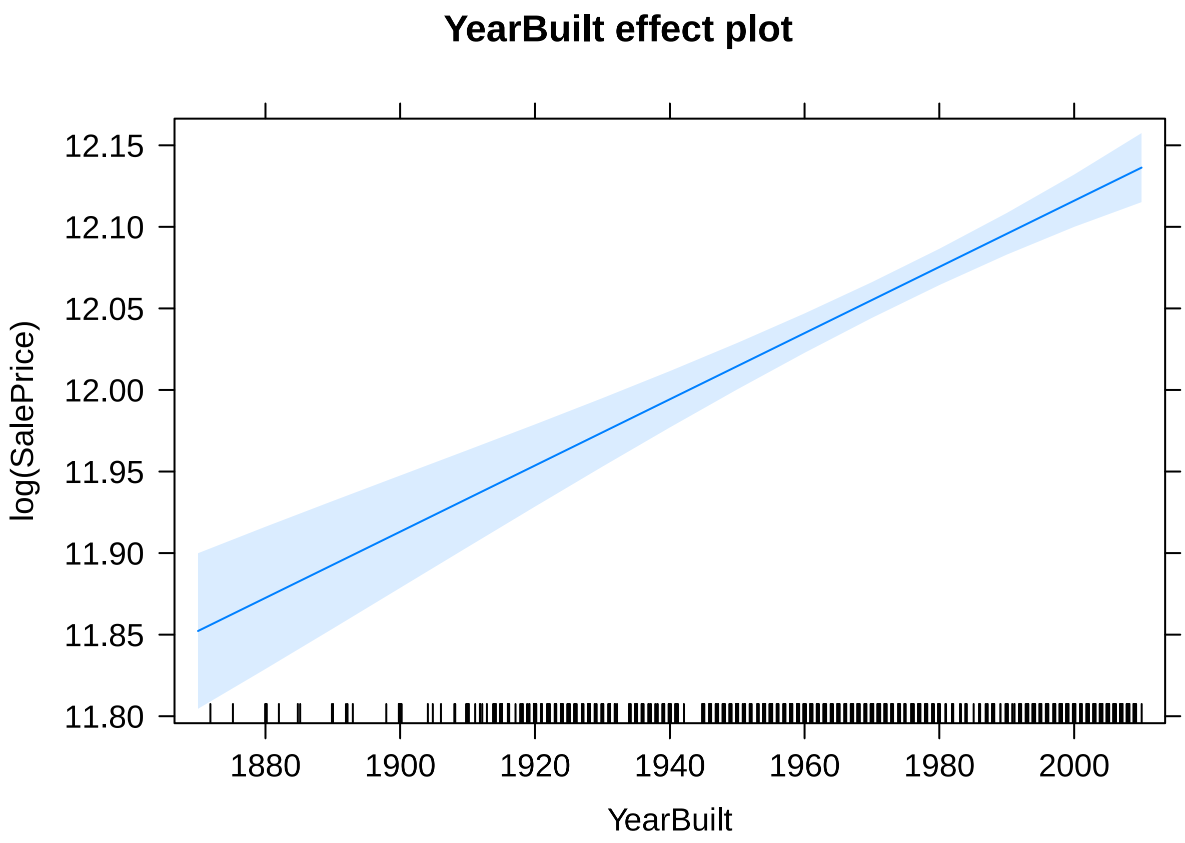
```
plot(effect(mod=Model1.1,"log(GrLivArea)"))
```



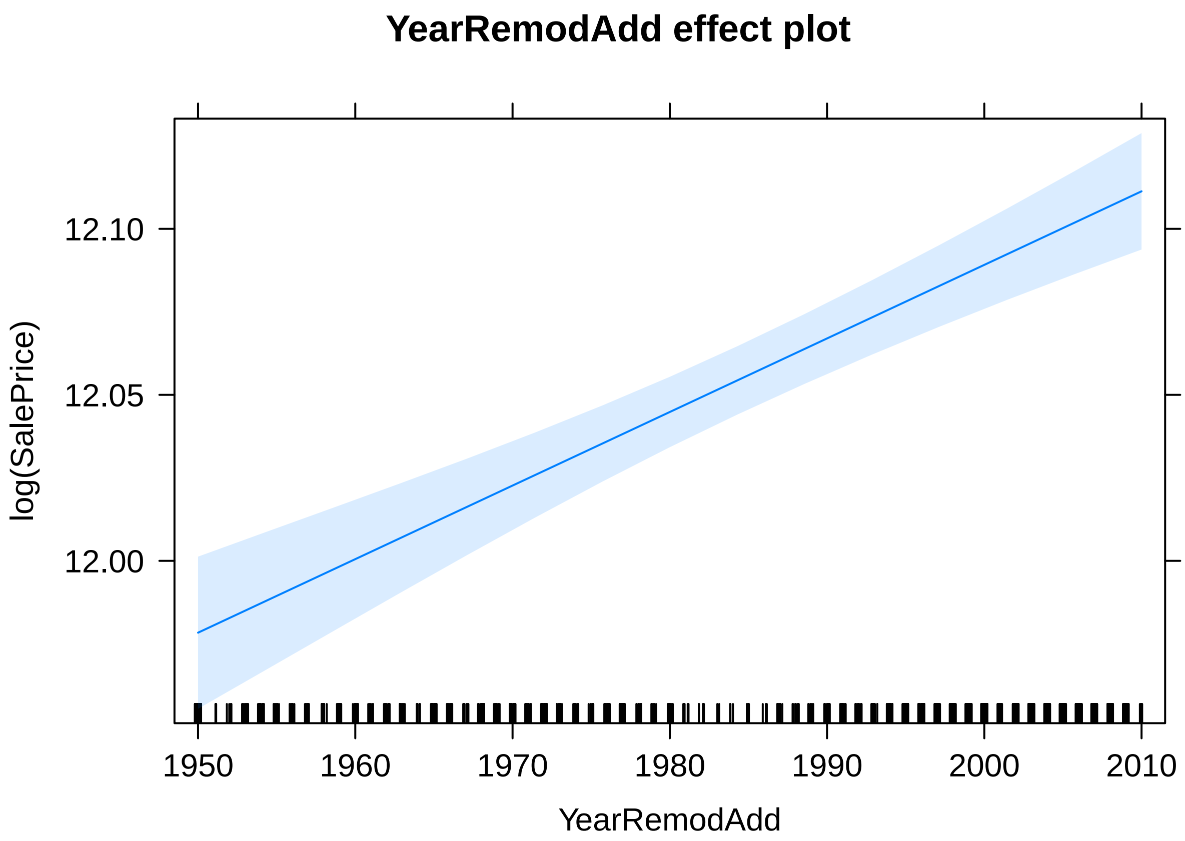
```
plot(effect(mod=Model1.1,"TotalBsmtSF"))
```



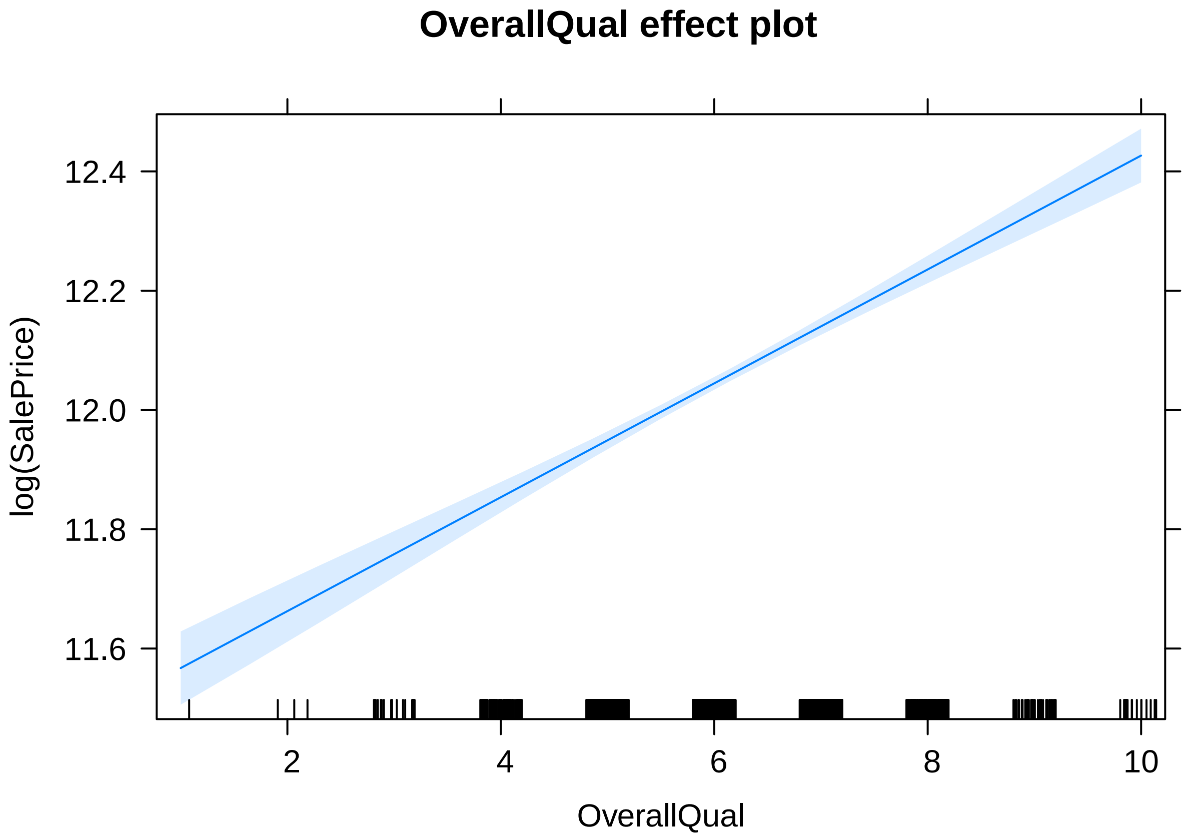
```r
plot(effect(mod=Model1.1,"YearBuilt"))
```



```
plot(effect(mod=Model1.1,"YearRemodAdd"))
```



```r
plot(effect(mod=Model1.1,"OverallQual"))
```



As we can see, the trends match our intuitive (economic reasoning). While garage area, year built and yearremoadd exhibit a linear trend, we see a concave downward graph for Lotarea and GrLiving Area.(Increasing at a decreasing rate). LotFrontage has a downward sloping linear graph. This suggests that with Increase in LotFrontage there is decrease in Sale Price.

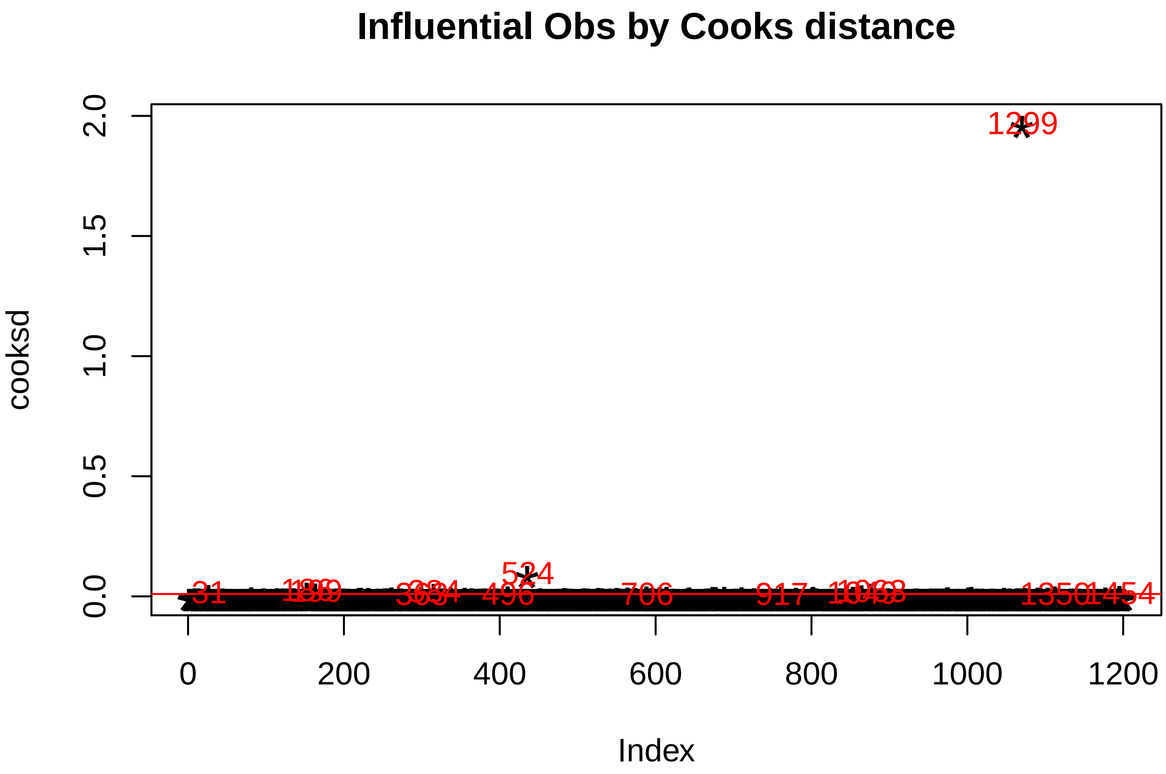
Now, I am going to comment statistical significance and Interpretation: The variables log(LotArea),GarageArea, log(GrLivArea) , YearBuilt, Yearremodadd, OverallQual,BldgTypeDuplex, HouseStyle1Story, HouseStyleS-Foyer are statistically significant at the 99% confidence level which we can see from the p-values. LotFrontage is significant at 5 % level. As $LotArea$ is increased by 1 %, then average value of sale price of house is increased by 1.163e-01 % (0.1163%) keeping other factors constant. Here the slope coefficient represents the elasticity. If the $Garage\ area$ increases by 1 sqft, then average value of sale price increases by $(1.786e-04 \times 100)\%$ or (0.1786%), holding all other variables constant. As $GrLivArea$ increases by 1 %, saleprice increases by 4.852e-01% or (0.48%) keeping other factors being constant. Here the slope coefficient represents the elasticity. $YearBuilt$: If the original construction year is one more recent year, then, on an average, with everything else constant, the sale price of the house is expected to increase by $(2.029e-03 \times 100)\%$ or (0.2029%) $YearRemodAdd$: If the remodel year is one year more recent, then, the Sale price of the house is expected to increase by $(2.216e-03 \times 100)\%$ or (0.22%) on average with everything else fixed. Coefficients of factors: The benchmark category for $Building\ type$ is $Single\ family\ detached$. So, the coefficient of $BldgType2fmCon$ is the differential intercept coefficient. It tells the difference in the sale price between Single Family detached building type and two family conversion building type keeping other factors constant which is $(-5.440e-02 \times 100)\%$ Likewise, the coefficient of $BldgType\ Duplex$ represents the difference in Sale Price between Single family detached and Duplex keeping other factors constant which is $(-1.342e-01 \times 100)\%$. The benchmark category for $House\ Style$ is $One\ and\ one-halfstory : 2nd\ level\ finished$. The coefficient of $Housestyle\ Split\ Foyer$ represents the difference in Sale price of of 1.5 story, second level finished and Split Foyer which is $(1.543e-01 \times 100)\%$ with everything else fixed. Likewise, everything fixed, the difference between sale price for 1.5 story 2nd level finished house style and $One\ and\ one-half\ story : 2nd\ level\ unfinished$

is $(9.146e-02 \times 100)\%$ keeping other factors fixed. The difference between $one\ story$ and 1.5story, 2nd level finished house style is $(7.054e-02 \times 100)\%$ keeping other factors constant. Coefficient of $Two\ and\ one-half\ story : 2nd\ level\ Unfinished$ tells us sale price difference b/w 2.5 story 2nd level unfinished and 1.5 story, 2nd level finished is $(-9.714e-02 \times 100)\%$, keeping other factors constant.

(d) In your model from part (c), identify if there are any outliers worth removing. If so, remove them but justify your reason for doing so and re-estimate your model from part(c).

We use Cook distance to find outliers. If the Cook distance of a value is 4 times larger than the average distance, we can consider this point as an outlier. We plot and locate outliers by showing cook distances. Note that I tried out the outliers test and also plotted the respective residuals vs y. Out of all these methods, I decided to remove the outliers listed below as suggested by Cook's distance since removing those improved my overall model.

```
#Cook's Distance:
cooksd <- cooks.distance(Model1.1)
plot(cooksd, pch="*", cex=2,
main="Influential Obs by Cooks distance")
abline(h = 4*mean(cooksd, na.rm=T), col="red")
text(x=1:length(cooksd)+1, y=cooksd,
     labels=ifelse(cooksd>4*mean(cooksd, na.rm=T),names(cooksd),""),
     col="red")
```



```
which(cooksd>4*mean(cooksd, na.rm=T))
```

```
##   31  186  199  363  384  496  524  706  917 1049 1063 1299 1350 1454 
##   26  152  163  299  315  410  435  588  761  864  876 1070 1112 1195
```

From the graph, we can observe that the outliers are located in row number 31,186, 199, 363 , 384, 496, 524 , 706 , 917, 1049, 1063, 1299, 1350, 1454, 26,152, 163, 299 , 315, 410 , 435, 588 , 761 , 864, 876, 1070, 1112 ,1195 from the Train data set. So, we update the model by the following code:

```
Model1.2 = update(Model1.1, subset = -c(31,  186,  199,  363 , 384,  496,  524 ,
                                        706 , 917, 1049, 1063, 1299, 1350, 1454,
                                        26,  152,  163,  299 , 315,  410 , 435,
                                        588 , 761 , 864,  876, 1070, 1112 ,1195))
summary(Model1.2)
```

```
## 
## Call:
## lm(formula = log(SalePrice) ~ log(LotArea) + GarageArea + sqrt(LotFrontage) + 
##     log(GrLivArea) + TotalBsmtSF + YearBuilt + YearRemodAdd + 
##     OverallQual + BldgType + HouseStyle, data = Train, subset = -c(31, 
##     186, 199, 363, 384, 496, 524, 706, 917, 1049, 1063, 1299, 
##     1350, 1454, 26, 152, 163, 299, 315, 410, 435, 588, 761, 864, 
##     876, 1070, 1112, 1195))
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.86721 -0.07624  0.01414  0.08150  0.52481 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       -8.250e-01  5.537e-01  -1.490 0.136532    
## log(LotArea)       1.308e-01  1.413e-02   9.257  < 2e-16 ***
## GarageArea         1.618e-04  2.519e-05   6.425 1.92e-10 ***
## sqrt(LotFrontage) -5.784e-04  4.567e-03  -0.127 0.899253    
## log(GrLivArea)     4.096e-01  2.735e-02  14.979  < 2e-16 ***
## TotalBsmtSF        1.568e-04  1.738e-05   9.018  < 2e-16 ***
## YearBuilt          1.690e-03  2.108e-04   8.019 2.60e-15 ***
## YearRemodAdd       2.301e-03  2.570e-04   8.952  < 2e-16 ***
## OverallQual        8.516e-02  5.217e-03  16.325  < 2e-16 ***
## BldgType2fmCon    -4.227e-02  2.835e-02  -1.491 0.136259    
## BldgTypeDuplex    -1.242e-01  2.283e-02  -5.442 6.44e-08 ***
## BldgTypeTwnhs     -8.862e-03  3.085e-02  -0.287 0.773950    
## BldgTypeTwnhsE     3.916e-02  2.010e-02   1.948 0.051625 .  
## HouseStyle1.5Unf   9.051e-02  4.368e-02   2.072 0.038452 *  
## HouseStyle1Story   2.813e-02  1.794e-02   1.568 0.117077    
## HouseStyle2.5Fin  -3.818e-03  5.987e-02  -0.064 0.949161    
## HouseStyle2.5Unf  -8.498e-02  4.649e-02  -1.828 0.067792 .  
## HouseStyle2Story   3.508e-02  1.615e-02   2.173 0.030018 *  
## HouseStyleSFoyer   1.287e-01  3.337e-02   3.857 0.000121 ***
## HouseStyleSLvl     6.664e-02  2.550e-02   2.614 0.009077 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1384 on 1156 degrees of freedom
## Multiple R-squared:  0.8864, Adjusted R-squared:  0.8845
## F-statistic: 474.7 on 19 and 1156 DF,  p-value: < 2.2e-16
```

```
AIC(Model1.2)
```

```
## [1] -1291.78
```

```
BIC(Model1.2)
```

```
## [1] -1185.313
```

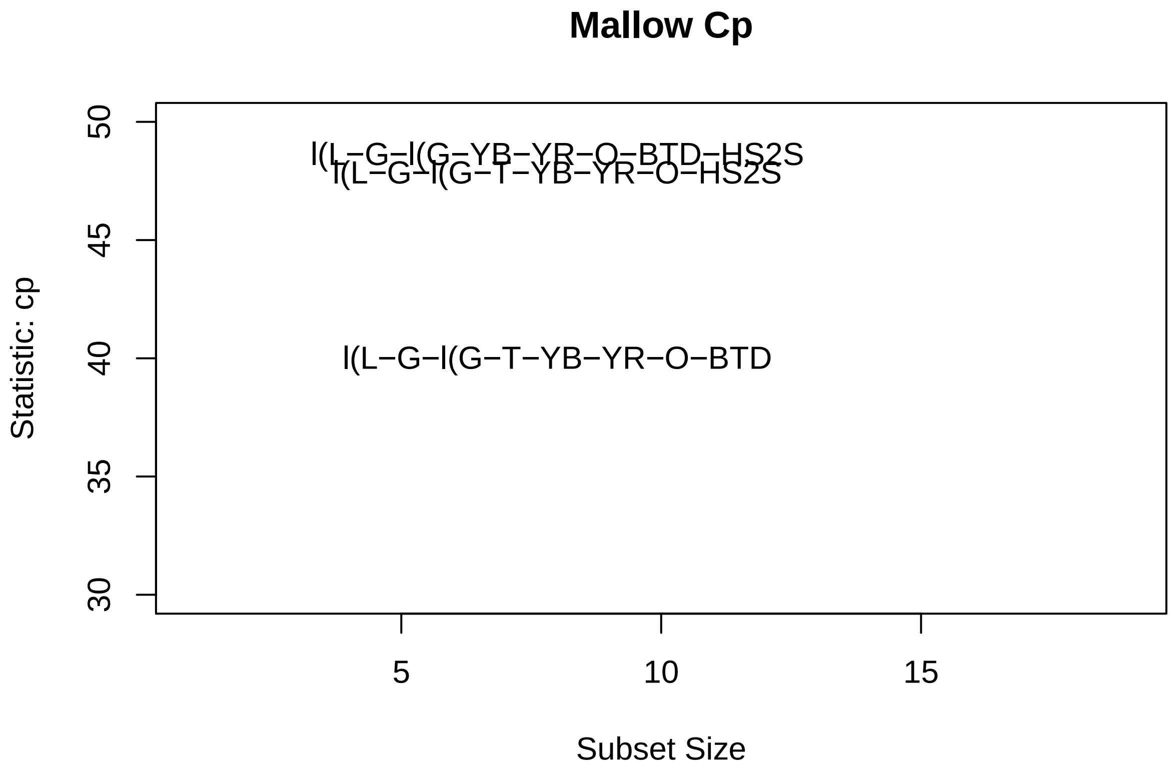
As we can see Model1.2 (Model after removal of outliers) is better than Model1.1 (Model with outliers). Our R^2 has improved from 0.8515 to 0.8864. This suggest that the model is a better fit. The adjusted R^2 has improved form 0.8491 to 0.8864. The AIC and BIC also suggest that Model1.2 is better than Model1.1. Outliers, in my opinion should be removed since they can have large influence on our estimates. Also, the coefficients for variables get significant at higher levels after removing the outliers.

(e) Use Mallows Cp for identifying which terms you will keep from the model in part (d) and also test for multi-collinearity. Based on your findings estimate a new model.

Mallows' Cp Criterion is a way to assess the fit of a multiple regression model. The technique then compares the full model with a smaller model with “p” parameters and determines how much error is left unexplained by the partial model. Or, more specifically, it estimates the standardized total mean square of estimation for the partial model with the formula (Hocking, 1976).

First, I will carry out the Mallow's CP test as follows:

```
library(broom)
library(leaps)
library(car)
ss = regsubsets(log(SalePrice) ~ log(LotArea)+ GarageArea +
                  sqrt(LotFrontage) + log(GrLivArea)+TotalBsmtSF +
                  YearBuilt + YearRemodAdd +  OverallQual +BldgType +
                  HouseStyle, method = c("exhaustive"),
                  nbest = 3, data = Train)
subsets(ss,statistic="cp", ylim = c(30,50) , legend = F,
        main = "Mallow Cp",
        col = "steelblue4")
```



```
##                     Abbreviation
## log(LotArea)                 l(L
## GarageArea                     G
## sqrt(LotFrontage)              s
## log(GrLivArea)               l(G
## TotalBsmtSF                    T
## YearBuilt                     YB
## YearRemodAdd                  YR
## OverallQual                    O
## BldgType2fmCon               BT2
## BldgTypeDuplex               BTD
## BldgTypeTwnhs               BlTT
## BldgTypeTwnhsE              BTTE
## HouseStyle1.5Unf            HS1.
## HouseStyle1Story            HS1S
## HouseStyle2.5Fin          HS2.5F
## HouseStyle2.5Unf          HS2.5U
## HouseStyle2Story            HS2S
## HouseStyleSFoyer            HSSF
## HouseStyleSLvl              HSSL
```

The Mallows's CP suggests us to keep only the following terms in the model: `log(LotArea)`, `GarageArea`, `log(GrLivArea)`, `TotalBsmtSF`, `YearBuilt`, `YearRemodAdd`,`OverallQual` and `Buildingtype_duplex`

Creating an new regression model as per Mallow's Cp:

```
#Note that we have to create a new vector Blgdtype_Duplex and add it
#to our data frame Train as follows:
#(It is one of the levels of BlgdType categorical variable)

Train = Train %>%
  mutate(BlgdType_Duplex = ifelse(BldgType == "Duplex", 1, 0))
Train[,"BlgdType_Duplex"] <- as.factor(Train[,"BlgdType_Duplex"])

Model1.3 <- lm(log(SalePrice) ~ log(LotArea)+ GarageArea  + log(GrLivArea)+
                  TotalBsmtSF + YearBuilt + YearRemodAdd +  OverallQual  +
                  BlgdType_Duplex, data = Train, subset = -c(31 , 186 , 199,
                                                             363, 384 , 496 ,
                                                             524  ,706, 917,
                                                             1049 ,1063, 1299,
                                                             1350 ,1454 ,26 ,
                                                             152 , 163, 299 ,
                                                             315 , 410, 435,
                                                             588 , 761, 864 ,
                                                             876, 1070 ,1112,1195))
summary(Model1.3)
```

```
##
## Call:
## lm(formula = log(SalePrice) ~ log(LotArea) + GarageArea + log(GrLivArea) +
##     TotalBsmtSF + YearBuilt + YearRemodAdd + OverallQual + BlgdType_Duplex,
##     data = Train, subset = -c(31, 186, 199, 363, 384, 496, 524,
##         706, 917, 1049, 1063, 1299, 1350, 1454, 26, 152, 163,
##         299, 315, 410, 435, 588, 761, 864, 876, 1070, 1112, 1195))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87180 -0.07449  0.01191  0.08288  0.50575
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.187e+00  5.238e-01  -2.266   0.0236 *
## log(LotArea)      1.210e-01  9.626e-03  12.573  < 2e-16 ***
## GarageArea        1.691e-04  2.521e-05   6.709 3.05e-11 ***
## log(GrLivArea)    3.838e-01  1.759e-02  21.811  < 2e-16 ***
## TotalBsmtSF       1.555e-04  1.283e-05  12.122  < 2e-16 ***
## YearBuilt         2.013e-03  1.930e-04  10.435  < 2e-16 ***
## YearRemodAdd      2.309e-03  2.565e-04   9.000  < 2e-16 ***
## OverallQual       8.677e-02  5.118e-03  16.955  < 2e-16 ***
## BlgdType_Duplex1 -1.049e-01  2.177e-02  -4.817 1.65e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1399 on 1167 degrees of freedom
## Multiple R-squared:  0.8829, Adjusted R-squared:  0.8821
## F-statistic:  1100 on 8 and 1167 DF,  p-value: < 2.2e-16
```

```
#Estimating for Multicollinearity:
car::vif(Model1.3)
```

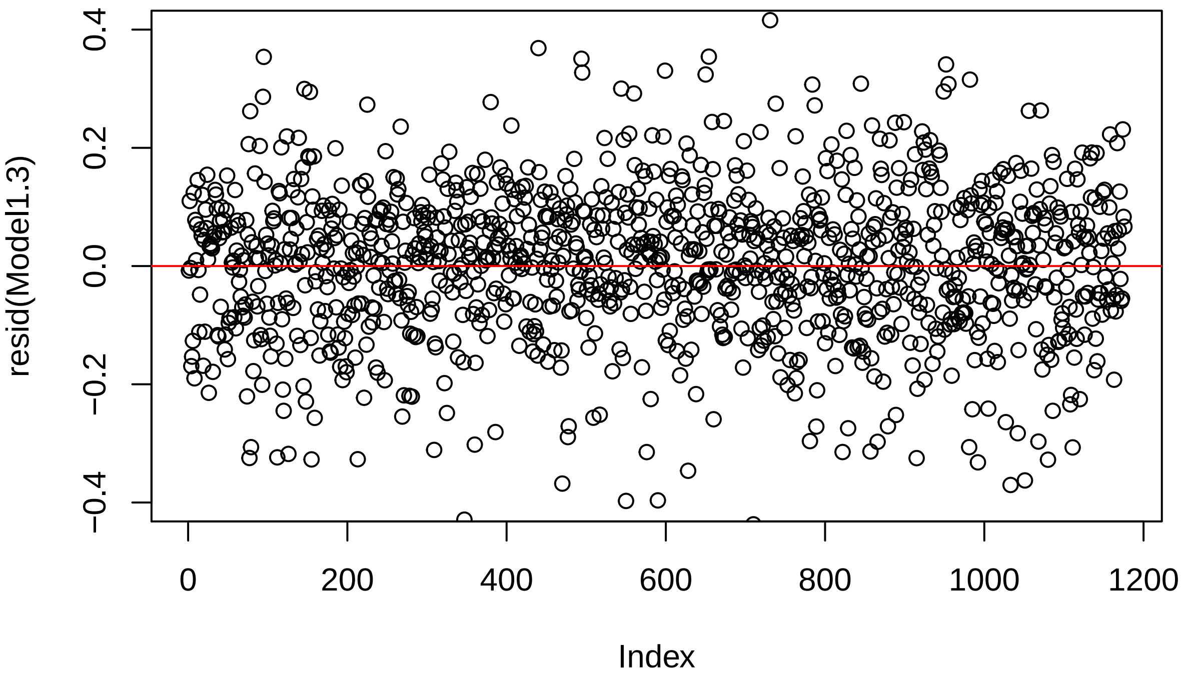
```
##    log(LotArea)      GarageArea  log(GrLivArea)     TotalBsmtSF       YearBuilt 
##        1.319612        1.812279        2.011735        1.723829        2.188296 
##    YearRemodAdd     OverallQual BlgdType_Duplex 
##        1.765287        3.108614        1.093615
```

Since the value of VIF is < 5, there is no problem of Multi-collinearity in our model. So, there is no further term worth removing.

(f) For your model in part (e) plot the respective residuals vs. y, and comment on your results.

We use log(LotArea), GarageArea, log(GrLivArea), TotalBsmtSF, YearBuilt, YearRemodAdd, OverallQual and Blgdtype_Duplex as variables for new model. We do regression to it and plot the residuals. First we plot residuals from `Model1.3`

```r
plot(resid(Model1.3), ylim =c (-0.4, 0.4))
abline(h = 0, col = "red")
```



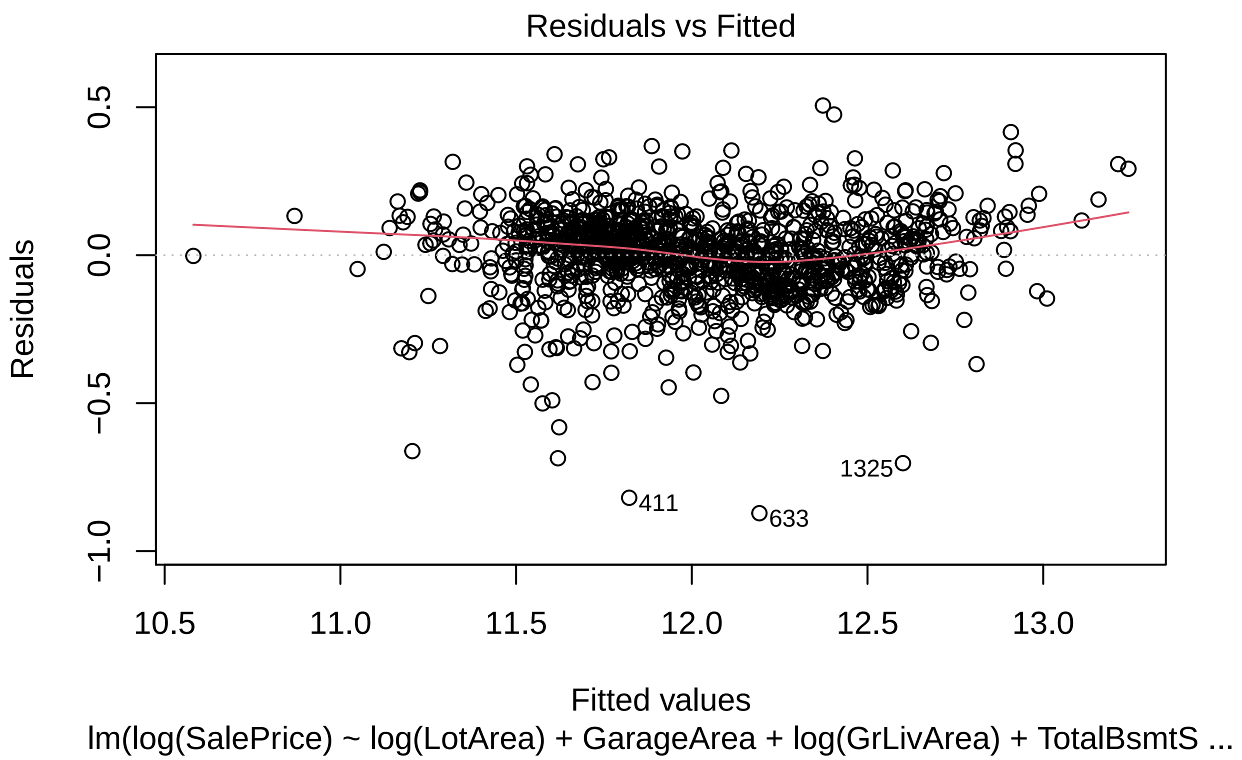
```r
mean(resid(Model1.3))
```

```
## [1] 2.331738e-18
```

Most residuals are centered around 0. The mean of residual is 2.331738e-18, which is close to 0.

Now, we plot the respective residuals vs. y:

```r
plot(Model1.3,1)
```



Residuals are closer to 0 when y is around 126753, our model might predicts houseprice close to $126753 better. But we still found 3 outliers, 411, 633 and 1325.

(g) Using AIC and BIC for model comparison, identify which model is better, (c) or (e).Why?

AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model.

```r
#library(AER)
AIC(Model1.1,Model1.3)
```

```
##          df        AIC
## Model1.1 21  -947.3456
## Model1.3 10 -1278.4335
```

```r
BIC(Model1.1,Model1.3)
```

```
##          df        BIC
## Model1.1 21  -840.4365
## Model1.3 10 -1227.7347
```

As we can see above, the AIC and BIC of the Model1.3 is better than that of Model1.1. Because lower the AIC and BIC, better the model. So as per this, Model1.3 is better than Model1.1. The reason why Model1.3 is better than Model1.1 can also be seen form a higher R^2 and a higher $Adjusted\ R^2$. The Model1.3 would be better because it doesnot have outliers and Mallow's Cp has been performed on it. On the other hand, Model1.1 has outliers and some explanatory variables which as suggested by Mallow's need to be removed.

(h) Estimate a model based on (g) that includes interaction terms and if needed, any higher power terms. Comment on the performance of this model compared to your other two models.

We carry out a Resest test.

```
resettest(Model1.3, power = 2)
```

```
##
##  RESET test
##
## data:  Model1.3
## RESET = 23.232, df1 = 1, df2 = 1166, p-value = 1.626e-06
```

Reset Test suggests that we change our model a bit. So, I add interaction term. The overall quality of houses can be related to the age of the house. So we add one interaction term `YearBuilt:OverallQual`.

```
Model1.4 <- lm(log(SalePrice) ~ log(LotArea)+ GarageArea  + log(GrLivArea)+
                 TotalBsmtSF +  YearBuilt + YearRemodAdd +  OverallQual+
                     BlgdType_Duplex + YearBuilt:OverallQual,
                     subset  = -c(31 , 186 , 199,  363,  384 , 496 , 524  ,706,
                                  917, 1049 ,1063, 1299, 1350 ,1454 ,26 , 152 ,
                                  163,  299 , 315 , 410,  435,  588 , 761,  864,
                                  876, 1070 ,1112 ,1195),
                    data = Train)
summary(Model1.4)
```

```
##
## Call:
## lm(formula = log(SalePrice) ~ log(LotArea) + GarageArea + log(GrLivArea) +
##     TotalBsmtSF + YearBuilt + YearRemodAdd + OverallQual + BlgdType_Duplex +
##     YearBuilt:OverallQual, data = Train, subset = -c(31, 186,
##     199, 363, 384, 496, 524, 706, 917, 1049, 1063, 1299, 1350,
##     1454, 26, 152, 163, 299, 315, 410, 435, 588, 761, 864, 876,
##     1070, 1112, 1195))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87447 -0.07573  0.01104  0.08323  0.50928
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.402e+00  1.353e+00  -1.776    0.076 .
## log(LotArea)           1.207e-01  9.633e-03  12.526  < 2e-16 ***
## GarageArea             1.718e-04  2.536e-05   6.775 1.97e-11 ***
## log(GrLivArea)         3.833e-01  1.760e-02  21.773  < 2e-16 ***
## TotalBsmtSF            1.573e-04  1.296e-05  12.139  < 2e-16 ***
## YearBuilt              2.618e-03  6.498e-04   4.029 5.97e-05 ***
## YearRemodAdd           2.324e-03  2.570e-04   9.042  < 2e-16 ***
## OverallQual            2.877e-01  2.063e-01   1.395    0.163    
## BlgdType_Duplex1      -1.060e-01  2.181e-02  -4.863 1.32e-06 ***
## YearBuilt:OverallQual -1.019e-04  1.046e-04  -0.974    0.330    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1399 on 1166 degrees of freedom
## Multiple R-squared:  0.883,  Adjusted R-squared:  0.8821 
## F-statistic:   978 on 9 and 1166 DF,  p-value: < 2.2e-16
```

```
AIC(Model1.4)
```

```
## [1] -1277.39
```

```
BIC(Model1.4)
```

```
## [1] -1221.622
```

I am not very satisfied with this since the coefficient of `YearBuilt:OverallQual` is negative which is counter-intuitive. To check if we should include any power terms, I add $Total\ Bsmt^2$. My intuition behind this is that Basement area might exhibit diminishing returns after a point.

```
Model1.5 <- lm(log(SalePrice) ~ log(LotArea)+ GarageArea  + log(GrLivArea)+
                     I(TotalBsmtSF^2) + YearBuilt + YearRemodAdd +OverallQual+
                     BlgdType_Duplex ,
                     subset  = -c(31 , 186 , 199,  363,  384 , 496 , 524 ,706,
                                  917, 1049 ,1063, 1299, 1350 ,1454 ,26 , 152 ,
                                  163,  299 , 315 , 410,  435,  588 , 761, 864,
                                  876, 1070 ,1112 ,1195),
                    data = Train)
summary(Model1.5)
```

```
## 
## Call:
## lm(formula = log(SalePrice) ~ log(LotArea) + GarageArea + log(GrLivArea) + 
##     I(TotalBsmtSF^2) + YearBuilt + YearRemodAdd + OverallQual + 
##     BlgdType_Duplex, data = Train, subset = -c(31, 186, 199, 
##     363, 384, 496, 524, 706, 917, 1049, 1063, 1299, 1350, 1454, 
##     26, 152, 163, 299, 315, 410, 435, 588, 761, 864, 876, 1070, 
##     1112, 1195))
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.86083 -0.07645  0.01317  0.08506  0.51537 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       -1.267e+00  5.268e-01  -2.406   0.0163 *  
## log(LotArea)       1.240e-01  9.658e-03  12.834  < 2e-16 ***
## GarageArea         1.683e-04  2.541e-05   6.623 5.37e-11 ***
## log(GrLivArea)    3.865e-01  1.770e-02  21.838  < 2e-16 ***
## I(TotalBsmtSF^2)  5.977e-08  5.237e-09  11.412  < 2e-16 ***
## YearBuilt         2.097e-03  1.935e-04  10.838  < 2e-16 ***
## YearRemodAdd      2.287e-03  2.581e-04   8.861  < 2e-16 ***
## OverallQual       8.700e-02  5.164e-03  16.846  < 2e-16 ***
## BlgdType_Duplex1 -1.248e-01  2.197e-02  -5.681 1.69e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1408 on 1167 degrees of freedom
## Multiple R-squared:  0.8814, Adjusted R-squared:  0.8806
## F-statistic:  1084 on 8 and 1167 DF,  p-value: < 2.2e-16
```

```
AIC(Model1.5)
```

```
## [1] -1263.391
```

```
BIC(Model1.5)
```

```
## [1] -1212.692
```

Although the coefficient of the new term is statistically significant, Model1.4 seems better than Model1.5 as per AIC and BIC. Trying to add interaction term between TotalBsmtSF and Overall Quality. This makes sense since a house with large Total basement area in Sqfoot and high over all quality will have high sale price.

```
Model1.6 <- lm(log(SalePrice) ~ log(LotArea)+ GarageArea  + log(GrLivArea) +
                      YearBuilt + YearRemodAdd +  OverallQual+
                      BlgdType_Duplex + TotalBsmtSF:OverallQual,
                      subset  = -c(31 , 186 , 199,  363,  384 , 496 , 524  ,706,
                                  917, 1049 ,1063, 1299, 1350 ,1454 ,26 , 152 ,
                                  163,  299 , 315 , 410,  435,  588 , 761,  864 ,
                                  876, 1070 ,1112 ,1195),
                     data = Train)
summary(Model1.6)
```

```
##
## Call:
## lm(formula = log(SalePrice) ~ log(LotArea) + GarageArea + log(GrLivArea) +
##     YearBuilt + YearRemodAdd + OverallQual + BlgdType_Duplex +
##     TotalBsmtSF:OverallQual, data = Train, subset = -c(31, 186,
##     199, 363, 384, 496, 524, 706, 917, 1049, 1063, 1299, 1350,
##     1454, 26, 152, 163, 299, 315, 410, 435, 588, 761, 864, 876,
##     1070, 1112, 1195))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86449 -0.07102  0.01427  0.08133  0.50496
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.204e+00  5.190e-01  -2.320   0.0205 *
## log(LotArea)             1.202e-01  9.529e-03  12.610  < 2e-16 ***
## GarageArea               1.578e-04  2.510e-05   6.288 4.53e-10 ***
## log(GrLivArea)           3.939e-01  1.742e-02  22.613  < 2e-16 ***
## YearBuilt                2.081e-03  1.905e-04  10.927  < 2e-16 ***
## YearRemodAdd             2.308e-03  2.543e-04   9.075  < 2e-16 ***
## OverallQual              5.829e-02  5.972e-03   9.760  < 2e-16 ***
## BlgdType_Duplex1        -1.097e-01  2.159e-02  -5.081 4.38e-07 ***
## OverallQual:TotalBsmtSF  2.398e-05  1.842e-06  13.020  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1387 on 1167 degrees of freedom
## Multiple R-squared:  0.8849, Adjusted R-squared:  0.8841
## F-statistic:  1122 on 8 and 1167 DF,  p-value: < 2.2e-16
```

```
AIC(Model1.2,Model1.3, Model1.4, Model1.5, Model1.6)
```

```
##          df       AIC
## Model1.2 21 -1291.780
## Model1.3 10 -1278.433
## Model1.4 11 -1277.390
## Model1.5 10 -1263.391
## Model1.6 10 -1298.463
```

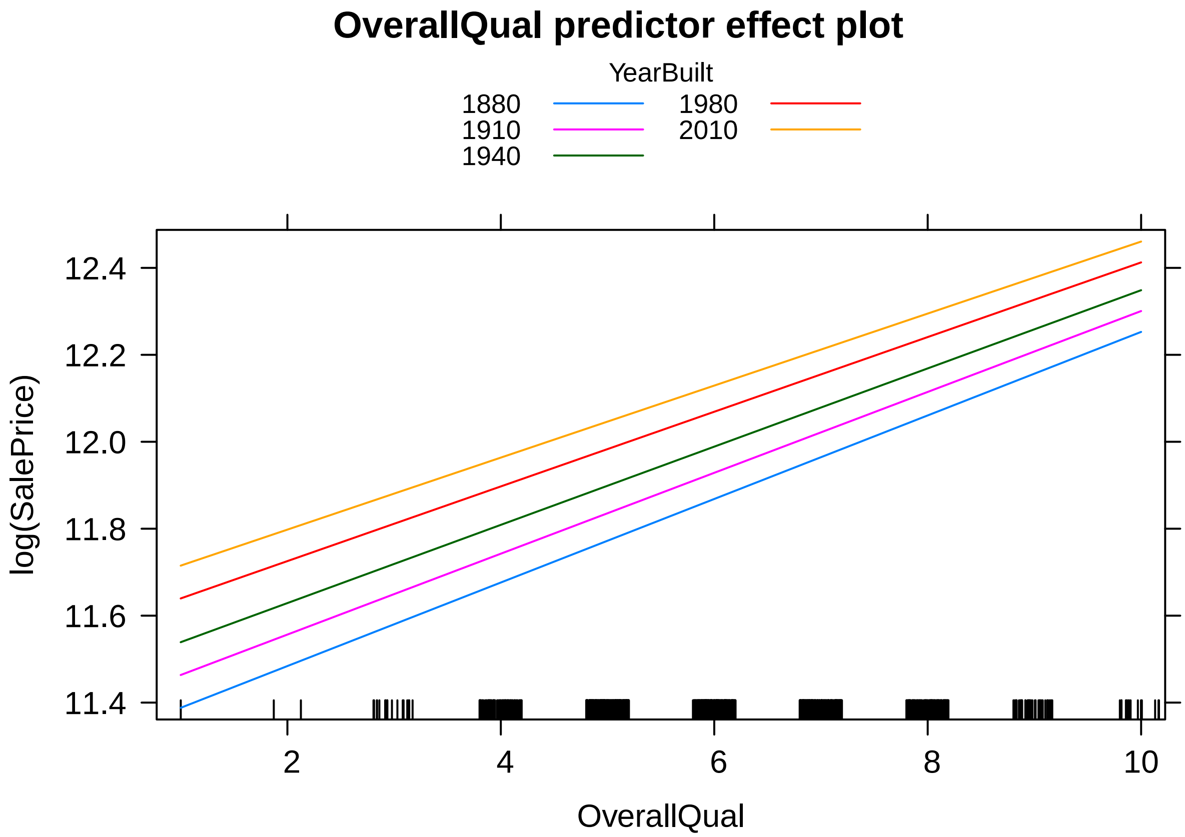
```
BIC(Model1.2,Model1.3,Model1.4,Model1.5,Model1.6)
```

```
##          df       BIC
## Model1.2 21 -1185.313
## Model1.3 10 -1227.735
## Model1.4 11 -1221.622
## Model1.5 10 -1212.692
## Model1.6 10 -1247.764
```

The AIC and BIC suggest that that Model1.6 is the better model. It has the lowest AIC and BIC. If we compare the R^2, it is 88.4% for Model1.6. And the Adjusted R^2 is the highest for this model although by a few decimal points, but still its the highest. The interaction term has a positive coefficient as well. This is in line with our intuition that a home with larger basement area and high overall quality will have a positive impact on sale price. The coefficient is statistically significant also. Model1.2 is the baseline model without outliers. It has a lower R^2 and insignificant coefficients values. Model 1.6 performs better than this in terms of R^2 and as suggested by AIC and BIC. The reason is that it has interaction term and Mallow's Cp has been performed on it. Model1.3 is the model on which Mallow's Cp has been carried. Model1.6 is better than it. It has a slightly better R^2 and Adjusted R^2 as well. Also AIC and BIC suggest in favor Model1.6. Although on grounds of Multi-collinearity, Overall qual has a high VIF. This problem dos not exist in the other two models. I considered Model1.6 only for comparison with Model.2 and Model1.3 since that seemed relevant. Although I may have some comments to make: If we compare Model1.4 to Model1.3 and Model1.2, we can see that R^2 and $Adjusted\ R^2$ are almost the same. According to AIC and BIC, Model1.3 is better than Model1.4. The AIC suggests that Model1.2 is better than Model1.4. Although the BIC suggests that Model1.4 is better than Model1.2. The reason could be because the lower the degree of freedom, the better the BIC estimate would be.

The effect of OverallQual on different ages are shown in the graph below:

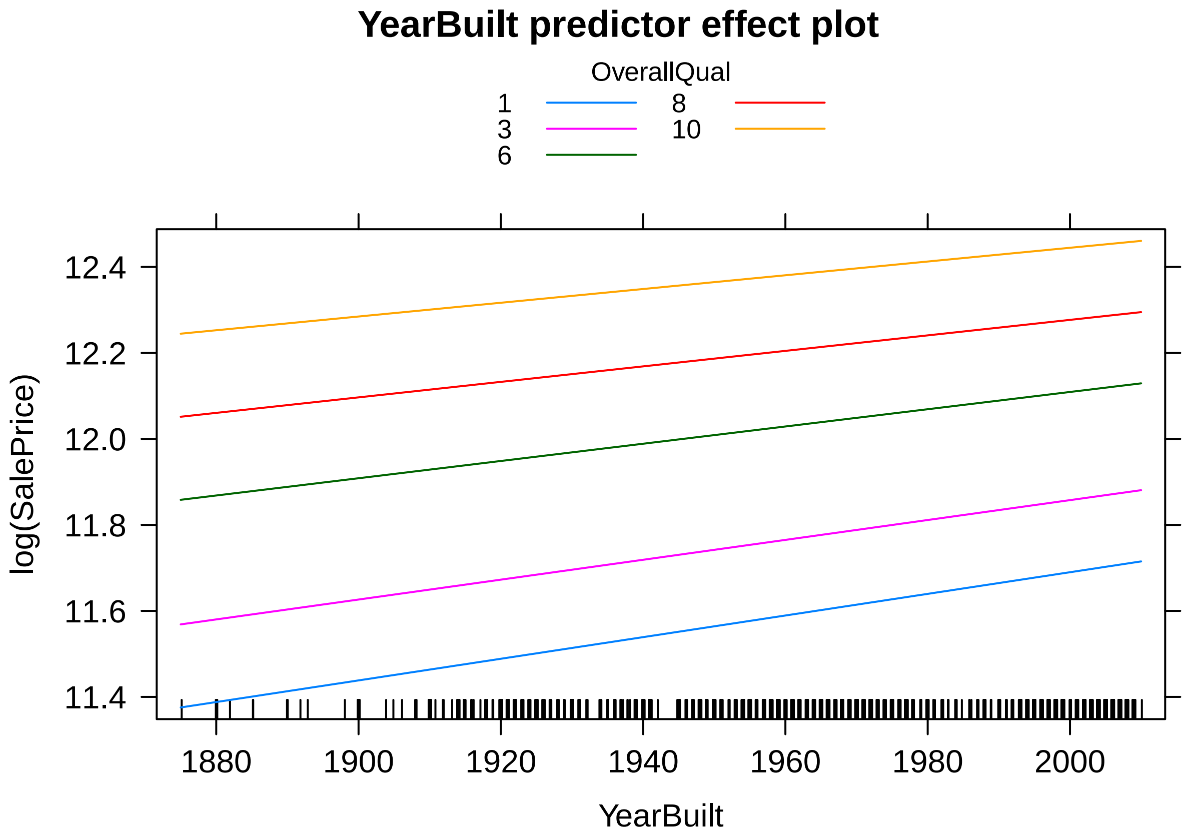
```
plot(predictorEffects(Model1.4, ~ OverallQual),
     lattice=list(key.args=list(cex=.8, cex.title=.8)),
     lines=list(multiline=TRUE))
```



This graph shows us that for each year, as `OverallQual` increases, the price of the house increases. With each year depicted in the graphs, the price of house goes up with an increase in Overallquality. The newer houses that are bulit later with better quality have a higher sale price.

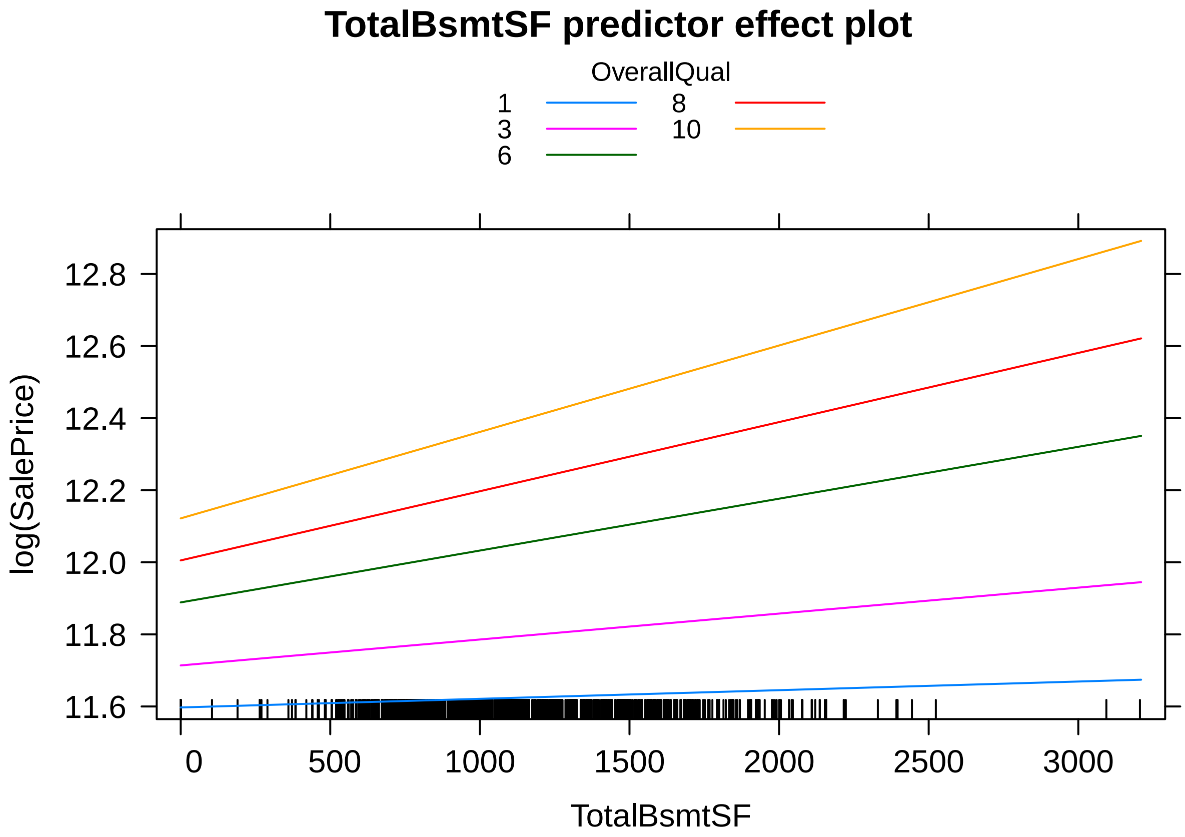
The Effect of YearBuilt on different qualities can be shown below:

```
plot(predictorEffects(Model1.4, ~ YearBuilt),
     lattice=list(key.args=list(cex=.8, cex.title=.8)),
     lines=list(multiline=TRUE))
```

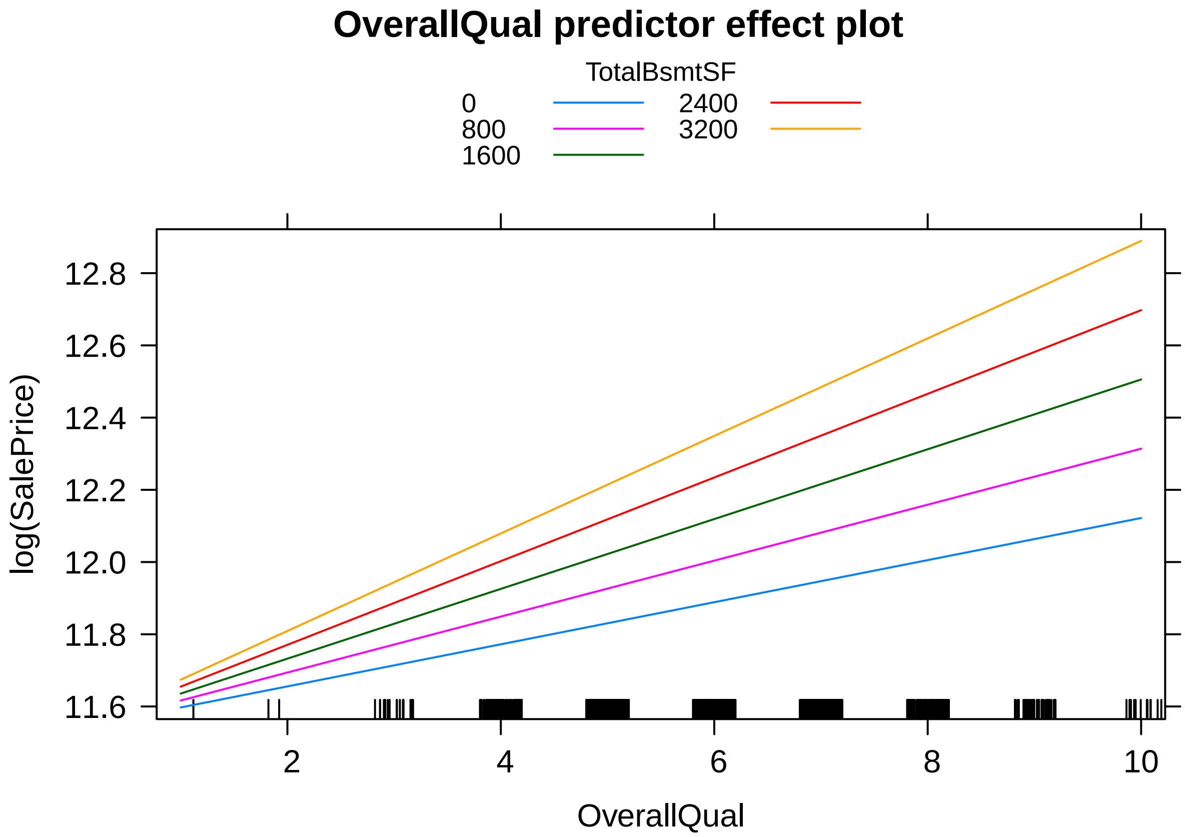


It is seen that houses with poor quality have lower sale price as compared to houses with better quality in the same year. The graph also illustrates that people prefer newer houses with high overall quality. The sale price of such houses is high.

```
plot(predictorEffects(Model1.6, ~ TotalBsmtSF),
     lattice=list(key.args=list(cex=.8, cex.title=.8)),
     lines=list(multiline=TRUE))
```



```
plot(predictorEffects(Model1.6, ~ OverallQual),
     lattice=list(key.args=list(cex=.8, cex.title=.8)),
     lines=list(multiline=TRUE))
```



The graphs show that a higher Overall Quality and large Total BsmtSF have a high sale price. It can be seen that most of the data lies between values of 500 to 2000 Sqft of Total basement area.

(i) Lastly, choose your favorite model from all the ones estimated and perform a five-fold cross validation test on it. Then use the dataset to evaluate how well your model predicts home prices for out of sample data, and comment on your overall findings.

```r
library(MASS)
library(tidyverse)
library(caret)
library(Metrics)
```

Now we carry out 5 fold cross validation using Model1.3. I chose Model1.3. Intuitively, the variables make economic significance and most of them make statistical significance. It has no multi-collinearity or heteroskedasticity whereas Model1.6 shows that Overallqual has multicollinearity. So, model1.3 is my favorite model which I choose for this part.

```r
# Define training control
set.seed(123)
train.control <- trainControl(method = "cv", number = 5)
# Train the model
#removing the outliers
Train.1 <- Train[-c(31 , 186 , 199,  363,  384 , 496 , 524  ,706,  917, 1049,
                    1063,1299, 1350 ,1454 ,26 , 152 , 163,  299 , 315 , 410,
                    435, 588 ,761,  864 , 876, 1070 ,1112 ,1195),]
#Cross-Validation
```

```
model <- train(log(SalePrice) ~ log(LotArea)+ GarageArea  + log(GrLivArea)+
                 TotalBsmtSF + YearBuilt + YearRemodAdd +  OverallQual  +
                 BlgdType_Duplex ,
                 data = Train.1,
                 method = "lm",
                 trControl = train.control)

# Summarize the results
print(model)
```

```
## Linear Regression
##
## 1176 samples
##    8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 941, 940, 942, 941, 940
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1401763  0.8821887  0.103028
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
predictions <- Model1.3 %>% predict(Train.1)
summary(predictions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.58   11.74   11.99   12.02   12.30   13.24
```

```
rmse(predictions,log(Train.1$SalePrice))
```

```
## [1] 0.1393184
```

Our model seems to predict the actual sale price of homes well because: The RMSE from the cross validation on the train data set is (0.1401763) and RMSE for Model1.3 is 0.1393184.These values are not so different. Also, both RMSE value are very low(< 0.5). Based on the R squared I derived it predicts the model based on 88.21%.Which supports the notion that my RMSE is small due to the model having a strong R squared.

Question 2

Assume a healthcare insurance company hired you as a consultant to develop an econometric model to estimate the number of doctor visits a patient has over a 3 month period. The rational behind this study is that patients with a higher number of doctors visits wold pose a higher liability in terms of insurance expenses, and therefore, this may be mitigated via a higher insurance premium. The panel data are from the German Health Care Usage Dataset, and consist of 7,293 individuals across varying numbers of periods with a total of 27,326 observations.

(a) Build a multiple regression model with a subset of 10 predictors (at most), including interaction and non-linear transformations if appropriate. For this part you only need to briefly discuss a justification for the model chosen, and discuss the respective regression output.

Since there are NA values, I first omit the NA. The procedure for the same is as below:

```
#reading the data set
library(readr)
german_healthcare_usage <- read_csv("UCLA MAE/ECON 430/R/Homework 2-20201020/german_healthcare_usage.cs
#checking for NAs
sapply(german_healthcare_usage, function(x) sum(is.na(x)))
```

```
##       ID   FEMALE     YEAR      AGE  HANDDUM      ALC  FAMHIST  HANDPER
##        1        2        1        3        1        4        2        1
##   HHKIDS     EDUC  MARRIED   HAUPTS    REALS   FACHHS   ABITUR     UNIV
##        2        2        2        1        1        1        2        1
##  WORKING    BLUEC   WHITEC     SELF    BEAMT   DOCVIS  HOSPVIS UNEMPLOY
##        1        1        1        1        1        1        1        1
##   PUBLIC    ADDON   NUMOBS     HSAT   DOCTOR  HEALTHY YEAR1984 YEAR1985
##        1        4        1        1        2        1        1        1
## YEAR1986 YEAR1987 YEAR1988 YEAR1991 YEAR1994   LOGINC       TI HOSPITAL
##        1        2        5        2        1        6        2        6
##   HHNINC  NEWHSAT PRESCRIP
##        2        1        1
```

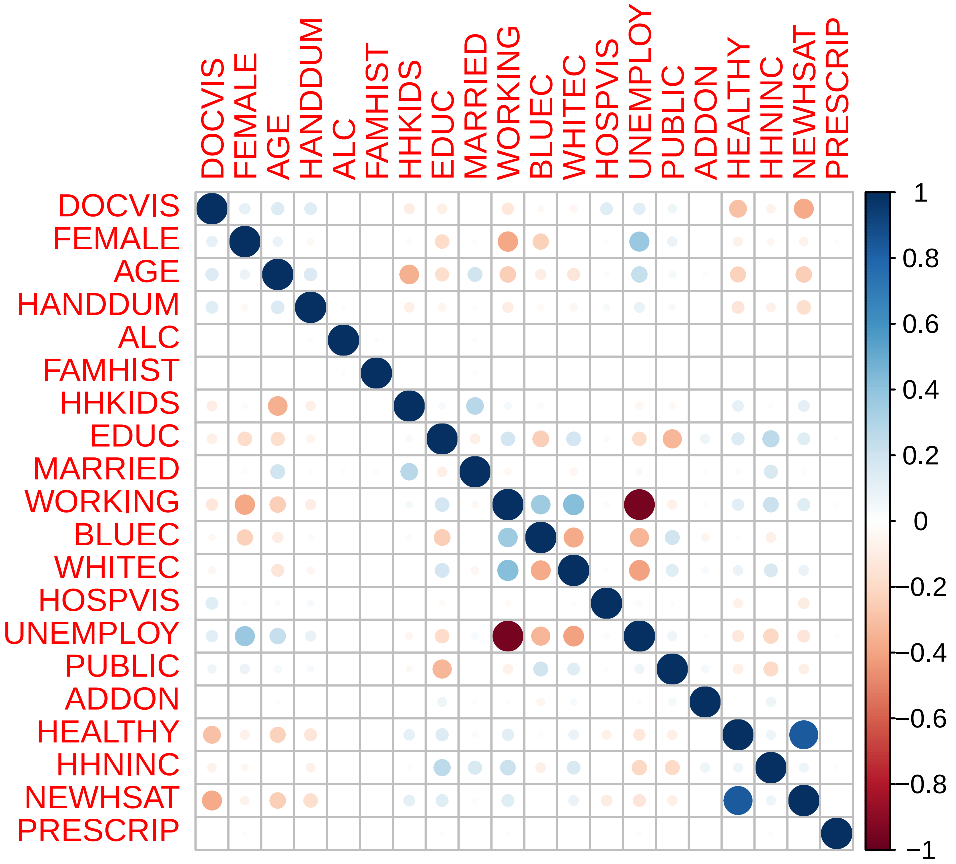
```
#Since this commnad tells us that our data has NAs, I will omit them as follows:
german_healthcare_usage <- na.omit(german_healthcare_usage)
#re-cheking for NAs
sapply(german_healthcare_usage, function(x) sum(is.na(x)))
```

```
##       ID   FEMALE     YEAR      AGE  HANDDUM      ALC  FAMHIST  HANDPER
##        0        0        0        0        0        0        0        0
##   HHKIDS     EDUC  MARRIED   HAUPTS    REALS   FACHHS   ABITUR     UNIV
##        0        0        0        0        0        0        0        0
##  WORKING    BLUEC   WHITEC     SELF    BEAMT   DOCVIS  HOSPVIS UNEMPLOY
##        0        0        0        0        0        0        0        0
##   PUBLIC    ADDON   NUMOBS     HSAT   DOCTOR  HEALTHY YEAR1984 YEAR1985
##        0        0        0        0        0        0        0        0
## YEAR1986 YEAR1987 YEAR1988 YEAR1991 YEAR1994   LOGINC       TI HOSPITAL
##        0        0        0        0        0        0        0        0
##   HHNINC  NEWHSAT PRESCRIP
##        0        0        0
```

```
#Now, our NAs have been successfully been removed so we can proceed ahead
```

Now, I select some variables from the data set, `german_healthcare_usage.csv` and I make a correlation plot to know which variables are correlated with doctor visits in the last three months. On the basis of the plot, I seek to choose the variables for my study. I believe that visual representation would help me choose the variables better

```
library(corrplot)
attach(german_healthcare_usage)
x=c(22,2,4,5,6,7,9,10,11,17,18,19,23,24,25,26,30,41,42,43)
corrplot(cor(german_healthcare_usage[,x]))
```



I Choose the following variables since they share a strong correlation with the doctor visits. I also added an interaction term to the model FEMALE:AGE Let us run the model and see the results.

```
Model2.1 = lm(DOCVIS~HANDDUM+HOSPVIS + WORKING +HHNINC+EDUC+NEWHSAT+FEMALE:AGE
+FEMALE+AGE, data=german_healthcare_usage)
summary(Model2.1)
```

```
##
## Call:
## lm(formula = DOCVIS ~ HANDDUM + HOSPVIS + WORKING + HHNINC +
##     EDUC + NEWHSAT + FEMALE:AGE + FEMALE + AGE, data = german_healthcare_usage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.606  -2.404  -0.920   0.833 111.947
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.654234   0.278295  27.504  < 2e-16 ***
## HANDDUM      0.822075   0.079322  10.364  < 2e-16 ***
## HOSPVIS      0.620569   0.035884  17.294  < 2e-16 ***
## WORKING     -0.342599   0.077562  -4.417 1.00e-05 ***
## HHNINC      -0.840006   0.189640  -4.429 9.48e-06 ***
## EDUC        -0.007864   0.014578  -0.539 0.589607
## NEWHSAT     -0.814790   0.014508 -56.161  < 2e-16 ***
## FEMALE       0.964607   0.252755   3.816 0.000136 ***
## AGE          0.024197   0.004010   6.034 1.62e-09 ***
## FEMALE:AGE  -0.004794   0.005631  -0.851 0.394561
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.213 on 27287 degrees of freedom
## Multiple R-squared:  0.1612, Adjusted R-squared:  0.1609
## F-statistic: 582.6 on 9 and 27287 DF,  p-value: < 2.2e-16
```

The coefficient of `female` and `age` are insignificant, hence I remove that term and re-run the regression with the updated model.

```
Model2.1 = lm(DOCVIS ~ HANDDUM+HOSPVIS + WORKING +HHNINC+EDUC+NEWHSAT
+FEMALE+AGE, data=german_healthcare_usage)
summary(Model2.1)
```

```
##
## Call:
## lm(formula = DOCVIS ~ HANDDUM + HOSPVIS + WORKING + HHNINC +
##     EDUC + NEWHSAT + FEMALE + AGE, data = german_healthcare_usage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.642  -2.399  -0.920   0.834 111.939
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.738094   0.260280  29.730  < 2e-16 ***
## HANDDUM      0.825877   0.079196  10.428  < 2e-16 ***
## HOSPVIS      0.620848   0.035882  17.303  < 2e-16 ***
## WORKING     -0.338529   0.077414  -4.373 1.23e-05 ***
## HHNINC      -0.834631   0.189534  -4.404 1.07e-05 ***
## EDUC        -0.007173   0.014555  -0.493    0.622
## NEWHSAT     -0.814946   0.014507 -56.176  < 2e-16 ***
## FEMALE       0.757708   0.069495  10.903  < 2e-16 ***
## AGE          0.021919   0.002987   7.337 2.24e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.213 on 27288 degrees of freedom
## Multiple R-squared:  0.1612, Adjusted R-squared:  0.1609
## F-statistic: 655.3 on 8 and 27288 DF,  p-value: < 2.2e-16
```

Now I carry the reset test to see if my model needs any power terms

```
resettest(Model2.1, power = 2, type = "regressor")
```

```
##
```

```
##  RESET test
##
## data:  Model2.1
## RESET = 68.824, df1 = 8, df2 = 27280, p-value < 2.2e-16
```

The test rejects the null hypothesis that there should not be any quadratic term. So, I will add a quadratic term to my model. I think I will add AGE^2 to my model.Age likely doesn't have a linear effect on doctor visits. So economically speaking, It makes more sense to add AGE^2.

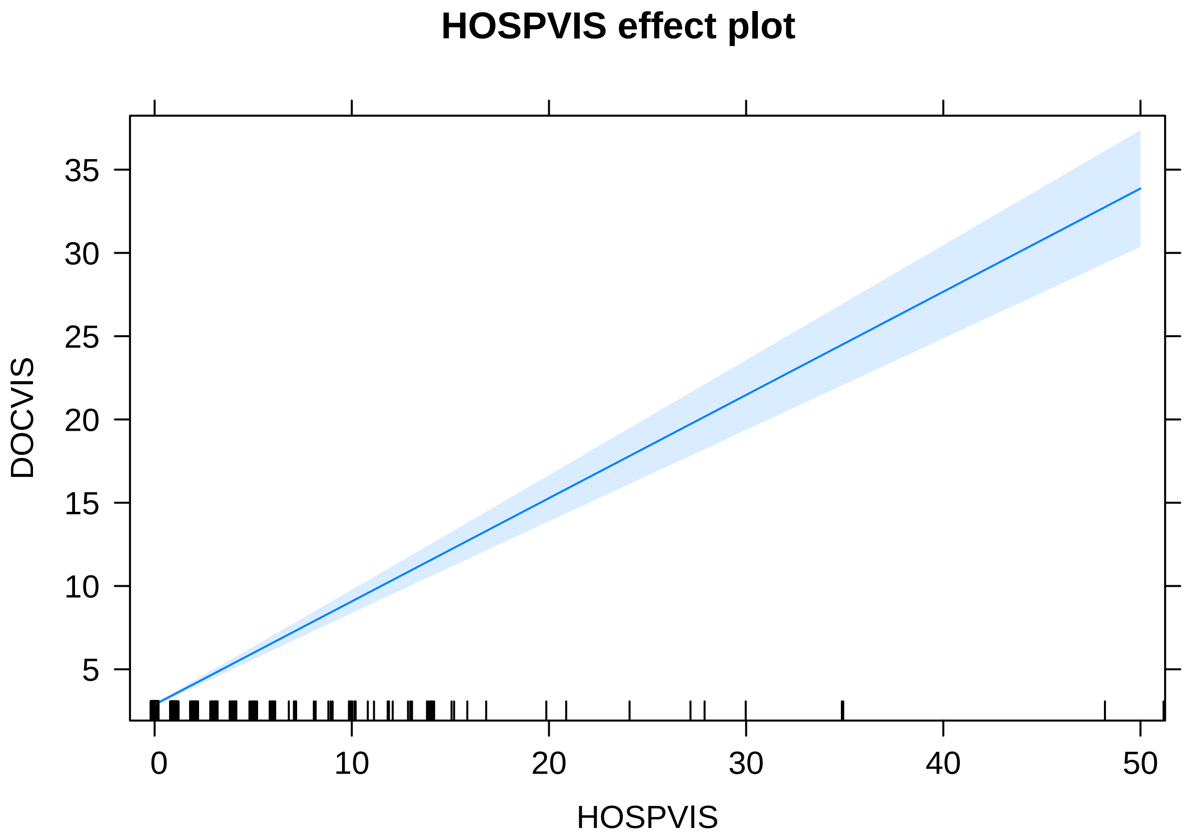
```
Model2.1 = lm(DOCVIS ~ HANDDUM+HOSPVIS + WORKING +HHNINC+EDUC+NEWHSAT
+FEMALE+AGE +I(AGE^2), data=german_healthcare_usage)
summary(Model2.1)
```

```
##
## Call:
## lm(formula = DOCVIS ~ HANDDUM + HOSPVIS + WORKING + HHNINC +
##     EDUC + NEWHSAT + FEMALE + AGE + I(AGE^2), data = german_healthcare_usage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.594  -2.373  -0.923   0.822 112.235
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.7064880  0.5477057  19.548  < 2e-16 ***
## HANDDUM      0.8054830  0.0792114  10.169  < 2e-16 ***
## HOSPVIS      0.6198359  0.0358580  17.286  < 2e-16 ***
## WORKING     -0.2336778  0.0792128  -2.950 0.003180 **
## HHNINC      -0.6661086  0.1913720  -3.481 0.000501 ***
## EDUC        -0.0087584  0.0145478  -0.602 0.547150
## NEWHSAT     -0.8172704  0.0145021 -56.355  < 2e-16 ***
## FEMALE       0.7947540  0.0697080  11.401  < 2e-16 ***
## AGE         -0.1289103  0.0246728  -5.225 1.76e-07 ***
## I(AGE^2)     0.0017237  0.0002799   6.158 7.45e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.21 on 27287 degrees of freedom
## Multiple R-squared:  0.1623, Adjusted R-squared:  0.162
## F-statistic: 587.5 on 9 and 27287 DF,  p-value: < 2.2e-16
```

As compared to the previous models, R^2 has improved and AIC and BIC is also lower. So we will go ahead with this model. Most of our coefficients are statistically significant at 5% level of significance. Individuals that are handicapped, female, or are older all tend to visit the doctor more within a three-month period and hence the positive coefficient. Increase in Hospital visists implies high doctor visits. Those who are educated tend to take care of their health and visit the doctor less often. Hence, the negative coefficient. Those who are working and those who have a high income visit the doctor less often. Intutively, this could be owing to the fact that these people can afford a healthy lifestyle and good nutritious meal. They must be enjoying a high standard of living. Those with higher health satisfaction also tend to visit the doctor less often.

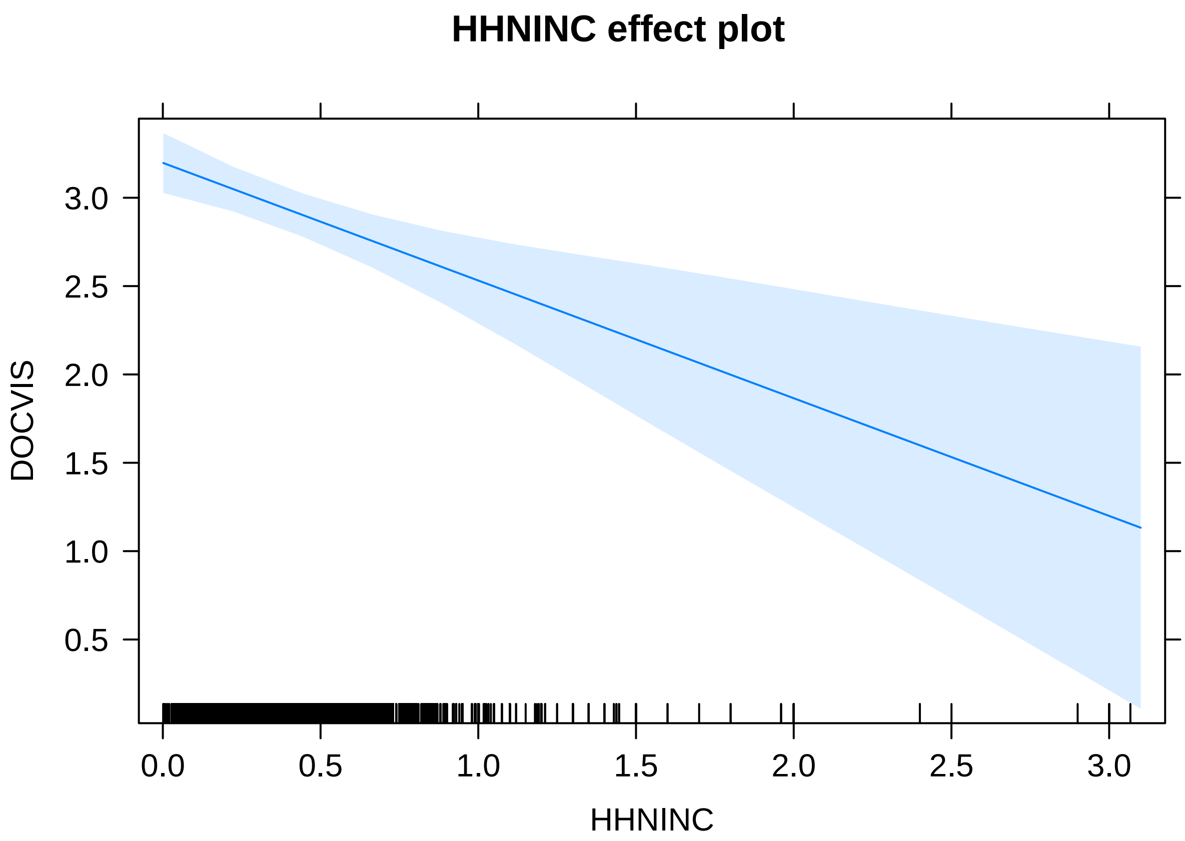
We can look at the marginal effects plots to understand the relationships visually.

```
library(effects)
plot(effect(mod=Model2.1,"HOSPVIS"))
```



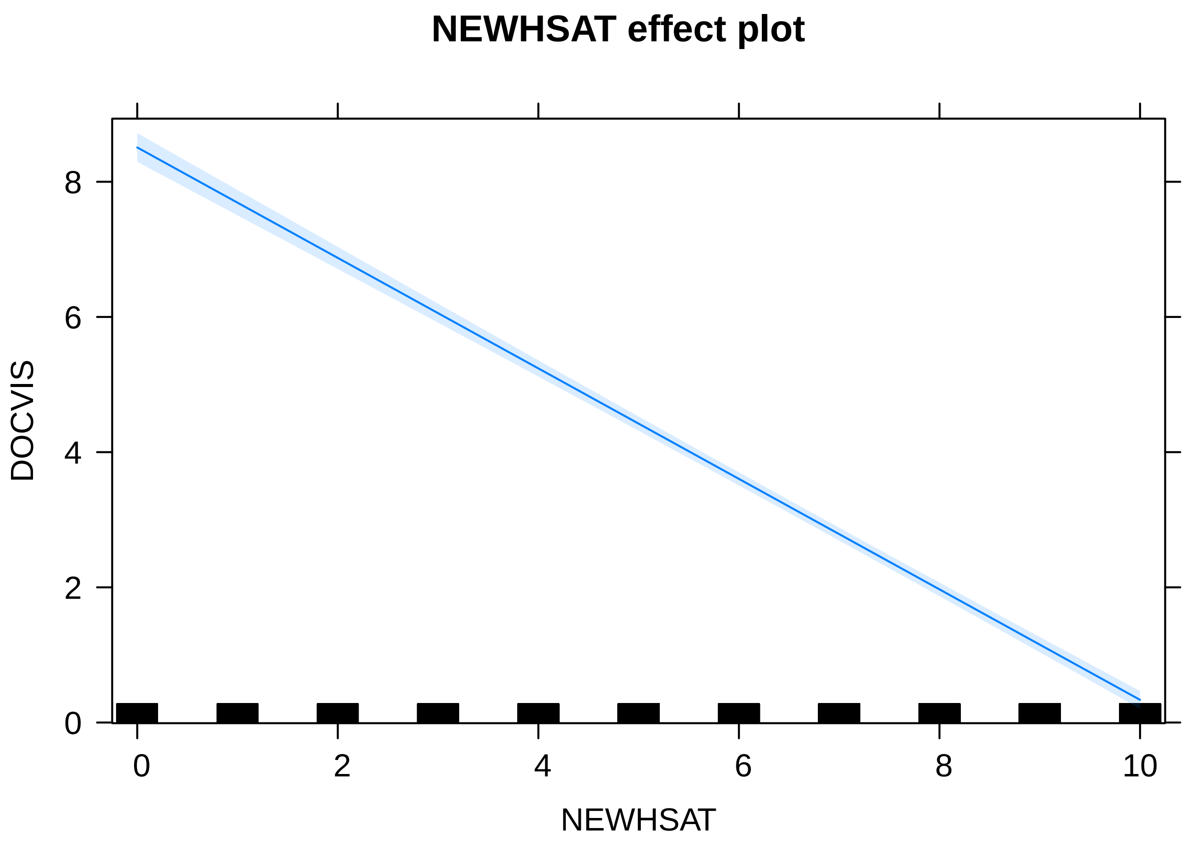
The result is similar to our intution.All other things being constant, with higher hospital visits, Doctor visits increase. An upward Linear trend is exhibited. Also, with increase in Hospital visits spread of Doctor visits becomes larger.

```
plot(effect(mod=Model2.1,"HHNINC"))
```



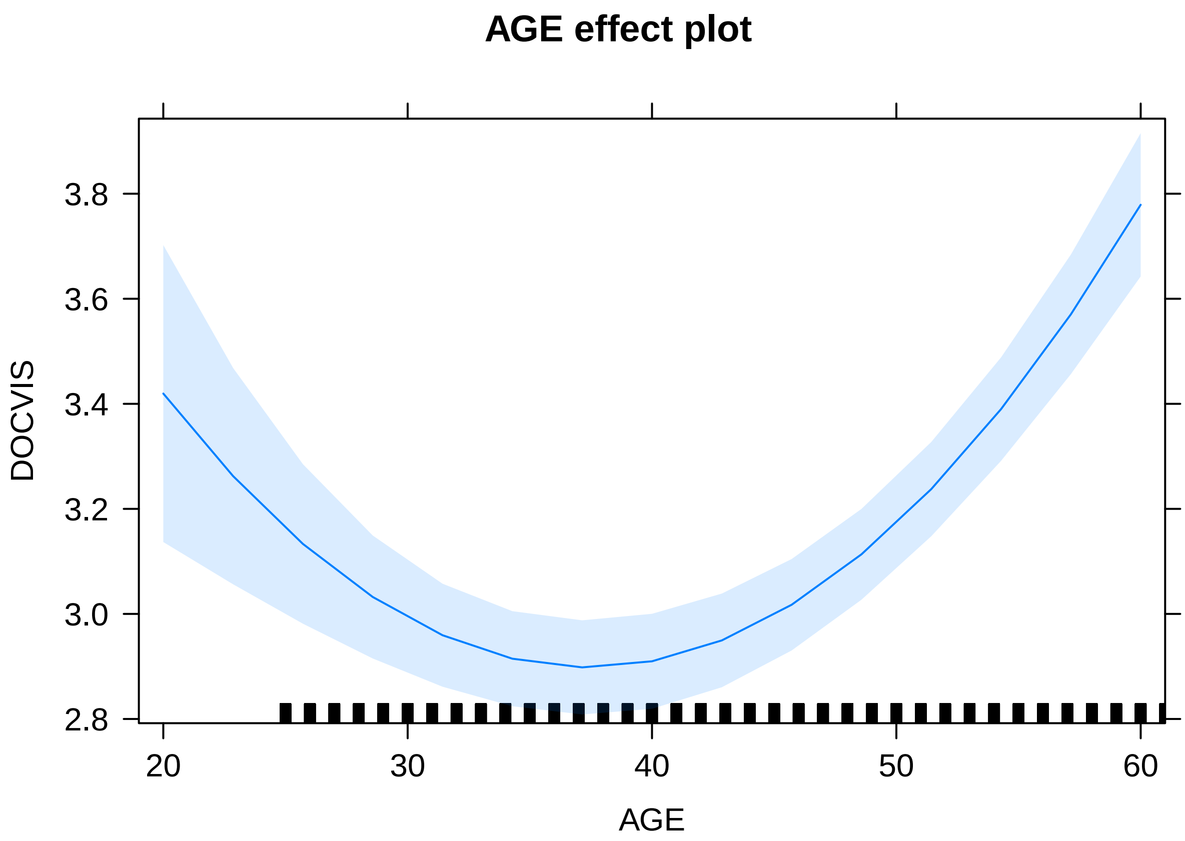
The result again matches our intution. With higher incomes, Doctor visits to tend to reduce. Again, At higher incomes, spread of Doctor visits is larger suggesting Heteroskedasticity. Also notice that data values are concentrated at lower income levels.

```
plot(effect(mod=Model2.1,"NEWHSAT"))
```



Those who have high health satisfaction tend to visit the doctor less often. A downward linear trend is shown.

```
plot(effect(mod=Model2.1,"AGE"))
```



This visual representation suggests that at lower age, children tend to visit the doctor more often as they tend to fall sick more often. At higher ages, due to health ailments, doctors visit tend to increase. They show a downward trend in young ages.

(b) Differences in Differences: In 1987 the German Government passed a series of legislations to improve healthcare access for unemployed people and women.

i. Determine whether or not the policy worked for women.

```
german_healthcare_usage$TIME= ifelse(german_healthcare_usage$YEAR>=1987,1,0)
german_healthcare_usage$DID = german_healthcare_usage$TIME * german_healthcare_usage$FEMALE
attach(german_healthcare_usage)
```

```
## The following objects are masked from german_healthcare_usage (pos = 3):
##
##     ABITUR, ADDON, AGE, ALC, BEAMT, BLUEC, DOCTOR, DOCVIS, EDUC,
##     FACHHS, FAMHIST, FEMALE, HANDDUM, HANDPER, HAUPTS, HEALTHY, HHKIDS,
##     HHNINC, HOSPITAL, HOSPVIS, HSAT, ID, LOGINC, MARRIED, NEWHSAT,
##     NUMOBS, PRESCRIP, PUBLIC, REALS, SELF, TI, UNEMPLOY, UNIV, WHITEC,
##     WORKING, YEAR, YEAR1984, YEAR1985, YEAR1986, YEAR1987, YEAR1988,
##     YEAR1991, YEAR1994
```

```
Model2.2.1 = lm(DOCVIS~FEMALE+TIME+DID,data=german_healthcare_usage)
summary(Model2.2.1)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + TIME + DID, data = german_healthcare_usage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -3.917  -2.643  -1.643   0.387 118.357
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.64255    0.07335  36.026   <2e-16 ***
## FEMALE       1.27462    0.10590  12.036   <2e-16 ***
## TIME        -0.02944    0.09620  -0.306    0.760
## DID         -0.18560    0.13899  -1.335    0.182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.661 on 27293 degrees of freedom
## Multiple R-squared:  0.01067,    Adjusted R-squared:  0.01056
## F-statistic: 98.08 on 3 and 27293 DF,  p-value: < 2.2e-16
```

We can see that the Difference in Difference estimate, `DID` is insignificant. Hence, the policy does not work for women.

ii. Determine whether or not the policy worked for unemployed.

```r
german_healthcare_usage$DID = german_healthcare_usage$TIME * german_healthcare_usage$UNEMPLOY
Model2.2.2 = lm(DOCVIS~UNEMPLOY+TIME+DID,data=german_healthcare_usage)
summary(Model2.2.2)
```

```
##
## Call:
## lm(formula = DOCVIS ~ UNEMPLOY + TIME + DID, data = german_healthcare_usage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -4.308  -2.722  -1.722   0.324 116.692
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.67609    0.06569  40.741   <2e-16 ***
## UNEMPLOY     1.63193    0.11037  14.786   <2e-16 ***
## TIME         0.04550    0.08480   0.537   0.5916
## DID         -0.25961    0.14752  -1.760   0.0784 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.648 on 27293 degrees of freedom
## Multiple R-squared:  0.01509,    Adjusted R-squared:  0.01498
## F-statistic: 139.4 on 3 and 27293 DF,  p-value: < 2.2e-16
```

Here the DID estimator is significant to the 10% level so it somewhat worked for the unemployed.

(c) Test the hypothesis that the number of doctor visits a patient has over a 3 month period is greater for women than for men.

```
femalevisits=0
malevisits=0
Model2.3 = lm(DOCVIS~FEMALE,data=german_healthcare_usage)
summary(Model2.3)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE, data = german_healthcare_usage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -3.793  -2.625  -1.625   0.375 118.375
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.62543    0.04746   55.32   <2e-16 ***
## FEMALE       1.16707    0.06859   17.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.661 on 27295 degrees of freedom
## Multiple R-squared:  0.0105, Adjusted R-squared:  0.01046
## F-statistic: 289.5 on 1 and 27295 DF,  p-value: < 2.2e-16
```

```
for(i in 1:27297){
if(FEMALE[i]==1){
femalevisits[i]=DOCVIS[i]
}else{
malevisits[i]=DOCVIS[i]
}
}
femalevisits=na.omit(femalevisits)
malevisits=na.omit(malevisits)
t.test(femalevisits,malevisits,alternative="greater", mu=0,paired = FALSE,
       var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  femalevisits and malevisits
## t = 16.898, df = 25787, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.0532    Inf
## sample estimates:
## mean of x mean of y
##  3.792212  2.625431
```

We fail to accept the null that the true difference in means is not greater than zero. Therefore we conclude that the number of doctorvisits is greater for females than for males.

(d) Based on your findings propose and test your own hypothesis of interest using the linear functional form: $\lambda = c_1\ \beta_1 + c_2\ \beta_2 + ...$

My Hypothesis of interest is as follows: whether an increase in age by one year and a decrease in level of health satisfaction by one unit will increase the number of doctor visits.

Let us create a model

```
library(multcomp)
```

```
## Warning: package 'multcomp' was built under R version 4.0.3
```

```
Model2.4 <- lm(DOCVIS ~ AGE + NEWHSAT, data = german_healthcare_usage)
summary(glht(Model2.4, linfct = c("1*AGE - 1*NEWHSAT = 0")))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = DOCVIS ~ AGE + NEWHSAT, data = german_healthcare_usage)
##
## Linear Hypotheses:
##                            Estimate Std. Error t value Pr(>|t|)
## 1 * AGE - 1 * NEWHSAT == 0  0.91302    0.01393   65.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Model2.4 results in approximately 1 doctor visit every 3 months. Now we will change this slightly to incorporate a decrease in health satisfaction by two units and an increase in age by two years.

```
Model2.4 <- lm(DOCVIS ~ AGE + NEWHSAT, data = german_healthcare_usage)
summary(glht(Model2.4, linfct = c("2*AGE - 2*NEWHSAT = 0")))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = DOCVIS ~ AGE + NEWHSAT, data = german_healthcare_usage)
##
## Linear Hypotheses:
##                            Estimate Std. Error t value Pr(>|t|)
## 2 * AGE - 2 * NEWHSAT == 0  1.82604    0.02787   65.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

As we can see, the number of doctor visits increased to about 2 visits. We can therefore accept the hypothesis that an increase in age and decrease in health satisfaction by one unit will increase the number of doctor visits.