Households in New Haven

```
setwd("/Users/esin/Documents/Spring./Stat245/hw9")
data = dget("NewHaven.txt")
```

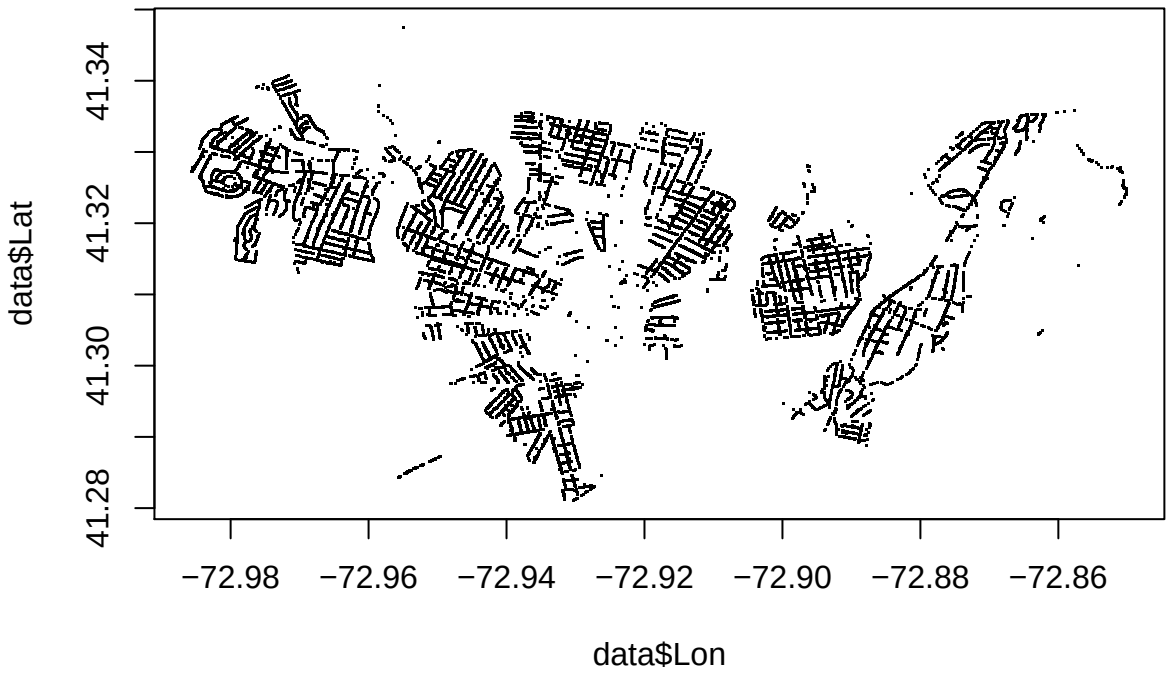
```
summary(data)
```

```
##     parcelID       Address               Lat             Long
##  Min.   :  291   Length:18104       Min.   :41.28   Min.   :-72.99
##  1st Qu.: 8739   Class :character   1st Qu.:41.31   1st Qu.:-72.95
##  Median :15496   Mode  :character   Median :41.32   Median :-72.93
##  Mean   :15377                      Mean   :41.32   Mean   :-72.93
##  3rd Qu.:22077                      3rd Qu.:41.33   3rd Qu.:-72.90
##  Max.   :27301                      Max.   :41.35   Max.   :-72.85
##
##     owner               CurVal              size          LivingArea
##  Length:18104       Min.   :   17640   Min.   : 0.000   Min.   :   330
##  Class :character   1st Qu.:  114870   1st Qu.: 0.080   1st Qu.:  1267
##  Mode  :character   Median :  150360   Median : 0.130   Median :  1914
##                     Mean   :  173646   Mean   : 0.143   Mean   :  2160
##                     3rd Qu.:  199010   3rd Qu.: 0.180   3rd Qu.:  2852
##                     Max.   :31814300   Max.   :20.500   Max.   :132935
##
##  TotalBedrooms     TotalBathrooms       ACtype              Grade
##  Min.   :  1.000   Min.   :  1.000   Length:18104       Min.   :1.000
##  1st Qu.:  3.000   1st Qu.:  1.000   Class :character   1st Qu.:2.000
##  Median :  4.000   Median :  2.000   Mode  :character   Median :2.000
##  Mean   :  3.983   Mean   :  2.029                      Mean   :2.253
##  3rd Qu.:  5.000   3rd Qu.:  3.000                      3rd Qu.:3.000
##  Max.   :205.000   Max.   :205.000                      Max.   :3.000
##                                                         NA's   :16
##   Depreciation          Year          Garage         Garage.area
##  Min.   :-0.5000   Min.   :1763   Min.   :0.00000   Min.   :   0.00
##  1st Qu.: 0.1500   1st Qu.:1900   1st Qu.:0.00000   1st Qu.:   0.00
##  Median : 0.2000   Median :1918   Median :0.00000   Median :   0.00
##  Mean   : 0.2293   Mean   :1926   Mean   :0.05502   Mean   :  16.56
##  3rd Qu.: 0.3000   3rd Qu.:1950   3rd Qu.:0.00000   3rd Qu.:   0.00
##  Max.   : 0.8900   Max.   :2010   Max.   :3.00000   Max.   :2210.00
##
##    condo           house
##  Mode :logical   Mode :logical
##  FALSE:15359     FALSE:2764
##  TRUE :2745      TRUE :15340
##
##
##
##
```

This is a dataset of addresses in New Haven. It is a condensed region as the lat and lon min max is (41.28-41.35), (-72.99,-72.85). There is a large range of current values. The size, living area and total bedroooms are all around similar in the 1st, mean and 3rd quartiles but have large outliers in the maximums. The

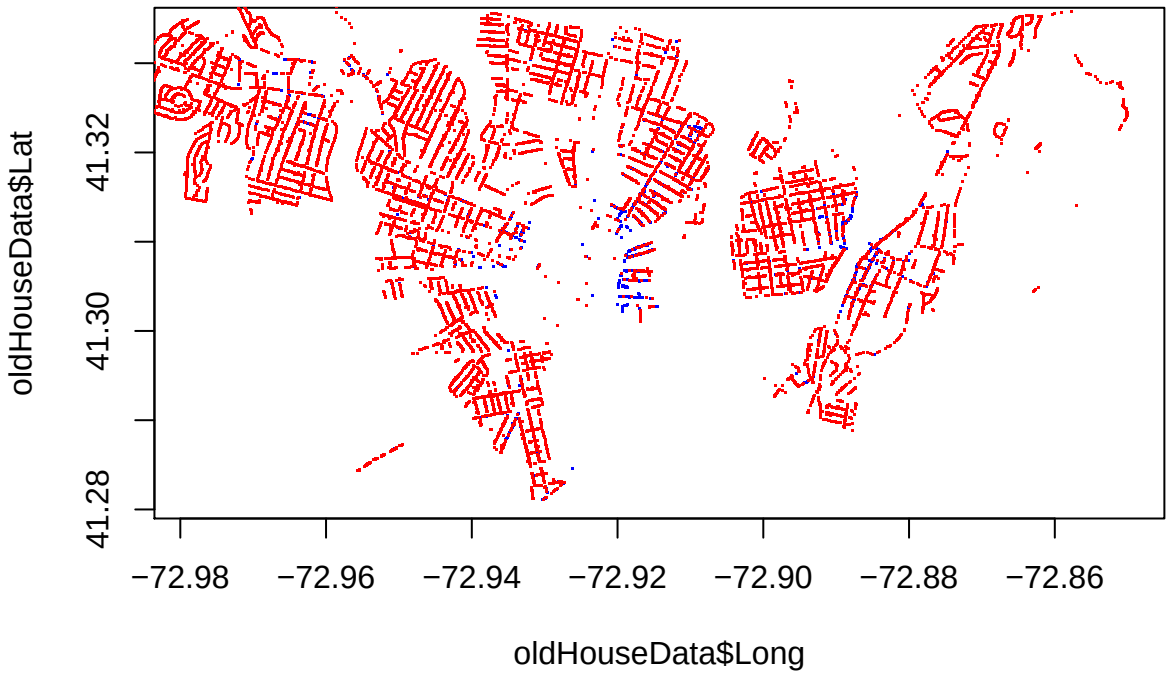
highest house price is 31 million, which we could imagine is for a house that is extremely extravagant. Some of these are not houses but are condos, which we can see that when the condo is false it mostly matches up to when the house is true.

```
plot(data$Lon, data$Lat, pch='.')
```

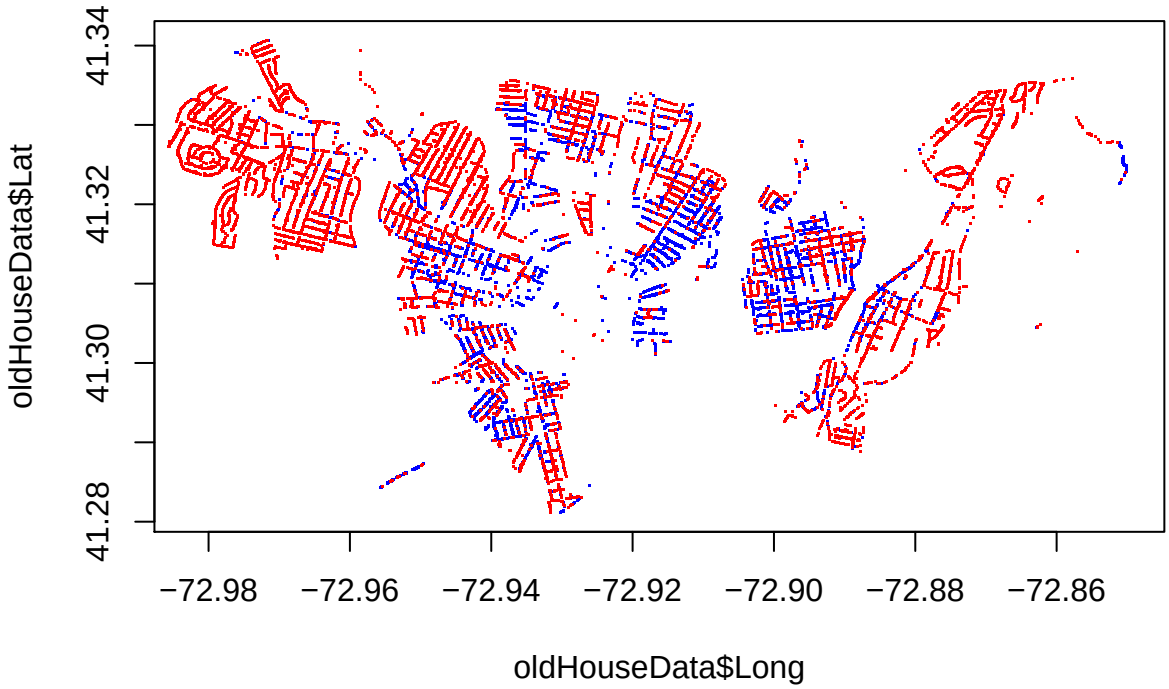


I will be studying the old houses. I am curious where the old houses in this region are located. I will be creating different maps to see what the placement of the old houses are like.

```
old=1900
oldHouseData = subset(data, Year<old)
newHouseData = subset(data, Year>=old)
plot(oldHouseData$Long, oldHouseData$Lat, pch='.', col='blue')
points(newHouseData$Long, newHouseData$Lat, pch='.', col='red')
```

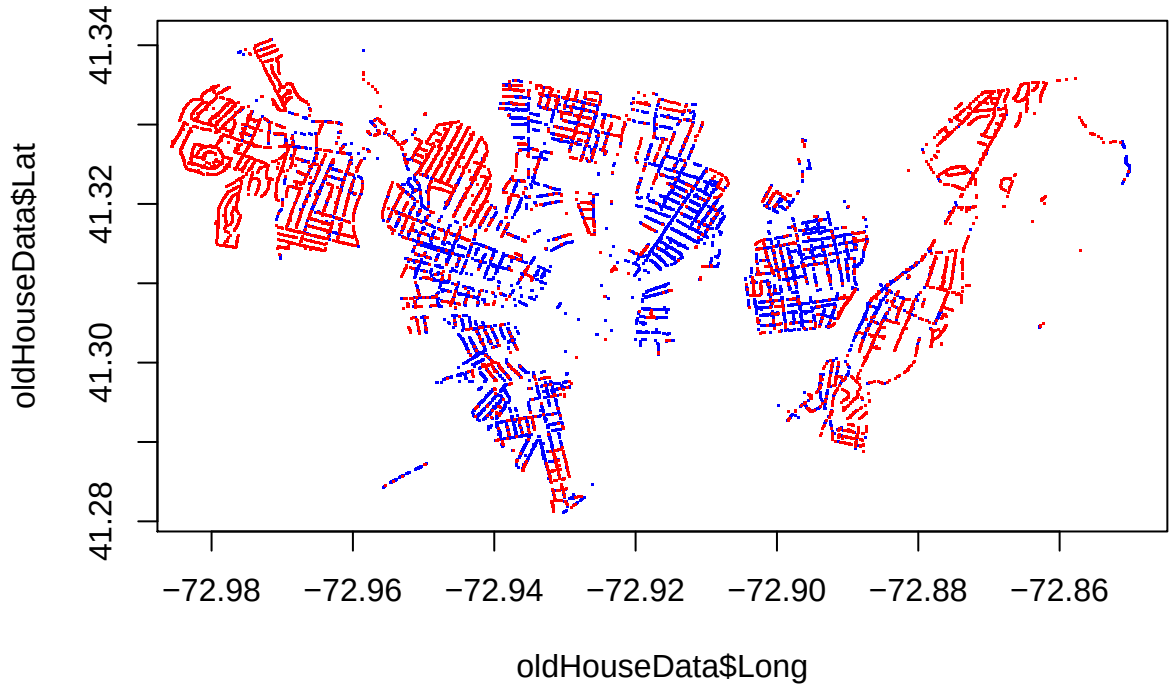



```
old=1910
oldHouseData = subset(data, Year<old)
newHouseData = subset(data, Year>=old)
plot(oldHouseData$Long, oldHouseData$Lat, pch='.', col='blue')
points(newHouseData$Long, newHouseData$Lat, pch='.', col='red')
```

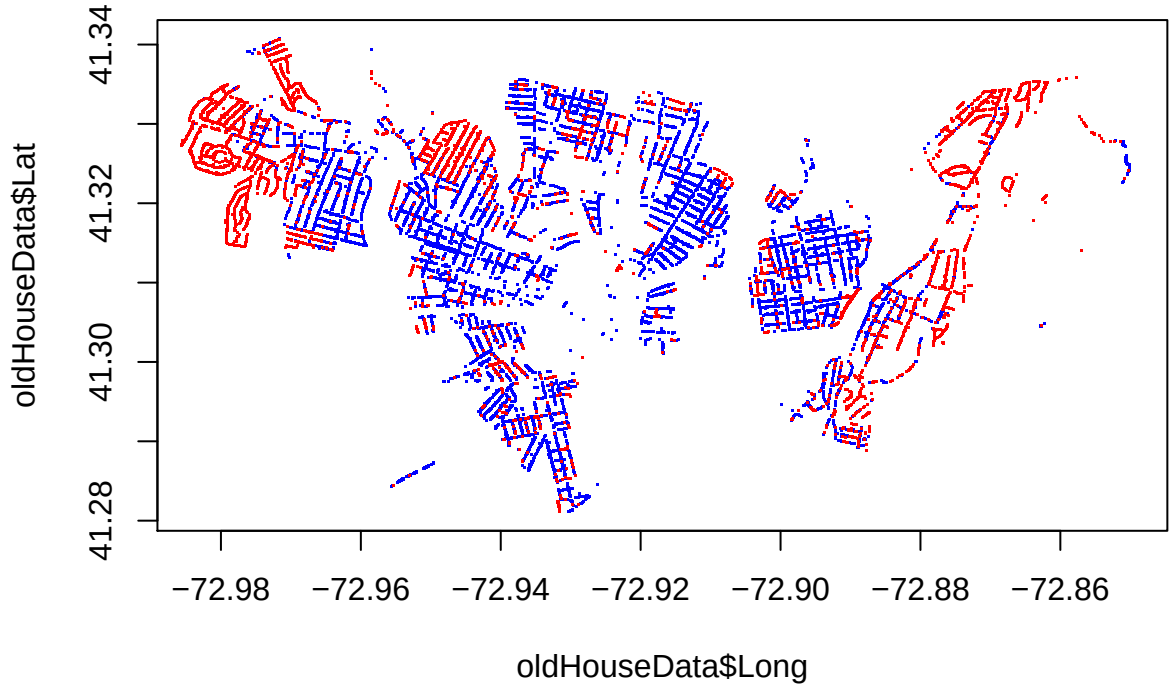



```
old=1920
oldHouseData = subset(data, Year<old)
newHouseData = subset(data, Year>=old)
```

```
plot(oldHouseData$Long, oldHouseData$Lat, pch='.', col='blue')
points(newHouseData$Long, newHouseData$Lat, pch='.', col='red')
```




```
old=1930
oldHouseData = subset(data, Year<old)
newHouseData = subset(data, Year>=old)
plot(oldHouseData$Long, oldHouseData$Lat, pch='.', col='blue')
points(newHouseData$Long, newHouseData$Lat, pch='.', col='red')
```



As we can see the old houses seem to be more concentrated along the inside of the city. It seems as though, therefore, the expansion of the city occurred out from the inside outwards as time passed. We will pick old = 1910 as representative of this moment in time as the expansion was happening.

```
old=1910
oldHouseData = subset(data, Year<old)
newHouseData = subset(data, Year>=old)
summary(oldHouseData)
```

```
##     parcelID       Address               Lat             Long
##  Min.   : 2739   Length:6964        Min.   :41.28   Min.   :-72.98
##  1st Qu.: 9487   Class :character   1st Qu.:41.31   1st Qu.:-72.94
##  Median :14278   Mode  :character   Median :41.31   Median :-72.92
##  Mean   :14275                      Mean   :41.31   Mean   :-72.92
##  3rd Qu.:19060                      3rd Qu.:41.32   3rd Qu.:-72.90
##  Max.   :26982                      Max.   :41.34   Max.   :-72.85
##
##     owner               CurVal             size          LivingArea
##  Length:6964        Min.   :  17640   Min.   :0.0000   Min.   :  376
##  Class :character   1st Qu.: 119140   1st Qu.:0.0800   1st Qu.: 1664
##  Mode  :character   Median : 154560   Median :0.1100   Median : 2362
##                     Mean   : 178212   Mean   :0.1346   Mean   : 2497
##                     3rd Qu.: 207638   3rd Qu.:0.1600   3rd Qu.: 3148
##                     Max.   :2013970   Max.   :2.6200   Max.   :22292
##
##  TotalBedrooms    TotalBathrooms      ACtype              Grade
##  Min.   : 1.000   Min.   : 1.000   Length:6964        Min.   :1.000
##  1st Qu.: 3.000   1st Qu.: 2.000   Class :character   1st Qu.:2.000
##  Median : 4.000   Median : 2.000   Mode  :character   Median :2.000
##  Mean   : 4.624   Mean   : 2.374                      Mean   :2.187
##  3rd Qu.: 6.000   3rd Qu.: 3.000                      3rd Qu.:2.000
##  Max.   :21.000   Max.   :12.000                      Max.   :3.000
##                                                       NA's   :4
##   Depreciation          Year          Garage         Garage.area
##  Min.   :-0.5000   Min.   :1763   Min.   :0.00000   Min.   :   0.000
##  1st Qu.: 0.2000   1st Qu.:1900   1st Qu.:0.00000   1st Qu.:   0.000
##  Median : 0.3000   Median :1900   Median :0.00000   Median :   0.000
##  Mean   : 0.2712   Mean   :1896   Mean   :0.00201   Mean   :   0.783
##  3rd Qu.: 0.3000   3rd Qu.:1900   3rd Qu.:0.00000   3rd Qu.:   0.000
##  Max.   : 0.8600   Max.   :1909   Max.   :3.00000   Max.   :1854.000
##
##    condo            house
##  Mode :logical   Mode :logical
##  FALSE:6831      FALSE:141
##  TRUE :133       TRUE :6823
##
##
##
##
```

```
summary(newHouseData)
```

```
##     parcelID       Address               Lat             Long
```

```
##  Min.   :  291   Length:11140       Min.   :41.28   Min.   :-72.99
##  1st Qu.: 7242   Class :character   1st Qu.:41.31   1st Qu.:-72.96
##  Median :17103   Mode  :character   Median :41.32   Median :-72.93
##  Mean   :16066                      Mean   :41.32   Mean   :-72.93
##  3rd Qu.:23849                      3rd Qu.:41.33   3rd Qu.:-72.89
##  Max.   :27301                      Max.   :41.35   Max.   :-72.85
##
##     owner              CurVal              size           LivingArea
##  Length:11140       Min.   :   22400   Min.   : 0.0000   Min.   :   330
##  Class :character   1st Qu.:  112542   1st Qu.: 0.0600   1st Qu.:  1152
##  Mode  :character   Median :  147560   Median : 0.1400   Median :  1616
##                     Mean   :  170792   Mean   : 0.1483   Mean   :  1949
##                     3rd Qu.:  194600   3rd Qu.: 0.1900   3rd Qu.:  2490
##                     Max.   :31814300   Max.   :20.5000   Max.   :132935
##
##  TotalBedrooms     TotalBathrooms       ACtype              Grade
##  Min.   :  1.000   Min.   :  1.000   Length:11140       Min.   :1.000
##  1st Qu.:  2.000   1st Qu.:  1.000   Class :character   1st Qu.:2.000
##  Median :  3.000   Median :  2.000   Mode  :character   Median :2.000
##  Mean   :  3.582   Mean   :  1.813                      Mean   :2.294
##  3rd Qu.:  4.000   3rd Qu.:  2.000                      3rd Qu.:3.000
##  Max.   :205.000   Max.   :205.000                      Max.   :3.000
##                                                         NA's   :12
##   Depreciation         Year          Garage         Garage.area
##  Min.   :0.0000   Min.   :1910   Min.   :0.00000   Min.   :   0.00
##  1st Qu.:0.1500   1st Qu.:1920   1st Qu.:0.00000   1st Qu.:   0.00
##  Median :0.2000   Median :1940   Median :0.00000   Median :   0.00
##  Mean   :0.2032   Mean   :1945   Mean   :0.08815   Mean   :  26.42
##  3rd Qu.:0.3000   3rd Qu.:1967   3rd Qu.:0.00000   3rd Qu.:   0.00
##  Max.   :0.8900   Max.   :2010   Max.   :3.00000   Max.   :2210.00
##
##    condo           house
##  Mode :logical   Mode :logical
##  FALSE:8528      FALSE:2623
##  TRUE :2612      TRUE :8517
##
##
##
##
```

We can see that there is a difference between the sizes of these old and new houses. The old houses have a maximum size of 2 while the new ones have a maximum (outlier) possibility of much greater. Interestingly, other than the outlier of the very large house, it seems like the old houses are in genereal larger (3148 of a 3rd quartile compared to 2490.)

They also have a higher number of bedrooms (3rd quartile = 6, while new houses have 3rd quartile of 4.) The bathrooms number is likewise higher. The garage however is smaller which makes sense, as cars did not exist. More of the new houses are condos rather than houses while the old ones are less likely to be condos.

The old houses have a higher median value.

The 5 variables are: size, bathrooms, bedrooms, living area, depreciation. Output: curval

```
fit <- lm(oldHouseData$CurVal ~ oldHouseData$size + oldHouseData$TotalBathrooms + oldHouseData$TotalBedr
summary(fit)
```

```
##
## Call:
## lm(formula = oldHouseData$CurVal ~ oldHouseData$size + oldHouseData$TotalBathrooms +
##     oldHouseData$TotalBedrooms + oldHouseData$LivingArea + oldHouseData$Depreciation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -544248  -34608  -13376   16139 1800135
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 112367.38    3270.66  34.356   <2e-16 ***
## oldHouseData$size           166265.40    7764.13  21.415   <2e-16 ***
## oldHouseData$TotalBathrooms   9298.43    1078.79   8.619   <2e-16 ***
## oldHouseData$TotalBedrooms  -13061.40     628.56 -20.780   <2e-16 ***
## oldHouseData$LivingArea         61.74       1.10  56.147   <2e-16 ***
## oldHouseData$Depreciation  -266957.07    9196.79 -29.027   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62480 on 6958 degrees of freedom
## Multiple R-squared:  0.5736, Adjusted R-squared:  0.5733
## F-statistic:  1872 on 5 and 6958 DF,  p-value: < 2.2e-16
```

```
fit2 <- lm(newHouseData$CurVal ~ newHouseData$size + newHouseData$TotalBathrooms + newHouseData$TotalBe
summary(fit2)
```

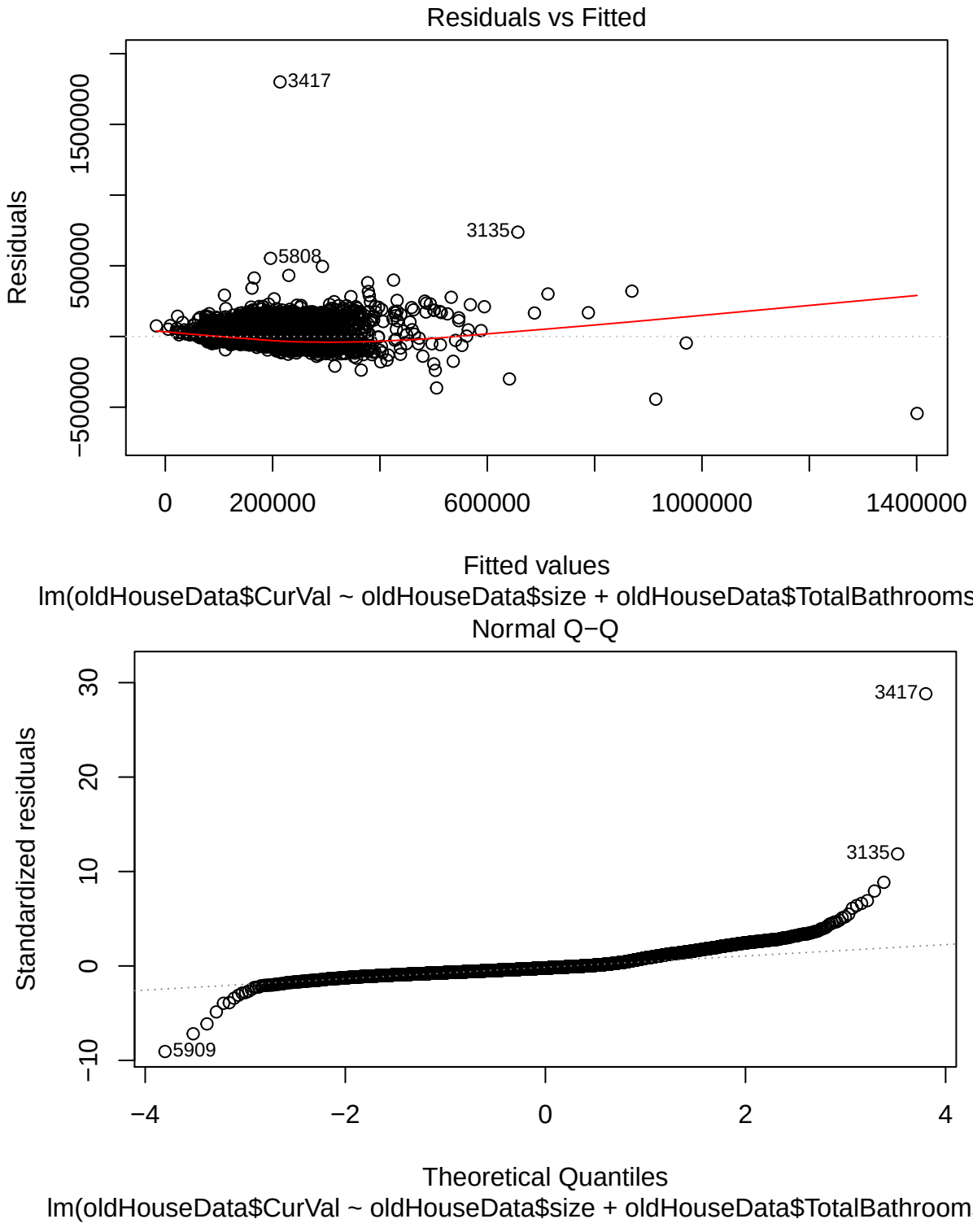
```
##
## Call:
## lm(formula = newHouseData$CurVal ~ newHouseData$size + newHouseData$TotalBathrooms +
##     newHouseData$TotalBedrooms + newHouseData$LivingArea + newHouseData$Depreciation)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8301342   -39555      -51    39309 20945820
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.462e+04  7.449e+03   1.963   0.0496 *
## newHouseData$size            2.572e+05  9.617e+03  26.747  < 2e-16 ***
## newHouseData$TotalBathrooms  7.086e+04  3.491e+03  20.294  < 2e-16 ***
## newHouseData$TotalBedrooms  -1.389e+03  2.907e+03  -0.478   0.6329
## newHouseData$LivingArea      2.580e+01  5.027e+00   5.133  2.9e-07 ***
## newHouseData$Depreciation   -2.744e+05  3.234e+04  -8.484  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 301900 on 11134 degrees of freedom
## Multiple R-squared:  0.3738, Adjusted R-squared:  0.3735
## F-statistic:  1329 on 5 and 11134 DF,  p-value: < 2.2e-16
```

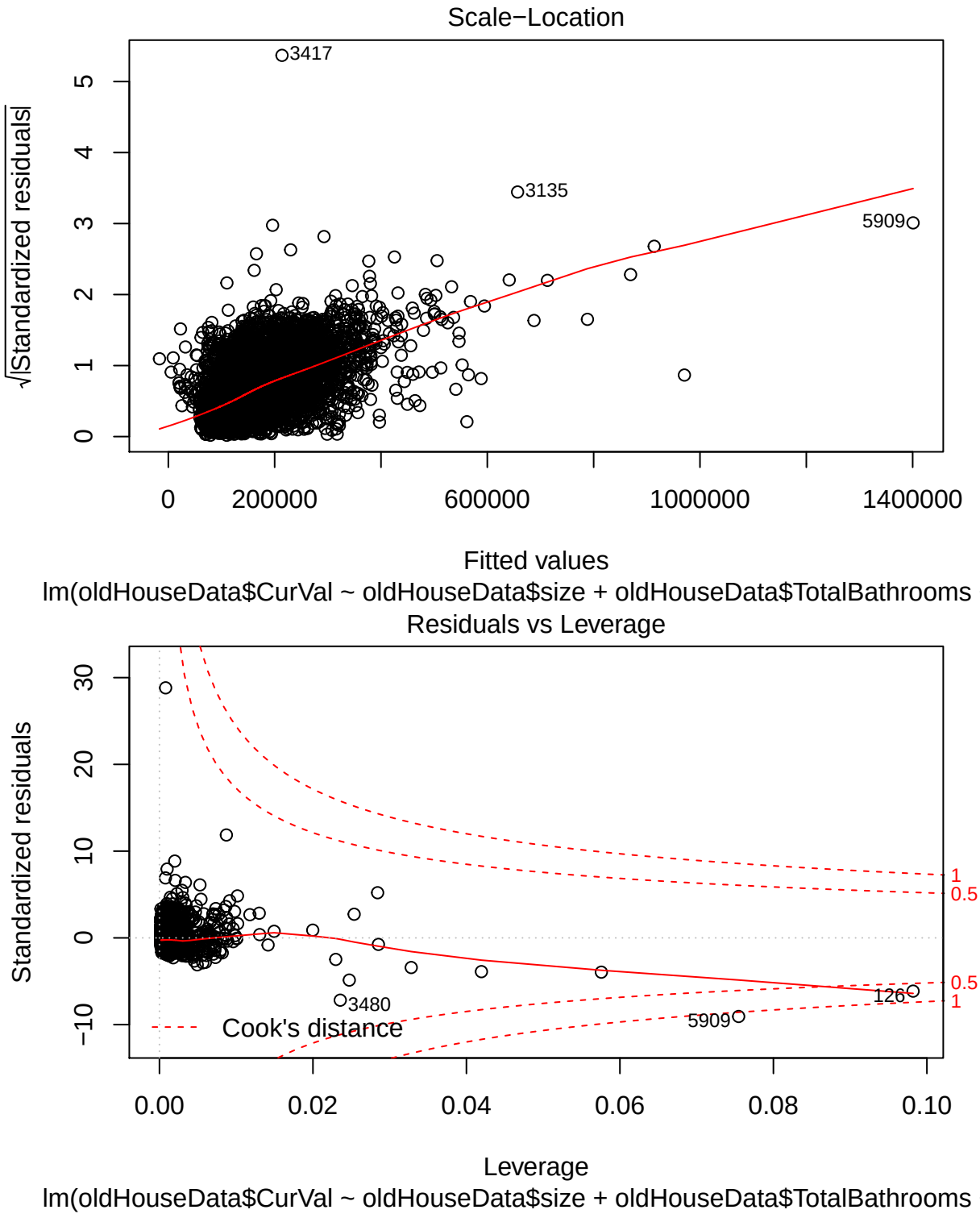
When predicting the price of old houses we can see that all of these factors I fit were highly significant. The size, the total bathrooms, the bedrooms, the living area, and the depreciation are all important factors in the price of an old house.

However, comparatively, the size, and number of bedrooms are not important for a new house. (At alpha = 0.05 significance)

Now, looking for outliers:

```
plot(fit)
```

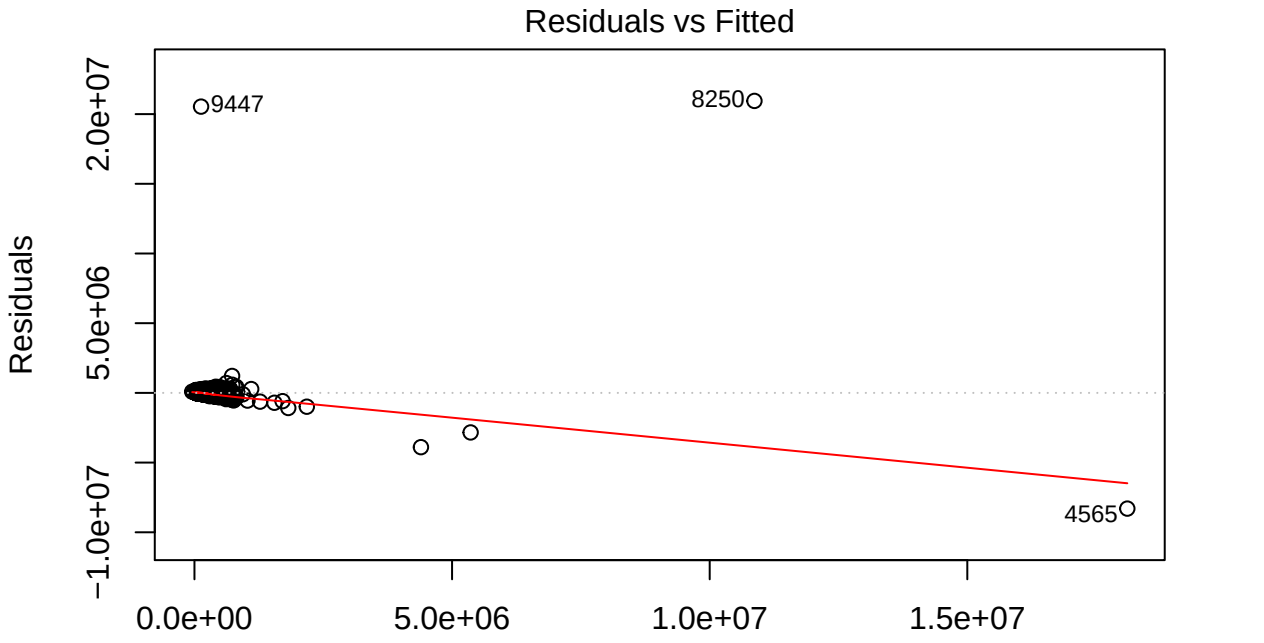


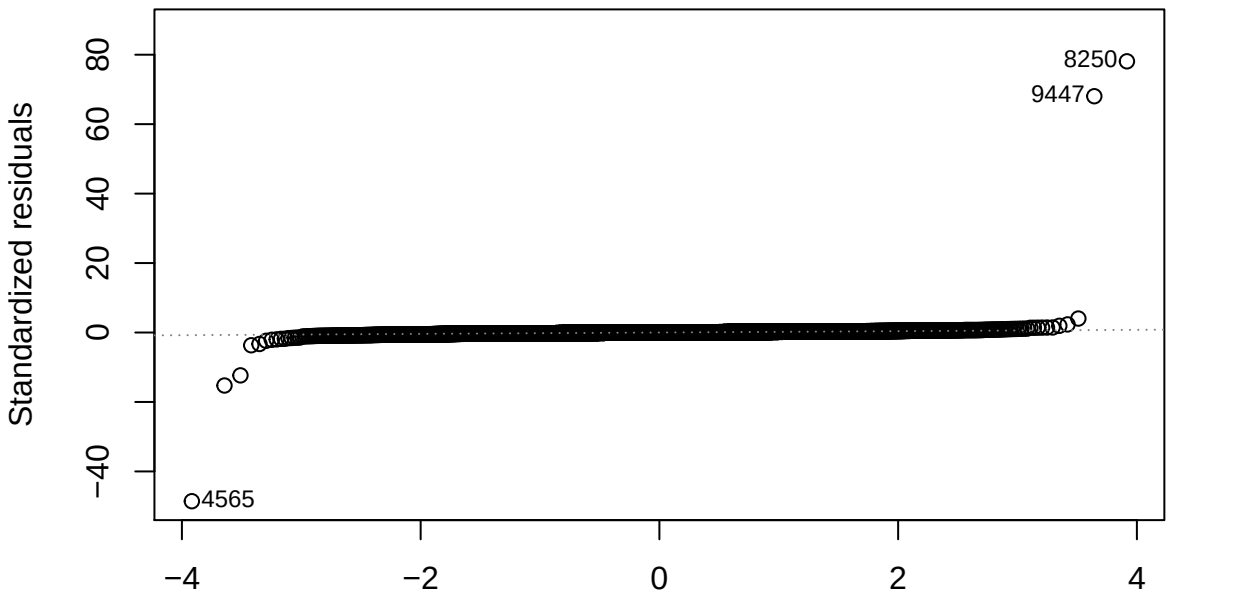


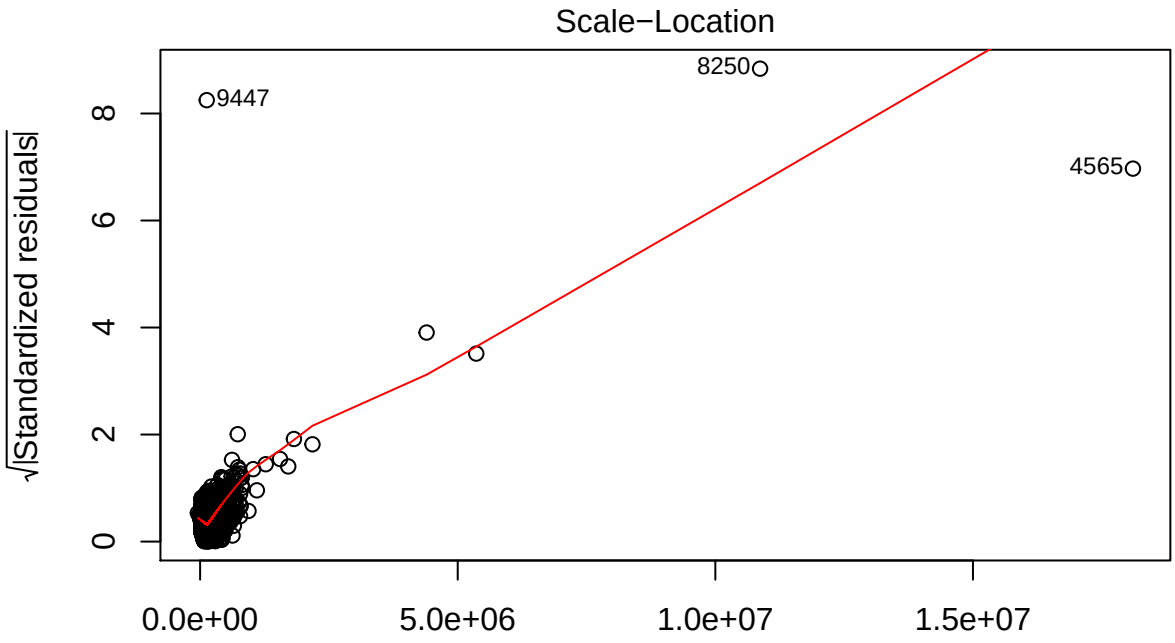
Looking at this cook's distance plot, any outliers are >1 . Here we can't see any dramatic outliers, though possibly 5909. However this looks like just a regular house so it is probably not an outlier.

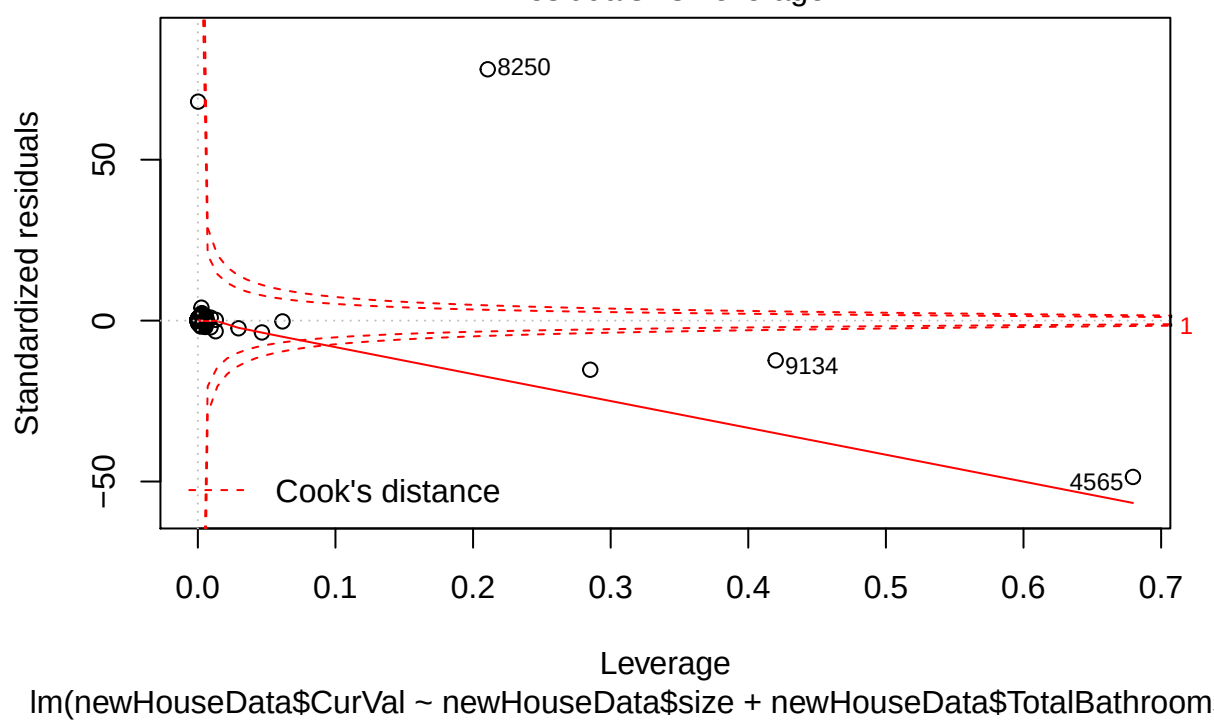
For the newHouseData:

```
plot(fit2)
```


Residuals vs Fitted
Residuals
9447
8250
4565
2.0e+07
5.0e+06
−1.0e+07
0.0e+00
5.0e+06
1.0e+07
1.5e+07
Fitted values
lm(newHouseData$CurVal ~ newHouseData$size + newHouseData$TotalBathrooms +

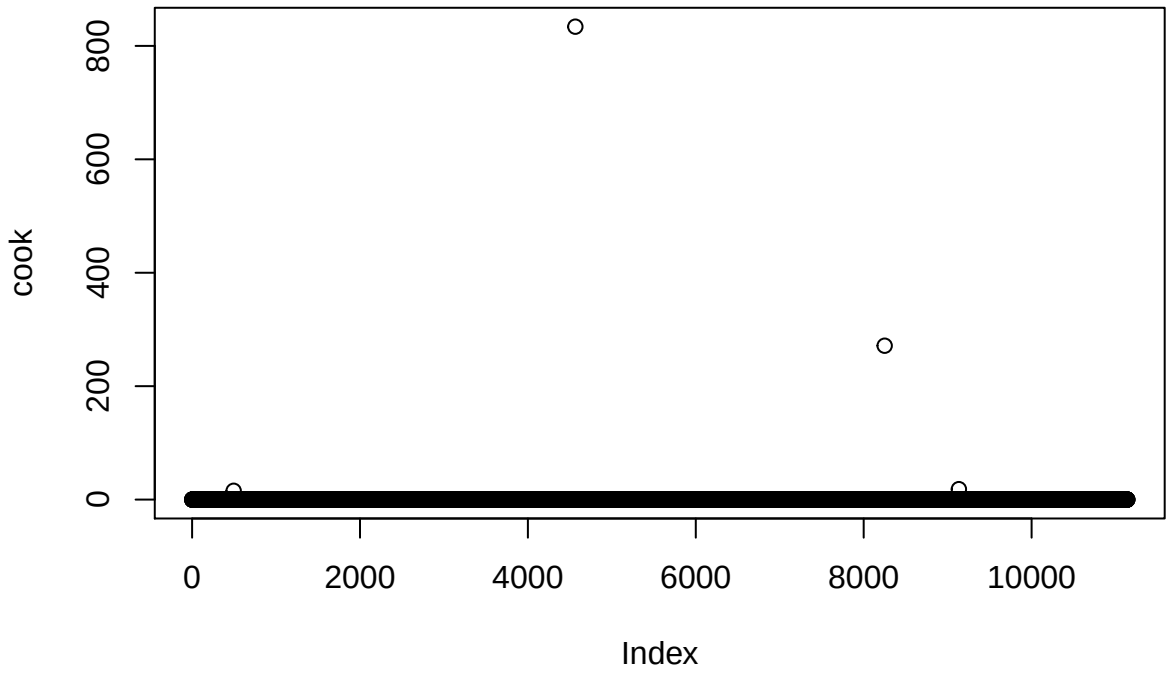

Normal Q−Q
Standardized residuals
8250
9447
4565
80
60
40
20
0
−40
−4
−2
0
2
4
Theoretical Quantiles
lm(newHouseData$CurVal ~ newHouseData$size + newHouseData$TotalBathrooms +





The outliers are 8250, 9134, and 4565. Looking them up, we can see that they might not be considered regular houses so they should be removed.

```
cook = cooks.distance(fit2)
plot(cook)
```




```
newHouseData[cook>0.04,]
```

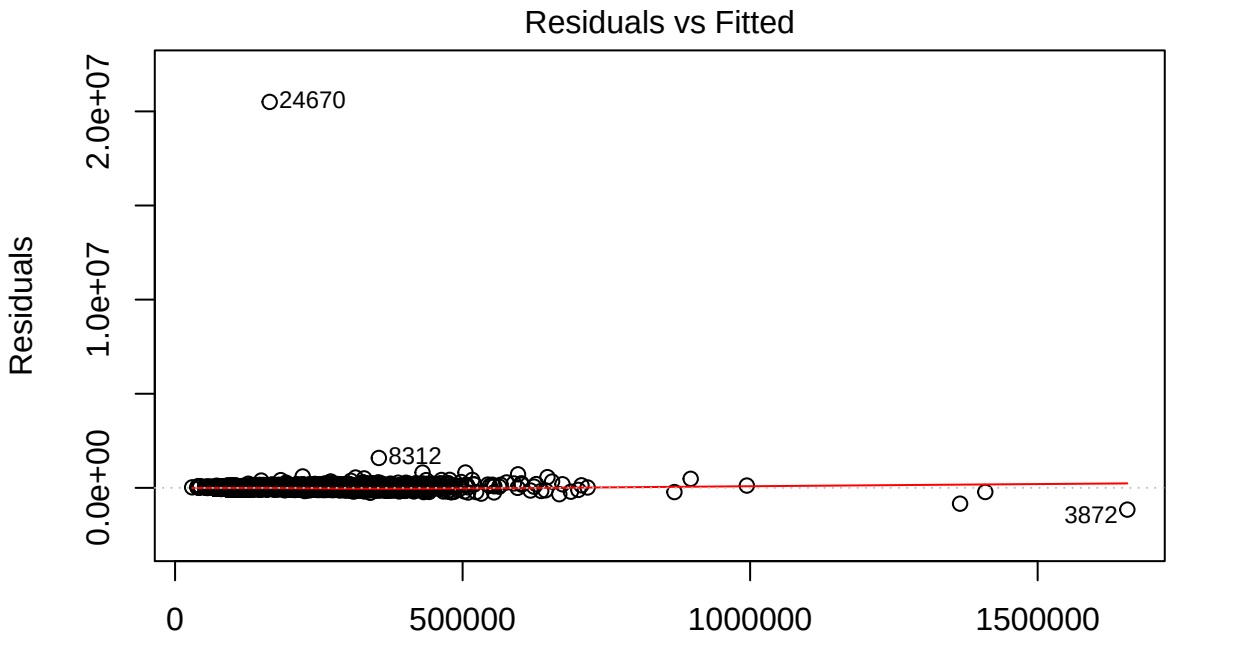
```
##       parcelID                            Address      Lat      Long
## 3872      3918     201 RUSSELL ST, New Haven, CT 41.30004 -72.88064
## 4882      4951      35 EASTERN ST, New Haven, CT 41.30806 -72.87383
## 13245    13549        18 TOWER LA, New Haven, CT 41.30062 -72.92831
## 23042    23676       400 BLAKE ST, New Haven, CT 41.32875 -72.95651
## 24221    24862 358 SPRINGSIDE AVE, New Haven, CT 41.33938 -72.95844
## 24670    25314         986 FOREST, New Haven, CT 41.31883 -72.97047
##                                      owner   CurVal  size LivingArea
## 3872  FITZMAURICE JAMES JR & KRAUSE BRUCE   500990 16.91       1768
## 4882               CITY OF NEW HAVEN GOLF   737660  6.90       1296
## 13245      NEW HAVEN JEWISH COMMUNITY COU  9804340  1.78     132935
## 23042            METROPOLITAN DEVELOPMENT 31814300  7.96      64302
## 24221              CITY OF NEW HAVEN PARK  2522660 20.50       2344
## 24670      HOPKINS COMMITTEE OF TRUSTEES  20666240  0.00       2322
##       TotalBedrooms TotalBathrooms ACtype Grade Depreciation Year Garage
## 3872              2              1      1     3         0.30 1924      0
## 4882              3              1      1     2         0.25 1960      0
## 13245           205            205      0     3         0.14 1969      0
## 23042           109            103      1     2         0.00 2006      0
## 24221             3              1      1     2         0.20 1910      0
## 24670             4              2      1     3         0.30 1920      0
##       Garage.area condo house
## 3872            0 FALSE  TRUE
## 4882            0 FALSE  TRUE
## 13245           0 FALSE FALSE
## 23042           0 FALSE FALSE
## 24221           0 FALSE  TRUE
## 24670           0  TRUE FALSE
```

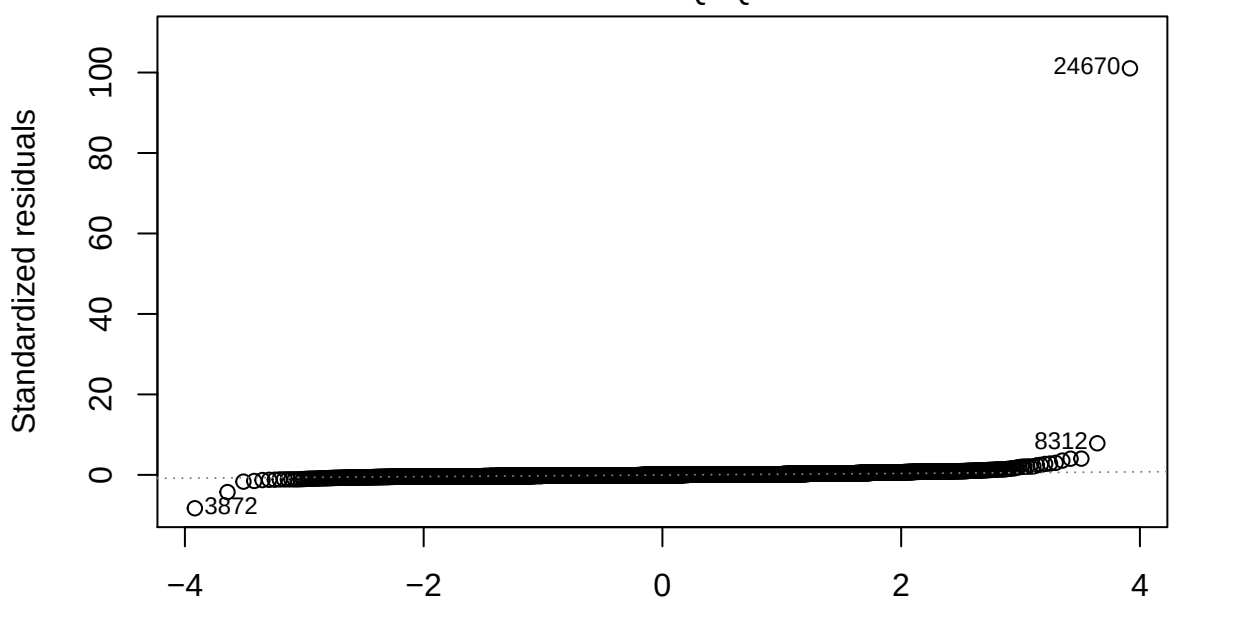
This finds some additional outliers, but these all appear to just be houses.

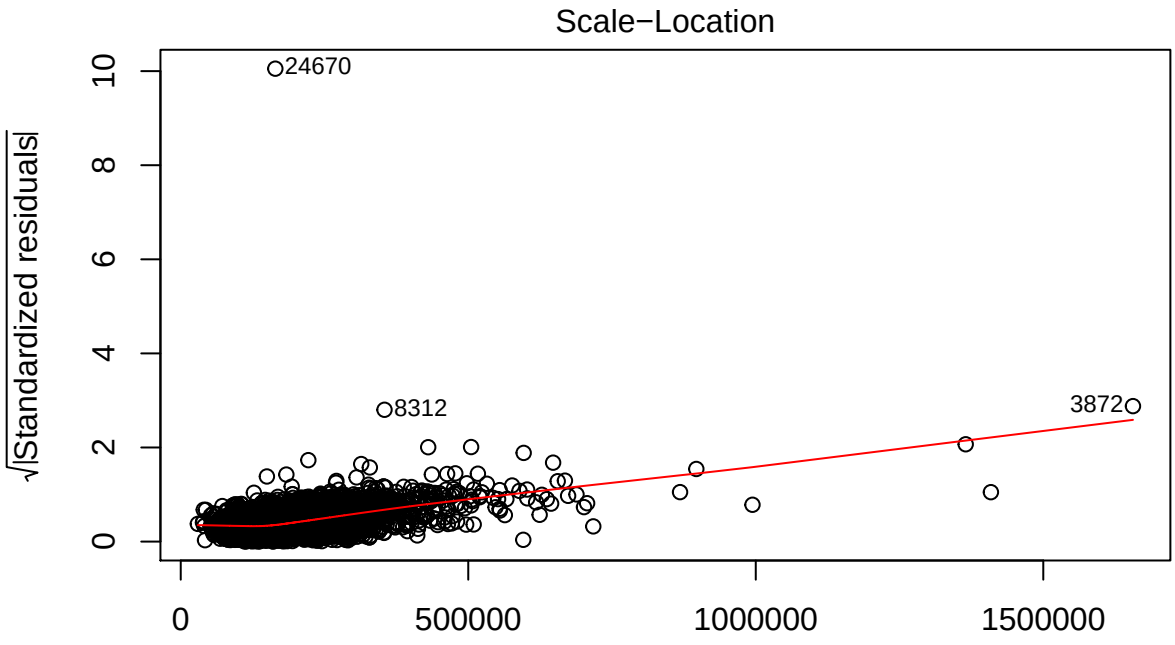
Removing the original outliers:

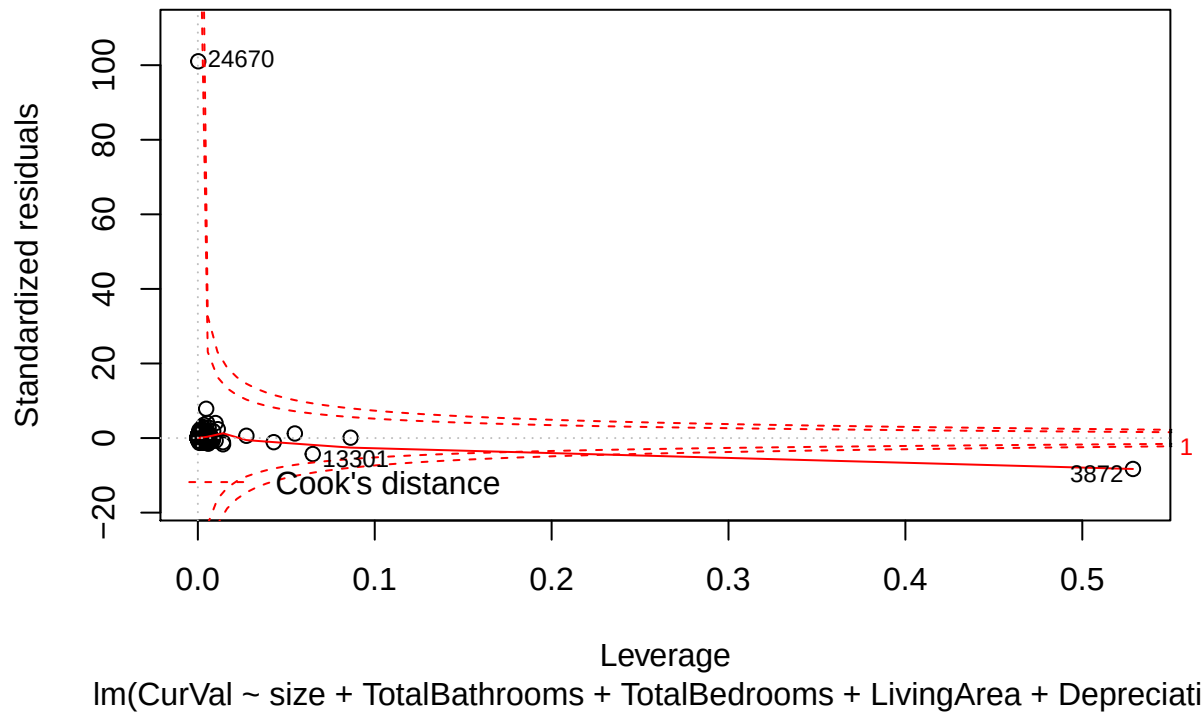
```
newHouseOutliersremoved = newHouseData[-c(8250, 9134, 4565),]
```

```
fit2 <- lm(CurVal ~ size + TotalBathrooms + TotalBedrooms + LivingArea + Depreciation, newHouseOutliersr
plot(fit2)
```










```
summary(fit2)
```

```
##
## Call:
## lm(formula = CurVal ~ size + TotalBathrooms + TotalBedrooms +
```

```
##     LivingArea + Depreciation, data = newHouseOutliersremoved)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1154989   -34010    -6168    19516 20501769
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.713e+04  5.490e+03  14.049  < 2e-16 ***
## size            8.928e+04  8.809e+03  10.136  < 2e-16 ***
## TotalBathrooms  4.843e+03  3.428e+03   1.413    0.158
## TotalBedrooms  -1.510e+04  1.959e+03  -7.708 1.38e-14 ***
## LivingArea      7.878e+01  3.509e+00  22.451  < 2e-16 ***
## Depreciation   -1.496e+05  2.187e+04  -6.839 8.37e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 202900 on 11131 degrees of freedom
## Multiple R-squared:  0.1194, Adjusted R-squared:  0.119
## F-statistic: 301.9 on 5 and 11131 DF,  p-value: < 2.2e-16
```

Hence, this is looking much better. Now that we have removed outliers, we can see that the bedrooms and size actually are important for a new house, and it is actually the bathrooms that are not important. (Though we could also remove the new outlier 3872 to be sure.)

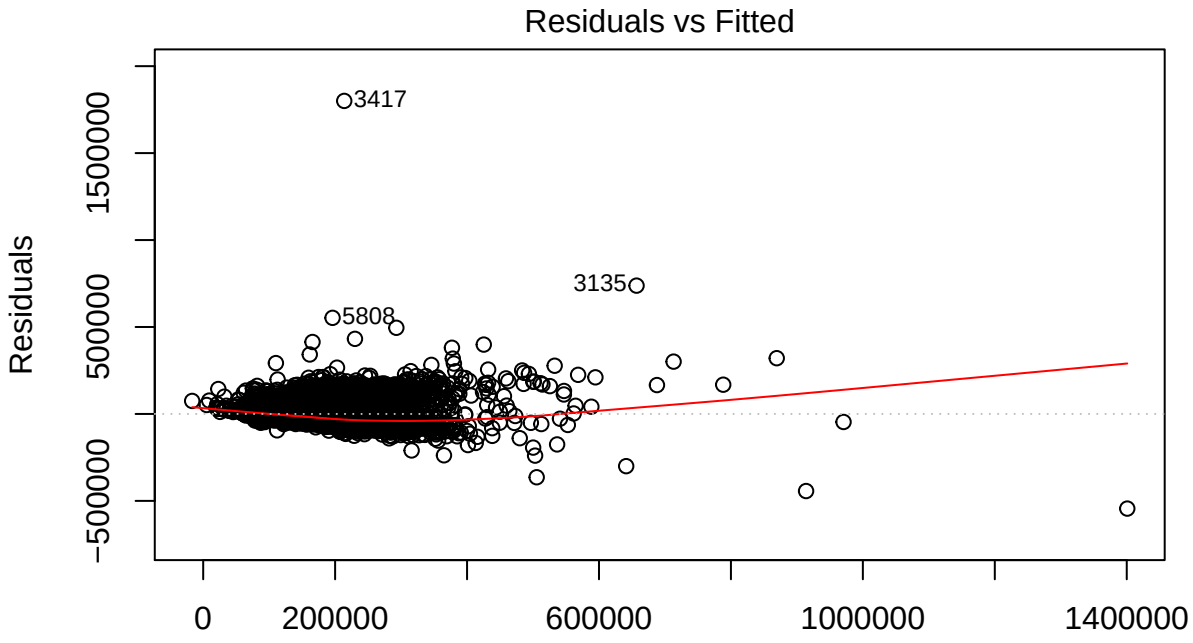
Continuing to look at the old data fit, we can see that all the variables are significant in the prediction.

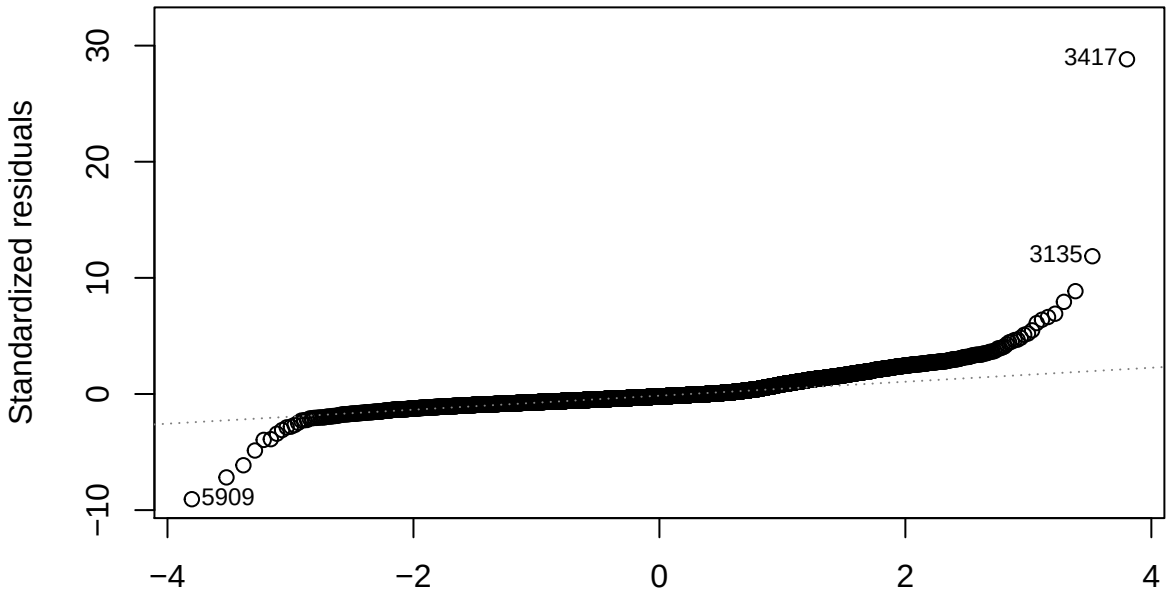
```
summary(fit)
```

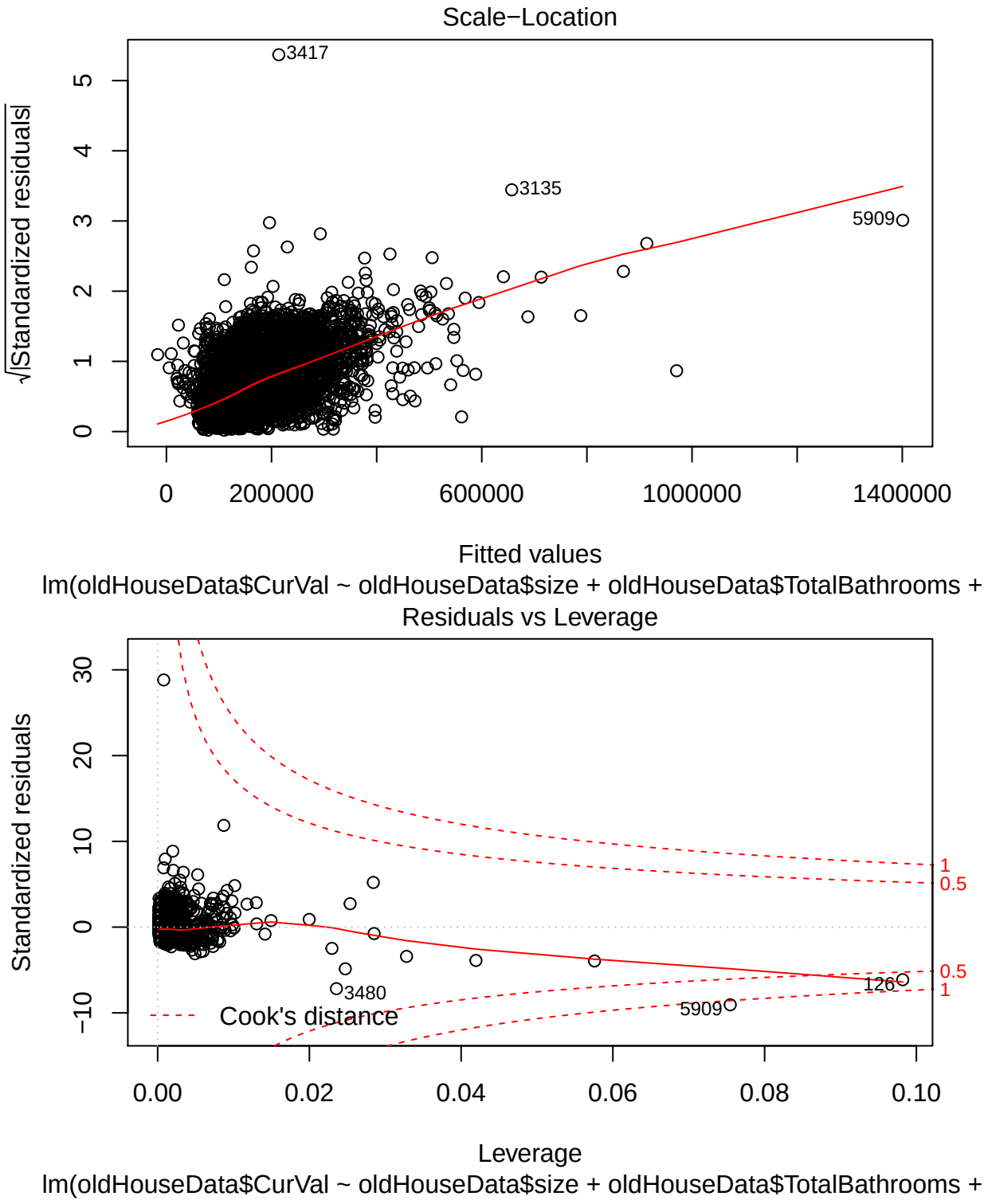
```
##
## Call:
## lm(formula = oldHouseData$CurVal ~ oldHouseData$size + oldHouseData$TotalBathrooms +
##     oldHouseData$TotalBedrooms + oldHouseData$LivingArea + oldHouseData$Depreciation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -544248  -34608  -13376   16139 1800135
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 112367.38    3270.66  34.356   <2e-16 ***
## oldHouseData$size           166265.40    7764.13  21.415   <2e-16 ***
## oldHouseData$TotalBathrooms   9298.43    1078.79   8.619   <2e-16 ***
## oldHouseData$TotalBedrooms  -13061.40     628.56 -20.780   <2e-16 ***
## oldHouseData$LivingArea         61.74       1.10  56.147   <2e-16 ***
## oldHouseData$Depreciation  -266957.07    9196.79 -29.027   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62480 on 6958 degrees of freedom
## Multiple R-squared:  0.5736, Adjusted R-squared:  0.5733
## F-statistic:  1872 on 5 and 6958 DF,  p-value: < 2.2e-16
```

So the price of a pre 1910 house can be modeled by: price = 112367.38 + 166265.40(size) + 9298.43(num bathroom) + -13061.40(num bedroom)+ 61.74(living space) -266957.07(depreciation).

```
plot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(oldHouseData$CurVal ~ oldHouseData$size + oldHouseData$TotalBathrooms +
3417
3135
5808
−500000
0
500000
1500000
0
200000
600000
1000000
1400000

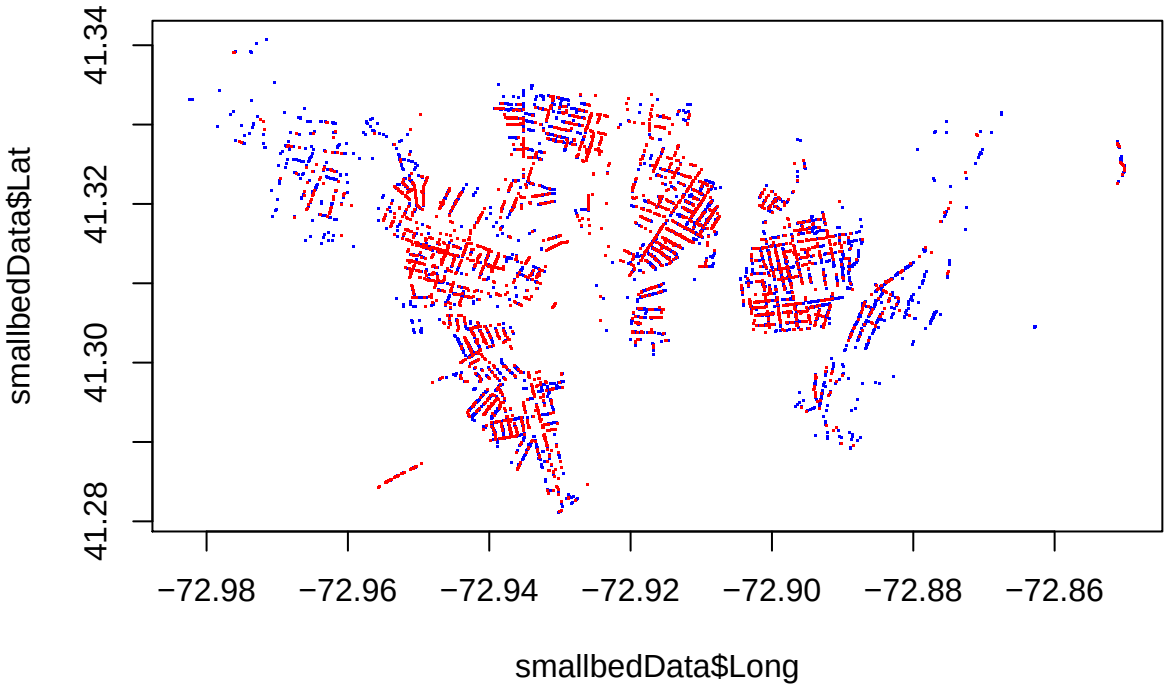

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(oldHouseData$CurVal ~ oldHouseData$size + oldHouseData$TotalBathrooms +
3417
3135
5909
−10
0
10
20
30
−4
−2
0
2
4



It is pretty interesting that increasing the price of a bedroom would decrease the price of an old house. We can look at this geographically at where the high and low number of bedroom apartments are.

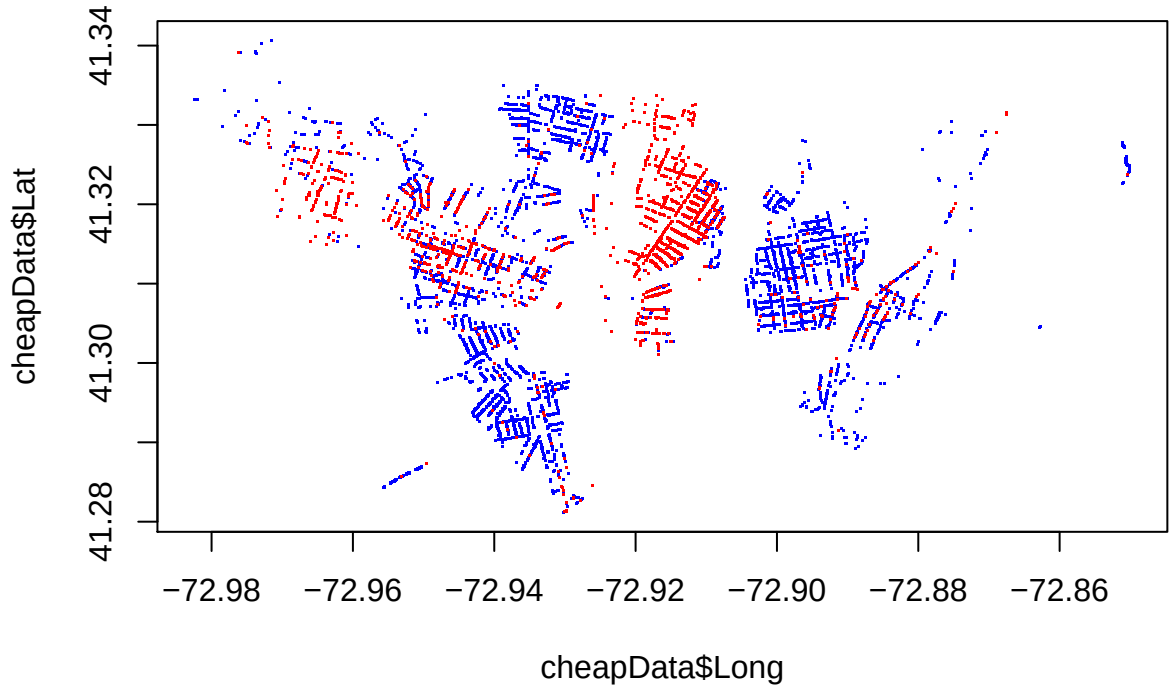
```
numBed = 5
smallbedData = subset(oldHouseData, TotalBedrooms<numBed)
largebedData = subset(oldHouseData, TotalBedrooms>=numBed)
```

```
plot(smallbedData$Long, smallbedData$Lat, pch='.', col='blue')
points(largebedData$Long, largebedData$Lat, pch='.', col='red')
```



The small number of bedrooms (less than 4) is shown in blue. The larger number of bedrooms and in red. There doesn't seem to be a difference in location between the number of bedrooms.

```
cheaper = 200000
cheapData = subset(oldHouseData, CurVal<cheaper)
priceyData = subset(oldHouseData, CurVal>=cheaper)
plot(cheapData$Long, cheapData$Lat, pch='.', col='blue')
points(priceyData$Long, priceyData$Lat, pch='.', col='red')
```



However, the more expensive houses are shown in red and are definitely in the North and west of town.

```
bedroomFit <- lm(CurVal ~ TotalBedrooms , oldHouseData)
summary(bedroomFit)
```

```
##
## Call:
## lm(formula = CurVal ~ TotalBedrooms, data = oldHouseData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -203212  -52554  -25823   26260 1809647
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    90484.6     3047.9   29.69   <2e-16 ***
## TotalBedrooms  18973.0      616.8   30.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 89750 on 6962 degrees of freedom
## Multiple R-squared:  0.1197, Adjusted R-squared:  0.1195
## F-statistic: 946.3 on 1 and 6962 DF,  p-value: < 2.2e-16
```

Since the bedroom is positive on its own, there has to be interactions with the other variables that are causing it to become negative when combined with the other variables. Overall, we have found that we can model the price of an old house with the following variables: size, bathrooms, bedrooms, living area, depreciation. The adjusted R^2 was 0.5733, showing it to be a reasonably good fit.