Final Project Scratch Notebook

```
## Helper to display regression function with n coefficients
dispRegFunc <- function(reg) {
  coefs <- reg$coefficients
  b0 = coefs[1]
  n <- length(coefs)
  my_formula <- paste0("Y = ", round(b0, digits = 6))
  for (i in 2:n) {
    my_formula <- paste0(my_formula, " + ", round(coefs[i],6), names(coefs)[i])
  }
  my_formula
}
```

Step 0: joined team

Step 1: Load Data and review variables

```
## corrplot 0.92 loaded
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:MASS':
##
##     cement
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
## Loading required package: carData
```

```
## 'data.frame':    21613 obs. of  21 variables:
##  $ id           : num  7129300520 6414100192 5631500400 2487200875 1954400510 ...
##  $ date         : chr  "20141013T000000" "20141209T000000" "20150225T000000" "20141209T000000" ...
##  $ price        : chr  " $221,900.00 " " $538,000.00 " " $180,000.00 " " $604,000.00 " ...
##  $ bedrooms     : int  3 3 2 4 3 4 3 3 3 3 ...
##  $ bathrooms    : num  1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
##  $ sqft_living  : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
##  $ sqft_lot     : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
##  $ floors       : num  1 2 1 1 1 1 2 1 1 2 ...
##  $ waterfront   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ view         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ condition    : int  3 3 3 5 3 3 3 3 3 3 ...
##  $ grade        : int  7 7 6 7 8 11 7 7 7 7 ...
```

```
##  $ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
##  $ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
##  $ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
##  $ yr_renovated : int  0 1991 0 0 0 0 0 0 0 0 ...
##  $ zipcode      : int  98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
##  $ lat          : num  47.5 47.7 47.7 47.5 47.6 ...
##  $ long         : num  -122 -122 -122 -122 -122 ...
##  $ sqft_living15: int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
##  $ sqft_lot15   : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
```

```
##            id          date         price      bedrooms     bathrooms
##             0             0             0             0             0
##   sqft_living      sqft_lot        floors    waterfront          view
##             0             0             0             0             0
##     condition         grade    sqft_above sqft_basement      yr_built
##             0             0             0             0             0
##  yr_renovated       zipcode           lat          long sqft_living15
##             0             0             0             0             0
##    sqft_lot15
##             0
```

```
##        id                 date              price              bedrooms
##  Min.   :   1000102   Length:21613       Length:21613       Min.   : 0.000
##  1st Qu.:2123049194   Class :character   Class :character   1st Qu.: 3.000
##  Median :3904930410   Mode  :character   Mode  :character   Median : 3.000
##  Mean   :4580301521                                        Mean   : 3.371
##  3rd Qu.:7308900445                                        3rd Qu.: 4.000
##  Max.   :9900000190                                        Max.   :33.000
##    bathrooms      sqft_living       sqft_lot           floors
##  Min.   :0.000   Min.   :  290   Min.   :    520   Min.   :1.000
##  1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040   1st Qu.:1.000
##  Median :2.250   Median : 1910   Median :   7618   Median :1.500
##  Mean   :2.115   Mean   : 2080   Mean   :  15107   Mean   :1.494
##  3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688   3rd Qu.:2.000
##  Max.   :8.000   Max.   :13540   Max.   :1651359   Max.   :3.500
##    waterfront            view          condition         grade
##  Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 1.000
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000
##  Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000
##  Mean   :0.007542   Mean   :0.2343   Mean   :3.409   Mean   : 7.657
##  3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000
##  Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000
##    sqft_above   sqft_basement       yr_built     yr_renovated
##  Min.   : 290   Min.   :   0.0   Min.   :1900   Min.   :   0.0
##  1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0
##  Median :1560   Median :   0.0   Median :1975   Median :   0.0
##  Mean   :1788   Mean   : 291.5   Mean   :1971   Mean   :  84.4
##  3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0
##  Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.0
##     zipcode           lat             long        sqft_living15
##  Min.   :98001   Min.   :47.16   Min.   :-122.5   Min.   : 399
##  1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490
##  Median :98065   Median :47.57   Median :-122.2   Median :1840
##  Mean   :98078   Mean   :47.56   Mean   :-122.2   Mean   :1987
```

```
##  3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360
##  Max.   :98199   Max.   :47.78   Max.   :-121.3   Max.   :6210
##    sqft_lot15
##  Min.   :   651
##  1st Qu.:  5100
##  Median :  7620
##  Mean   : 12768
##  3rd Qu.: 10083
##  Max.   :871200
```

```
## [1] " $221,900.00 "   " $538,000.00 "   " $180,000.00 "   " $604,000.00 "
## [5] " $510,000.00 "   " $1,225,000.00 "
```

```
## [1] "$221,900.00"
```

```
## [1] "221,900.00"
```

```
## [1] "221900.00"
```

```
##        id                 date               price            bedrooms
##  Min.   :   1000102   Length:21613       Min.   :  75000   Min.   : 0.000
##  1st Qu.:2123049194   Class :character   1st Qu.: 321950   1st Qu.: 3.000
##  Median :3904930410   Mode  :character   Median : 450000   Median : 3.000
##  Mean   :4580301521                      Mean   : 540088   Mean   : 3.371
##  3rd Qu.:7308900445                      3rd Qu.: 645000   3rd Qu.: 4.000
##  Max.   :9900000190                      Max.   :7700000   Max.   :33.000
##    bathrooms      sqft_living       sqft_lot           floors
##  Min.   :0.000   Min.   :  290   Min.   :    520   Min.   :1.000
##  1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040   1st Qu.:1.000
##  Median :2.250   Median : 1910   Median :   7618   Median :1.500
##  Mean   :2.115   Mean   : 2080   Mean   :  15107   Mean   :1.494
##  3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688   3rd Qu.:2.000
##  Max.   :8.000   Max.   :13540   Max.   :1651359   Max.   :3.500
##    waterfront            view          condition         grade
##  Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 1.000
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000
##  Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000
##  Mean   :0.007542   Mean   :0.2343   Mean   :3.409   Mean   : 7.657
##  3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000
##  Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000
##    sqft_above   sqft_basement       yr_built     yr_renovated
##  Min.   : 290   Min.   :   0.0   Min.   :1900   Min.   :   0.0
##  1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0
##  Median :1560   Median :   0.0   Median :1975   Median :   0.0
##  Mean   :1788   Mean   : 291.5   Mean   :1971   Mean   :  84.4
##  3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0
##  Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.0
##     zipcode           lat             long        sqft_living15
##  Min.   :98001   Min.   :47.16   Min.   :-122.5   Min.   : 399
##  1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490
##  Median :98065   Median :47.57   Median :-122.2   Median :1840
##  Mean   :98078   Mean   :47.56   Mean   :-122.2   Mean   :1987
##  3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360
```

```
##  Max.   :98199   Max.   :47.78   Max.   :-121.3   Max.   :6210
##    sqft_lot15
##  Min.   :   651
##  1st Qu.:  5100
##  Median :  7620
##  Mean   : 12768
##  3rd Qu.: 10083
##  Max.   :871200
```

```
## [1] "20141013T000000" "20141209T000000" "20150225T000000" "20141209T000000"
## [5] "20150218T000000" "20140512T000000"
```

```
## [1] "20141013"
```

Step 2: Divide Data into train and test sets, 70/30 split

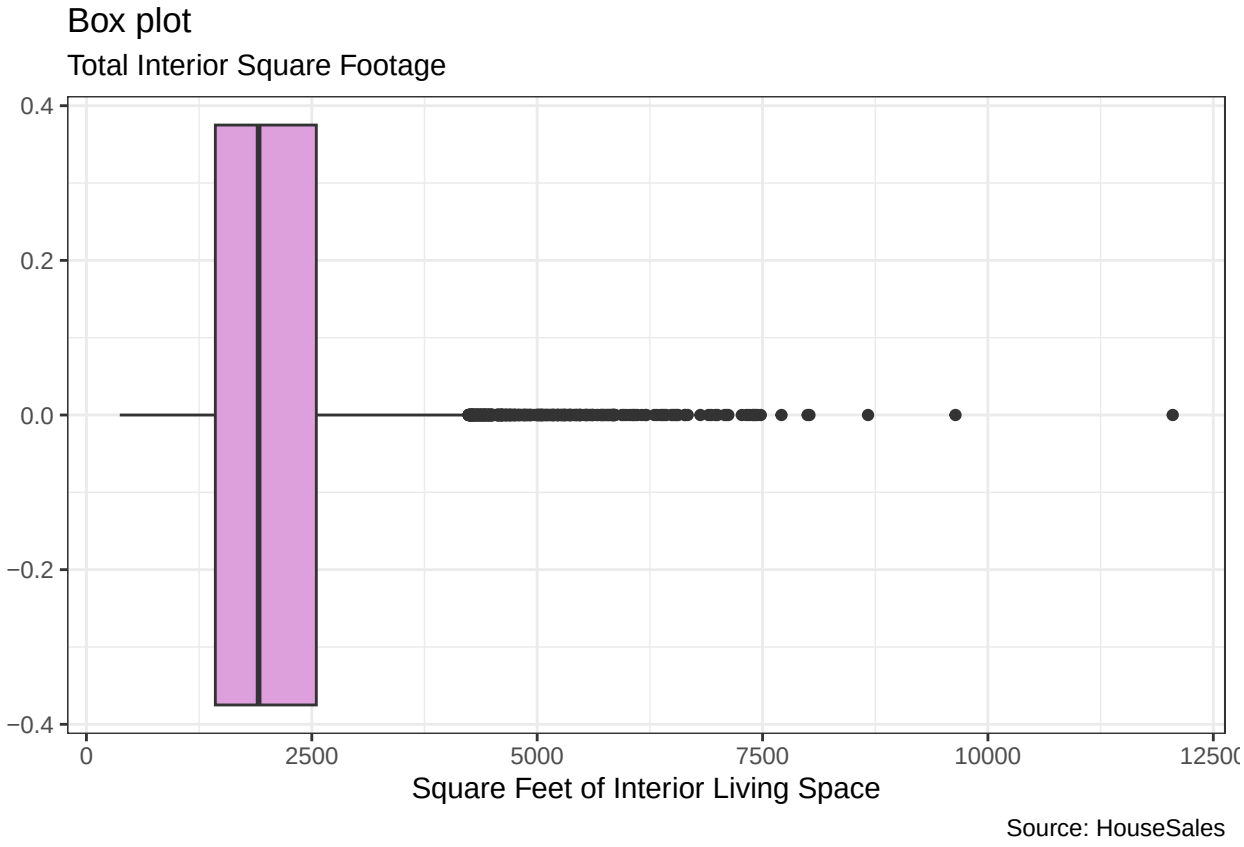
Step 3: Inspect Data and Drop variables, create categorical variables

```
#summary(H.train)

# Dropping unneeded variables
drop <- c("id","lat","long", "date", "zipcode", "sqft_basement")
df = H.train[,!(names(H.train) %in% drop)]
df.test = H.test[,!(names(H.test) %in% drop)]
summary(df)
```

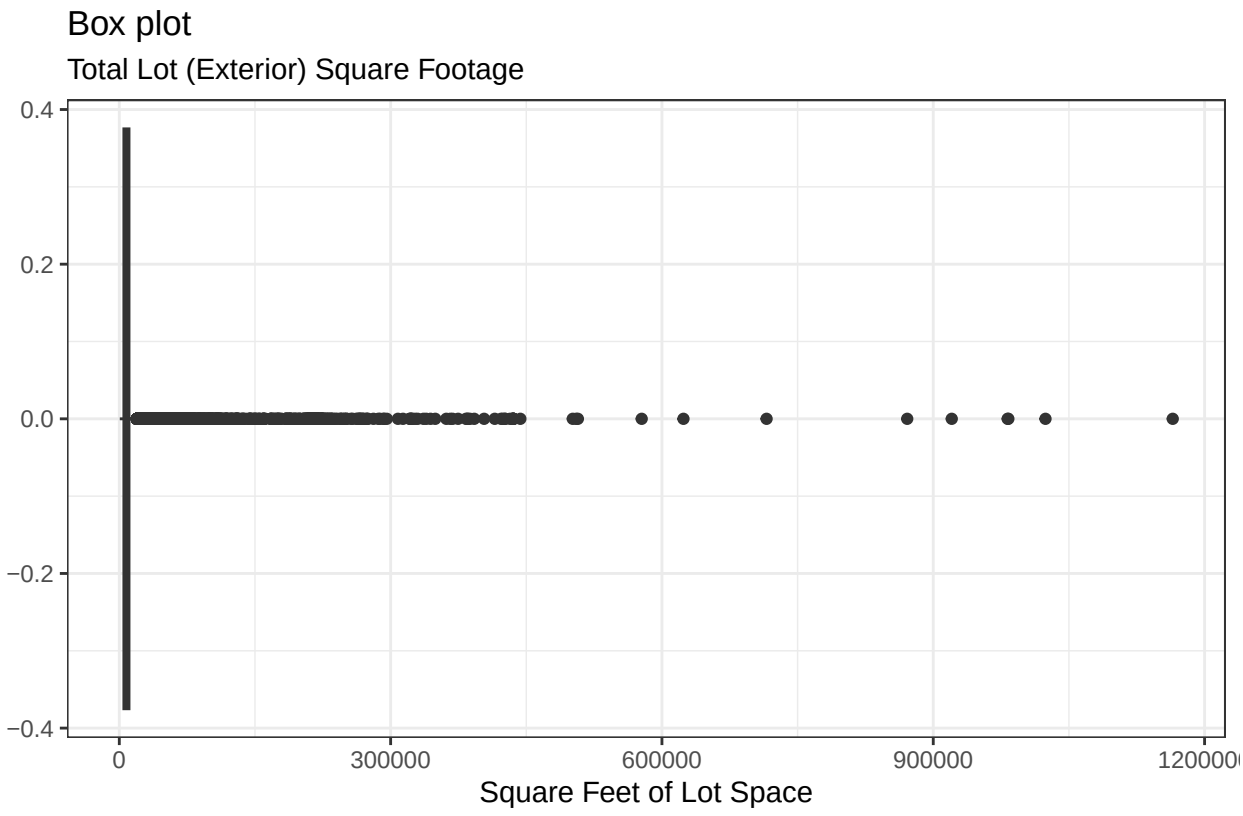
```
##      price            bedrooms        bathrooms      sqft_living
##  Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  370
##  1st Qu.: 320000   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1430
##  Median : 450000   Median : 3.000   Median :2.250   Median : 1910
##  Mean   : 538714   Mean   : 3.373   Mean   :2.117   Mean   : 2080
##  3rd Qu.: 641250   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550
##  Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.   :12050
##     sqft_lot           floors        waterfront           view
##  Min.   :    520   Min.   :1.000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.:   5027   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median :   7600   Median :1.500   Median :0.000000   Median :0.0000
##  Mean   :  14914   Mean   :1.498   Mean   :0.006808   Mean   :0.2298
##  3rd Qu.:  10735   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000
##  Max.   :1164794   Max.   :3.500   Max.   :1.000000   Max.   :4.0000
##    condition         grade          sqft_above      yr_built
##  Min.   :1.000   Min.   : 3.000   Min.   : 370   Min.   :1900
##  1st Qu.:3.000   1st Qu.: 7.000   1st Qu.:1190   1st Qu.:1951
##  Median :3.000   Median : 7.000   Median :1560   Median :1975
##  Mean   :3.413   Mean   : 7.657   Mean   :1787   Mean   :1971
##  3rd Qu.:4.000   3rd Qu.: 8.000   3rd Qu.:2217   3rd Qu.:1997
##  Max.   :5.000   Max.   :13.000   Max.   :8570   Max.   :2015
##   yr_renovated     sqft_living15    sqft_lot15
##  Min.   :   0.00   Min.   : 399   Min.   :   651
##  1st Qu.:   0.00   1st Qu.:1490   1st Qu.:  5100
##  Median :   0.00   Median :1840   Median :  7620
##  Mean   :  85.06   Mean   :1988   Mean   : 12728
##  3rd Qu.:   0.00   3rd Qu.:2370   3rd Qu.: 10100
##  Max.   :2015.00   Max.   :6110   Max.   :858132
```

```r
attach(df)
```

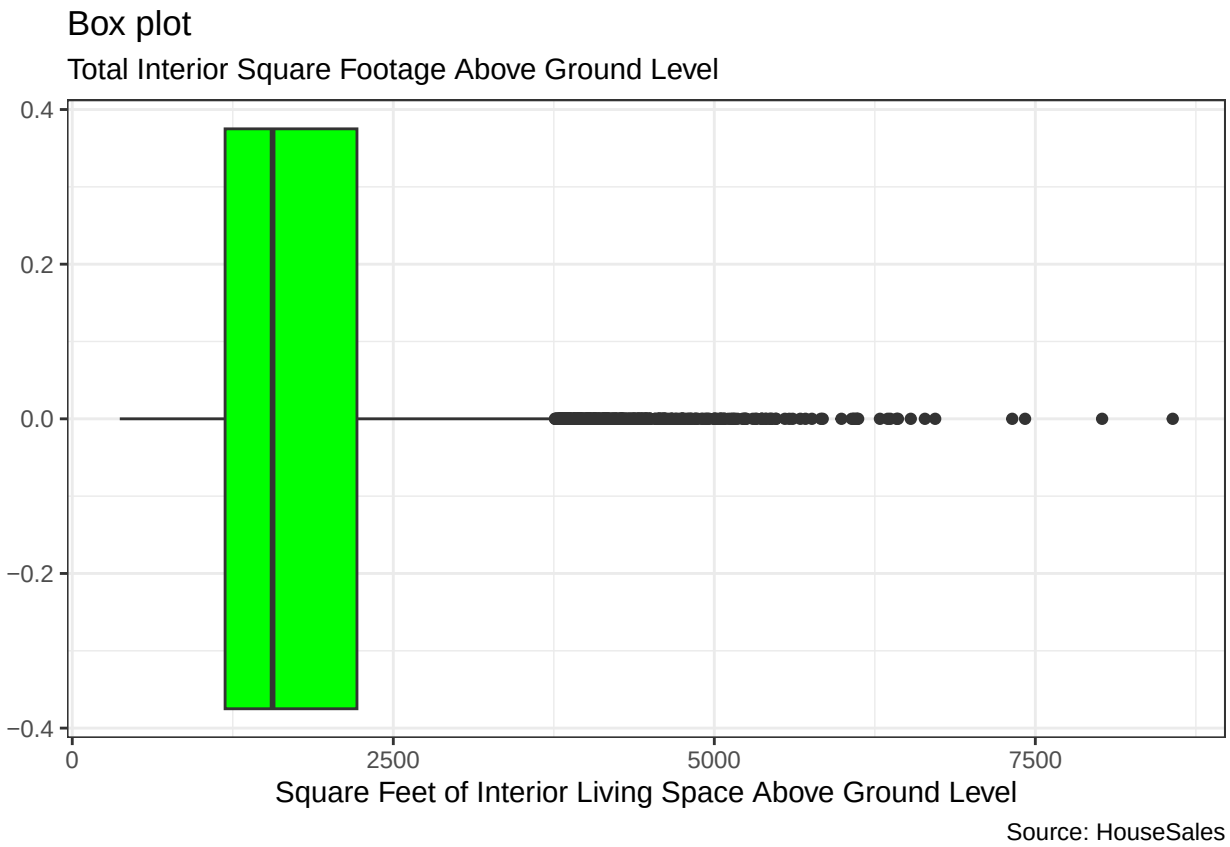
```r
# boxplots for starting reference
par(mfrow=c(2,2))
g1 <- ggplot(H.train, aes(sqft_living))
g1 + geom_boxplot(varwidth=T, fill="plum") +
    labs(title="Box plot",
         subtitle="Total Interior Square Footage",
         caption="Source: HouseSales",
         x="Square Feet of Interior Living Space")
```



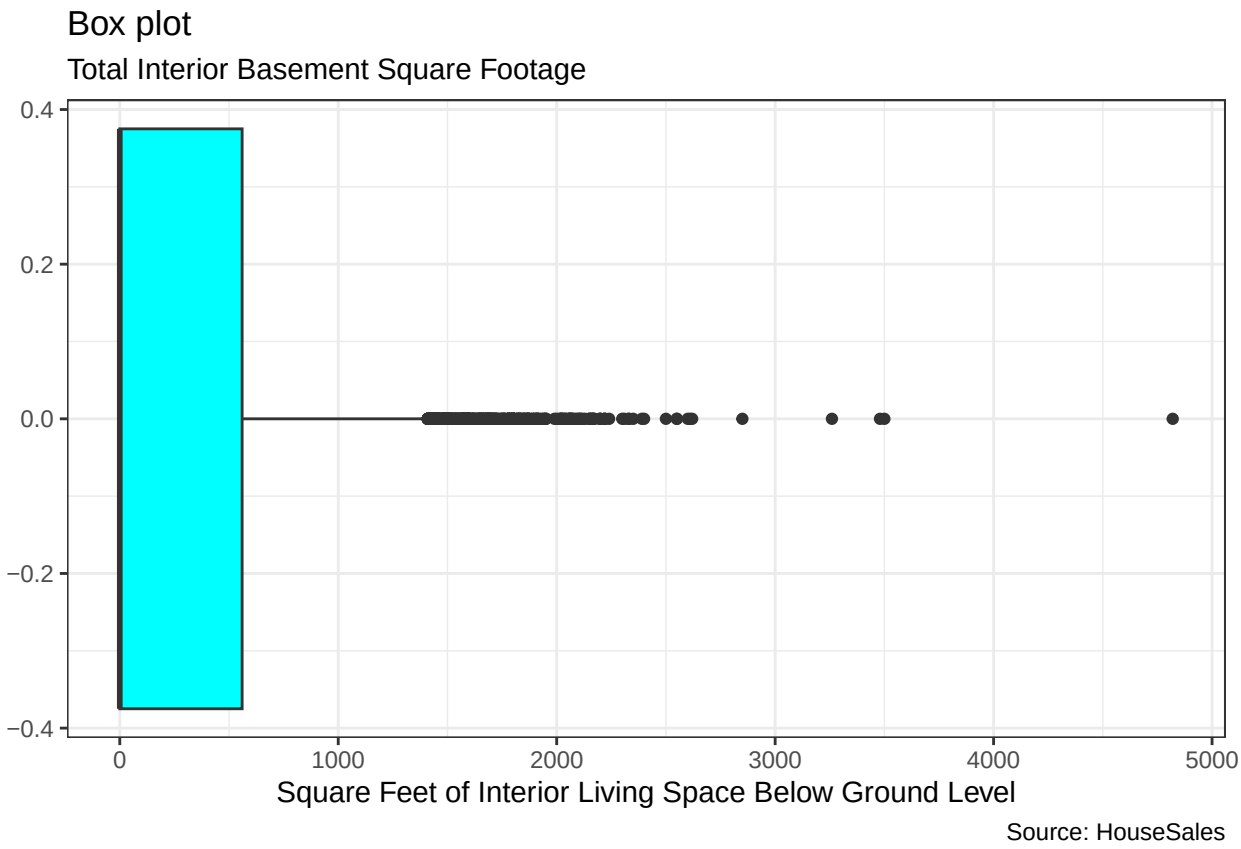

```r
g2 <- ggplot(H.train, aes(sqft_lot))
g2 + geom_boxplot(varwidth=T, fill="orange") +
    labs(title="Box plot",
         subtitle="Total Lot (Exterior) Square Footage",
         caption="Source: HouseSales",
         x="Square Feet of Lot Space")
```



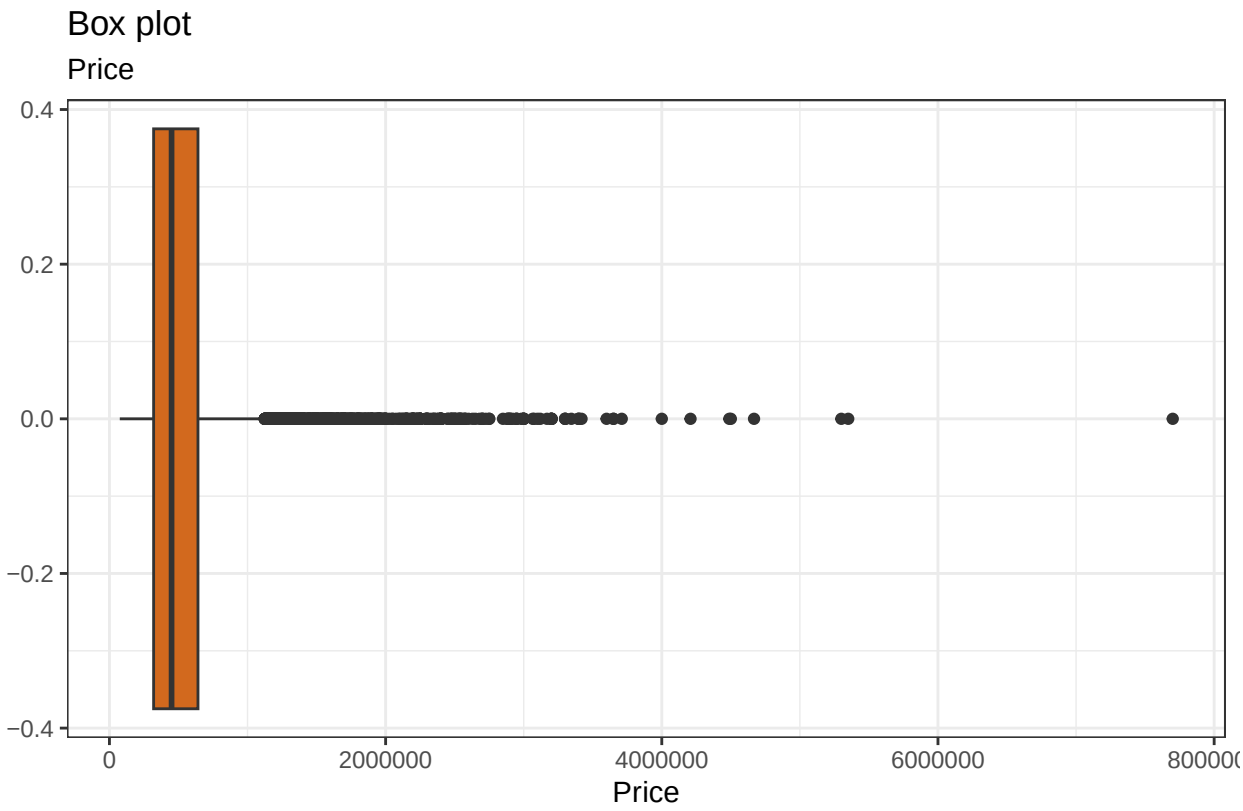

```
g3 <- ggplot(H.train, aes(sqft_above))
g3 + geom_boxplot(varwidth=T, fill="green") +
    labs(title="Box plot",
         subtitle="Total Interior Square Footage Above Ground Level",
         caption="Source: HouseSales",
         x="Square Feet of Interior Living Space Above Ground Level")
```



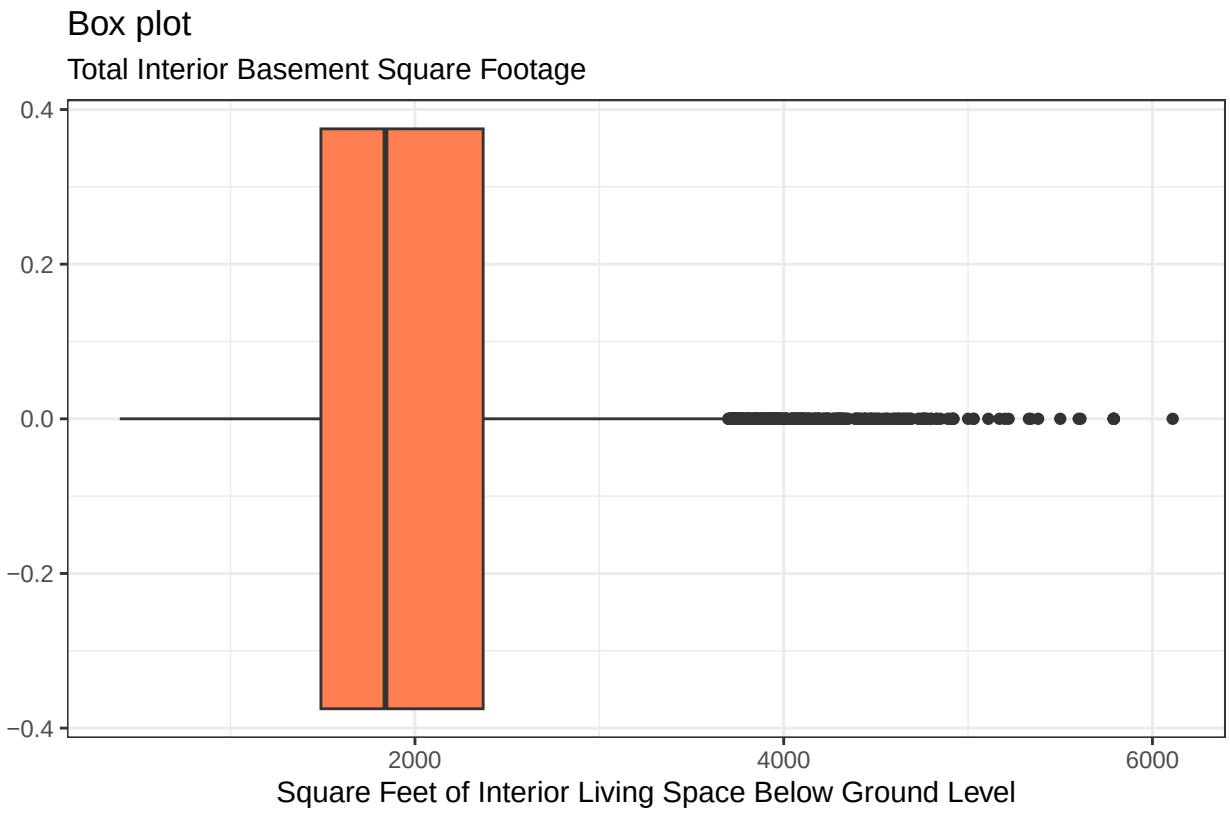

```
g4 <- ggplot(H.train, aes(sqft_basement))
g4 + geom_boxplot(varwidth=T, fill="cyan") +
    labs(title="Box plot",
         subtitle="Total Interior Basement Square Footage",
         caption="Source: HouseSales",
         x="Square Feet of Interior Living Space Below Ground Level")
```



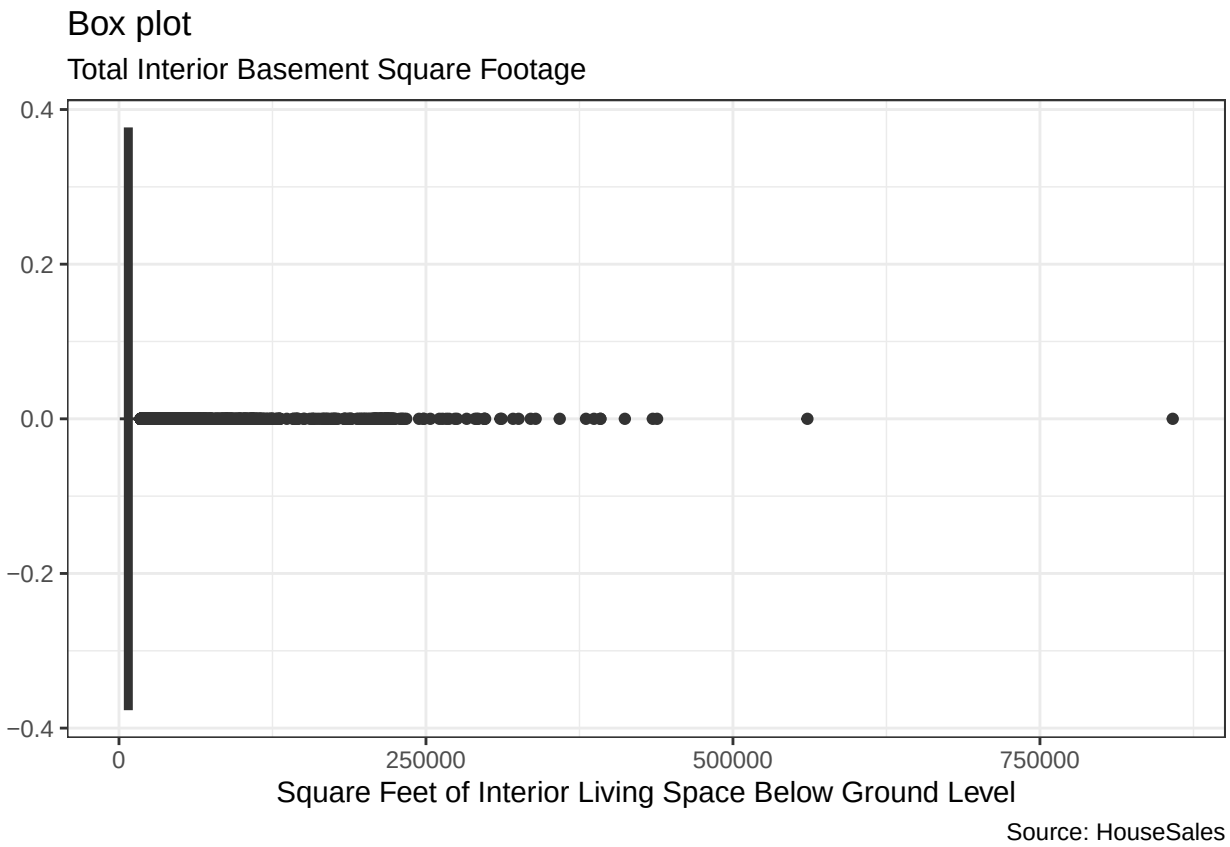

```
g5 <- ggplot(H.train, aes(price))
g5 + geom_boxplot(varwidth=T, fill="chocolate") +
    labs(title="Box plot",
         subtitle="Price",
         caption="Source: HouseSales",
         x="Price")
```




```
g5 <- ggplot(H.train, aes(sqft_living15))
g5 + geom_boxplot(varwidth=T, fill="coral") +
    labs(title="Box plot",
         subtitle="Total Interior Basement Square Footage",
         caption="Source: HouseSales",
         x="Square Feet of Interior Living Space Below Ground Level")
```




```
g6 <- ggplot(H.train, aes(sqft_lot15))
g6 + geom_boxplot(varwidth=T, fill="azure2") +
    labs(title="Box plot",
         subtitle="Total Interior Basement Square Footage",
         caption="Source: HouseSales",
         x="Square Feet of Interior Living Space Below Ground Level")
```



Step 4: Build regression model to predict price

```
summary(df)
```

```
##      price            bedrooms        bathrooms      sqft_living
##  Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  370
##  1st Qu.: 320000   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1430
##  Median : 450000   Median : 3.000   Median :2.250   Median : 1910
##  Mean   : 538714   Mean   : 3.373   Mean   :2.117   Mean   : 2080
##  3rd Qu.: 641250   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550
##  Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.   :12050
##     sqft_lot           floors        waterfront           view
##  Min.   :    520   Min.   :1.000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.:   5027   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median :   7600   Median :1.500   Median :0.000000   Median :0.0000
##  Mean   :  14914   Mean   :1.498   Mean   :0.006808   Mean   :0.2298
##  3rd Qu.:  10735   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000
##  Max.   :1164794   Max.   :3.500   Max.   :1.000000   Max.   :4.0000
##    condition         grade          sqft_above      yr_built
##  Min.   :1.000   Min.   : 3.000   Min.   : 370   Min.   :1900
##  1st Qu.:3.000   1st Qu.: 7.000   1st Qu.:1190   1st Qu.:1951
##  Median :3.000   Median : 7.000   Median :1560   Median :1975
##  Mean   :3.413   Mean   : 7.657   Mean   :1787   Mean   :1971
##  3rd Qu.:4.000   3rd Qu.: 8.000   3rd Qu.:2217   3rd Qu.:1997
##  Max.   :5.000   Max.   :13.000   Max.   :8570   Max.   :2015
##   yr_renovated    sqft_living15    sqft_lot15
```

```
##  Min.   :   0.00   Min.   : 399   Min.   :   651
##  1st Qu.:   0.00   1st Qu.:1490   1st Qu.:  5100
##  Median :   0.00   Median :1840   Median :  7620
##  Mean   :  85.06   Mean   :1988   Mean   : 12728
##  3rd Qu.:   0.00   3rd Qu.:2370   3rd Qu.: 10100
##  Max.   :2015.00   Max.   :6110   Max.   :858132
```

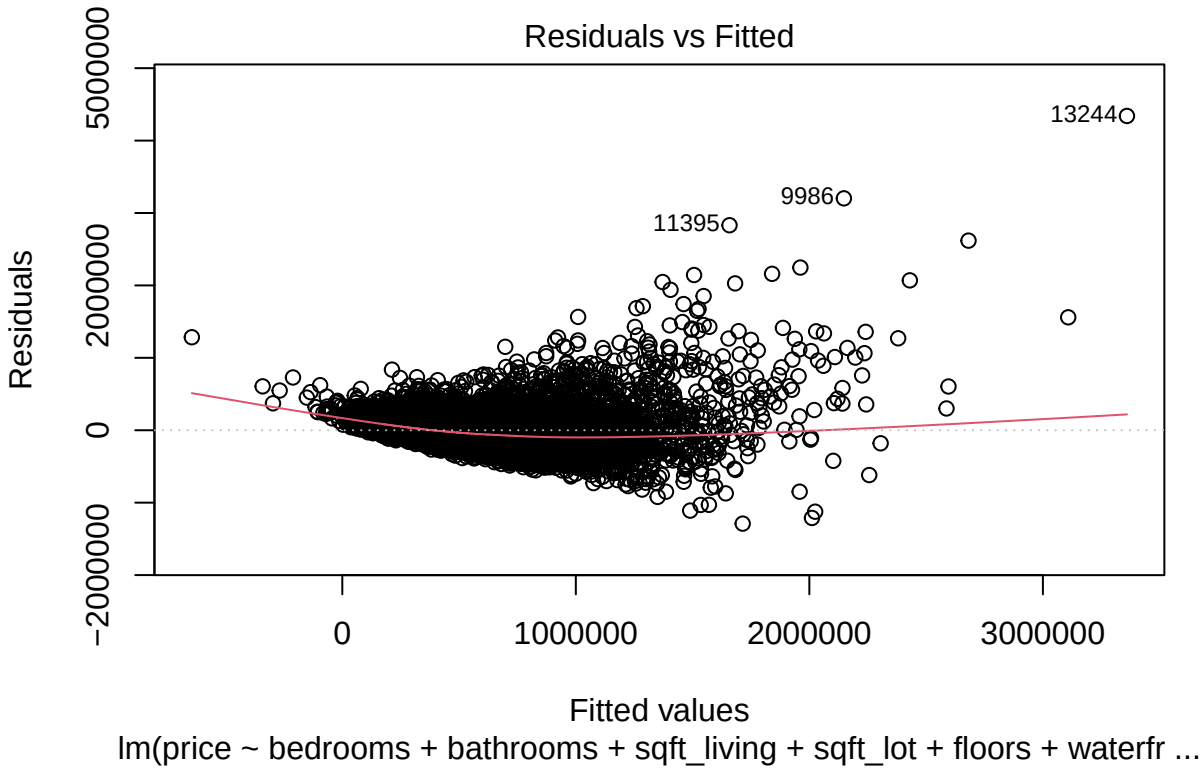
```
class(df)
```

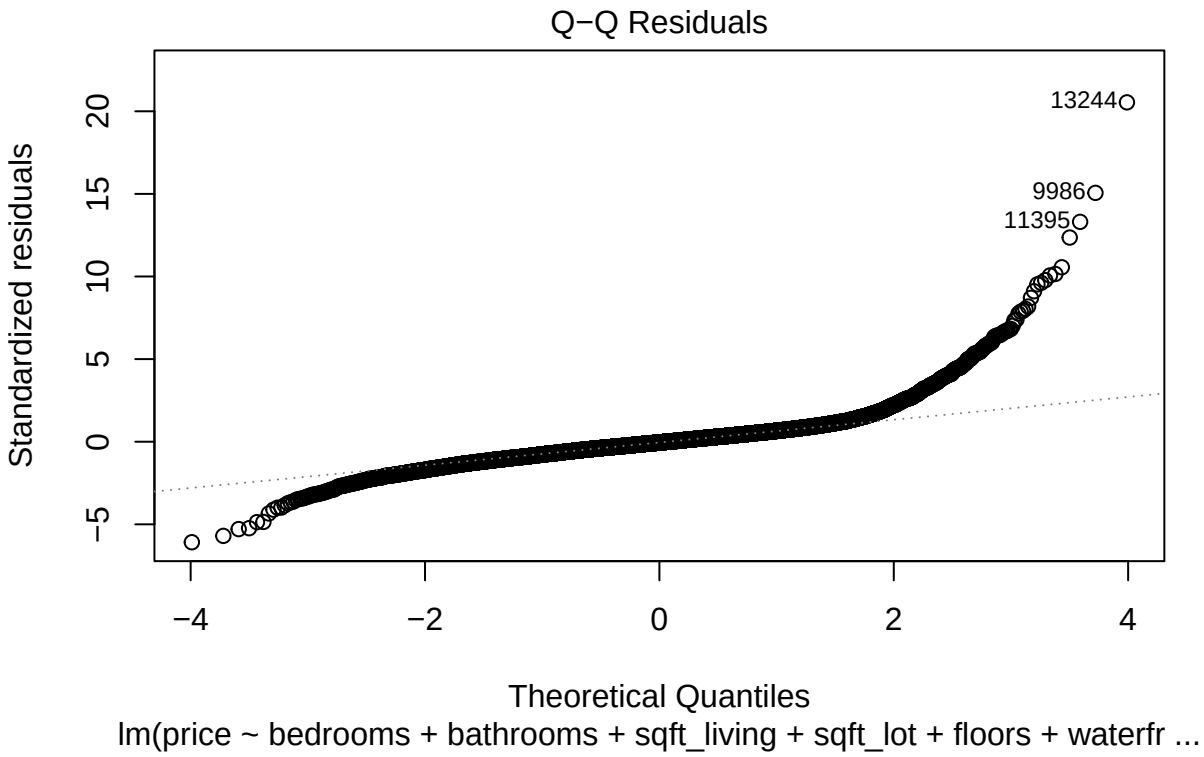
```
## [1] "data.frame"
```

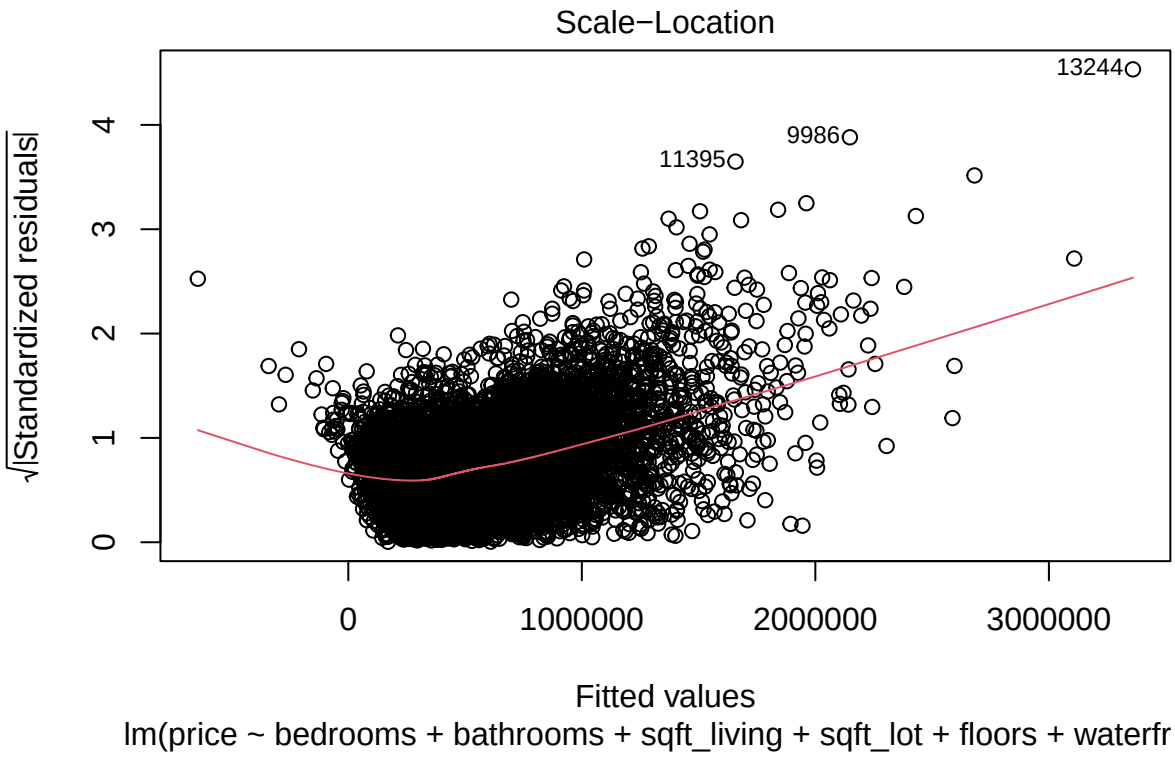
```
#H.full <- lm(price ~ ., data=df)
H.full <- lm(price ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors + waterfront + view + condi
summary(H.full)
```

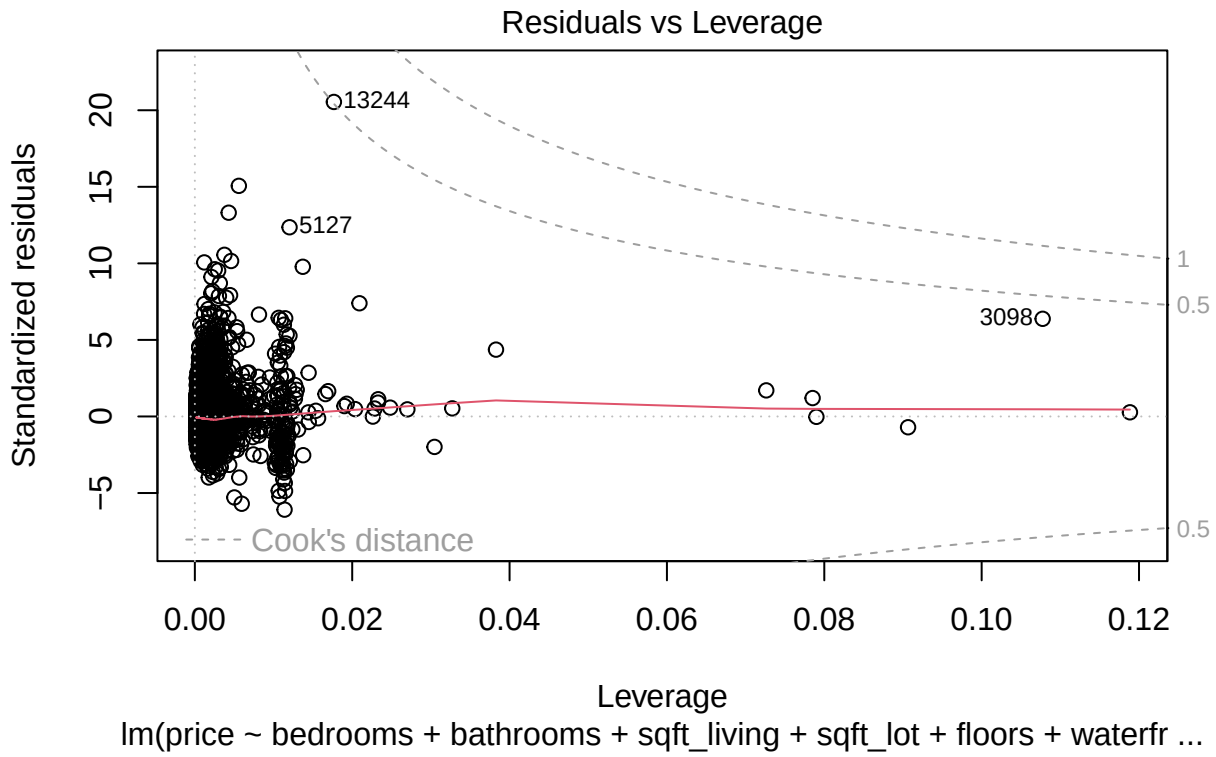
```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
##     floors + waterfront + view + condition + grade + sqft_above +
##     yr_built + yr_renovated + sqft_living15 + sqft_lot15)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1289637  -108762    -9857    89075  4339983
##
## Coefficients:
##                   Estimate    Std. Error t value Pr(>|t|)
## (Intercept)   6076824.66465  162760.31078  37.336  < 2e-16 ***
## bedrooms       -36840.55742    2348.67803 -15.686  < 2e-16 ***
## bathrooms       40944.31964    4078.87382  10.038  < 2e-16 ***
## sqft_living       170.77978       5.49225  31.095  < 2e-16 ***
## sqft_lot           -0.00298       0.06367  -0.047   0.9627
## floors          37359.24731    4476.83220   8.345  < 2e-16 ***
## waterfront     524005.87561   22924.85295  22.858  < 2e-16 ***
## view            45917.98162    2700.16253  17.006  < 2e-16 ***
## condition       20741.26066    2934.06263   7.069 1.63e-12 ***
## grade          118396.04869    2643.54460  44.787  < 2e-16 ***
## sqft_above        -21.32162       5.35033  -3.985 6.78e-05 ***
## yr_built        -3515.80565      83.47793 -42.117  < 2e-16 ***
## yr_renovated       15.54184       4.58403   3.390   0.0007 ***
## sqft_living15      34.68791       4.27199   8.120 5.03e-16 ***
## sqft_lot15         -0.45859       0.09213  -4.978 6.51e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 213200 on 15114 degrees of freedom
## Multiple R-squared:  0.6529, Adjusted R-squared:  0.6526
## F-statistic:  2031 on 14 and 15114 DF,  p-value: < 2.2e-16
```

```
plot(H.full)
```


Residuals vs Fitted
Residuals
Fitted values
lm(price ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors + waterfr ...
13244
9986
11395
5000000
2000000
0
−2000000
0
1000000
2000000
3000000


Q−Q Residuals
Standardized residuals
−5
0
5
10
15
20
13244
9986
11395
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors + waterfr ...


Scale−Location
13244
9986
11395
√|Standardized residuals|
0
1
2
3
4
0
1000000
2000000
3000000
Fitted values
lm(price ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors + waterfr ...



Step 5: Creating scatter plots and correlation matrix for the train data + result interpretation

Interpretation:

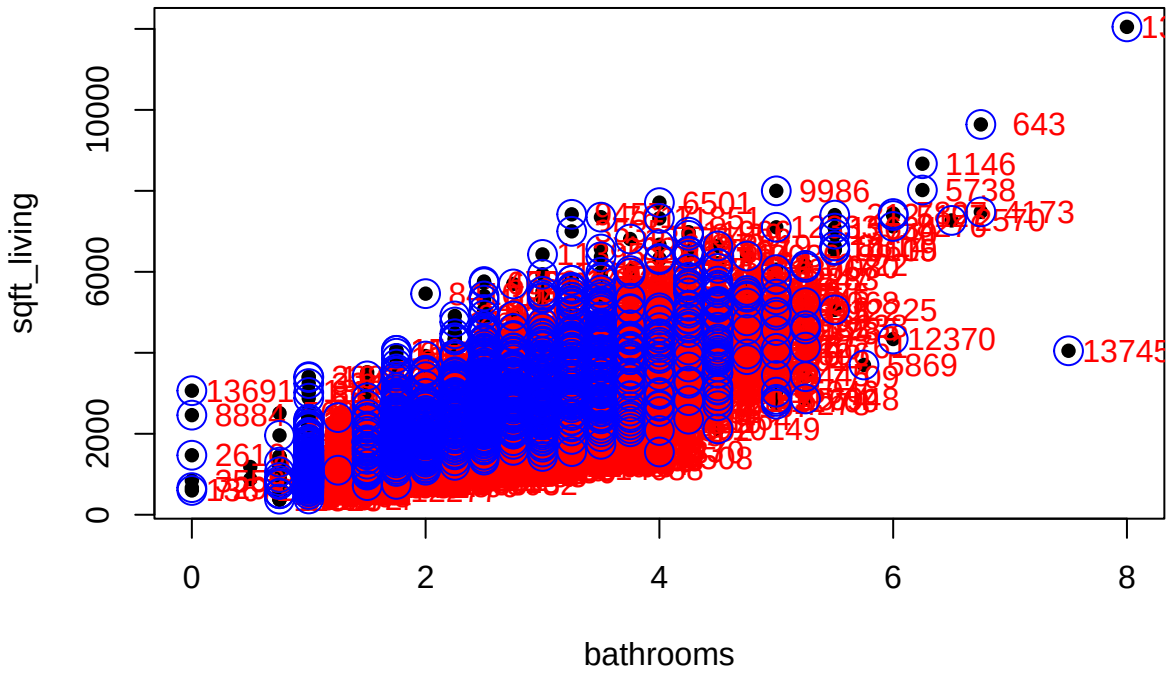
Scatter Plots:

Correlation Matrix: The matrix shows some issues of multi-collinearity are appearing with several of the variables. We can see that sqft_living is strongly correlated with sqft_above (0.87). This makes sense as these two values are closely related with each other. Usually much of the livable space of a house is also located above ground level, hence the relationship. Additional string correlations are seen between sqft_above & grade(0.76), sqft_above & sqft_above15(0.76), sqft_living & price(0.7), as well as several others. As a result of these correlations, we will have to address issues of multi-collinearity within our model.

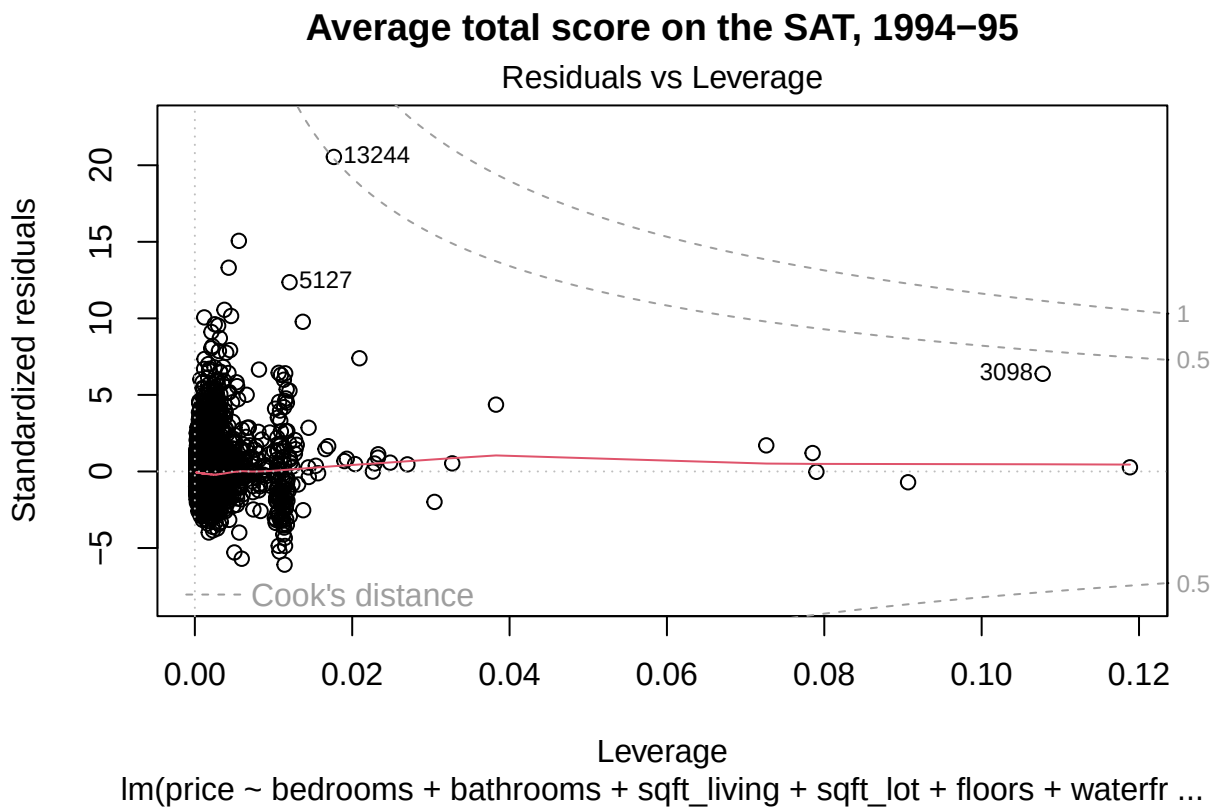
VIF helps determine the severity of multicollinearity. Previous concerns with sqft_living and sqft_above seem to hold some weight as these two predictors have the heighest VIFs, though still below our decision point of $VIF < 10$.

```
# Scatter plots needed: Leverages vs. Index,  Residuals vs Leverage, hat matrix, Studentized Residuals
n <- nrow(df)
p <- length(H.full$coefficients)
hii<- hatvalues(H.full)


# Hat Matrix
plot(sqft_living~bathrooms, pch=16 )
text(bathrooms +0.5, sqft_living,
```

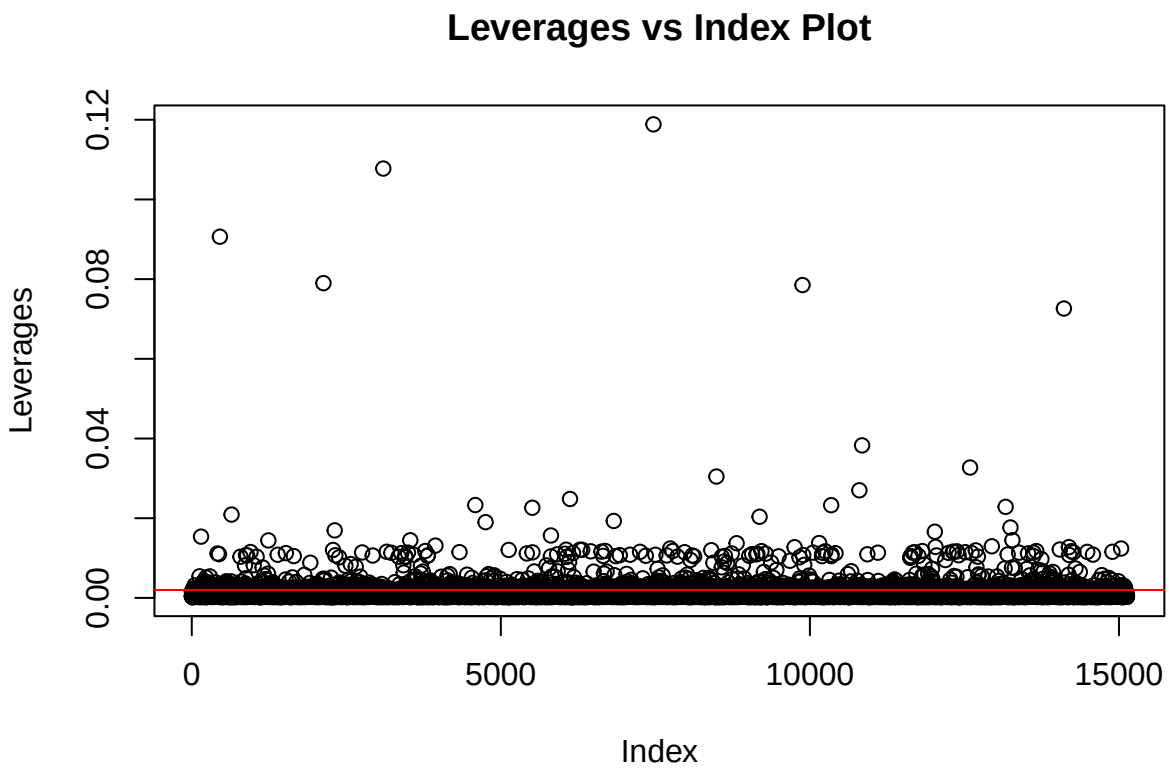
```
labels=as.character(1:length(bathrooms)),col="red")
index<-hii>2*p/n
points( bathrooms[index], sqft_living[index], cex=2.0, col="blue")
```



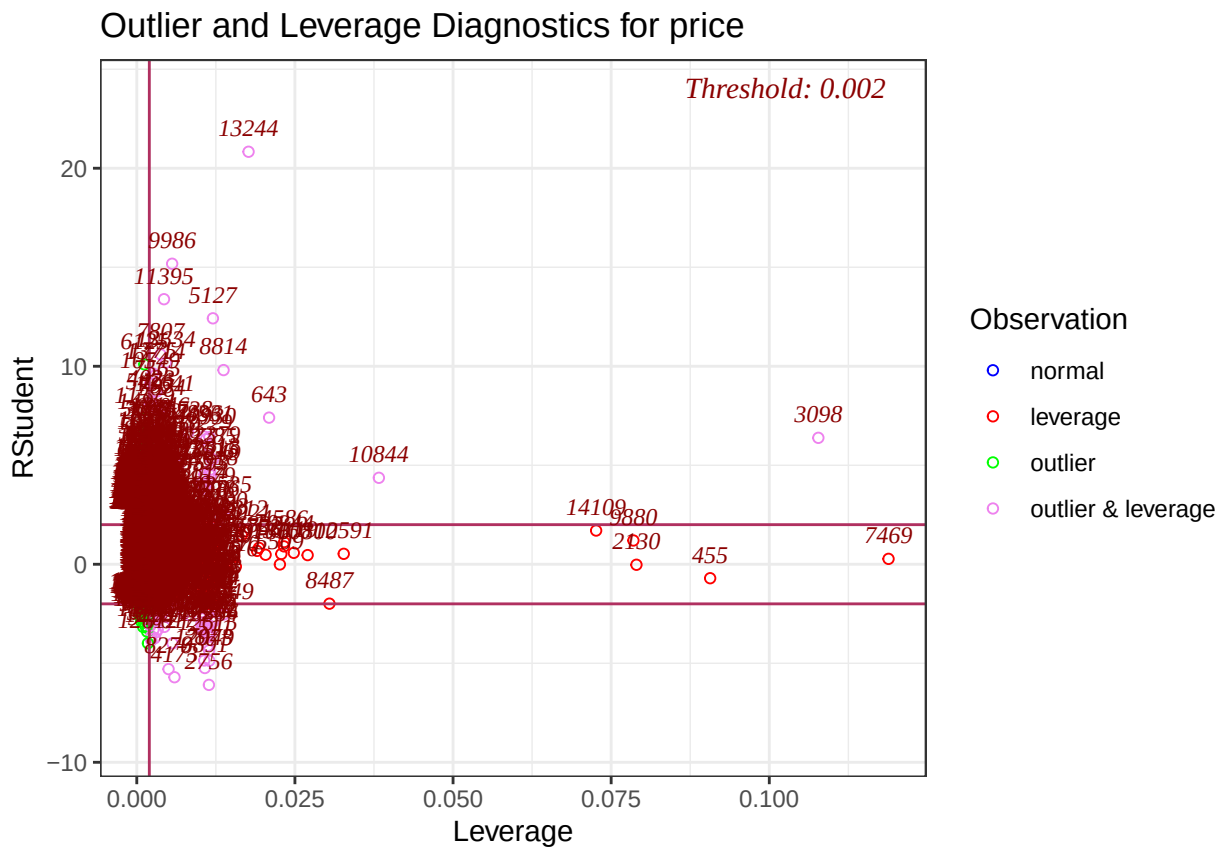

```
# Residuals vs. Leverage
plot(H.full, which=5, main='Average total score on the SAT, 1994-95')
```



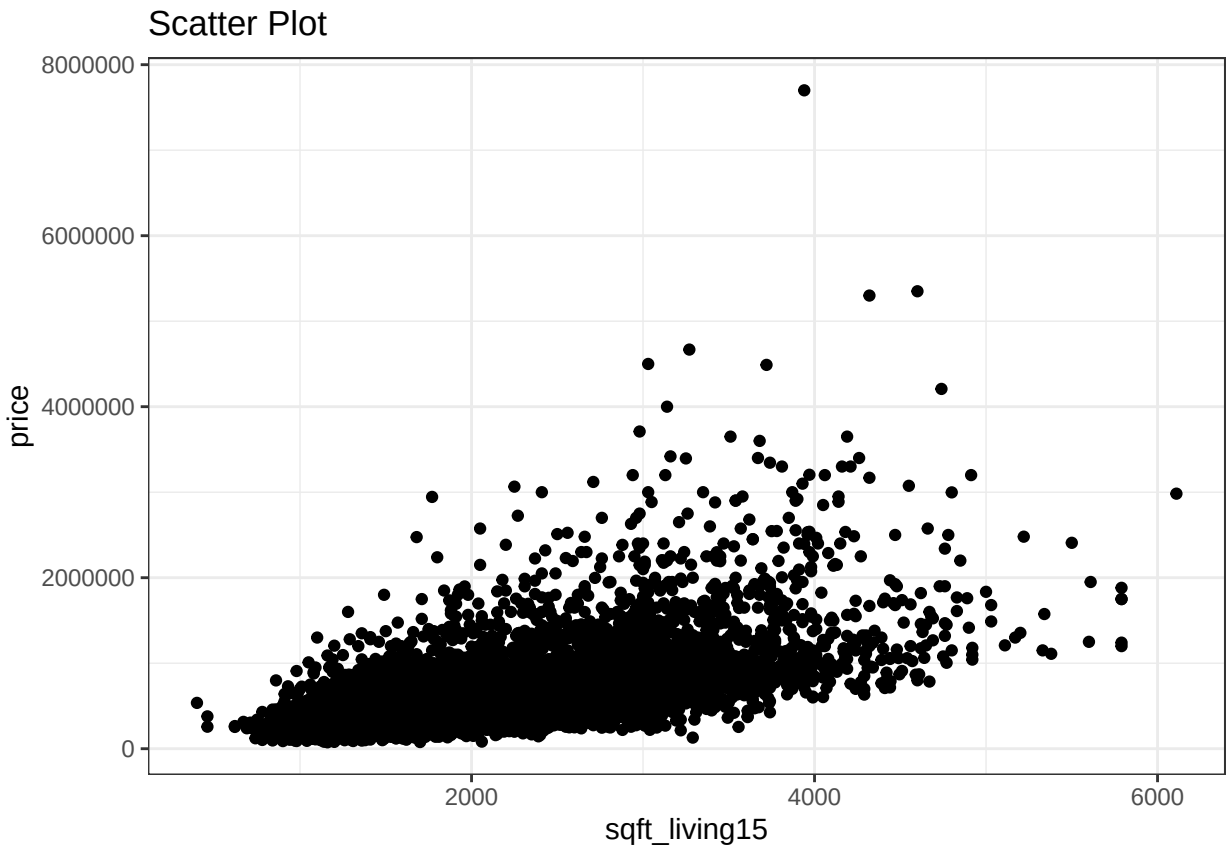

```r
# Leverages vs. Index

plot(hii, ylab="Leverages",main="Leverages vs Index Plot")
abline(h=2*p/n,col="red")
```



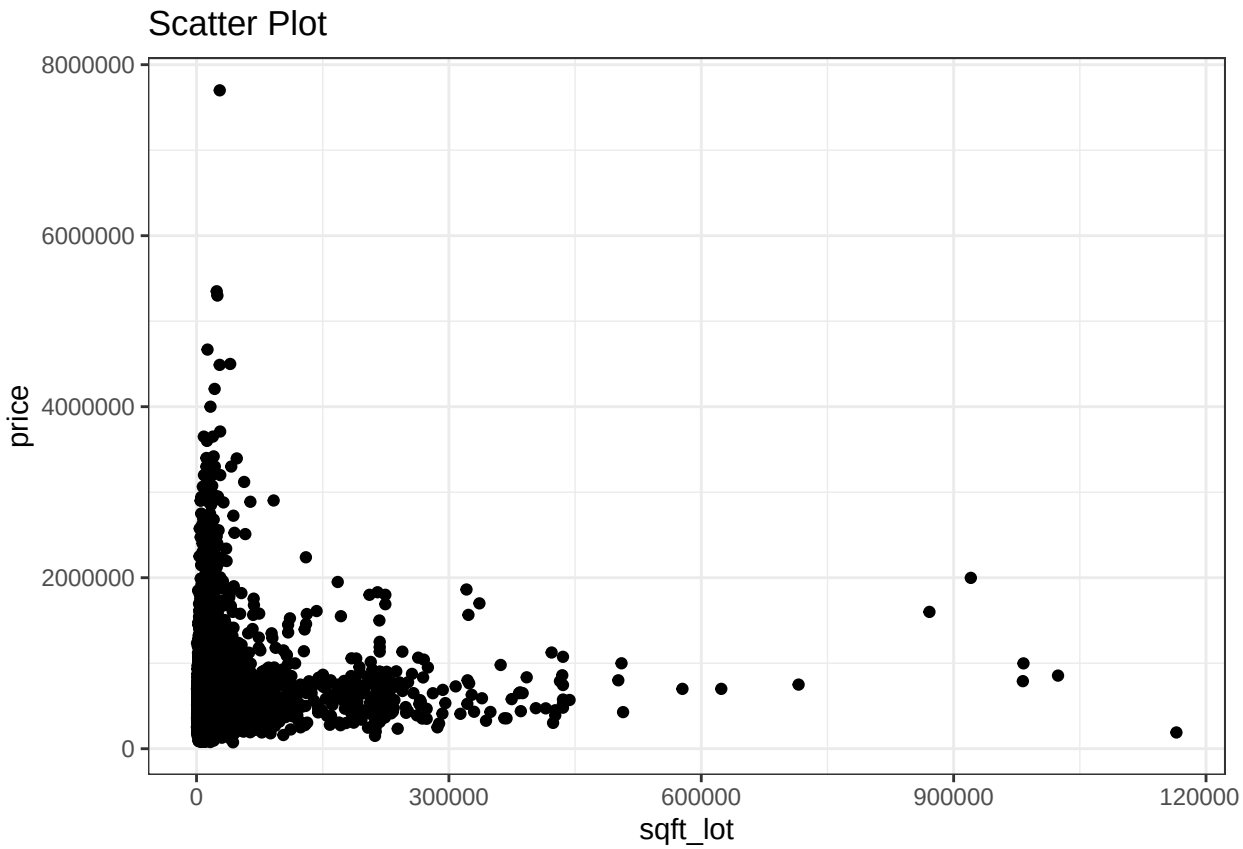

```
####Studentized Residuals vs Leverage Plot ####
#Graph for detecting influential observations.

ols_plot_resid_lev(H.full)
```



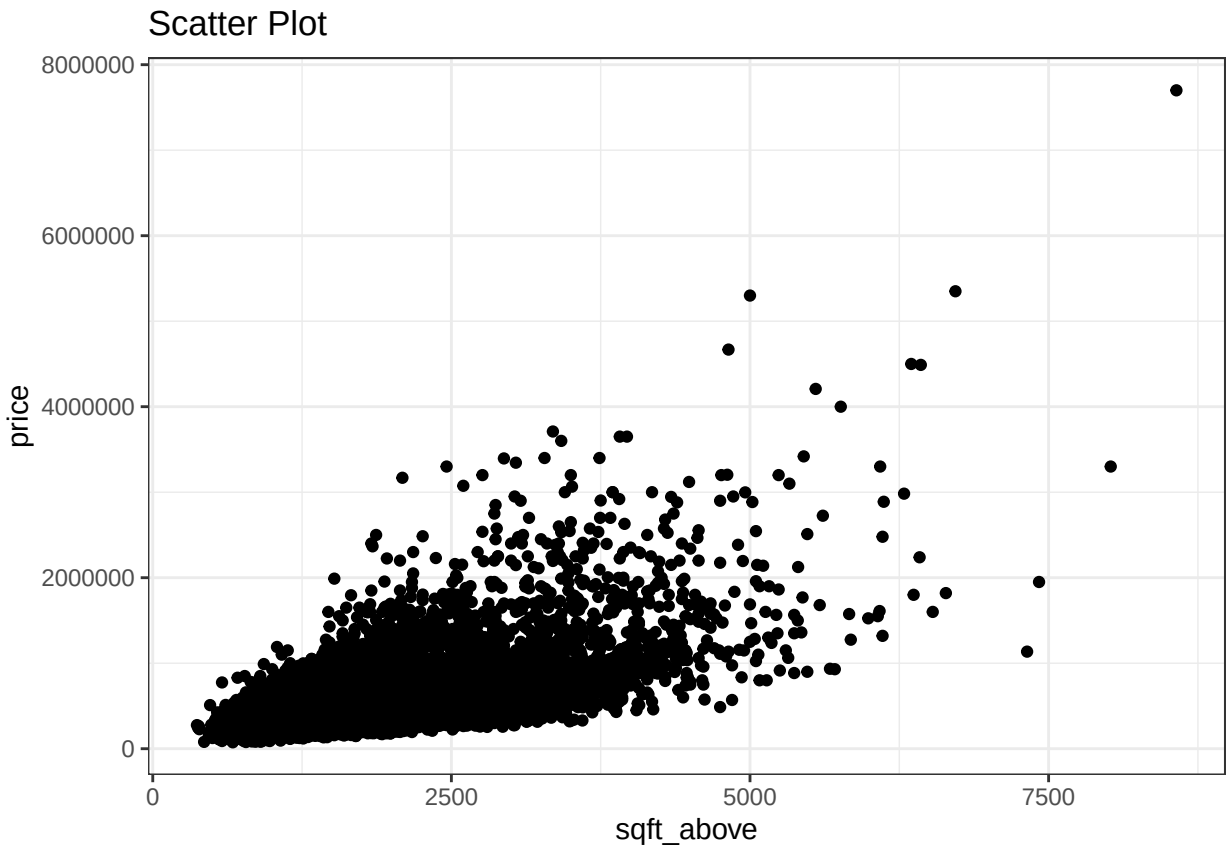

```
# Scatter plots
ggplot(df, aes(sqft_living15, price)) +
   geom_point() +
   ggtitle("Scatter Plot")
```



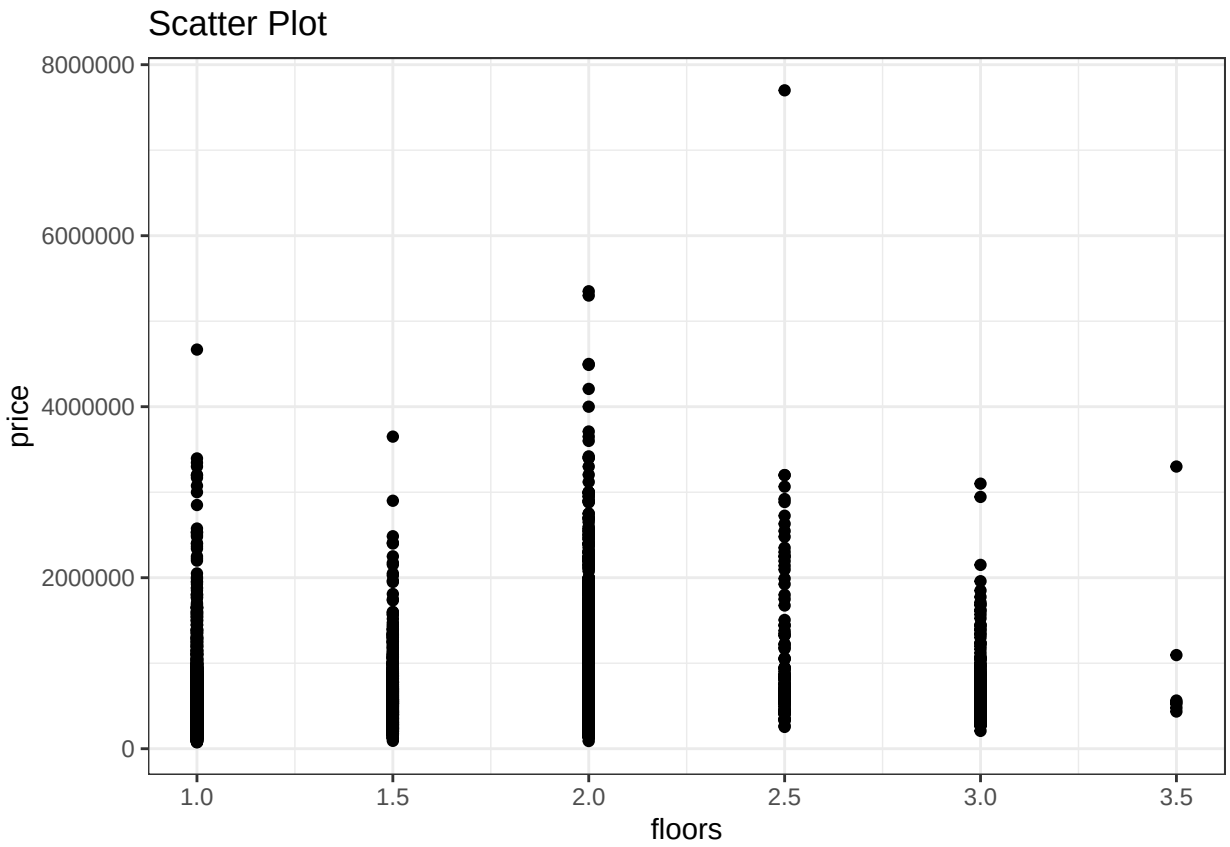

```
ggplot(df, aes(sqft_lot, price)) +
   geom_point() +
   ggtitle("Scatter Plot")
```



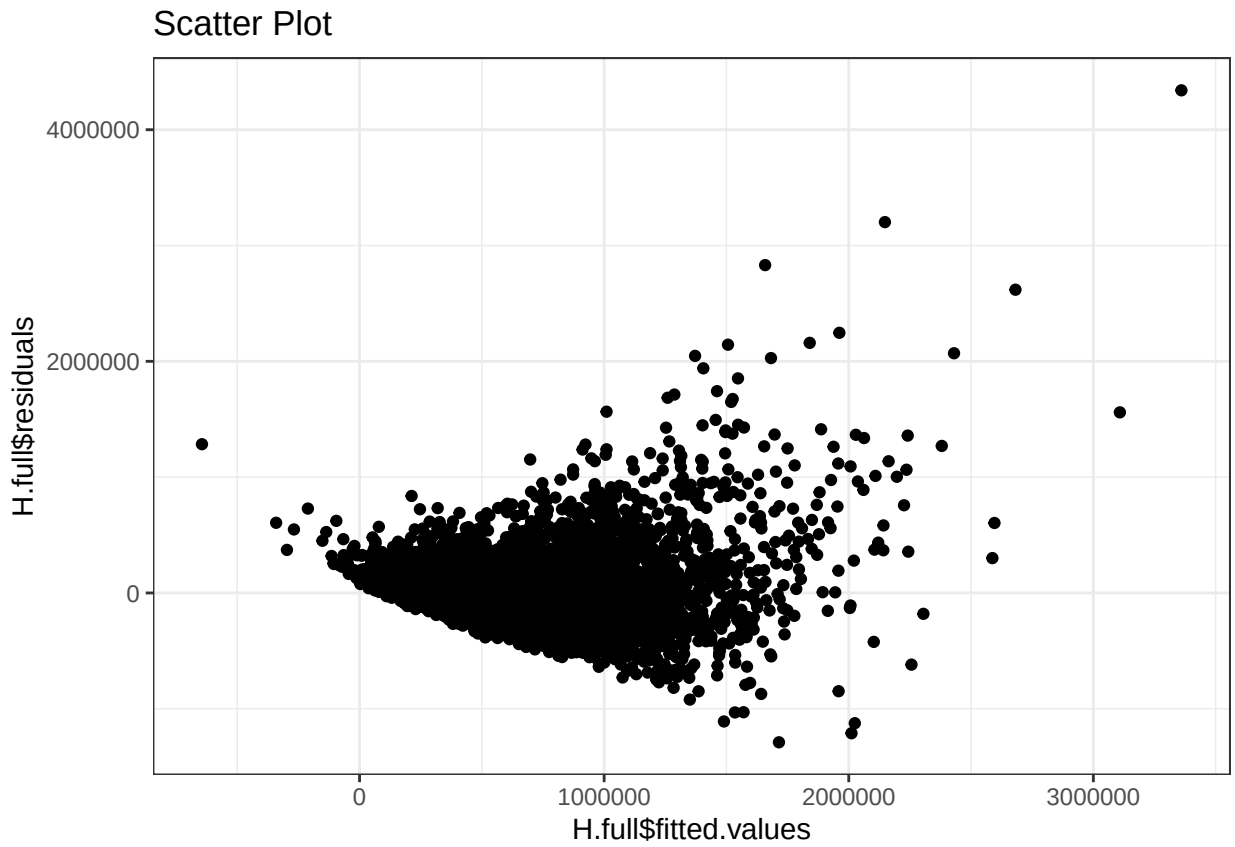

```
ggplot(df, aes(sqft_above, price)) +
   geom_point() +
   ggtitle("Scatter Plot")
```




```
ggplot(df, aes(floors, price)) +
   geom_point() +
   ggtitle("Scatter Plot")
```




```
ggplot(df, aes(H.full$fitted.values, H.full$residuals)) +
   geom_point() +
   ggtitle("Scatter Plot")
```

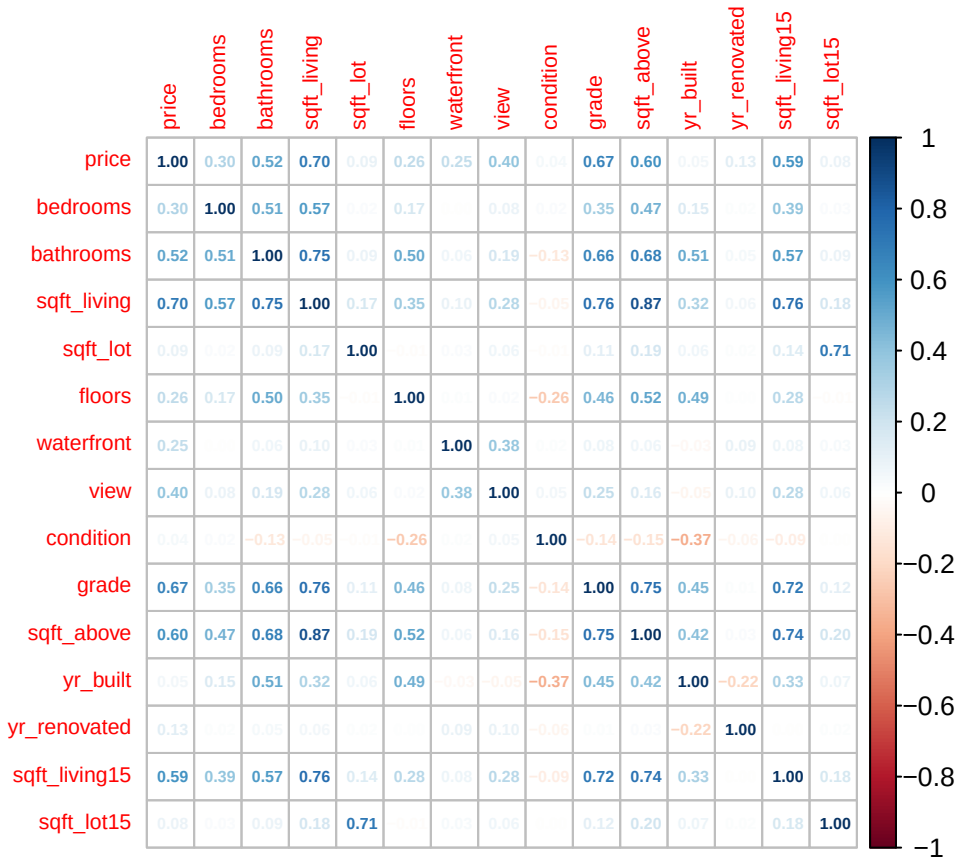



```
# Correlation Matrix
cor_matrix <- round(cor(df),3)

cor_matrix
```

```
##                price bedrooms bathrooms sqft_living sqft_lot floors waterfront
## price          1.000    0.305     0.520       0.702    0.090  0.257      0.246
## bedrooms       0.305    1.000     0.508       0.570    0.022  0.173     -0.001
## bathrooms      0.520    0.508     1.000       0.749    0.087  0.500      0.061
## sqft_living    0.702    0.570     0.749       1.000    0.172  0.351      0.096
## sqft_lot       0.090    0.022     0.087       0.172    1.000 -0.012      0.027
## floors         0.257    0.173     0.500       0.351   -0.012  1.000      0.010
## waterfront     0.246   -0.001     0.061       0.096    0.027  0.010      1.000
## view           0.396    0.083     0.189       0.285    0.065  0.022      0.385
## condition      0.044    0.024    -0.127      -0.054   -0.013 -0.264      0.018
## grade          0.667    0.349     0.659       0.759    0.112  0.457      0.078
## sqft_above     0.600    0.473     0.680       0.874    0.186  0.523      0.063
## yr_built       0.050    0.153     0.507       0.315    0.060  0.488     -0.031
## yr_renovated   0.128    0.016     0.048       0.056    0.018  0.003      0.090
## sqft_living15  0.593    0.390     0.568       0.762    0.145  0.279      0.077
## sqft_lot15     0.084    0.025     0.086       0.181    0.713 -0.012      0.030
##                  view condition  grade sqft_above yr_built yr_renovated
## price           0.396     0.044  0.667      0.600    0.050        0.128
## bedrooms        0.083     0.024  0.349      0.473    0.153        0.016
## bathrooms       0.189    -0.127  0.659      0.680    0.507        0.048
## sqft_living     0.285    -0.054  0.759      0.874    0.315        0.056
```

```
## sqft_lot       0.065    -0.013  0.112      0.186    0.060        0.018
## floors         0.022    -0.264  0.457      0.523    0.488        0.003
## waterfront     0.385     0.018  0.078      0.063   -0.031        0.090
## view           1.000     0.052  0.246      0.160   -0.054        0.096
## condition      0.052     1.000 -0.143     -0.152   -0.367       -0.059
## grade          0.246    -0.143  1.000      0.752    0.447        0.013
## sqft_above     0.160    -0.152  0.752      1.000    0.418        0.025
## yr_built      -0.054    -0.367  0.447      0.418    1.000       -0.220
## yr_renovated   0.096    -0.059  0.013      0.025   -0.220        1.000
## sqft_living15  0.279    -0.091  0.715      0.735    0.329       -0.001
## sqft_lot15     0.065    -0.005  0.119      0.196    0.075        0.019
##               sqft_living15 sqft_lot15
## price                 0.593      0.084
## bedrooms              0.390      0.025
## bathrooms             0.568      0.086
## sqft_living           0.762      0.181
## sqft_lot              0.145      0.713
## floors                0.279     -0.012
## waterfront            0.077      0.030
## view                  0.279      0.065
## condition            -0.091     -0.005
## grade                 0.715      0.119
## sqft_above            0.735      0.196
## yr_built              0.329      0.075
## yr_renovated         -0.001      0.019
## sqft_living15         1.000      0.178
## sqft_lot15            0.178      1.000
```

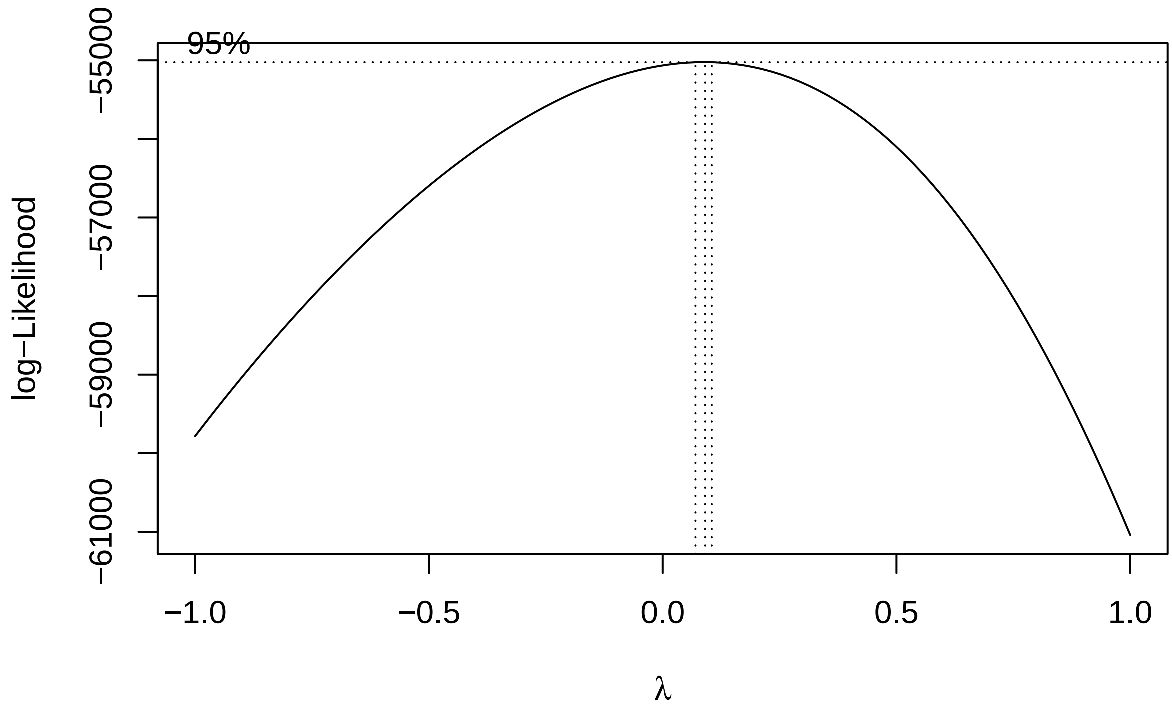
```
corrplot(cor_matrix, method="number", addCoef.col = 1,number.cex = 0.5, tl.cex = 0.7)
```




```
# Supporting VIF for multicollinearity
vif(H.full)
```

```
##      bedrooms     bathrooms   sqft_living      sqft_lot        floors
##      1.621672      3.283748      8.360479      2.055772      1.942427
##    waterfront          view     condition         grade    sqft_above
##      1.183095      1.387480      1.225034      3.199261      6.462112
##      yr_built  yr_renovated sqft_living15    sqft_lot15
##      2.007975      1.136626      2.879043      2.077777
```

```
boxcox(H.full,seq(-1,1,0.1))
```



Step 6: Build the best multiple linear models by using the stepwise selection method. Compare the performance of the best two linear models.

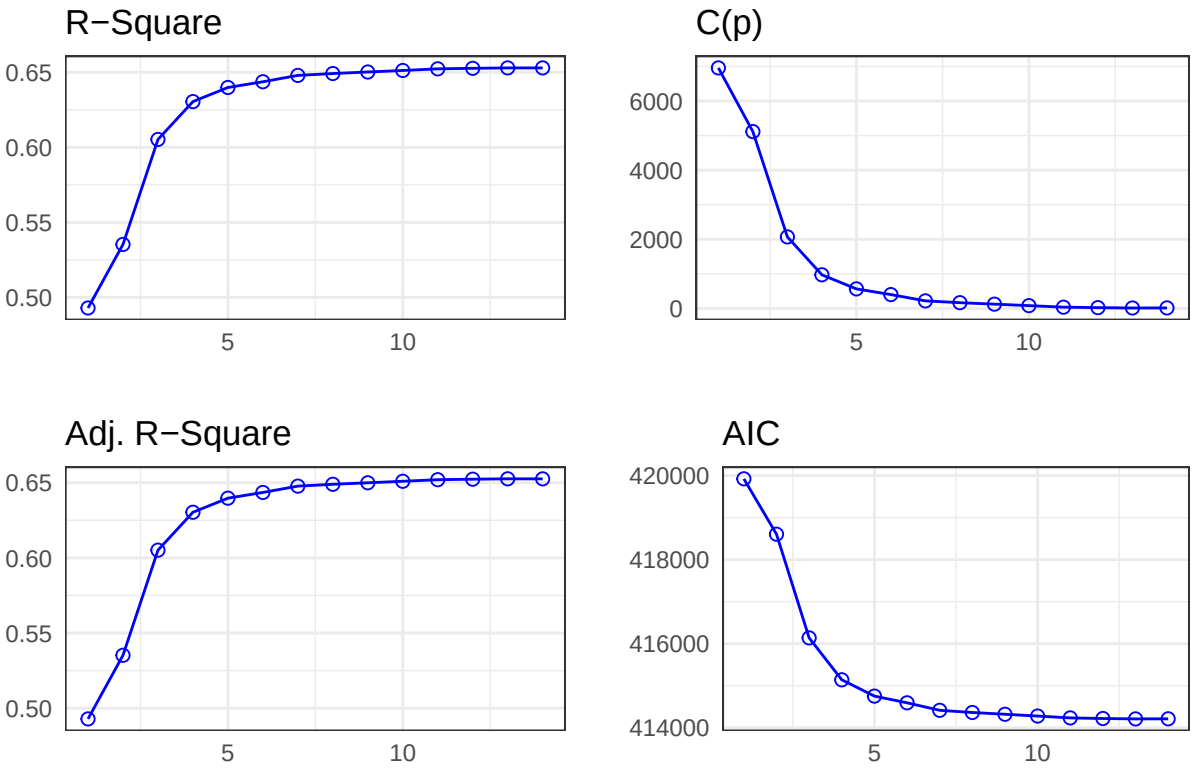
First model is made using the best subset selection method

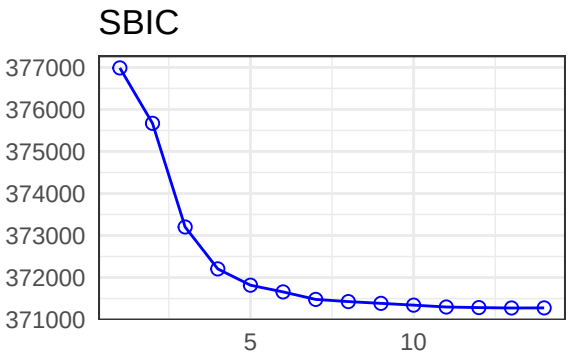
```
# Step 1: Variable validation using: best_subset,

# BE CAREFUL WHEN RUNNING THIS BLOCK AS IT IS COMPUTE INTENSIVE
k1<-ols_step_best_subset(H.full)
k1
```

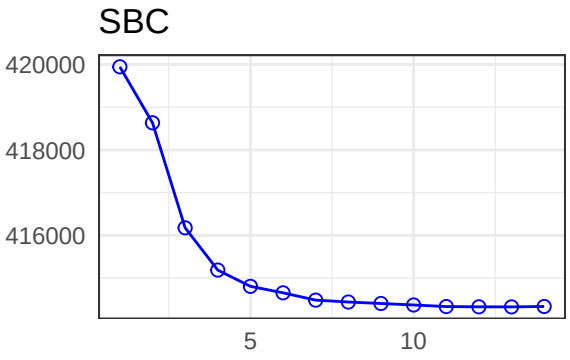
```
##                                                                  Best Subsets Regression
## -----------------------------------------------------------------------------------------------------------------
## Model Index    Predictors
## -----------------------------------------------------------------------------------------------------------------
##      1         sqft_living
##      2         sqft_living grade
##      3         sqft_living grade yr_built
##      4         sqft_living waterfront grade yr_built
##      5         sqft_living waterfront view grade yr_built
##      6         bedrooms sqft_living waterfront view grade yr_built
##      7         bedrooms bathrooms sqft_living waterfront view grade yr_built
##      8         bedrooms bathrooms sqft_living waterfront view grade yr_built sqft_lot15
##      9         bedrooms bathrooms sqft_living waterfront view grade yr_built sqft_living15 sqft_lot15
##     10         bedrooms bathrooms sqft_living floors waterfront view grade yr_built sqft_living15 sq
##     11         bedrooms bathrooms sqft_living floors waterfront view condition grade yr_built sqft_li
##     12         bedrooms bathrooms sqft_living floors waterfront view condition grade sqft_above yr_bu
```

```
##     13         bedrooms bathrooms sqft_living floors waterfront view condition grade sqft_above yr_bu
##     14         bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition grade sqft_al
## ----------------------------------------------------------------------------------------------------
##
##                                                                                 Subsets Regression Summary
## ----------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)           AIC           SBIC            SBC
## ----------------------------------------------------------------------------------------------------
##   1        0.4929      0.4929      0.4924    6957.1650    419922.7636    376987.0637    419945.6368
##   2        0.5353      0.5352      0.5345    5114.4341    418604.9671    375669.0830    418635.4646
##   3        0.6052      0.6052      0.6045    2069.3598    416138.1365    373202.9063    416176.2583
##   4        0.6305      0.6304      0.6293     972.6365    415140.8799    372206.0324    415186.6261
##   5        0.6399      0.6398      0.6386     565.3048    414753.0559    371818.3879    414806.4265
##   6        0.6437      0.6436      0.6422     400.6355    414593.4065    371658.8143    414654.4015
##   7        0.6479      0.6478      0.6463     218.5180    414414.8034    371480.3496    414483.4227
##   8        0.6492      0.6490      0.6475     165.9725    414362.8824    371428.4658    414439.1261
##   9        0.6502      0.6500      0.6483     123.5738    414320.8497    371386.4715    414404.7178
##  10        0.6512      0.6510      0.6492      81.0093    414278.5234    371344.1958    414370.0158
##  11        0.6523      0.6521      0.6503      36.2409    414233.8649    371299.6033    414332.9816
##  12        0.6527      0.6524      0.6505      22.4979    414220.1263    371285.8902    414326.8675
##  13        0.6529      0.6526      0.6507      13.0022    414210.6236    371276.4093    414324.9892
##  14        0.6529      0.6526      0.6506      15.0000    414212.6214    371278.4091    414334.6113
## ----------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

```
plot(k1,guide="none")
```


R−Square
0.65
0.60
0.55
0.50
5
10
C(p)
6000
4000
2000
0
5
10
Adj. R−Square
0.65
0.60
0.55
0.50
5
10
AIC
420000
418000
416000
414000
5
10






```
summary(k1)
```

```
##      mindex            n          predictors          rsquare
##  Min.   : 1.00   Min.   : 1.00   Length:14          Min.   :0.4929
##  1st Qu.: 4.25   1st Qu.: 4.25   Class :character   1st Qu.:0.6328
##  Median : 7.50   Median : 7.50   Mode  :character   Median :0.6486
##  Mean   : 7.50   Mean   : 7.50                      Mean   :0.6255
##  3rd Qu.:10.75   3rd Qu.:10.75                      3rd Qu.:0.6520
##  Max.   :14.00   Max.   :14.00                      Max.   :0.6529
##       adjr           predrsq             cp               aic
##  Min.   :0.4929   Min.   :0.4924   Min.   :  13.00   Min.   :414211
##  1st Qu.:0.6327   1st Qu.:0.6316   1st Qu.:  47.43   1st Qu.:414245
##  Median :0.6484   Median :0.6469   Median : 192.25   Median :414389
##  Mean   :0.6253   Mean   :0.6239   Mean   :1196.81   Mean   :415243
##  3rd Qu.:0.6518   3rd Qu.:0.6500   3rd Qu.: 870.80   3rd Qu.:415044
##  Max.   :0.6526   Max.   :0.6507   Max.   :6957.16   Max.   :419923
##       sbic             sbc              msep
##  Min.   :371276   Min.   :414325   Min.   : 687456926401000
##  1st Qu.:371311   1st Qu.:414344   1st Qu.: 689125072055000
##  Median :371454   Median :414461   Median : 695856826129000
##  Mean   :372309   Mean   :415316   Mean   : 741513875969000
##  3rd Qu.:372109   3rd Qu.:415092   3rd Qu.: 726848783066000
##  Max.   :376987   Max.   :419946   Max.   :1003605385610000
##       fpe                  apc              hsp
##  Min.   :45481726870   Min.   :0.3477   Min.   :3006463
```

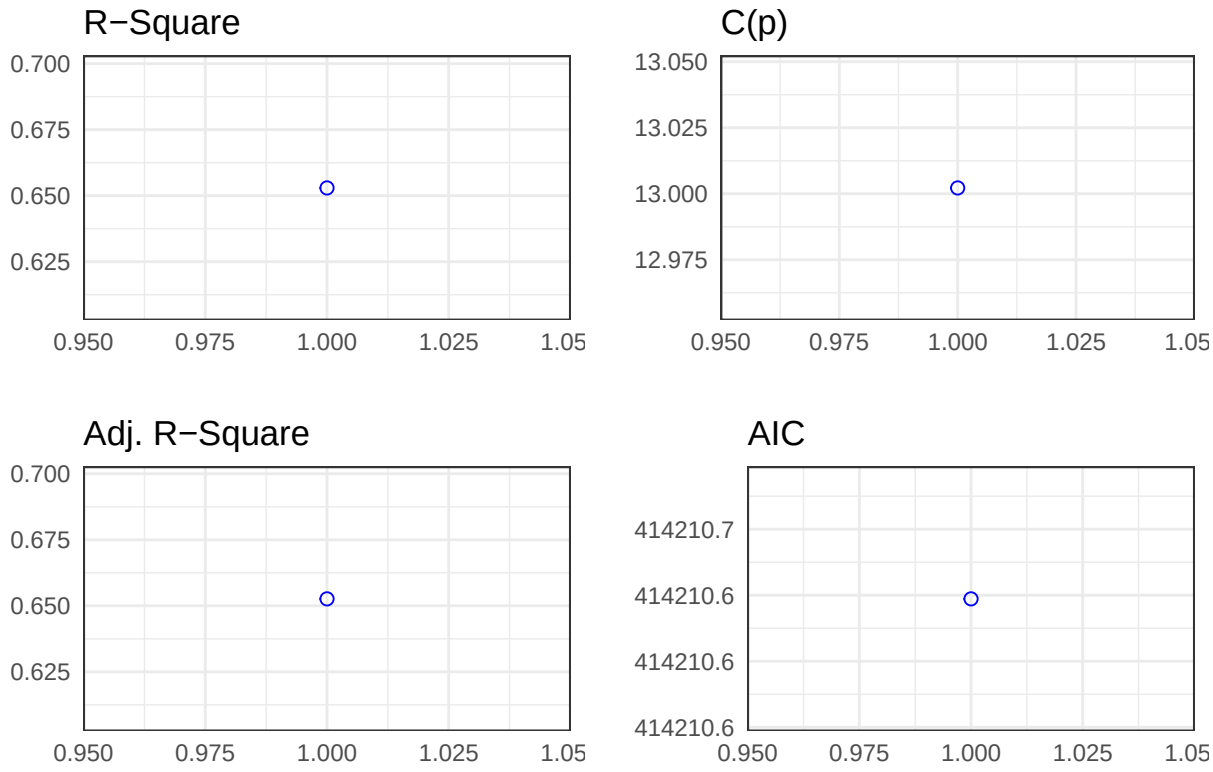
```
##  1st Qu.:45585314870   1st Qu.:0.3485   1st Qu.:3013309
##  Median :46020736517   Median :0.3518   Median :3042091
##  Mean   :49039042722   Mean   :0.3749   Mean   :3241609
##  3rd Qu.:48060068124   3rd Qu.:0.3674   3rd Qu.:3176895
##  Max.   :66345300990   Max.   :0.5072   Max.   :4385596
```

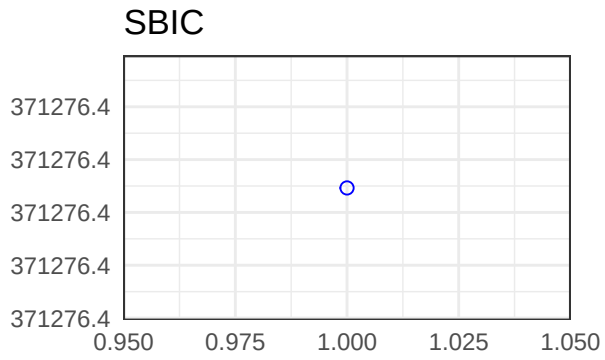
```
k2 <- ols_step_backward_p(H.full, prem = 0.05)
k2
```

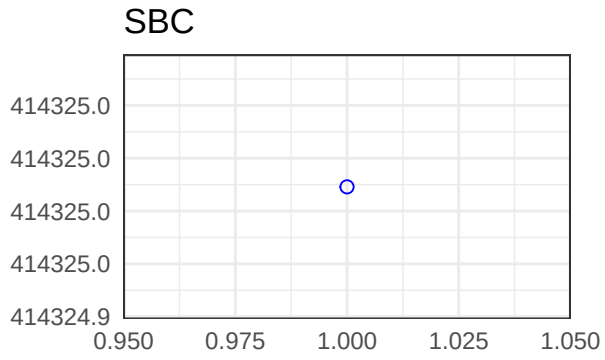
```
##
##
##                                Elimination Summary
## ------------------------------------------------------------------------------------
##         Variable                    Adj.
## Step    Removed     R-Square    R-Square     C(p)          AIC           RMSE
## ------------------------------------------------------------------------------------
##    1    sqft_lot      0.6529      0.6526    13.0022    414210.6236    213165.8465
## ------------------------------------------------------------------------------------
```

```
plot(k2)
```

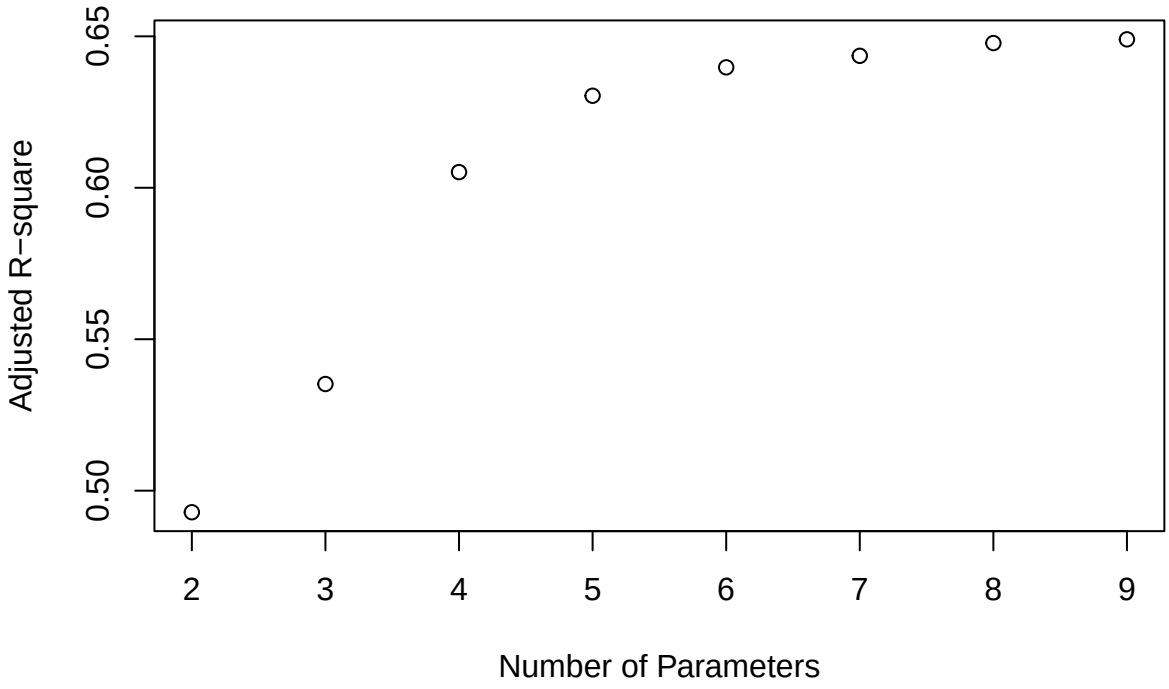
```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```


R−Square
0.700
0.675
0.650
0.625
0.950
0.975
1.000
1.025
1.05
C(p)
13.050
13.025
13.000
12.975
0.950
0.975
1.000
1.025
1.05
Adj. R−Square
0.700
0.675
0.650
0.625
0.950
0.975
1.000
1.025
1.05
AIC
414210.7
414210.6
414210.6
414210.6
0.950
0.975
1.000
1.025
1.05





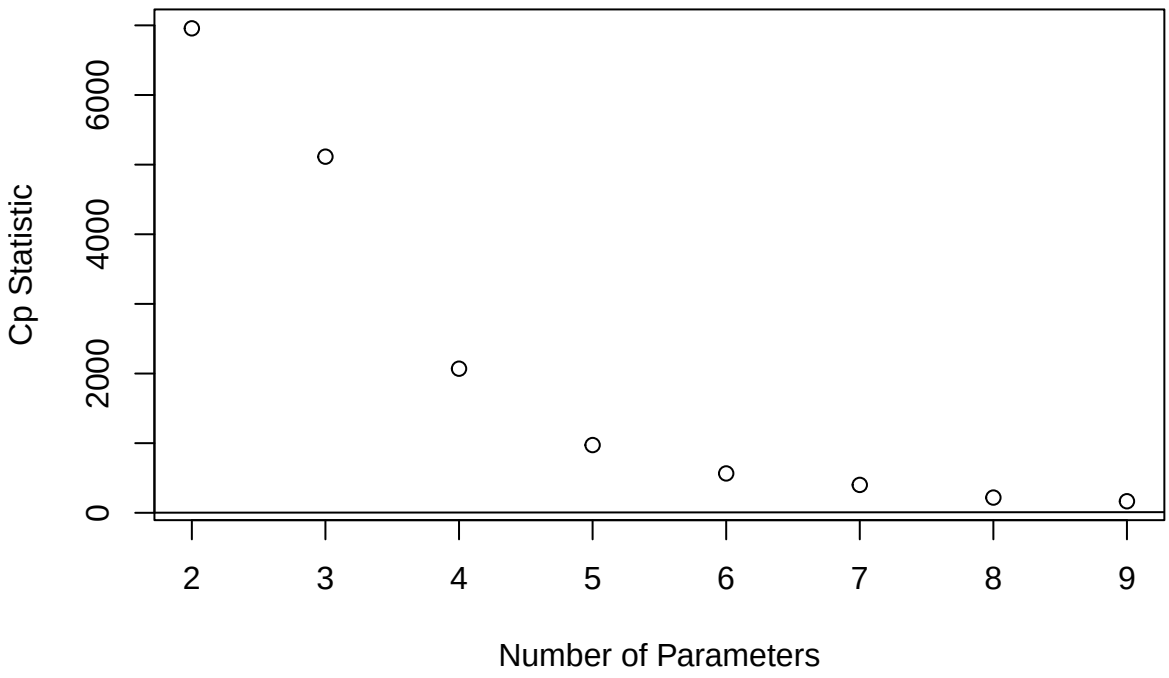

```
library(leaps)
k3 <- regsubsets(price~., data=df)
rs<-summary(k3)

# Plotting the adjusted R-square values against number of parameters
plot(2:9, rs$adjr2, xlab="Number of Parameters", ylab="Adjusted R-square")
```




```
which.max(rs$adjr2)
```

```
## [1] 8
```

```
# Plotting the CP values against number of parameters
plot(2:9, rs$cp, xlab="Number of Parameters", ylab="Cp Statistic")
abline(0,1)
```




```
rs$which
```

```
##   (Intercept) bedrooms bathrooms sqft_living sqft_lot floors waterfront  view
## 1        TRUE    FALSE     FALSE        TRUE    FALSE  FALSE      FALSE FALSE
## 2        TRUE    FALSE     FALSE        TRUE    FALSE  FALSE      FALSE FALSE
## 3        TRUE    FALSE     FALSE        TRUE    FALSE  FALSE      FALSE FALSE
## 4        TRUE    FALSE     FALSE        TRUE    FALSE  FALSE       TRUE FALSE
## 5        TRUE    FALSE     FALSE        TRUE    FALSE  FALSE       TRUE  TRUE
## 6        TRUE     TRUE     FALSE        TRUE    FALSE  FALSE       TRUE  TRUE
## 7        TRUE     TRUE      TRUE        TRUE    FALSE  FALSE       TRUE  TRUE
## 8        TRUE     TRUE      TRUE        TRUE    FALSE  FALSE       TRUE  TRUE
##   condition grade sqft_above yr_built yr_renovated sqft_living15 sqft_lot15
## 1     FALSE FALSE      FALSE    FALSE         FALSE          FALSE      FALSE
## 2     FALSE  TRUE      FALSE    FALSE         FALSE          FALSE      FALSE
## 3     FALSE  TRUE      FALSE     TRUE         FALSE          FALSE      FALSE
## 4     FALSE  TRUE      FALSE     TRUE         FALSE          FALSE      FALSE
## 5     FALSE  TRUE      FALSE     TRUE         FALSE          FALSE      FALSE
## 6     FALSE  TRUE      FALSE     TRUE         FALSE          FALSE      FALSE
## 7     FALSE  TRUE      FALSE     TRUE         FALSE          FALSE      FALSE
## 8     FALSE  TRUE      FALSE     TRUE         FALSE          FALSE       TRUE
```

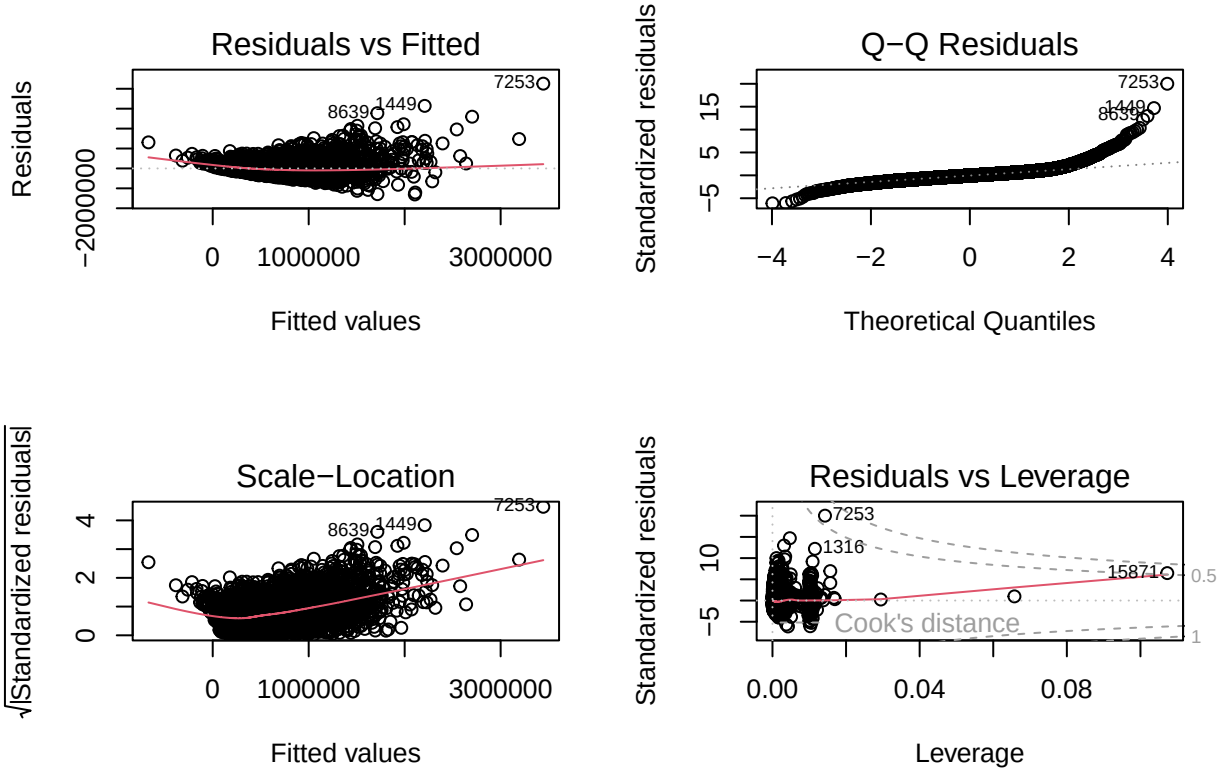
```
k4 <- lm(price ~ sqft_living + grade + yr_built + waterfront + view + bedrooms + bathrooms + sqft_lot15
ols_mallows_cp(k4, H.full)
```

```
## [1] 165.9725
```

```r
summary(k4)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + grade + yr_built + waterfront +
##     view + bedrooms + bathrooms + sqft_lot15, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1305860  -110584    -9516    89711  4255838
##
## Coefficients:
##                  Estimate    Std. Error t value          Pr(>|t|)
## (Intercept) 6343160.70155  139346.79030  45.521           < 2e-16 ***
## sqft_living     168.46613       3.86995  43.532           < 2e-16 ***
## grade        125978.93609    2480.82343  50.781           < 2e-16 ***
## yr_built      -3606.76474      73.53159 -49.051           < 2e-16 ***
## waterfront   520550.96351   22977.32486  22.655           < 2e-16 ***
## view          49487.62143    2643.94535  18.717           < 2e-16 ***
## bedrooms     -37291.79504    2352.47631 -15.852           < 2e-16 ***
## bathrooms     50216.20773    3885.28405  12.925           < 2e-16 ***
## sqft_lot15       -0.48437       0.06592  -7.347 0.000000000000212 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 214300 on 15120 degrees of freedom
## Multiple R-squared:  0.6492, Adjusted R-squared:  0.649
## F-statistic:  3497 on 8 and 15120 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(k4)
```




```
dispRegFunc(k4)
```

```
## [1] "Y = 6343160.70155 + 168.46613sqft_living + 125978.93609grade + -3606.764735yr_built + 520550.963
```