Project

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

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

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.0.4
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.0.5
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-3
```

```
library(pls)
```

```
## Warning: package 'pls' was built under R version 4.0.5
```

```
##
## Attaching package: 'pls'
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

```
library(mgcv)
```

```
## Loading required package: nlme
## This is mgcv 1.8-33. For overview type 'help("mgcv-package")'.
```

```
data <- read.csv("kc_house_data.csv", header = TRUE)
head(data)
```

```
##           id            date   price bedrooms bathrooms sqft_living sqft_lot
## 1 7129300520 20141013T000000  221900        3      1.00        1180     5650
## 2 6414100192 20141209T000000  538000        3      2.25        2570     7242
## 3 5631500400 20150225T000000  180000        2      1.00         770    10000
## 4 2487200875 20141209T000000  604000        4      3.00        1960     5000
## 5 1954400510 20150218T000000  510000        3      2.00        1680     8080
## 6 7237550310 20140512T000000 1225000        4      4.50        5420   101930
##   floors waterfront view condition grade sqft_above sqft_basement yr_built
## 1      1          0    0         3     7       1180             0     1955
## 2      2          0    0         3     7       2170           400     1951
## 3      1          0    0         3     6        770             0     1933
## 4      1          0    0         5     7       1050           910     1965
## 5      1          0    0         3     8       1680             0     1987
## 6      1          0    0         3    11       3890          1530     2001
##   yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 1            0   98178 47.5112 -122.257          1340       5650
## 2         1991   98125 47.7210 -122.319          1690       7639
## 3            0   98028 47.7379 -122.233          2720       8062
## 4            0   98136 47.5208 -122.393          1360       5000
## 5            0   98074 47.6168 -122.045          1800       7503
## 6            0   98053 47.6561 -122.005          4760     101930
```

```
dim(data)
```

```
## [1] 21613    21
```

```
# Summary statistics of the data
summary(data)
```

```
##        id                date              price            bedrooms
##  Min.   :1.000e+06   Length:21613       Min.   :  75000   Min.   : 0.000
##  1st Qu.:2.123e+09   Class :character   1st Qu.: 321950   1st Qu.: 3.000
##  Median :3.905e+09   Mode  :character   Median : 450000   Median : 3.000
##  Mean   :4.580e+09                      Mean   : 540088   Mean   : 3.371
##  3rd Qu.:7.309e+09                      3rd Qu.: 645000   3rd Qu.: 4.000
##  Max.   :9.900e+09                      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
```

```
# Determining whether any data is missing
sum(is.na(data))
```

```
## [1] 0
```

```
# Investigating the 33 bedroom property as this seems to be an outlier
data[data$bedrooms==33,]
```

```
##               id            date  price bedrooms bathrooms sqft_living sqft_lot
## 15871 2402100895 20140625T000000 640000       33      1.75        1620     6000
##       floors waterfront view condition grade sqft_above sqft_basement yr_built
## 15871      1          0    0         5     7       1040           580     1947
##       yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 15871            0   98103 47.6878 -122.331          1330       4700
```

```
data[data$bedrooms == 0,]
```

```
##               id            date   price bedrooms bathrooms sqft_living
## 876   6306400140 20140612T000000 1095000        0      0.00        3064
## 3120  3918400017 20150205T000000  380000        0      0.00        1470
## 3468  1453602309 20140805T000000  288000        0      1.50        1430
## 4869  6896300380 20141002T000000  228000        0      1.00         390
## 6995  2954400190 20140624T000000 1295650        0      0.00        4810
## 8478  2569500210 20141117T000000  339950        0      2.50        2290
## 8485  2310060040 20140925T000000  240000        0      2.50        1810
## 9774  3374500520 20150429T000000  355000        0      0.00        2460
## 9855  7849202190 20141223T000000  235000        0      0.00        1470
## 12654 7849202299 20150218T000000  320000        0      2.50        1490
## 14424 9543000205 20150413T000000  139950        0      0.00         844
## 18380 1222029077 20141029T000000  265000        0      0.75         384
## 19453 3980300371 20140926T000000  142000        0      0.00         290
##       sqft_lot floors waterfront view condition grade sqft_above sqft_basement
## 876       4764    3.5          0    2         3     7       3064             0
## 3120       979    3.0          0    2         3     8       1470             0
## 3468      1650    3.0          0    0         3     7       1430             0
## 4869      5900    1.0          0    0         2     4        390             0
## 6995     28008    2.0          0    0         3    12       4810             0
## 8478      8319    2.0          0    0         3     8       2290             0
## 8485      5669    2.0          0    0         3     7       1810             0
## 9774      8049    2.0          0    0         3     8       2460             0
## 9855      4800    2.0          0    0         3     7       1470             0
## 12654     7111    2.0          0    0         3     7       1490             0
## 14424     4269    1.0          0    0         4     7        844             0
## 18380   213444    1.0          0    0         3     4        384             0
## 19453    20875    1.0          0    0         1     1        290             0
##       yr_built yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 876       1990            0   98102 47.6362 -122.322          2360       4000
## 3120      2006            0   98133 47.7145 -122.356          1470       1399
## 3468      1999            0   98125 47.7222 -122.290          1430       1650
## 4869      1953            0   98118 47.5260 -122.261          2170       6000
## 6995      1990            0   98053 47.6642 -122.069          4740      35061
## 8478      1985            0   98042 47.3473 -122.151          2500       8751
## 8485      2003            0   98038 47.3493 -122.053          1810       5685
## 9774      1990            0   98031 47.4095 -122.168          2520       8050
## 9855      1996            0   98065 47.5265 -121.828          1060       7200
## 12654     1999            0   98065 47.5261 -121.826          1500       4675
## 14424     1913            0   98001 47.2781 -122.250          1380       9600
## 18380     2003            0   98070 47.4177 -122.491          1920     224341
## 19453     1963            0   98024 47.5308 -121.888          1620      22850
```

```
data[data$bathrooms == 0, ]
```

```
##               id            date   price bedrooms bathrooms sqft_living
## 876   6306400140 20140612T000000 1095000        0         0        3064
## 1150  3421079032 20150217T000000   75000        1         0         670
## 3120  3918400017 20150205T000000  380000        0         0        1470
## 5833  5702500050 20141104T000000  280000        1         0         600
## 6995  2954400190 20140624T000000 1295650        0         0        4810
## 9774  3374500520 20150429T000000  355000        0         0        2460
## 9855  7849202190 20141223T000000  235000        0         0        1470
## 10482  203100435 20140918T000000  484000        1         0         690
## 14424 9543000205 20150413T000000  139950        0         0         844
## 19453 3980300371 20140926T000000  142000        0         0         290
##       sqft_lot floors waterfront view condition grade sqft_above sqft_basement
## 876       4764    3.5          0    2         3     7       3064             0
## 1150     43377    1.0          0    0         3     3        670             0
## 3120       979    3.0          0    2         3     8       1470             0
## 5833     24501    1.0          0    0         2     3        600             0
## 6995     28008    2.0          0    0         3    12       4810             0
## 9774      8049    2.0          0    0         3     8       2460             0
## 9855      4800    2.0          0    0         3     7       1470             0
## 10482    23244    1.0          0    0         4     7        690             0
## 14424     4269    1.0          0    0         4     7        844             0
## 19453    20875    1.0          0    0         1     1        290             0
##       yr_built yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 876       1990            0   98102 47.6362 -122.322          2360       4000
## 1150      1966            0   98022 47.2638 -121.906          1160      42882
## 3120      2006            0   98133 47.7145 -122.356          1470       1399
## 5833      1950            0   98045 47.5316 -121.749           990      22549
## 6995      1990            0   98053 47.6642 -122.069          4740      35061
## 9774      1990            0   98031 47.4095 -122.168          2520       8050
## 9855      1996            0   98065 47.5265 -121.828          1060       7200
## 10482     1948            0   98053 47.6429 -121.955          1690      19290
## 14424     1913            0   98001 47.2781 -122.250          1380       9600
## 19453     1963            0   98024 47.5308 -121.888          1620      22850
```

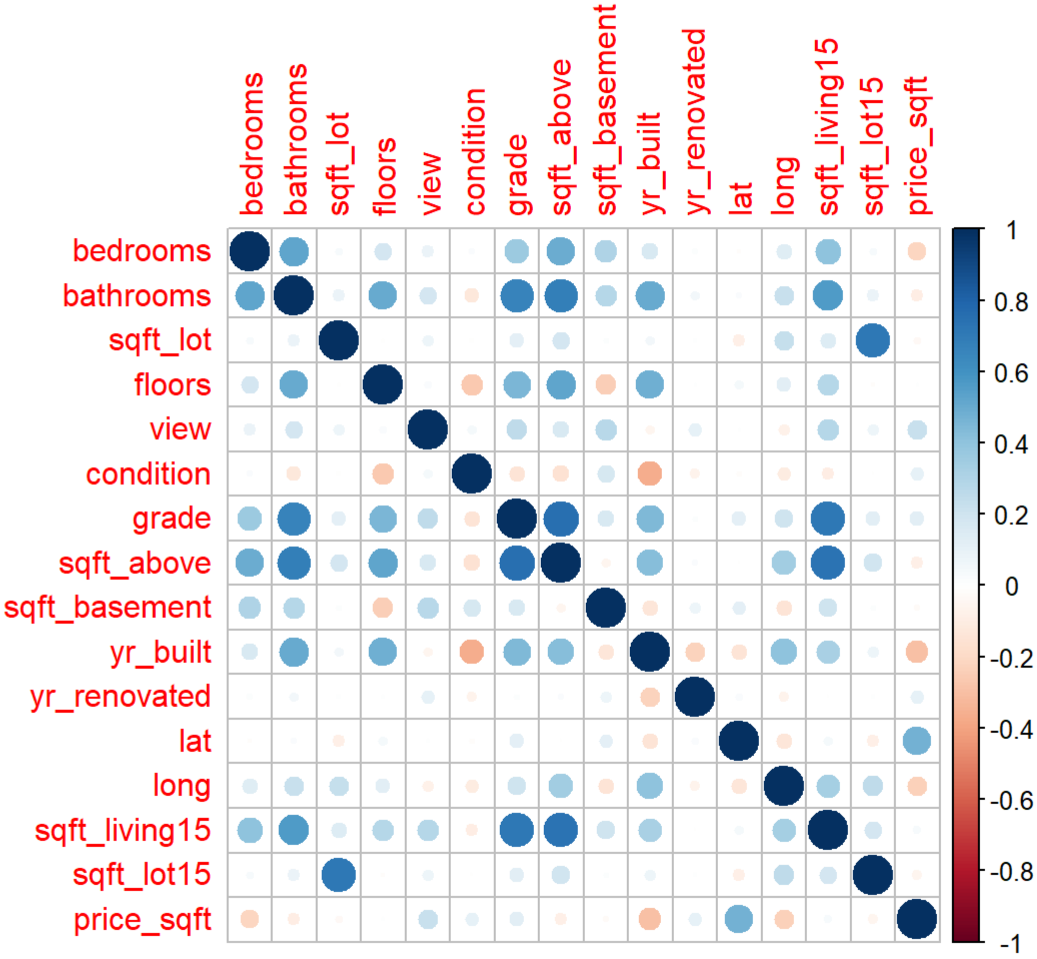
```
# Removing properties that are unusual (33 bedrooms, no bedrooms or bathrooms)
data <- data[-which.max(data$bedrooms),]
data <- data[-which(data$bedrooms == 0),]
data <- data[-which(data$bathrooms == 0),]

# Removing the ID column as this is just a unique identifier for each property
data <- data[,-1]

# Converting categorical variables to factors
data$zipcode <- as.factor(data$zipcode)

# Converting to price per square foot
data$price_sqft <- data$price / data$sqft_living
data <- data[,c(-2, -5)]

data$waterfront <- as.factor(data$waterfront)

summary(data)
```

```
##      date              bedrooms        bathrooms        sqft_lot
##  Length:21596       Min.   : 1.000   Min.   :0.500   Min.   :    520
##  Class :character   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.:   5040
##  Mode  :character   Median : 3.000   Median :2.250   Median :   7619
##                     Mean   : 3.372   Mean   :2.116   Mean   :  15100
##                     3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.:  10686
##                     Max.   :11.000   Max.   :8.000   Max.   :1651359
##
##     floors      waterfront      view         condition        grade
##  Min.   :1.000   0:21433    Min.   :0.0000   Min.   :1.00   Min.   : 3.000
##  1st Qu.:1.000   1:  163    1st Qu.:0.0000   1st Qu.:3.00   1st Qu.: 7.000
##  Median :1.500              Median :0.0000   Median :3.00   Median : 7.000
##  Mean   :1.494              Mean   :0.2343   Mean   :3.41   Mean   : 7.658
##  3rd Qu.:2.000              3rd Qu.:0.0000   3rd Qu.:4.00   3rd Qu.: 8.000
##  Max.   :3.500              Max.   :4.0000   Max.   :5.00   Max.   :13.000
##
##    sqft_above   sqft_basement       yr_built     yr_renovated
##  Min.   : 370   Min.   :   0.0   Min.   :1900   Min.   :   0.00
##  1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.00
##  Median :1560   Median :   0.0   Median :1975   Median :   0.00
##  Mean   :1789   Mean   : 291.7   Mean   :1971   Mean   :  84.47
##  3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.00
##  Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.00
##
##     zipcode           lat             long        sqft_living15
##  98103  :  601   Min.   :47.16   Min.   :-122.5   Min.   : 399
##  98038  :  589   1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490
##  98115  :  583   Median :47.57   Median :-122.2   Median :1840
##  98052  :  574   Mean   :47.56   Mean   :-122.2   Mean   :1987
##  98117  :  553   3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360
##  98042  :  547   Max.   :47.78   Max.   :-121.3   Max.   :6210
##  (Other):18149
##    sqft_lot15       price_sqft
##  Min.   :   651   Min.   : 87.59
##  1st Qu.:  5100   1st Qu.:182.29
##  Median :  7620   Median :244.63
##  Mean   : 12759   Mean   :264.11
##  3rd Qu.: 10083   3rd Qu.:318.27
##  Max.   :871200   Max.   :810.14
##
```

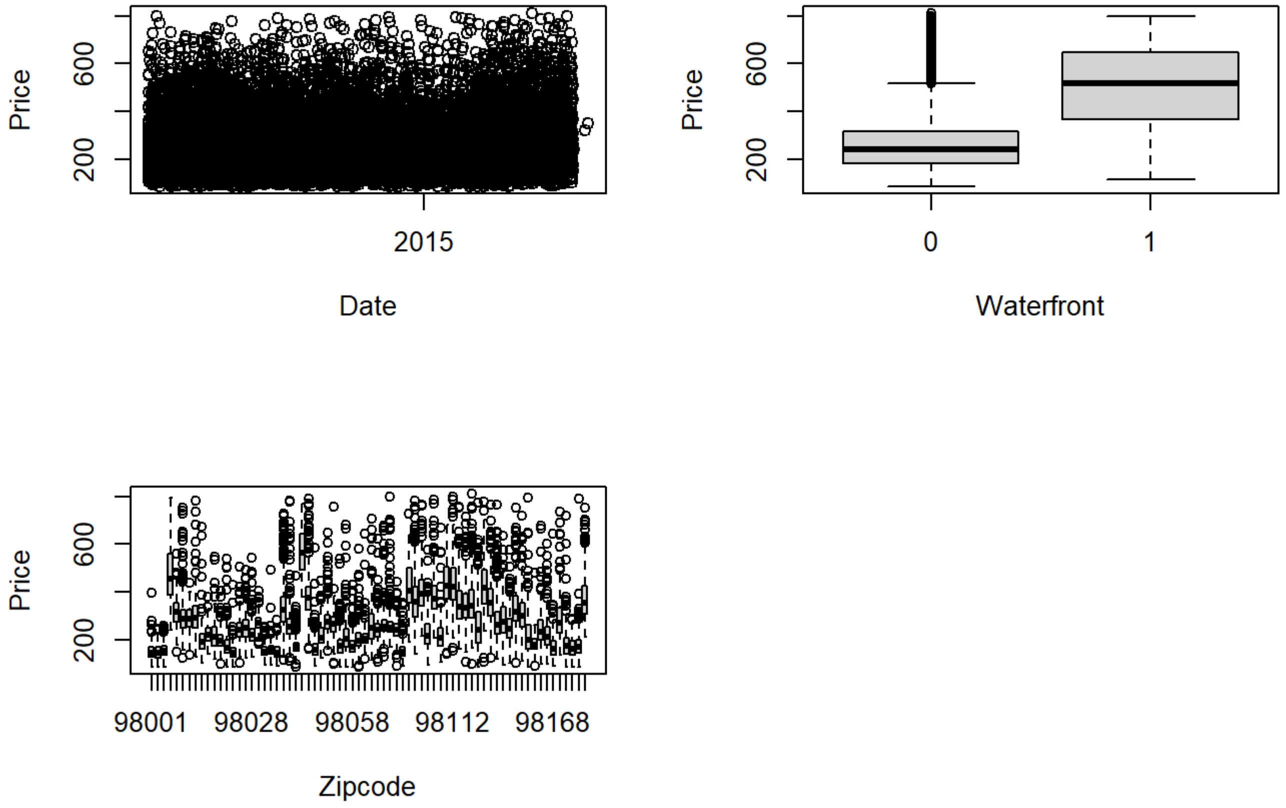
```
# Converting the date data to dates in R
Dates <- NULL
for (i in data$date) {
  # Grabbing the date in format yyyymmdd
  d <- substr(i, 1, 8)
  Dates <- rbind(Dates, d)
}

# Converting strings to dates
Dates <- as.Date(Dates, "%Y%m%d")

# Updating the date column to the new date values
data$date <- Dates

head(data$date)
```

```
## [1] "2014-10-13" "2014-12-09" "2015-02-25" "2014-12-09" "2015-02-18"
## [6] "2014-05-12"
```

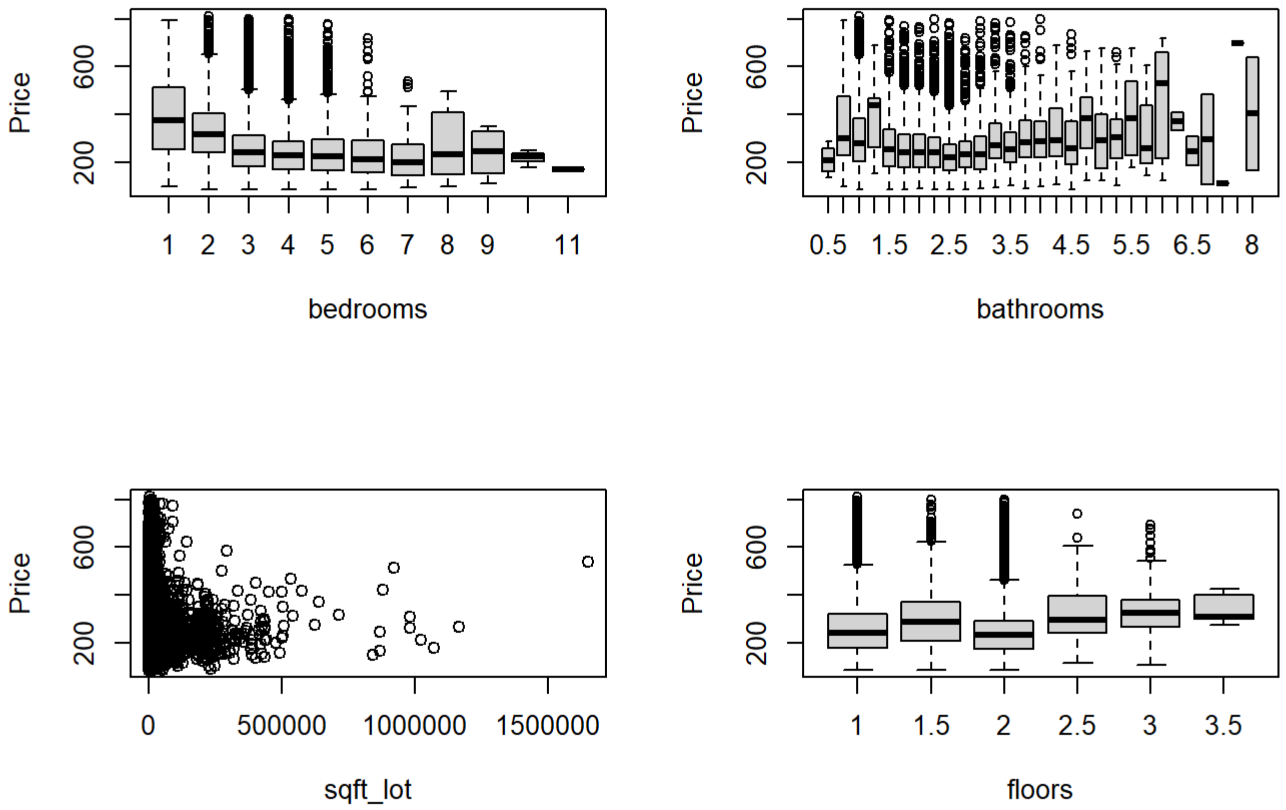
```
# Correlation between the numeric predictors
corrplot::corrplot(cor(data[,c(-1, -6, -14)]))
```



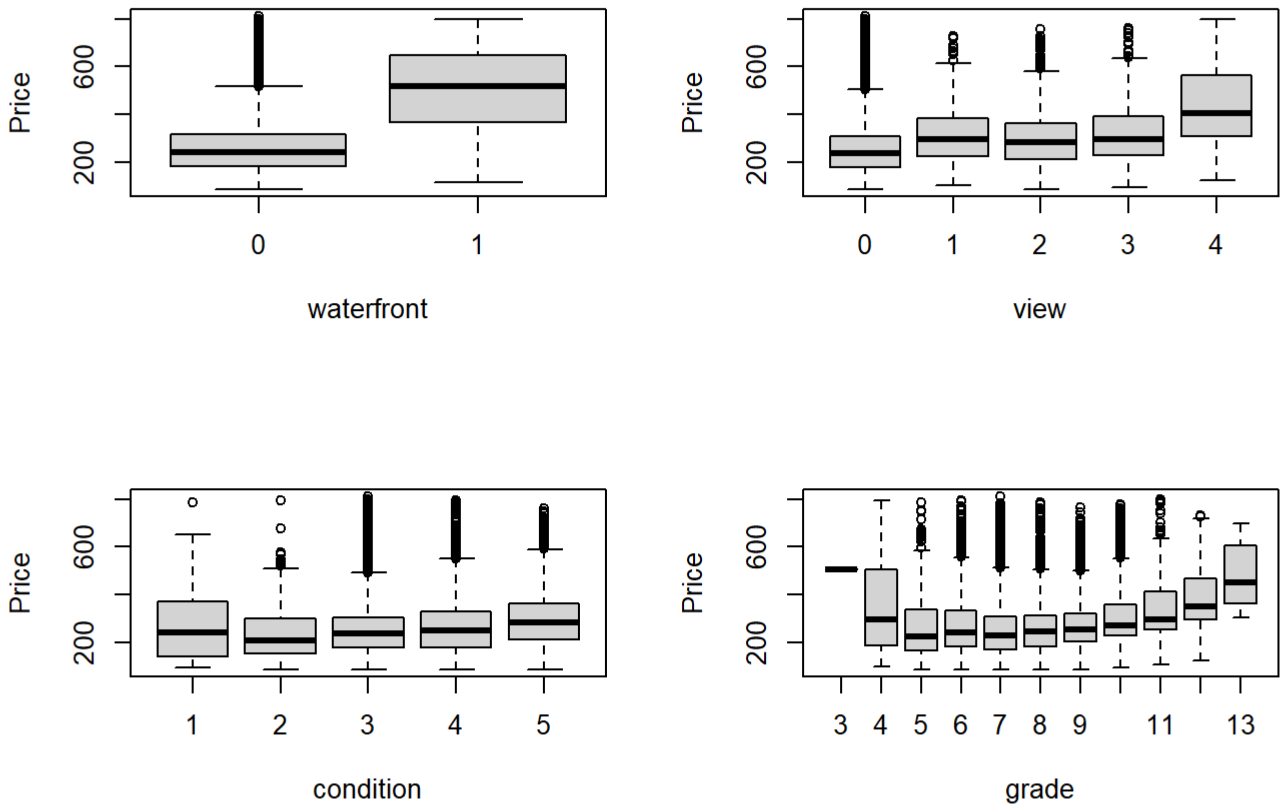
```
par(mfrow=c(2,2))
plot(data$date, data$price, xlab = "Date", ylab = "Price")
plot(data$waterfront, data$price, xlab = "Waterfront", ylab = "Price")
plot(data$zipcode, data$price, xlab = "Zipcode", ylab = "Price")
```



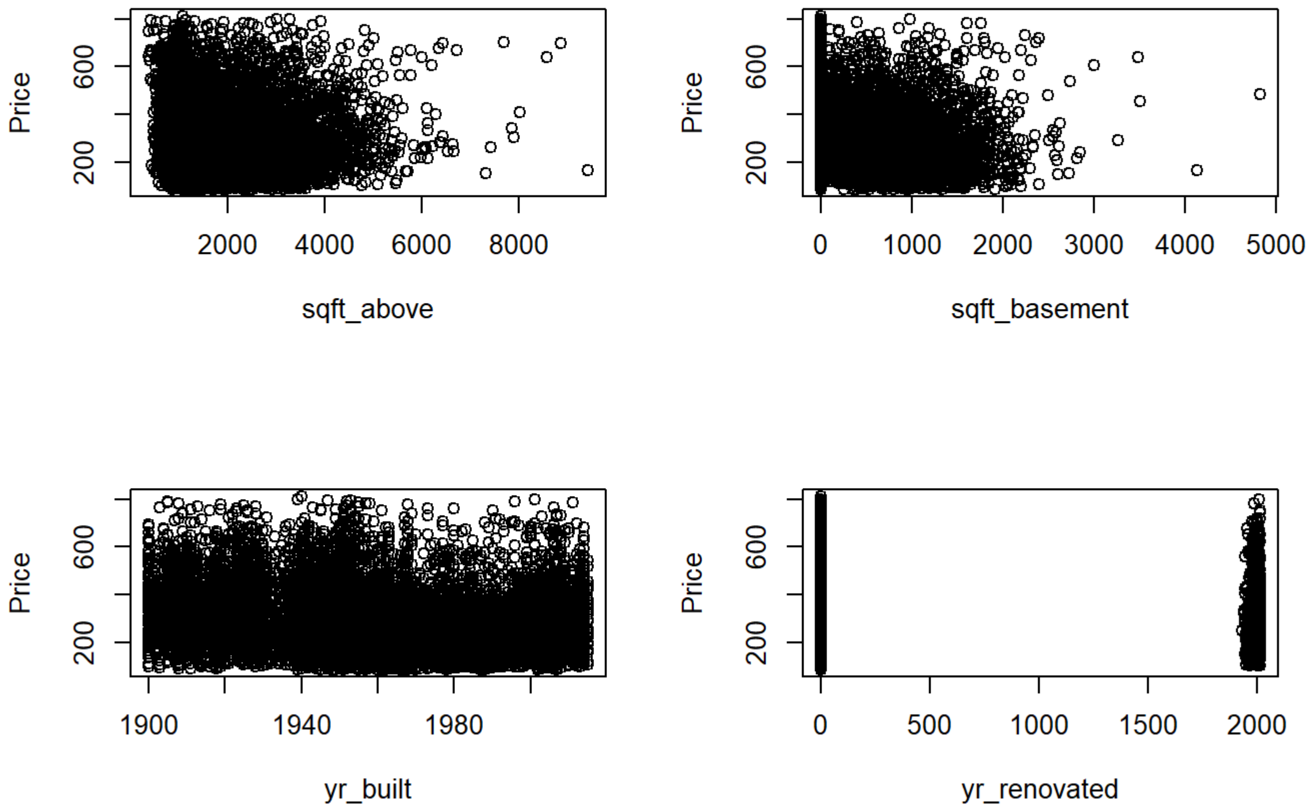
```
num.cols <- colnames(data[,c(-1, -14)])

par(mfrow=c(2,2))
for (i in num.cols[1:4]) {
  if (i == num.cols[3]) {
    plot(data[,i], data$price, xlab = i, ylab = "Price")
  } else {
    plot(as.factor(data[,i]), data$price, xlab = i, ylab = "Price")
  }
}
```



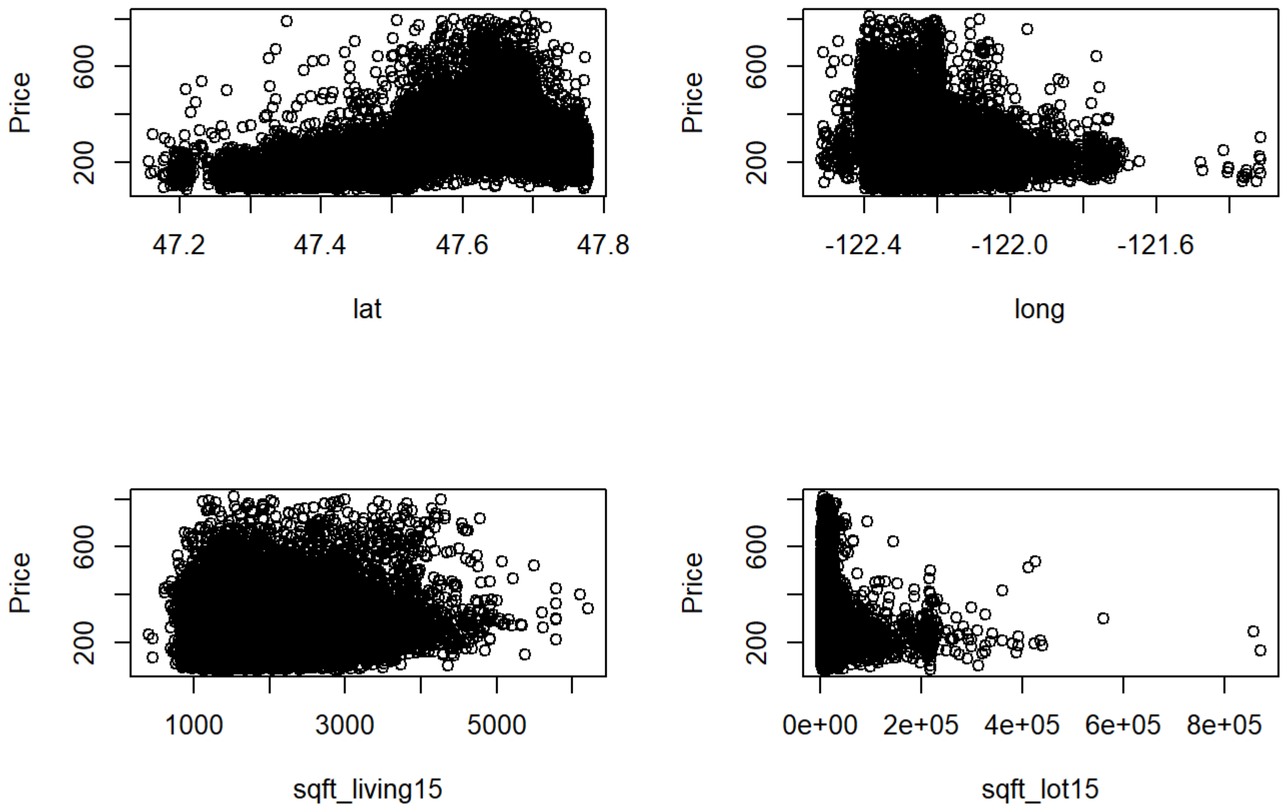
```
par(mfrow=c(2,2))
for (i in num.cols[5:8]) {
  plot(as.factor(data[,i]), data$price, xlab = i, ylab = "Price")
}
```



```
par(mfrow=c(2,2))
for (i in num.cols[9:12]) {
  plot(data[,i], data$price, xlab = i, ylab = "Price")
}
```



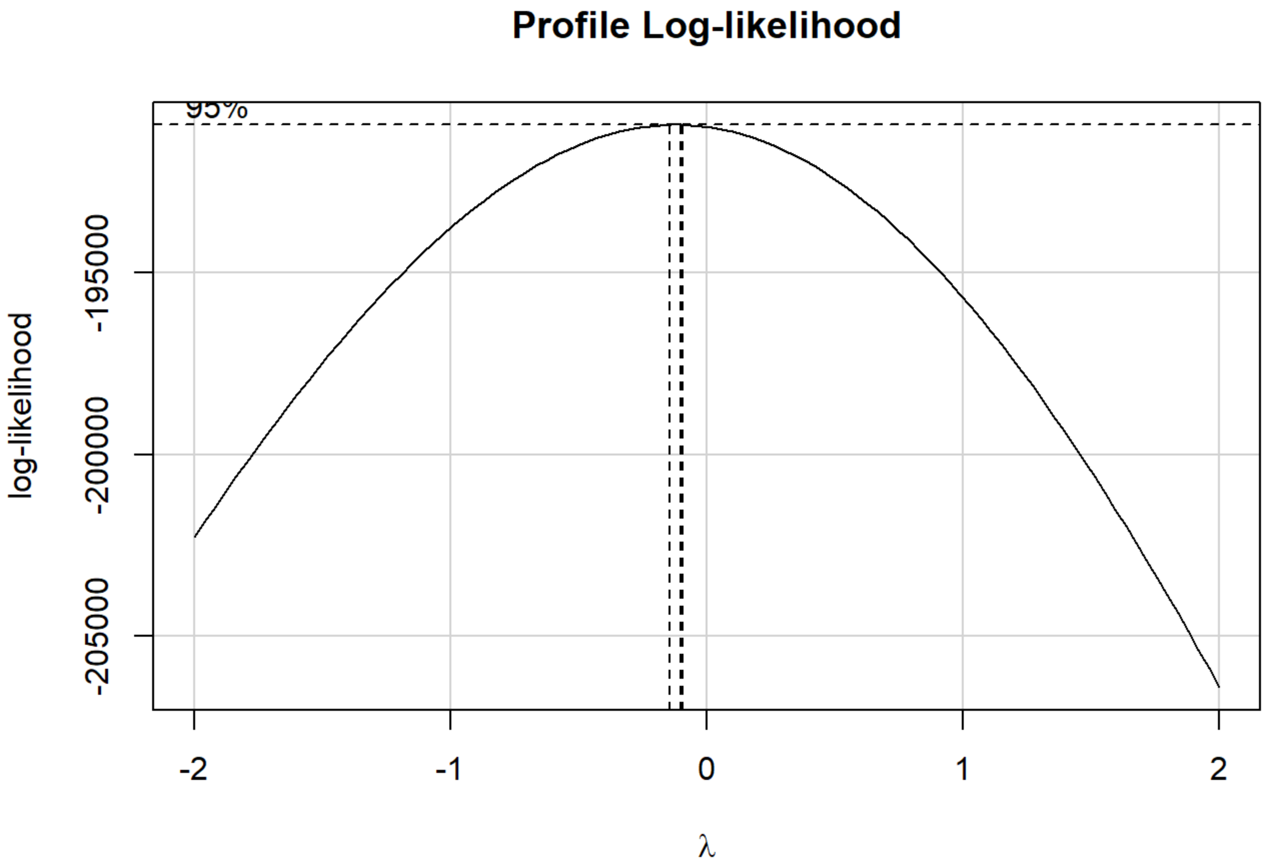
```
par(mfrow=c(2,2))
for (i in num.cols[13:16]) {
  plot(data[,i], data$price, xlab = i, ylab = "Price")
}
```



```
# Building the full model

lm1 <- lm(price_sqft ~ ., data = data)
summary(lm1)
```

```
##
## Call:
## lm(formula = price_sqft ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -344.37  -30.92   -3.56   23.75  564.97
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.555e+04  2.245e+03  -6.924 4.51e-12 ***
## date           6.210e-02  3.555e-03  17.467  < 2e-16 ***
## bedrooms      -1.318e+01  5.867e-01 -22.463  < 2e-16 ***
## bathrooms      4.360e+00  9.644e-01   4.521 6.18e-06 ***
## sqft_lot       1.738e-04  1.404e-05  12.380  < 2e-16 ***
## floors        -1.946e+01  1.156e+00 -16.830  < 2e-16 ***
## waterfront1    1.972e+02  5.166e+00  38.173  < 2e-16 ***
## view           1.964e+01  6.416e-01  30.613  < 2e-16 ***
## condition      1.019e+01  7.057e-01  14.432  < 2e-16 ***
## grade          1.853e+01  6.668e-01  27.785  < 2e-16 ***
## sqft_above    -3.769e-02  1.111e-03 -33.918  < 2e-16 ***
## sqft_basement -7.571e-02  1.303e-03 -58.118  < 2e-16 ***
## yr_built      -2.491e-01  2.376e-02 -10.481  < 2e-16 ***
## yr_renovated   6.515e-03  1.077e-03   6.050 1.47e-09 ***
## zipcode98002   3.456e+00  5.287e+00   0.654 0.513295
## zipcode98003  -5.960e+00  4.727e+00  -1.261 0.207366
## zipcode98004   3.072e+02  8.590e+00  35.756  < 2e-16 ***
## zipcode98005   1.466e+02  9.183e+00  15.970  < 2e-16 ***
## zipcode98006   1.335e+02  7.508e+00  17.783  < 2e-16 ***
## zipcode98007   1.255e+02  9.476e+00  13.245  < 2e-16 ***
## zipcode98008   1.278e+02  9.001e+00  14.197  < 2e-16 ***
## zipcode98010   7.238e+01  8.060e+00   8.979  < 2e-16 ***
## zipcode98011   5.167e+01  1.171e+01   4.413 1.02e-05 ***
## zipcode98014   6.823e+01  1.286e+01   5.304 1.14e-07 ***
## zipcode98019   5.327e+01  1.269e+01   4.199 2.70e-05 ***
## zipcode98022   2.850e+01  7.004e+00   4.069 4.74e-05 ***
## zipcode98023  -1.619e+01  4.349e+00  -3.722 0.000198 ***
## zipcode98024   9.702e+01  1.132e+01   8.572  < 2e-16 ***
## zipcode98027   1.097e+02  7.706e+00  14.240  < 2e-16 ***
## zipcode98028   4.333e+01  1.137e+01   3.810 0.000140 ***
## zipcode98029   1.262e+02  8.804e+00  14.337  < 2e-16 ***
## zipcode98030   8.807e+00  5.198e+00   1.694 0.090231 .
## zipcode98031   1.133e+01  5.414e+00   2.092 0.036459 *
## zipcode98032  -6.628e+00  6.283e+00  -1.055 0.291485
## zipcode98033   1.664e+02  9.756e+00  17.053  < 2e-16 ***
## zipcode98034   8.474e+01  1.046e+01   8.101 5.75e-16 ***
## zipcode98038   4.329e+01  5.837e+00   7.416 1.25e-13 ***
## zipcode98039   4.004e+02  1.161e+01  34.498  < 2e-16 ***
## zipcode98040   2.076e+02  7.598e+00  27.329  < 2e-16 ***
## zipcode98042   1.973e+01  4.975e+00   3.966 7.33e-05 ***
## zipcode98045   9.305e+01  1.078e+01   8.628  < 2e-16 ***
## zipcode98052   1.145e+02  9.960e+00  11.497  < 2e-16 ***
## zipcode98053   1.080e+02  1.067e+01  10.116  < 2e-16 ***
## zipcode98055   1.854e+01  6.029e+00   3.075 0.002105 **
## zipcode98056   5.373e+01  6.550e+00   8.203 2.47e-16 ***
## zipcode98058   2.649e+01  5.696e+00   4.650 3.33e-06 ***
## zipcode98059   5.860e+01  6.426e+00   9.119  < 2e-16 ***
## zipcode98065   8.977e+01  9.942e+00   9.029  < 2e-16 ***
## zipcode98070   1.336e+01  7.591e+00   1.760 0.078442 .
## zipcode98072   6.987e+01  1.165e+01   5.999 2.02e-09 ***
## zipcode98074   1.055e+02  9.432e+00  11.187  < 2e-16 ***
## zipcode98075   1.097e+02  9.070e+00  12.099  < 2e-16 ***
## zipcode98077   6.872e+01  1.212e+01   5.670 1.45e-08 ***
## zipcode98092   9.197e+00  4.727e+00   1.946 0.051723 .
## zipcode98102   2.401e+02  1.007e+01  23.850  < 2e-16 ***
## zipcode98103   1.763e+02  9.434e+00  18.684  < 2e-16 ***
## zipcode98105   2.159e+02  9.685e+00  22.289  < 2e-16 ***
## zipcode98106   5.525e+01  6.988e+00   7.907 2.77e-15 ***
## zipcode98107   1.846e+02  9.725e+00  18.977  < 2e-16 ***
## zipcode98108   5.568e+01  7.716e+00   7.215 5.56e-13 ***
## zipcode98109   2.353e+02  1.002e+01  23.484  < 2e-16 ***
## zipcode98112   2.603e+02  8.891e+00  29.276  < 2e-16 ***
## zipcode98115   1.663e+02  9.589e+00  17.343  < 2e-16 ***
## zipcode98116   1.494e+02  7.803e+00  19.143  < 2e-16 ***
## zipcode98117   1.647e+02  9.711e+00  16.956  < 2e-16 ***
## zipcode98118   8.552e+01  6.816e+00  12.548  < 2e-16 ***
## zipcode98119   2.331e+02  9.464e+00  24.630  < 2e-16 ***
## zipcode98122   1.793e+02  8.443e+00  21.231  < 2e-16 ***
## zipcode98125   8.862e+01  1.036e+01   8.556  < 2e-16 ***
## zipcode98126   9.821e+01  7.166e+00  13.706  < 2e-16 ***
## zipcode98133   5.668e+01  1.069e+01   5.301 1.17e-07 ***
## zipcode98136   1.341e+02  7.346e+00  18.262  < 2e-16 ***
## zipcode98144   1.326e+02  7.848e+00  16.897  < 2e-16 ***
## zipcode98146   3.411e+01  6.558e+00   5.201 2.00e-07 ***
## zipcode98148   1.183e+01  8.920e+00   1.327 0.184654
## zipcode98155   4.839e+01  1.112e+01   4.352 1.36e-05 ***
## zipcode98166   3.135e+01  6.002e+00   5.223 1.77e-07 ***
## zipcode98168   3.029e+00  6.343e+00   0.477 0.633012
## zipcode98177   8.181e+01  1.116e+01   7.329 2.40e-13 ***
## zipcode98178   1.459e+01  6.553e+00   2.227 0.025945 *
## zipcode98188   8.551e+00  6.725e+00   1.272 0.203535
## zipcode98198  -6.999e+00  5.096e+00  -1.374 0.169575
## zipcode98199   1.760e+02  9.219e+00  19.095  < 2e-16 ***
## lat            7.763e+01  2.318e+01   3.349 0.000812 ***
## long          -9.349e+01  1.665e+01  -5.617 1.97e-08 ***
## sqft_living15  2.098e-02  1.056e-03  19.858  < 2e-16 ***
## sqft_lot15    -1.130e-06  2.211e-05  -0.051 0.959250
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.71 on 21509 degrees of freedom
## Multiple R-squared:  0.716,  Adjusted R-squared:  0.7149
## F-statistic: 630.7 on 86 and 21509 DF,  p-value: < 2.2e-16
```

```
bc <- boxCox(lm1)
```

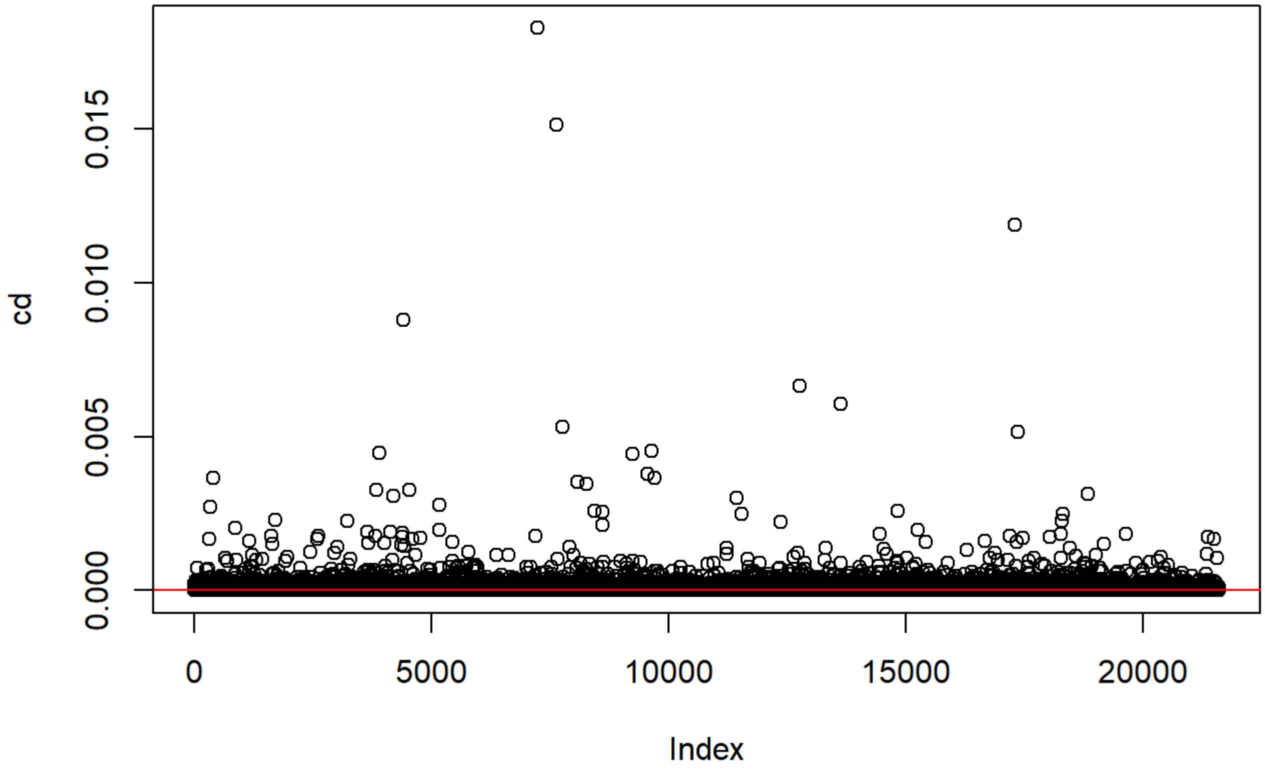


```
opt_lambda <- bc$x[which.max(bc$y)]
```

```
# Log transformation of dependent variable

lm2 <- lm(log(price_sqft) ~ ., data = data)
summary(lm2)
```

```
##
## Call:
## lm(formula = log(price_sqft) ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.19773 -0.10794 -0.00113  0.10475  1.60278
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.465e+01  7.398e+00  -7.387 1.56e-13 ***
## date           2.097e-04  1.172e-05  17.901  < 2e-16 ***
## bedrooms      -4.504e-02  1.933e-03 -23.296  < 2e-16 ***
## bathrooms      2.007e-02  3.178e-03   6.314 2.76e-10 ***
## sqft_lot       6.820e-07  4.626e-08  14.742  < 2e-16 ***
## floors        -6.277e-02  3.811e-03 -16.470  < 2e-16 ***
## waterfront1    5.007e-01  1.703e-02  29.409  < 2e-16 ***
## view           6.652e-02  2.114e-03  31.462  < 2e-16 ***
## condition      4.704e-02  2.326e-03  20.228  < 2e-16 ***
## grade          7.928e-02  2.198e-03  36.076  < 2e-16 ***
## sqft_above    -1.555e-04  3.662e-06 -42.475  < 2e-16 ***
## sqft_basement -2.827e-04  4.293e-06 -65.862  < 2e-16 ***
## yr_built      -6.389e-04  7.832e-05  -8.158 3.59e-16 ***
## yr_renovated   2.873e-05  3.548e-06   8.097 5.92e-16 ***
## zipcode98002   8.574e-03  1.742e-02   0.492 0.622666
## zipcode98003  -2.222e-03  1.558e-02  -0.143 0.886573
## zipcode98004   1.019e+00  2.831e-02  35.990  < 2e-16 ***
## zipcode98005   6.121e-01  3.026e-02  20.227  < 2e-16 ***
## zipcode98006   5.762e-01  2.474e-02  23.287  < 2e-16 ***
## zipcode98007   5.413e-01  3.123e-02  17.335  < 2e-16 ***
## zipcode98008   5.440e-01  2.966e-02  18.338  < 2e-16 ***
## zipcode98010   3.159e-01  2.656e-02  11.894  < 2e-16 ***
## zipcode98011   2.390e-01  3.859e-02   6.194 5.96e-10 ***
## zipcode98014   2.596e-01  4.239e-02   6.124 9.30e-10 ***
## zipcode98019   2.007e-01  4.181e-02   4.802 1.59e-06 ***
## zipcode98022   1.538e-01  2.308e-02   6.664 2.74e-11 ***
## zipcode98023  -5.905e-02  1.433e-02  -4.120 3.80e-05 ***
## zipcode98024   4.181e-01  3.730e-02  11.208  < 2e-16 ***
## zipcode98027   4.854e-01  2.540e-02  19.116  < 2e-16 ***
## zipcode98028   2.036e-01  3.748e-02   5.432 5.64e-08 ***
## zipcode98029   5.513e-01  2.901e-02  19.003  < 2e-16 ***
## zipcode98030   3.301e-02  1.713e-02   1.927 0.054023 .
## zipcode98031   4.568e-02  1.784e-02   2.560 0.010472 *
## zipcode98032  -3.542e-02  2.071e-02  -1.711 0.087167 .
## zipcode98033   6.385e-01  3.215e-02  19.860  < 2e-16 ***
## zipcode98034   3.674e-01  3.447e-02  10.659  < 2e-16 ***
## zipcode98038   1.906e-01  1.924e-02   9.909  < 2e-16 ***
## zipcode98039   1.228e+00  3.825e-02  32.099  < 2e-16 ***
## zipcode98040   7.961e-01  2.504e-02  31.796  < 2e-16 ***
## zipcode98042   8.060e-02  1.640e-02   4.916 8.89e-07 ***
## zipcode98045   3.993e-01  3.554e-02  11.234  < 2e-16 ***
## zipcode98052   4.914e-01  3.282e-02  14.971  < 2e-16 ***
## zipcode98053   4.632e-01  3.517e-02  13.169  < 2e-16 ***
## zipcode98055   9.486e-02  1.987e-02   4.774 1.81e-06 ***
## zipcode98056   2.559e-01  2.159e-02  11.855  < 2e-16 ***
## zipcode98058   1.278e-01  1.877e-02   6.809 1.01e-11 ***
## zipcode98059   2.821e-01  2.118e-02  13.320  < 2e-16 ***
## zipcode98065   3.944e-01  3.276e-02  12.038  < 2e-16 ***
## zipcode98070   1.832e-01  2.502e-02   7.322 2.53e-13 ***
## zipcode98072   3.047e-01  3.838e-02   7.939 2.14e-15 ***
## zipcode98074   4.637e-01  3.108e-02  14.919  < 2e-16 ***
## zipcode98075   4.869e-01  2.989e-02  16.292  < 2e-16 ***
## zipcode98077   3.035e-01  3.994e-02   7.599 3.10e-14 ***
## zipcode98092   3.869e-02  1.558e-02   2.483 0.013017 *
## zipcode98102   8.434e-01  3.318e-02  25.419  < 2e-16 ***
## zipcode98103   6.572e-01  3.109e-02  21.139  < 2e-16 ***
## zipcode98105   7.833e-01  3.192e-02  24.543  < 2e-16 ***
## zipcode98106   2.723e-01  2.303e-02  11.822  < 2e-16 ***
## zipcode98107   6.841e-01  3.205e-02  21.346  < 2e-16 ***
## zipcode98108   2.602e-01  2.543e-02  10.231  < 2e-16 ***
## zipcode98109   8.189e-01  3.302e-02  24.802  < 2e-16 ***
## zipcode98112   8.998e-01  2.930e-02  30.709  < 2e-16 ***
## zipcode98115   6.416e-01  3.160e-02  20.304  < 2e-16 ***
## zipcode98116   6.098e-01  2.571e-02  23.716  < 2e-16 ***
## zipcode98117   6.267e-01  3.200e-02  19.584  < 2e-16 ***
## zipcode98118   3.779e-01  2.246e-02  16.827  < 2e-16 ***
## zipcode98119   8.029e-01  3.119e-02  25.746  < 2e-16 ***
## zipcode98122   6.787e-01  2.782e-02  24.394  < 2e-16 ***
## zipcode98125   3.808e-01  3.413e-02  11.156  < 2e-16 ***
## zipcode98126   4.503e-01  2.361e-02  19.068  < 2e-16 ***
## zipcode98133   2.609e-01  3.524e-02   7.403 1.38e-13 ***
## zipcode98136   5.657e-01  2.421e-02  23.372  < 2e-16 ***
## zipcode98144   5.573e-01  2.586e-02  21.549  < 2e-16 ***
## zipcode98146   1.891e-01  2.161e-02   8.749  < 2e-16 ***
## zipcode98148   9.729e-02  2.940e-02   3.310 0.000935 ***
## zipcode98155   2.292e-01  3.664e-02   6.254 4.08e-10 ***
## zipcode98166   2.147e-01  1.978e-02  10.855  < 2e-16 ***
## zipcode98168   2.663e-02  2.090e-02   1.274 0.202698
## zipcode98177   3.562e-01  3.678e-02   9.683  < 2e-16 ***
## zipcode98178   8.576e-02  2.159e-02   3.972 7.16e-05 ***
## zipcode98188   5.001e-02  2.216e-02   2.257 0.024041 *
## zipcode98198   1.873e-02  1.679e-02   1.115 0.264808
## zipcode98199   6.690e-01  3.038e-02  22.022  < 2e-16 ***
## lat            4.718e-01  7.638e-02   6.177 6.64e-10 ***
## long          -2.847e-01  5.485e-02  -5.190 2.12e-07 ***
## sqft_living15  6.183e-05  3.481e-06  17.760  < 2e-16 ***
## sqft_lot15     1.069e-07  7.286e-08   1.467 0.142473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1935 on 21509 degrees of freedom
## Multiple R-squared:  0.7605, Adjusted R-squared:  0.7596
## F-statistic: 794.2 on 86 and 21509 DF,  p-value: < 2.2e-16
```

```
# Cook's distance to find outliers

cd <- cooks.distance(lm2)

plot(cd)
abline(h = quantile(cd, 4/nrow(data)), col = 'red')
```



```
# Removing outliers in the 99th percentile of cook's distance

#outliers <- which(cd > quantile(cd, .995))
outliers <- which(cd > 4/nrow(data))

data2 <- data[-outliers,]

summary(data2)
```

```
##       date               bedrooms        bathrooms        sqft_lot     
##  Min.   :2014-05-02   Min.   : 1.000   Min.   :0.500   Min.   :   520  
##  1st Qu.:2014-07-21   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.:  5001  
##  Median :2014-10-15   Median : 3.000   Median :2.250   Median :  7553  
##  Mean   :2014-10-28   Mean   : 3.378   Mean   :2.115   Mean   : 13619  
##  3rd Qu.:2015-02-17   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 10440  
##  Max.   :2015-05-27   Max.   :10.000   Max.   :6.750   Max.   :982998  
##                                                                        
##      floors      waterfront      view          condition         grade       
##  Min.   :1.000   0:20277    Min.   :0.0000   Min.   :1.000   Min.   : 4.000  
##  1st Qu.:1.000   1:   97    1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000  
##  Median :1.500              Median :0.0000   Median :3.000   Median : 7.000  
##  Mean   :1.498              Mean   :0.2121   Mean   :3.415   Mean   : 7.661  
##  3rd Qu.:2.000              3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000  
##  Max.   :3.500              Max.   :4.0000   Max.   :5.000   Max.   :13.000  
##                                                                               
##    sqft_above   sqft_basement       yr_built     yr_renovated    
##  Min.   : 390   Min.   :   0.0   Min.   :1900   Min.   :   0.00  
##  1st Qu.:1200   1st Qu.:   0.0   1st Qu.:1953   1st Qu.:   0.00  
##  Median :1560   Median :   0.0   Median :1976   Median :   0.00  
##  Mean   :1779   Mean   : 285.5   Mean   :1972   Mean   :  78.49  
##  3rd Qu.:2200   3rd Qu.: 550.0   3rd Qu.:1997   3rd Qu.:   0.00  
##  Max.   :7420   Max.   :3260.0   Max.   :2015   Max.   :2015.00  
##                                                                  
##     zipcode           lat             long        sqft_living15 
##  98103  :  595   Min.   :47.16   Min.   :-122.5   Min.   : 460  
##  98038  :  576   1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490  
##  98052  :  568   Median :47.57   Median :-122.2   Median :1840  
##  98115  :  567   Mean   :47.56   Mean   :-122.2   Mean   :1983  
##  98117  :  547   3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360  
##  98042  :  536   Max.   :47.78   Max.   :-121.3   Max.   :5790  
##  (Other):16985                                                  
##    sqft_lot15       price_sqft    
##  Min.   :   651   Min.   : 88.08  
##  1st Qu.:  5080   1st Qu.:183.49  
##  Median :  7565   Median :243.98  
##  Mean   : 11926   Mean   :260.95  
##  3rd Qu.:  9976   3rd Qu.:314.63  
##  Max.   :560617   Max.   :800.00  
## 
```

```
lm3 <- lm(log(price_sqft) ~ ., data = data2)
summary(lm3)
```

```
## 
## Call:
## lm(formula = log(price_sqft) ~ ., data = data2)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.53336 -0.09760 -0.00049  0.09603  0.52377 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   -3.902e+01  6.641e+00  -5.876 4.27e-09 ***
## date           1.943e-04  9.542e-06  20.367  < 2e-16 ***
## bedrooms      -3.274e-02  1.620e-03 -20.213  < 2e-16 ***
## bathrooms      1.831e-02  2.643e-03   6.929 4.37e-12 ***
## sqft_lot       8.735e-07  5.384e-08  16.225  < 2e-16 ***
## floors        -6.210e-02  3.141e-03 -19.771  < 2e-16 ***
## waterfront1    5.464e-01  1.681e-02  32.496  < 2e-16 ***
## view           6.558e-02  1.776e-03  36.931  < 2e-16 ***
## condition      4.421e-02  1.929e-03  22.919  < 2e-16 ***
## grade          8.409e-02  1.834e-03  45.840  < 2e-16 ***
## sqft_above    -1.851e-04  3.175e-06 -58.287  < 2e-16 ***
## sqft_basement -3.156e-04  3.658e-06 -86.265  < 2e-16 ***
## yr_built      -5.645e-04  6.547e-05  -8.622  < 2e-16 ***
## yr_renovated   3.448e-05  2.986e-06  11.549  < 2e-16 ***
## zipcode98002   1.293e-02  1.419e-02   0.911 0.362388    
## zipcode98003   1.125e-02  1.262e-02   0.891 0.372848    
## zipcode98004   1.055e+00  2.341e-02  45.074  < 2e-16 ***
## zipcode98005   6.570e-01  2.509e-02  26.182  < 2e-16 ***
## zipcode98006   5.974e-01  2.066e-02  28.917  < 2e-16 ***
## zipcode98007   5.784e-01  2.592e-02  22.319  < 2e-16 ***
## zipcode98008   5.660e-01  2.481e-02  22.811  < 2e-16 ***
## zipcode98010   2.520e-01  2.398e-02  10.511  < 2e-16 ***
## zipcode98011   2.989e-01  3.179e-02   9.401  < 2e-16 ***
## zipcode98014   2.717e-01  3.703e-02   7.337 2.26e-13 ***
## zipcode98019   2.338e-01  3.549e-02   6.588 4.57e-11 ***
## zipcode98022   1.174e-01  1.966e-02   5.969 2.43e-09 ***
## zipcode98023  -4.360e-02  1.175e-02  -3.712 0.000207 ***
## zipcode98024   3.976e-01  3.442e-02  11.549  < 2e-16 ***
## zipcode98027   5.065e-01  2.173e-02  23.314  < 2e-16 ***
## zipcode98028   2.608e-01  3.082e-02   8.463  < 2e-16 ***
## zipcode98029   5.590e-01  2.486e-02  22.485  < 2e-16 ***
## zipcode98030   4.477e-02  1.400e-02   3.198 0.001385 ** 
## zipcode98031   5.616e-02  1.461e-02   3.843 0.000122 ***
## zipcode98032  -2.093e-02  1.754e-02  -1.193 0.232747    
## zipcode98033   6.620e-01  2.664e-02  24.844  < 2e-16 ***
## zipcode98034   4.059e-01  2.849e-02  14.248  < 2e-16 ***
## zipcode98038   1.813e-01  1.668e-02  10.866  < 2e-16 ***
## zipcode98039   1.279e+00  3.486e-02  36.700  < 2e-16 ***
## zipcode98040   8.226e-01  2.063e-02  39.869  < 2e-16 ***
## zipcode98042   7.332e-02  1.384e-02   5.296 1.20e-07 ***
## zipcode98045   3.885e-01  3.181e-02  12.213  < 2e-16 ***
## zipcode98052   5.266e-01  2.743e-02  19.197  < 2e-16 ***
## zipcode98053   5.019e-01  2.978e-02  16.855  < 2e-16 ***
## zipcode98055   1.085e-01  1.647e-02   6.590 4.49e-11 ***
## zipcode98056   2.775e-01  1.786e-02  15.540  < 2e-16 ***
## zipcode98058   1.278e-01  1.567e-02   8.156 3.67e-16 ***
## zipcode98059   2.960e-01  1.769e-02  16.736  < 2e-16 ***
## zipcode98065   3.972e-01  2.883e-02  13.778  < 2e-16 ***
## zipcode98070   2.318e-01  2.247e-02  10.316  < 2e-16 ***
## zipcode98072   3.476e-01  3.190e-02  10.895  < 2e-16 ***
## zipcode98074   4.878e-01  2.636e-02  18.505  < 2e-16 ***
## zipcode98075   5.036e-01  2.547e-02  19.777  < 2e-16 ***
## zipcode98077   3.393e-01  3.344e-02  10.146  < 2e-16 ***
## zipcode98092   2.638e-02  1.277e-02   2.066 0.038797 *
## zipcode98102   8.785e-01  2.767e-02  31.745  < 2e-16 ***
## zipcode98103   7.079e-01  2.546e-02  27.808  < 2e-16 ***
## zipcode98105   8.209e-01  2.620e-02  31.327  < 2e-16 ***
## zipcode98106   3.115e-01  1.890e-02  16.475  < 2e-16 ***
## zipcode98107   7.255e-01  2.621e-02  27.675  < 2e-16 ***
## zipcode98108   3.114e-01  2.112e-02  14.746  < 2e-16 ***
## zipcode98109   8.601e-01  2.765e-02  31.104  < 2e-16 ***
## zipcode98112   9.361e-01  2.424e-02  38.615  < 2e-16 ***
## zipcode98115   6.977e-01  2.592e-02  26.920  < 2e-16 ***
## zipcode98116   6.753e-01  2.105e-02  32.083  < 2e-16 ***
## zipcode98117   6.860e-01  2.621e-02  26.179  < 2e-16 ***
## zipcode98118   4.075e-01  1.845e-02  22.090  < 2e-16 ***
## zipcode98119   8.593e-01  2.579e-02  33.314  < 2e-16 ***
## zipcode98122   7.048e-01  2.285e-02  30.846  < 2e-16 ***
## zipcode98125   4.325e-01  2.802e-02  15.435  < 2e-16 ***
## zipcode98126   4.939e-01  1.936e-02  25.510  < 2e-16 ***
## zipcode98133   3.229e-01  2.886e-02  11.186  < 2e-16 ***
## zipcode98136   6.131e-01  1.987e-02  30.853  < 2e-16 ***
## zipcode98144   5.891e-01  2.123e-02  27.747  < 2e-16 ***
## zipcode98146   2.224e-01  1.793e-02  12.401  < 2e-16 ***
## zipcode98148   1.495e-01  2.771e-02   5.395 6.93e-08 ***
## zipcode98155   2.939e-01  3.008e-02   9.771  < 2e-16 ***
## zipcode98166   2.635e-01  1.634e-02  16.131  < 2e-16 ***
## zipcode98168   6.402e-02  1.716e-02   3.732 0.000191 ***
## zipcode98177   3.928e-01  3.029e-02  12.967  < 2e-16 ***
## zipcode98178   1.127e-01  1.779e-02   6.335 2.43e-10 ***
## zipcode98188   6.673e-02  1.855e-02   3.596 0.000324 ***
## zipcode98198   4.129e-02  1.383e-02   2.985 0.002844 **
## zipcode98199   7.254e-01  2.488e-02  29.155  < 2e-16 ***
## lat            3.630e-01  6.301e-02   5.760 8.52e-09 ***
## long          -1.997e-01  5.095e-02  -3.919 8.92e-05 ***
## sqft_living15  6.622e-05  2.948e-06  22.465  < 2e-16 ***
## sqft_lot15     5.253e-08  7.546e-08   0.696 0.486318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1531 on 20287 degrees of freedom
## Multiple R-squared:  0.8346, Adjusted R-squared:  0.8339
## F-statistic:  1191 on 86 and 20287 DF,  p-value: < 2.2e-16
```

```
# Checking for multicolinearity
vif(lm3)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## date              1.011040  1        1.005505
## bedrooms          1.771346  1        1.330919
## bathrooms         3.280819  1        1.811303
## sqft_lot          2.578861  1        1.605883
## floors            2.511135  1        1.584656
## waterfront        1.164729  1        1.079226
## view              1.429693  1        1.195698
## condition         1.347910  1        1.160995
## grade             3.682230  1        1.918914
## sqft_above        5.480691  1        2.341088
## sqft_basement     2.143457  1        1.464055
## yr_built          3.147902  1        1.774233
## yr_renovated      1.166822  1        1.080195
## zipcode        8923.672817 69        1.068137
## lat              67.094646  1        8.191132
## long             43.426209  1        6.589857
## sqft_living15     3.417277  1        1.848588
## sqft_lot15        2.725768  1        1.650990
```

```
# Dropping zipcode as it is colinear with lat and long
lm4 <- lm(log(price_sqft) ~ . -zipcode, data = data2)

summary(lm4)
```

```
##
## Call:
## lm(formula = log(price_sqft) ~ . - zipcode, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89301 -0.16046 -0.00222  0.15000  0.99761
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.590e+01  1.979e+00 -33.298  < 2e-16 ***
## date           1.820e-04  1.470e-05  12.380  < 2e-16 ***
## bedrooms      -4.925e-02  2.459e-03 -20.024  < 2e-16 ***
## bathrooms      4.739e-02  4.042e-03  11.724  < 2e-16 ***
## sqft_lot       5.970e-07  8.278e-08   7.212 5.71e-13 ***
## floors         4.867e-02  4.375e-03  11.123  < 2e-16 ***
## waterfront1    4.447e-01  2.571e-02  17.298  < 2e-16 ***
## view           6.705e-02  2.661e-03  25.196  < 2e-16 ***
## condition      5.312e-02  2.885e-03  18.416  < 2e-16 ***
## grade          1.568e-01  2.673e-03  58.648  < 2e-16 ***
## sqft_above    -2.552e-04  4.749e-06 -53.742  < 2e-16 ***
## sqft_basement -2.880e-04  5.574e-06 -51.672  < 2e-16 ***
## yr_built      -3.736e-03  8.961e-05 -41.690  < 2e-16 ***
## yr_renovated   3.280e-05  4.567e-06   7.182 7.12e-13 ***
## lat            1.372e+00  1.263e-02 108.629  < 2e-16 ***
## long          -7.751e-02  1.481e-02  -5.232 1.69e-07 ***
## sqft_living15  8.202e-05  4.331e-06  18.937  < 2e-16 ***
## sqft_lot15    -3.004e-07  1.134e-07  -2.648   0.0081 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2363 on 20356 degrees of freedom
## Multiple R-squared:  0.6044, Adjusted R-squared:  0.6041
## F-statistic:  1829 on 17 and 20356 DF,  p-value: < 2.2e-16
```

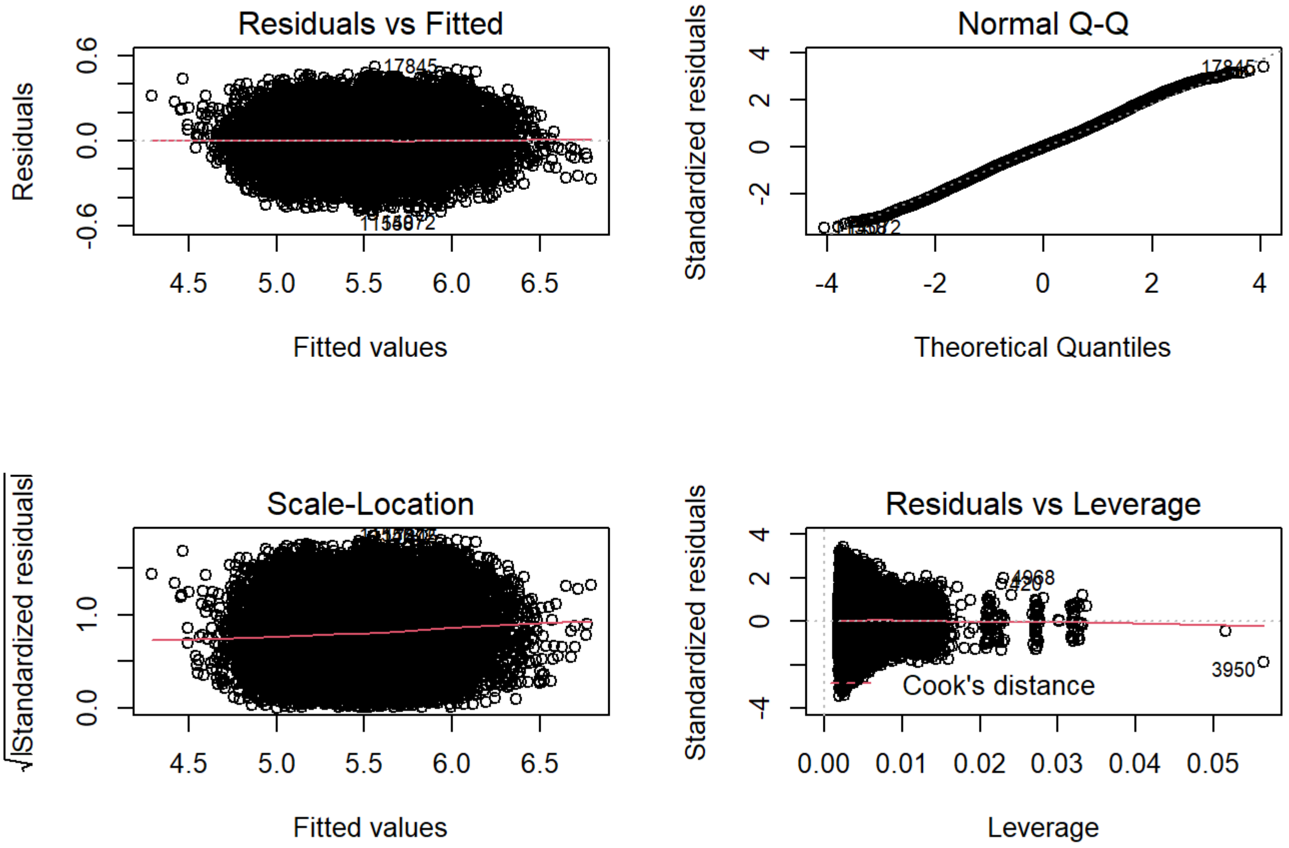
```
# Replacing lat and long with zipcode since the R2 value dropped drastically
lm5 <- lm(log(price_sqft) ~ . - lat - long, data=data2)
summary(lm5)
```

```
##
## Call:
## lm(formula = log(price_sqft) ~ . - lat - long, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52812 -0.09746 -0.00036  0.09593  0.52124
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.571e+00  2.028e-01  12.678  < 2e-16 ***
## date           1.950e-04  9.553e-06  20.414  < 2e-16 ***
## bedrooms      -3.283e-02  1.622e-03 -20.242  < 2e-16 ***
## bathrooms      1.833e-02  2.646e-03   6.928 4.41e-12 ***
## sqft_lot       8.649e-07  5.388e-08  16.052  < 2e-16 ***
## floors        -6.217e-02  3.142e-03 -19.787  < 2e-16 ***
## waterfront1    5.450e-01  1.683e-02  32.389  < 2e-16 ***
## view           6.552e-02  1.778e-03  36.855  < 2e-16 ***
## condition      4.387e-02  1.931e-03  22.718  < 2e-16 ***
## grade          8.454e-02  1.835e-03  46.061  < 2e-16 ***
## sqft_above    -1.857e-04  3.177e-06 -58.439  < 2e-16 ***
## sqft_basement -3.154e-04  3.663e-06 -86.113  < 2e-16 ***
## yr_built      -5.737e-04  6.549e-05  -8.759  < 2e-16 ***
## yr_renovated   3.445e-05  2.989e-06  11.524  < 2e-16 ***
## zipcode98002   1.515e-03  1.390e-02   0.109   0.9132
## zipcode98003   2.098e-02  1.248e-02   1.680   0.0929 .
## zipcode98004   1.153e+00  1.241e-02  92.897  < 2e-16 ***
## zipcode98005   7.459e-01  1.488e-02  50.136  < 2e-16 ***
## zipcode98006   6.624e-01  1.106e-02  59.893  < 2e-16 ***
## zipcode98007   6.619e-01  1.562e-02  42.378  < 2e-16 ***
## zipcode98008   6.453e-01  1.246e-02  51.789  < 2e-16 ***
## zipcode98010   2.074e-01  1.967e-02  10.548  < 2e-16 ***
## zipcode98011   4.465e-01  1.387e-02  32.193  < 2e-16 ***
## zipcode98014   3.252e-01  1.913e-02  17.004  < 2e-16 ***
## zipcode98019   3.274e-01  1.429e-02  22.910  < 2e-16 ***
## zipcode98022   2.834e-02  1.339e-02   2.117   0.0343 *
## zipcode98023  -2.548e-02  1.076e-02  -2.369   0.0178 *
## zipcode98024   4.178e-01  2.364e-02  17.675  < 2e-16 ***
## zipcode98027   5.387e-01  1.143e-02  47.123  < 2e-16 ***
## zipcode98028   4.174e-01  1.241e-02  33.625  < 2e-16 ***
## zipcode98029   5.966e-01  1.201e-02  49.690  < 2e-16 ***
## zipcode98030   5.023e-02  1.275e-02   3.940 8.19e-05 ***
## zipcode98031   7.408e-02  1.250e-02   5.924 3.19e-09 ***
## zipcode98032   3.704e-03  1.711e-02   0.217   0.8286
## zipcode98033   7.796e-01  1.128e-02  69.122  < 2e-16 ***
## zipcode98034   5.423e-01  1.060e-02  51.157  < 2e-16 ***
## zipcode98038   1.559e-01  1.047e-02  14.897  < 2e-16 ***
## zipcode98039   1.386e+00  2.851e-02  48.624  < 2e-16 ***
## zipcode98040   9.042e-01  1.284e-02  70.436  < 2e-16 ***
## zipcode98042   6.298e-02  1.058e-02   5.951 2.70e-09 ***
## zipcode98045   3.473e-01  1.362e-02  25.499  < 2e-16 ***
## zipcode98052   6.308e-01  1.053e-02  59.928  < 2e-16 ***
## zipcode98053   5.880e-01  1.145e-02  51.374  < 2e-16 ***
## zipcode98055   1.483e-01  1.289e-02  11.506  < 2e-16 ***
## zipcode98056   3.325e-01  1.134e-02  29.312  < 2e-16 ***
## zipcode98058   1.527e-01  1.101e-02  13.868  < 2e-16 ***
## zipcode98059   3.365e-01  1.099e-02  30.608  < 2e-16 ***
## zipcode98065   3.967e-01  1.217e-02  32.605  < 2e-16 ***
## zipcode98070   3.113e-01  1.921e-02  16.206  < 2e-16 ***
## zipcode98072   4.806e-01  1.258e-02  38.203  < 2e-16 ***
## zipcode98074   5.566e-01  1.121e-02  49.657  < 2e-16 ***
## zipcode98075   5.549e-01  1.183e-02  46.905  < 2e-16 ***
## zipcode98077   4.569e-01  1.413e-02  32.336  < 2e-16 ***
## zipcode98092   4.155e-03  1.177e-02   0.353   0.7240    
## zipcode98102   1.006e+00  1.868e-02  53.850  < 2e-16 ***
## zipcode98103   8.551e-01  1.080e-02  79.177  < 2e-16 ***
## zipcode98105   9.542e-01  1.369e-02  69.695  < 2e-16 ***
## zipcode98106   4.107e-01  1.211e-02  33.917  < 2e-16 ***
## zipcode98107   8.766e-01  1.294e-02  67.759  < 2e-16 ***
## zipcode98108   4.056e-01  1.474e-02  27.521  < 2e-16 ***
## zipcode98109   9.941e-01  1.843e-02  53.952  < 2e-16 ***
## zipcode98112   1.057e+00  1.357e-02  77.884  < 2e-16 ***
## zipcode98115   8.397e-01  1.072e-02  78.323  < 2e-16 ***
## zipcode98116   7.952e-01  1.219e-02  65.254  < 2e-16 ***
## zipcode98117   8.440e-01  1.083e-02  77.911  < 2e-16 ***
## zipcode98118   4.931e-01  1.107e-02  44.532  < 2e-16 ***
## zipcode98119   9.977e-01  1.511e-02  66.041  < 2e-16 ***
## zipcode98122   8.190e-01  1.290e-02  63.462  < 2e-16 ***
## zipcode98125   5.861e-01  1.142e-02  51.338  < 2e-16 ***
## zipcode98126   5.999e-01  1.195e-02  50.215  < 2e-16 ***
## zipcode98133   4.918e-01  1.088e-02  45.192  < 2e-16 ***
## zipcode98136   7.193e-01  1.299e-02  55.362  < 2e-16 ***
## zipcode98144   6.947e-01  1.217e-02  57.062  < 2e-16 ***
## zipcode98146   3.096e-01  1.288e-02  24.041  < 2e-16 ***
## zipcode98148   2.061e-01  2.654e-02   7.767 8.39e-15 ***
## zipcode98155   4.621e-01  1.109e-02  41.660  < 2e-16 ***
## zipcode98166   3.320e-01  1.321e-02  25.142  < 2e-16 ***
## zipcode98168   1.370e-01  1.288e-02  10.635  < 2e-16 ***
## zipcode98177   5.697e-01  1.313e-02  43.375  < 2e-16 ***
## zipcode98178   1.765e-01  1.301e-02  13.569  < 2e-16 ***
## zipcode98188   1.188e-01  1.642e-02   7.239 4.67e-13 ***
## zipcode98198   7.981e-02  1.274e-02   6.263 3.86e-10 ***
## zipcode98199   8.730e-01  1.231e-02  70.909  < 2e-16 ***
## sqft_living15  6.641e-05  2.951e-06  22.504  < 2e-16 ***
## sqft_lot15     2.940e-08  7.531e-08   0.390   0.6963    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1533 on 20289 degrees of freedom
## Multiple R-squared:  0.8342, Adjusted R-squared:  0.8335 
## F-statistic:  1215 on 84 and 20289 DF,  p-value: < 2.2e-16
```

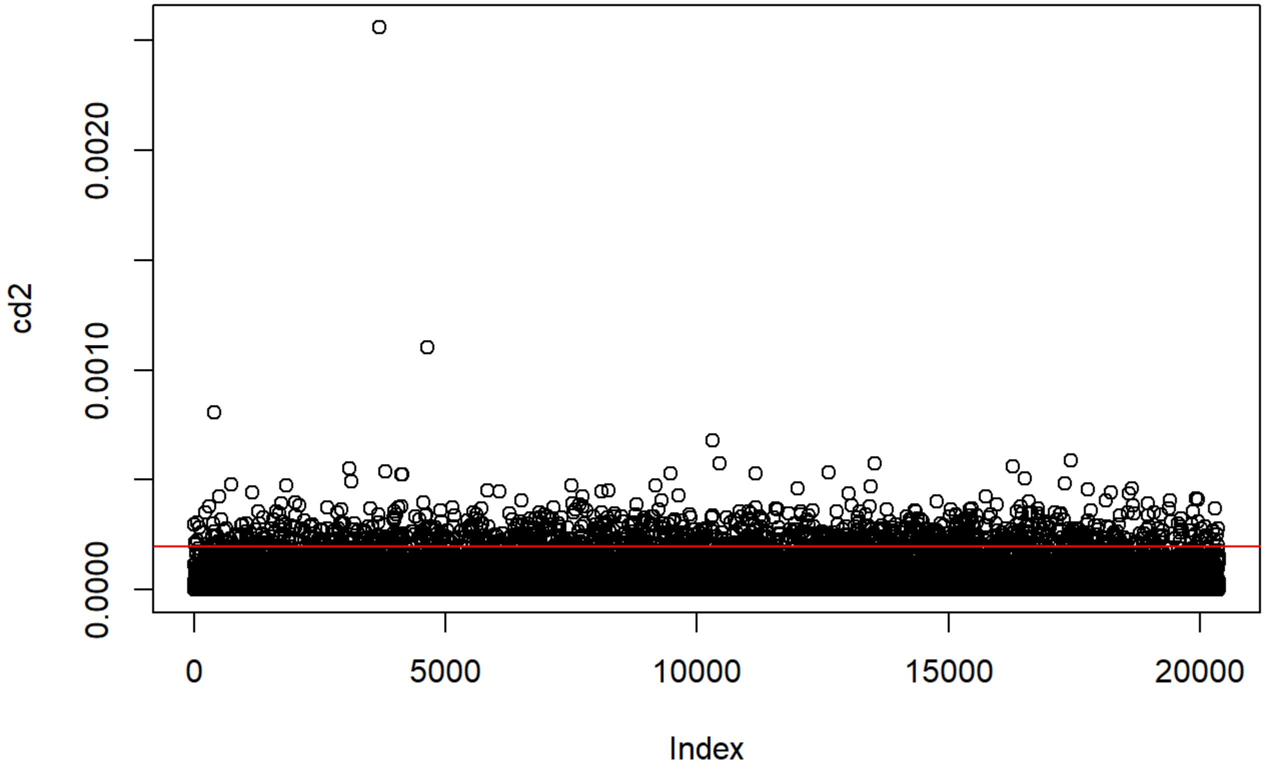
```
data3 <- data2[,c(-15, -16, -18)]

lm6 <- lm(log(price_sqft)~ ., data = data3)

summary(lm6)
```

```
## 
## Call:
## lm(formula = log(price_sqft) ~ ., data = data3)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.52816 -0.09740 -0.00042  0.09591  0.52129 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    2.571e+00  2.028e-01  12.679  < 2e-16 ***
## date           1.950e-04  9.553e-06  20.414  < 2e-16 ***
## bedrooms      -3.284e-02  1.622e-03 -20.254  < 2e-16 ***
## bathrooms      1.831e-02  2.646e-03   6.922 4.60e-12 ***
## sqft_lot       8.801e-07  3.720e-08  23.659  < 2e-16 ***
## floors        -6.220e-02  3.141e-03 -19.801  < 2e-16 ***
## waterfront1    5.450e-01  1.683e-02  32.393  < 2e-16 ***
## view           6.552e-02  1.778e-03  36.855  < 2e-16 ***
## condition      4.387e-02  1.931e-03  22.724  < 2e-16 ***
## grade          8.453e-02  1.835e-03  46.060  < 2e-16 ***
## sqft_above    -1.856e-04  3.177e-06 -58.442  < 2e-16 ***
## sqft_basement -3.154e-04  3.662e-06 -86.118  < 2e-16 ***
## yr_built      -5.737e-04  6.549e-05  -8.760  < 2e-16 ***
## yr_renovated   3.445e-05  2.989e-06  11.527  < 2e-16 ***
## zipcode98002   1.529e-03  1.390e-02   0.110   0.9124    
## zipcode98003   2.099e-02  1.248e-02   1.681   0.0928 .  
## zipcode98004   1.153e+00  1.241e-02  92.900  < 2e-16 ***
## zipcode98005   7.460e-01  1.488e-02  50.145  < 2e-16 ***
## zipcode98006   6.624e-01  1.106e-02  59.893  < 2e-16 ***
## zipcode98007   6.618e-01  1.562e-02  42.378  < 2e-16 ***
## zipcode98008   6.453e-01  1.246e-02  51.789  < 2e-16 ***
## zipcode98010   2.080e-01  1.962e-02  10.599  < 2e-16 ***
## zipcode98011   4.465e-01  1.387e-02  32.192  < 2e-16 ***
## zipcode98014   3.257e-01  1.910e-02  17.055  < 2e-16 ***
## zipcode98019   3.278e-01  1.427e-02  22.977  < 2e-16 ***
## zipcode98022   2.871e-02  1.336e-02   2.149   0.0316 *  
## zipcode98023  -2.548e-02  1.076e-02  -2.369   0.0179 *  
## zipcode98024   4.186e-01  2.355e-02  17.777  < 2e-16 ***
## zipcode98027   5.390e-01  1.142e-02  47.196  < 2e-16 ***
## zipcode98028   4.174e-01  1.241e-02  33.624  < 2e-16 ***
## zipcode98029   5.966e-01  1.201e-02  49.689  < 2e-16 ***
## zipcode98030   5.019e-02  1.275e-02   3.936 8.30e-05 ***
## zipcode98031   7.403e-02  1.250e-02   5.921 3.25e-09 ***
## zipcode98032   3.706e-03  1.711e-02   0.217   0.8285    
## zipcode98033   7.796e-01  1.128e-02  69.123  < 2e-16 ***
## zipcode98034   5.423e-01  1.060e-02  51.157  < 2e-16 ***
## zipcode98038   1.560e-01  1.047e-02  14.906  < 2e-16 ***
## zipcode98039   1.386e+00  2.851e-02  48.626  < 2e-16 ***
## zipcode98040   9.042e-01  1.284e-02  70.437  < 2e-16 ***
## zipcode98042   6.300e-02  1.058e-02   5.954 2.67e-09 ***
## zipcode98045   3.475e-01  1.361e-02  25.524  < 2e-16 ***
## zipcode98052   6.308e-01  1.053e-02  59.929  < 2e-16 ***
## zipcode98053   5.881e-01  1.144e-02  51.418  < 2e-16 ***
## zipcode98055   1.483e-01  1.289e-02  11.504  < 2e-16 ***
## zipcode98056   3.325e-01  1.134e-02  29.310  < 2e-16 ***
## zipcode98058   1.527e-01  1.101e-02  13.867  < 2e-16 ***
## zipcode98059   3.365e-01  1.099e-02  30.609  < 2e-16 ***
## zipcode98065   3.966e-01  1.216e-02  32.604  < 2e-16 ***
## zipcode98070   3.120e-01  1.912e-02  16.316  < 2e-16 ***
## zipcode98072   4.807e-01  1.258e-02  38.215  < 2e-16 ***
## zipcode98074   5.566e-01  1.121e-02  49.659  < 2e-16 ***
## zipcode98075   5.549e-01  1.183e-02  46.905  < 2e-16 ***
## zipcode98077   4.572e-01  1.410e-02  32.420  < 2e-16 ***
## zipcode98092   4.392e-03  1.175e-02   0.374   0.7086    
## zipcode98102   1.006e+00  1.868e-02  53.850  < 2e-16 ***
## zipcode98103   8.550e-01  1.080e-02  79.178  < 2e-16 ***
## zipcode98105   9.541e-01  1.369e-02  69.696  < 2e-16 ***
## zipcode98106   4.107e-01  1.211e-02  33.917  < 2e-16 ***
## zipcode98107   8.765e-01  1.294e-02  67.759  < 2e-16 ***
## zipcode98108   4.056e-01  1.474e-02  27.520  < 2e-16 ***
## zipcode98109   9.940e-01  1.842e-02  53.951  < 2e-16 ***
## zipcode98112   1.057e+00  1.357e-02  77.886  < 2e-16 ***
## zipcode98115   8.397e-01  1.072e-02  78.324  < 2e-16 ***
## zipcode98116   7.952e-01  1.219e-02  65.254  < 2e-16 ***
## zipcode98117   8.439e-01  1.083e-02  77.912  < 2e-16 ***
## zipcode98118   4.931e-01  1.107e-02  44.531  < 2e-16 ***
## zipcode98119   9.977e-01  1.511e-02  66.042  < 2e-16 ***
## zipcode98122   8.189e-01  1.290e-02  63.462  < 2e-16 ***
## zipcode98125   5.861e-01  1.142e-02  51.338  < 2e-16 ***
## zipcode98126   5.999e-01  1.195e-02  50.215  < 2e-16 ***
## zipcode98133   4.918e-01  1.088e-02  45.193  < 2e-16 ***
## zipcode98136   7.193e-01  1.299e-02  55.362  < 2e-16 ***
## zipcode98144   6.947e-01  1.217e-02  57.063  < 2e-16 ***
## zipcode98146   3.096e-01  1.288e-02  24.042  < 2e-16 ***
## zipcode98148   2.061e-01  2.654e-02   7.769 8.31e-15 ***
## zipcode98155   4.621e-01  1.109e-02  41.663  < 2e-16 ***
## zipcode98166   3.321e-01  1.321e-02  25.145  < 2e-16 ***
## zipcode98168   1.370e-01  1.288e-02  10.636  < 2e-16 ***
## zipcode98177   5.697e-01  1.313e-02  43.377  < 2e-16 ***
## zipcode98178   1.765e-01  1.301e-02  13.569  < 2e-16 ***
## zipcode98188   1.189e-01  1.641e-02   7.241 4.60e-13 ***
## zipcode98198   7.982e-02  1.274e-02   6.263 3.85e-10 ***
## zipcode98199   8.729e-01  1.231e-02  70.911  < 2e-16 ***
## sqft_living15  6.649e-05  2.944e-06  22.587  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1533 on 20290 degrees of freedom
## Multiple R-squared:  0.8342, Adjusted R-squared:  0.8335 
## F-statistic:  1230 on 83 and 20290 DF,  p-value: < 2.2e-16
```

```
# Checking the assumptions of linear models
par(mfrow=c(2,2))
plot(lm6)
```



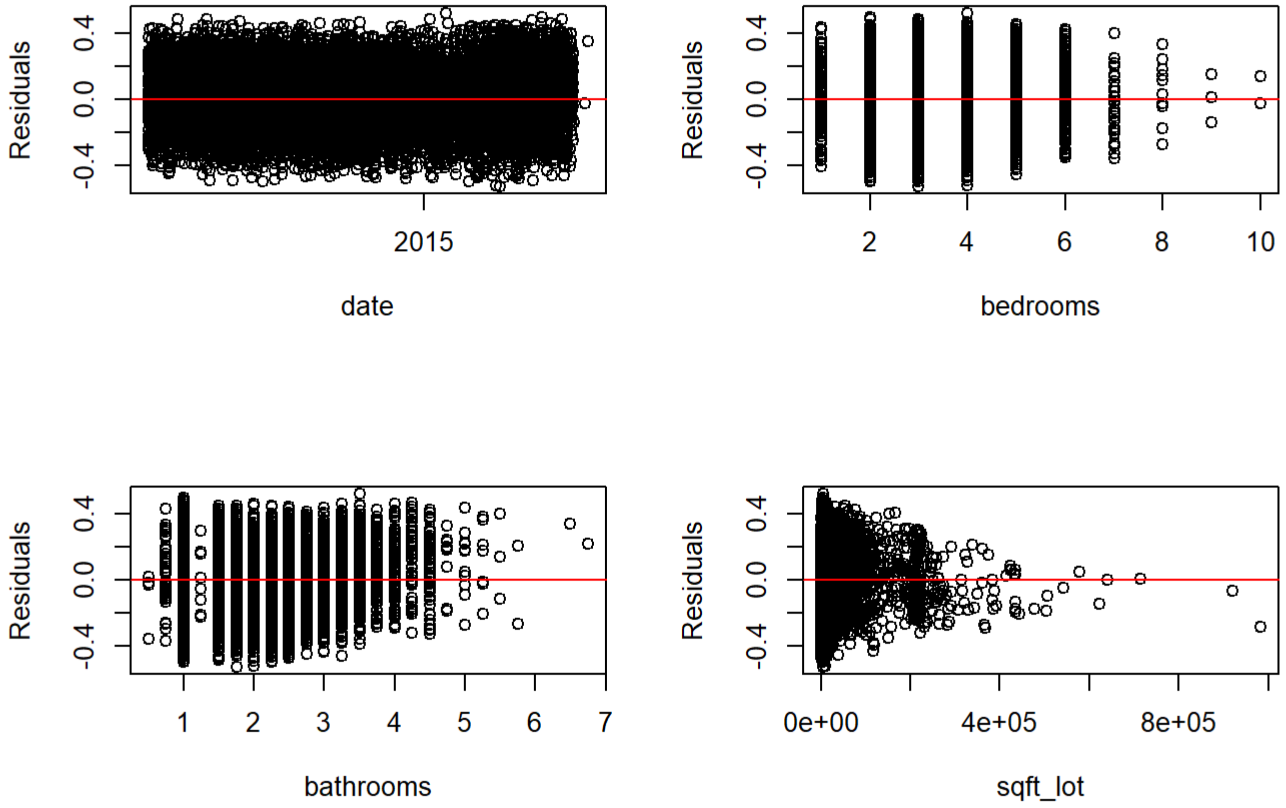
```
cd2 <- cooks.distance(lm6)

plot(cd2)
abline(h = 4/nrow(data3), col='red')
```



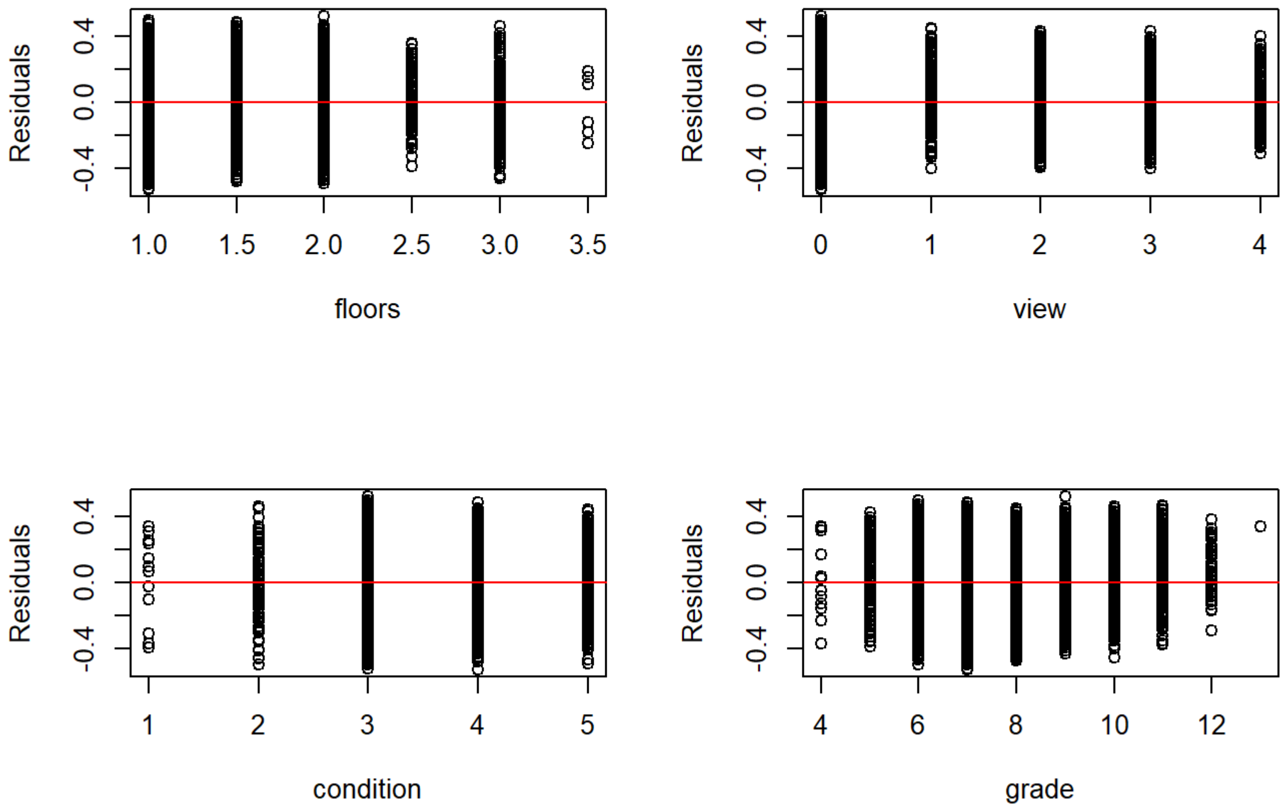
```
data4 <- data2[-which(cd > 4/nrow(data3)),]

lm7 <- lm(log(price_sqft) ~ ., data = data4)
summary(lm7)
```

```
##
## Call:
## lm(formula = log(price_sqft) ~ ., data = data4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53581 -0.09798 -0.00043  0.09616  0.52106
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.126e+01  6.959e+00  -5.929 3.10e-09 ***
## date           1.918e-04  9.800e-06  19.567  < 2e-16 ***
## bedrooms      -3.277e-02  1.663e-03 -19.705  < 2e-16 ***
## bathrooms      1.885e-02  2.717e-03   6.938 4.12e-12 ***
## sqft_lot       8.754e-07  5.486e-08  15.956  < 2e-16 ***
## floors        -6.395e-02  3.235e-03 -19.770  < 2e-16 ***
## waterfront1    5.461e-01  1.717e-02  31.807  < 2e-16 ***
## view           6.551e-02  1.823e-03  35.927  < 2e-16 ***
## condition      4.374e-02  1.981e-03  22.073  < 2e-16 ***
## grade          8.443e-02  1.884e-03  44.820  < 2e-16 ***
## sqft_above    -1.848e-04  3.269e-06 -56.545  < 2e-16 ***
## sqft_basement -3.167e-04  3.768e-06 -84.051  < 2e-16 ***
## yr_built      -5.594e-04  6.728e-05  -8.315  < 2e-16 ***
## yr_renovated   3.393e-05  3.064e-06  11.074  < 2e-16 ***
## zipcode98002   1.598e-02  1.474e-02   1.084 0.278320
## zipcode98003   9.759e-03  1.286e-02   0.759 0.447783
## zipcode98004   1.050e+00  2.405e-02  43.650  < 2e-16 ***
## zipcode98005   6.557e-01  2.578e-02  25.436  < 2e-16 ***
## zipcode98006   5.985e-01  2.133e-02  28.055  < 2e-16 ***
## zipcode98007   5.774e-01  2.672e-02  21.606  < 2e-16 ***
## zipcode98008   5.654e-01  2.563e-02  22.055  < 2e-16 ***
## zipcode98010   2.567e-01  2.485e-02  10.332  < 2e-16 ***
## zipcode98011   3.016e-01  3.272e-02   9.219  < 2e-16 ***
## zipcode98014   2.820e-01  3.803e-02   7.415 1.27e-13 ***
## zipcode98019   2.380e-01  3.678e-02   6.470 1.00e-10 ***
## zipcode98022   1.229e-01  2.033e-02   6.049 1.49e-09 ***
## zipcode98023  -4.506e-02  1.212e-02  -3.718 0.000201 ***
## zipcode98024   3.995e-01  3.574e-02  11.178  < 2e-16 ***
## zipcode98027   5.093e-01  2.253e-02  22.605  < 2e-16 ***
## zipcode98028   2.549e-01  3.164e-02   8.057 8.24e-16 ***
## zipcode98029   5.639e-01  2.586e-02  21.804  < 2e-16 ***
## zipcode98030   4.530e-02  1.441e-02   3.145 0.001663 **
## zipcode98031   5.710e-02  1.500e-02   3.808 0.000140 ***
## zipcode98032  -1.876e-02  1.801e-02  -1.042 0.297460
## zipcode98033   6.581e-01  2.740e-02  24.019  < 2e-16 ***
## zipcode98034   4.032e-01  2.927e-02  13.776  < 2e-16 ***
## zipcode98038   1.837e-01  1.739e-02  10.568  < 2e-16 ***
## zipcode98039   1.284e+00  3.626e-02  35.399  < 2e-16 ***
## zipcode98040   8.228e-01  2.123e-02  38.751  < 2e-16 ***
## zipcode98042   7.434e-02  1.437e-02   5.174 2.32e-07 ***
## zipcode98045   3.913e-01  3.340e-02  11.718  < 2e-16 ***
## zipcode98052   5.242e-01  2.826e-02  18.549  < 2e-16 ***
## zipcode98053   5.035e-01  3.082e-02  16.339  < 2e-16 ***
## zipcode98055   1.060e-01  1.693e-02   6.260 3.93e-10 ***
## zipcode98056   2.799e-01  1.840e-02  15.210  < 2e-16 ***
## zipcode98058   1.283e-01  1.616e-02   7.936 2.20e-15 ***
## zipcode98059   2.964e-01  1.829e-02  16.206  < 2e-16 ***
## zipcode98065   4.019e-01  3.009e-02  13.358  < 2e-16 ***
## zipcode98070   2.304e-01  2.330e-02   9.889  < 2e-16 ***
## zipcode98072   3.437e-01  3.281e-02  10.474  < 2e-16 ***
## zipcode98074   4.884e-01  2.726e-02  17.919  < 2e-16 ***
## zipcode98075   5.050e-01  2.636e-02  19.155  < 2e-16 ***
## zipcode98077   3.340e-01  3.459e-02   9.656  < 2e-16 ***
## zipcode98092   2.635e-02  1.315e-02   2.004 0.045094 *
## zipcode98102   8.777e-01  2.825e-02  31.071  < 2e-16 ***
## zipcode98103   7.068e-01  2.610e-02  27.081  < 2e-16 ***
## zipcode98105   8.145e-01  2.687e-02  30.311  < 2e-16 ***
## zipcode98106   3.052e-01  1.937e-02  15.757  < 2e-16 ***
## zipcode98107   7.240e-01  2.687e-02  26.946  < 2e-16 ***
## zipcode98108   3.093e-01  2.160e-02  14.316  < 2e-16 ***
## zipcode98109   8.514e-01  2.832e-02  30.068  < 2e-16 ***
## zipcode98112   9.332e-01  2.486e-02  37.538  < 2e-16 ***
## zipcode98115   6.951e-01  2.658e-02  26.148  < 2e-16 ***
## zipcode98116   6.732e-01  2.162e-02  31.144  < 2e-16 ***
## zipcode98117   6.809e-01  2.688e-02  25.337  < 2e-16 ***
## zipcode98118   4.044e-01  1.892e-02  21.381  < 2e-16 ***
## zipcode98119   8.545e-01  2.648e-02  32.264  < 2e-16 ***
## zipcode98122   7.058e-01  2.345e-02  30.093  < 2e-16 ***
## zipcode98125   4.295e-01  2.870e-02  14.965  < 2e-16 ***
## zipcode98126   4.920e-01  1.985e-02  24.790  < 2e-16 ***
## zipcode98133   3.172e-01  2.959e-02  10.718  < 2e-16 ***
## zipcode98136   6.101e-01  2.040e-02  29.907  < 2e-16 ***
## zipcode98144   5.906e-01  2.178e-02  27.119  < 2e-16 ***
## zipcode98146   2.237e-01  1.841e-02  12.149  < 2e-16 ***
## zipcode98148   1.474e-01  2.780e-02   5.301 1.16e-07 ***
## zipcode98155   2.890e-01  3.086e-02   9.364  < 2e-16 ***
## zipcode98166   2.622e-01  1.679e-02  15.619  < 2e-16 ***
## zipcode98168   5.788e-02  1.763e-02   3.283 0.001027 **
## zipcode98177   3.841e-01  3.106e-02  12.367  < 2e-16 ***
## zipcode98178   1.106e-01  1.823e-02   6.070 1.30e-09 ***
## zipcode98188   6.258e-02  1.897e-02   3.298 0.000974 ***
## zipcode98198   4.229e-02  1.425e-02   2.968 0.003004 **
## zipcode98199   7.218e-01  2.550e-02  28.310  < 2e-16 ***
## lat            3.723e-01  6.467e-02   5.756 8.73e-09 ***
## long          -2.146e-01  5.380e-02  -3.990 6.64e-05 ***
## sqft_living15  6.621e-05  3.035e-06  21.812  < 2e-16 ***
## sqft_lot15     2.967e-08  7.751e-08   0.383 0.701909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1529 on 19179 degrees of freedom
## Multiple R-squared:  0.8353, Adjusted R-squared:  0.8346
## F-statistic:  1131 on 86 and 19179 DF,  p-value: < 2.2e-16
```

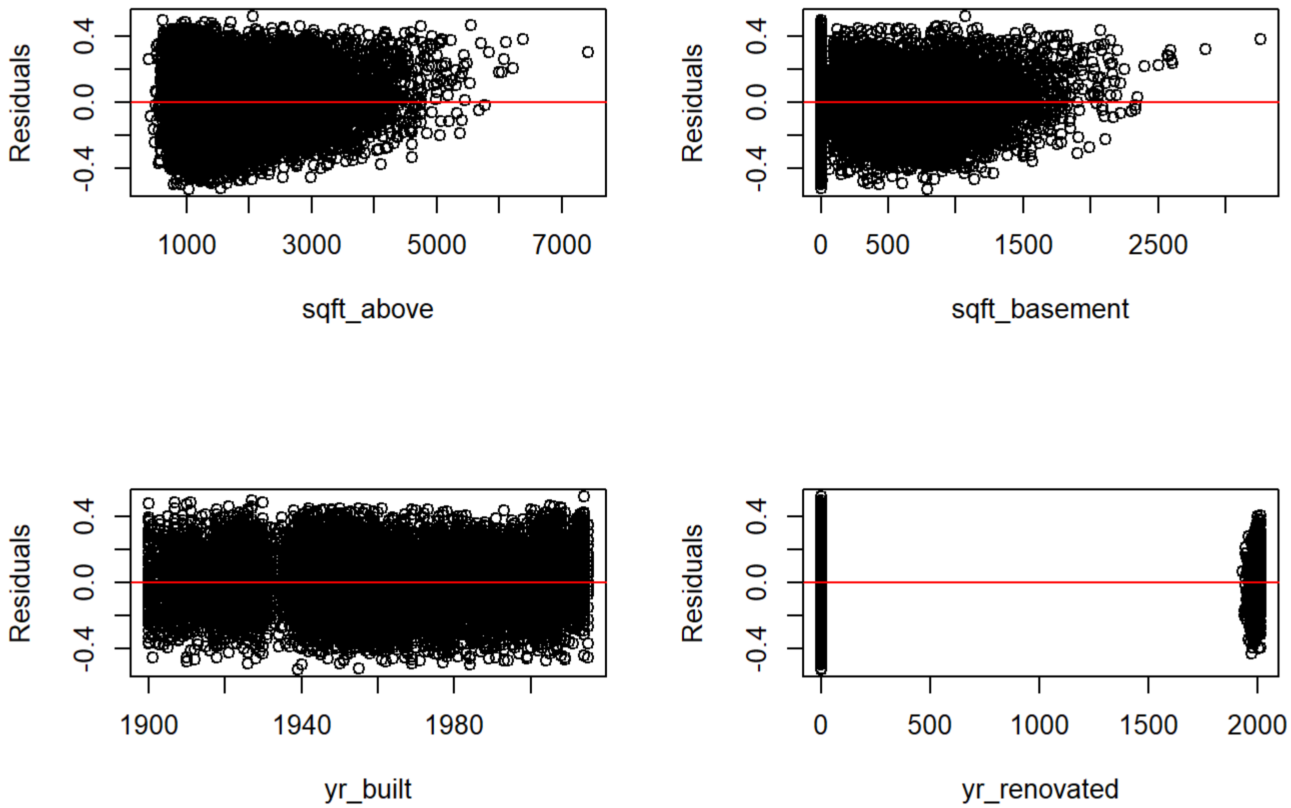
```
num.cols2 <- colnames(data3[,c(-6, -14, -16)])

par(mfrow = c(2,2))

for (i in num.cols2[1:4]) {
  plot(data3[,i], lm6$residuals, xlab = i, ylab = "Residuals")
  abline(h = 0, col='red')
}
```



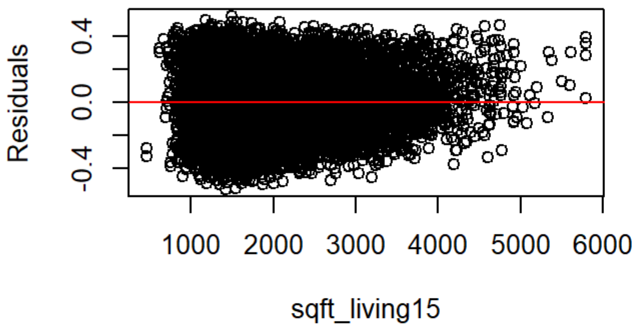
```
par(mfrow = c(2,2))

for (i in num.cols2[5:8]) {
  plot(data3[,i], lm6$residuals, xlab = i, ylab = "Residuals")
  abline(h = 0, col='red')
}
```



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

for (i in num.cols2[9:12]) {
  plot(data3[,i], lm6$residuals, xlab = i, ylab = "Residuals")
  abline(h = 0, col='red')
}
```



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

for (i in num.cols2[13]) {
  plot(data3[,i], lm6$residuals, xlab = i, ylab = "Residuals")
  abline(h = 0, col='red')
}
```



```
# Random Forest Regression to determine feature importance

rf <- randomForest(log(price_sqft) ~ ., data = data2[,c(-6,-14)] +
                   as.numeric(data2$zipcode) + as.numeric(data2$waterfront),
                   importance = TRUE)

rf
```

```
##
## Call:
##  randomForest(formula = log(price_sqft) ~ ., data = data2[, c(-6,      -14)] + as.numeric(dat
a2$zipcode) + as.numeric(data2$waterfront),      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##           Mean of squared residuals: 0.01780617
##                     % Var explained: 85.75
```

```
rf$importance
```

```
##                    %IncMSE IncNodePurity
## date          0.0006865678      41.60164
## bedrooms      0.0276198014     114.98868
## bathrooms     0.0253466280     106.59856
## sqft_lot      0.0066962638      63.94748
## floors        0.0386040547     143.06593
## view          0.1131407218     287.78255
## condition     0.0197439512      81.70609
## grade         0.0162194001      46.87869
## sqft_above    0.0120594979      94.81218
## sqft_basement 0.0203245059     135.24056
## yr_built      0.0114057350      70.13222
## yr_renovated  0.0809705164     356.69799
## lat           0.1126746622     402.55743
## long          0.1147148846     406.32394
## sqft_living15 0.0069931417      72.05548
## sqft_lot15    0.0103657598      89.32404
```

```
# Predictive models

# Creating a test and training set
set.seed(1)

flag <- sample(nrow(data2), nrow(data2)/10, replace = FALSE)
train <- data2[-flag,]
test <- data2[flag,]
```

```
lm6.p <- lm(log(price_sqft)~ . -lat -long -sqft_lot15, data = train)

summary(lm6.p)
```

```
## 
## Call:
## lm(formula = log(price_sqft) ~ . - lat - long - sqft_lot15, data = train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.52300 -0.09728  0.00010  0.09639  0.52842 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    2.536e+00  2.134e-01  11.882  < 2e-16 ***
## date           1.995e-04  1.005e-05  19.846  < 2e-16 ***
## bedrooms      -3.296e-02  1.704e-03 -19.348  < 2e-16 ***
## bathrooms      1.746e-02  2.780e-03   6.281 3.44e-10 ***
## sqft_lot       8.863e-07  4.006e-08  22.121  < 2e-16 ***
## floors        -5.995e-02  3.300e-03 -18.168  < 2e-16 ***
## waterfront1    5.342e-01  1.801e-02  29.652  < 2e-16 ***
## view           6.545e-02  1.876e-03  34.889  < 2e-16 ***
## condition      4.431e-02  2.033e-03  21.799  < 2e-16 ***
## grade          8.490e-02  1.934e-03  43.904  < 2e-16 ***
## sqft_above    -1.871e-04  3.355e-06 -55.777  < 2e-16 ***
## sqft_basement -3.142e-04  3.856e-06 -81.469  < 2e-16 ***
## yr_built      -5.943e-04  6.871e-05  -8.649  < 2e-16 ***
## yr_renovated   3.513e-05  3.135e-06  11.205  < 2e-16 ***
## zipcode98002  -2.019e-03  1.447e-02  -0.140 0.889013    
## zipcode98003   1.548e-02  1.310e-02   1.182 0.237161    
## zipcode98004   1.152e+00  1.292e-02  89.149  < 2e-16 ***
## zipcode98005   7.449e-01  1.536e-02  48.489  < 2e-16 ***
## zipcode98006   6.636e-01  1.158e-02  57.324  < 2e-16 ***
## zipcode98007   6.643e-01  1.646e-02  40.365  < 2e-16 ***
## zipcode98008   6.465e-01  1.311e-02  49.327  < 2e-16 ***
## zipcode98010   2.034e-01  2.056e-02   9.893  < 2e-16 ***
## zipcode98011   4.510e-01  1.468e-02  30.715  < 2e-16 ***
## zipcode98014   3.233e-01  1.993e-02  16.222  < 2e-16 ***
## zipcode98019   3.238e-01  1.482e-02  21.849  < 2e-16 ***
## zipcode98022   2.743e-02  1.381e-02   1.986 0.046999 *  
## zipcode98023  -2.817e-02  1.132e-02  -2.489 0.012820 *  
## zipcode98024   4.172e-01  2.430e-02  17.168  < 2e-16 ***
## zipcode98027   5.366e-01  1.200e-02  44.701  < 2e-16 ***
## zipcode98028   4.183e-01  1.294e-02  32.334  < 2e-16 ***
## zipcode98029   5.956e-01  1.260e-02  47.291  < 2e-16 ***
## zipcode98030   5.108e-02  1.338e-02   3.817 0.000135 ***
## zipcode98031   6.985e-02  1.312e-02   5.324 1.03e-07 ***
## zipcode98032   8.146e-04  1.785e-02   0.046 0.963597    
## zipcode98033   7.779e-01  1.183e-02  65.774  < 2e-16 ***
## zipcode98034   5.397e-01  1.104e-02  48.892  < 2e-16 ***
## zipcode98038   1.537e-01  1.094e-02  14.044  < 2e-16 ***
## zipcode98039   1.385e+00  3.037e-02  45.613  < 2e-16 ***
## zipcode98040   9.067e-01  1.353e-02  67.029  < 2e-16 ***
## zipcode98042   6.163e-02  1.114e-02   5.534 3.18e-08 ***
## zipcode98045   3.472e-01  1.437e-02  24.161  < 2e-16 ***
## zipcode98052   6.271e-01  1.103e-02  56.836  < 2e-16 ***
## zipcode98053   5.882e-01  1.199e-02  49.038  < 2e-16 ***
## zipcode98055   1.447e-01  1.350e-02  10.722  < 2e-16 ***
## zipcode98056   3.270e-01  1.190e-02  27.472  < 2e-16 ***
## zipcode98058   1.552e-01  1.155e-02  13.437  < 2e-16 ***
## zipcode98059   3.365e-01  1.151e-02  29.248  < 2e-16 ***
## zipcode98065   3.977e-01  1.262e-02  31.523  < 2e-16 ***
## zipcode98070   3.128e-01  2.030e-02  15.404  < 2e-16 ***
## zipcode98072   4.818e-01  1.313e-02  36.706  < 2e-16 ***
## zipcode98074   5.506e-01  1.174e-02  46.917  < 2e-16 ***
## zipcode98075   5.520e-01  1.247e-02  44.258  < 2e-16 ***
## zipcode98077   4.542e-01  1.463e-02  31.045  < 2e-16 ***
## zipcode98092   1.960e-03  1.222e-02   0.160 0.872543
## zipcode98102   1.003e+00  1.973e-02  50.834  < 2e-16 ***
## zipcode98103   8.458e-01  1.131e-02  74.779  < 2e-16 ***
## zipcode98105   9.556e-01  1.437e-02  66.495  < 2e-16 ***
## zipcode98106   4.185e-01  1.272e-02  32.902  < 2e-16 ***
## zipcode98107   8.711e-01  1.362e-02  63.959  < 2e-16 ***
## zipcode98108   4.048e-01  1.535e-02  26.372  < 2e-16 ***
## zipcode98109   9.881e-01  1.954e-02  50.580  < 2e-16 ***
## zipcode98112   1.055e+00  1.407e-02  74.942  < 2e-16 ***
## zipcode98115   8.339e-01  1.117e-02  74.639  < 2e-16 ***
## zipcode98116   7.906e-01  1.276e-02  61.945  < 2e-16 ***
## zipcode98117   8.398e-01  1.134e-02  74.058  < 2e-16 ***
## zipcode98118   4.969e-01  1.160e-02  42.844  < 2e-16 ***
## zipcode98119   9.911e-01  1.593e-02  62.222  < 2e-16 ***
## zipcode98122   8.162e-01  1.342e-02  60.819  < 2e-16 ***
## zipcode98125   5.809e-01  1.187e-02  48.943  < 2e-16 ***
## zipcode98126   5.945e-01  1.250e-02  47.570  < 2e-16 ***
## zipcode98133   4.902e-01  1.140e-02  43.008  < 2e-16 ***
## zipcode98136   7.145e-01  1.359e-02  52.585  < 2e-16 ***
## zipcode98144   6.926e-01  1.270e-02  54.548  < 2e-16 ***
## zipcode98146   3.043e-01  1.338e-02  22.742  < 2e-16 ***
## zipcode98148   2.046e-01  2.661e-02   7.689 1.56e-14 ***
## zipcode98155   4.570e-01  1.157e-02  39.487  < 2e-16 ***
## zipcode98166   3.327e-01  1.394e-02  23.867  < 2e-16 ***
## zipcode98168   1.370e-01  1.345e-02  10.185  < 2e-16 ***
## zipcode98177   5.669e-01  1.387e-02  40.864  < 2e-16 ***
## zipcode98178   1.762e-01  1.369e-02  12.871  < 2e-16 ***
## zipcode98188   1.256e-01  1.718e-02   7.309 2.80e-13 ***
## zipcode98198   7.552e-02  1.351e-02   5.591 2.30e-08 ***
## zipcode98199   8.754e-01  1.297e-02  67.503  < 2e-16 ***
## sqft_living15  6.661e-05  3.097e-06  21.510  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.153 on 18253 degrees of freedom
## Multiple R-squared:  0.8344, Adjusted R-squared:  0.8337
## F-statistic:  1108 on 83 and 18253 DF,  p-value: < 2.2e-16
```

```
pred.lm6 <- exp(predict(lm6.p, test))
lm6.mse <- mean((test$price_sqft - pred.lm6)^2)

lm6.mse
```

```
## [1] 2061.513
```

```
step.p <- step(lm(log(price_sqft)~., data = train), direction = "both")
```

```
## Start:  AIC=-68804.47
## log(price_sqft) ~ date + bedrooms + bathrooms + sqft_lot + floors + 
##     waterfront + view + condition + grade + sqft_above + sqft_basement + 
##     yr_built + yr_renovated + zipcode + lat + long + sqft_living15 + 
##     sqft_lot15
## 
##                 Df Sum of Sq     RSS    AIC
## - sqft_lot15     1      0.01  426.24 -68806
## <none>                        426.22 -68804
## - long           1      0.24  426.47 -68796
## - lat            1      0.75  426.97 -68774
## - bathrooms      1      0.93  427.15 -68767
## - yr_built       1      1.71  427.93 -68733
## - yr_renovated   1      2.94  429.16 -68681
## - sqft_lot       1      5.16  431.39 -68586
## - floors         1      7.67  433.90 -68479
## - bedrooms       1      8.71  434.94 -68435
## - date           1      9.16  435.39 -68417
## - sqft_living15  1     10.62  436.85 -68355
## - condition      1     11.27  437.49 -68328
## - waterfront     1     20.71  446.93 -67936
## - view           1     28.56  454.78 -67617
## - grade          1     44.61  470.84 -66981
## - sqft_above     1     72.29  498.51 -65934
## - sqft_basement  1    155.48  581.70 -63104
## - zipcode       69    595.67 1021.89 -52908
## 
## Step:  AIC=-68805.88
## log(price_sqft) ~ date + bedrooms + bathrooms + sqft_lot + floors + 
##     waterfront + view + condition + grade + sqft_above + sqft_basement + 
##     yr_built + yr_renovated + zipcode + lat + long + sqft_living15
## 
##                 Df Sum of Sq     RSS    AIC
## <none>                        426.24 -68806
## + sqft_lot15     1      0.01  426.22 -68804
## - long           1      0.23  426.47 -68798
## - lat            1      0.75  426.98 -68776
## - bathrooms      1      0.92  427.16 -68768
## - yr_built       1      1.71  427.95 -68734
## - yr_renovated   1      2.94  429.18 -68682
## - floors         1      7.69  433.92 -68480
## - bedrooms       1      8.73  434.97 -68436
## - date           1      9.16  435.40 -68418
## - sqft_living15  1     10.75  436.99 -68351
## - condition      1     11.28  437.52 -68329
## - sqft_lot       1     11.78  438.02 -68308
## - waterfront     1     20.72  446.96 -67937
## - view           1     28.55  454.79 -67619
## - grade          1     44.60  470.84 -66983
## - sqft_above     1     72.27  498.51 -65936
## - sqft_basement  1    155.49  581.73 -63105
## - zipcode       69    595.95 1022.19 -52904
```

```
summary(step.p)
```

```
##
## Call:
## lm(formula = log(price_sqft) ~ date + bedrooms + bathrooms +
##     sqft_lot + floors + waterfront + view + condition + grade +
##     sqft_above + sqft_basement + yr_built + yr_renovated + zipcode +
##     lat + long + sqft_living15, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52799 -0.09750 -0.00017  0.09589  0.53071
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.561e+01  6.904e+00  -5.158 2.53e-07 ***
## date           1.989e-04  1.004e-05  19.808  < 2e-16 ***
## bedrooms      -3.290e-02  1.702e-03 -19.335  < 2e-16 ***
## bathrooms      1.746e-02  2.777e-03   6.287 3.31e-10 ***
## sqft_lot       9.056e-07  4.032e-08  22.459  < 2e-16 ***
## floors        -5.984e-02  3.298e-03 -18.141  < 2e-16 ***
## waterfront1    5.362e-01  1.800e-02  29.788  < 2e-16 ***
## view           6.552e-02  1.874e-03  34.966  < 2e-16 ***
## condition      4.463e-02  2.031e-03  21.975  < 2e-16 ***
## grade          8.447e-02  1.933e-03  43.701  < 2e-16 ***
## sqft_above    -1.865e-04  3.353e-06 -55.629  < 2e-16 ***
## sqft_basement -3.143e-04  3.852e-06 -81.596  < 2e-16 ***
## yr_built      -5.878e-04  6.868e-05  -8.559  < 2e-16 ***
## yr_renovated   3.512e-05  3.131e-06  11.216  < 2e-16 ***
## zipcode98002   7.403e-03  1.477e-02   0.501 0.616177
## zipcode98003   7.173e-03  1.324e-02   0.542 0.587894
## zipcode98004   1.048e+00  2.457e-02  42.668  < 2e-16 ***
## zipcode98005   6.491e-01  2.621e-02  24.766  < 2e-16 ***
## zipcode98006   5.916e-01  2.164e-02  27.344  < 2e-16 ***
## zipcode98007   5.734e-01  2.720e-02  21.084  < 2e-16 ***
## zipcode98008   5.583e-01  2.605e-02  21.432  < 2e-16 ***
## zipcode98010   2.390e-01  2.504e-02   9.544  < 2e-16 ***
## zipcode98011   2.960e-01  3.343e-02   8.855  < 2e-16 ***
## zipcode98014   2.535e-01  3.889e-02   6.518 7.29e-11 ***
## zipcode98019   2.153e-01  3.713e-02   5.798 6.82e-09 ***
## zipcode98022   1.089e-01  2.041e-02   5.336 9.60e-08 ***
## zipcode98023  -4.300e-02  1.233e-02  -3.488 0.000489 ***
## zipcode98024   3.829e-01  3.582e-02  10.688  < 2e-16 ***
## zipcode98027   4.947e-01  2.272e-02  21.778  < 2e-16 ***
## zipcode98028   2.556e-01  3.238e-02   7.895 3.06e-15 ***
## zipcode98029   5.466e-01  2.599e-02  21.030  < 2e-16 ***
## zipcode98030   4.246e-02  1.467e-02   2.895 0.003792 **
## zipcode98031   4.846e-02  1.531e-02   3.166 0.001549 **
## zipcode98032  -2.413e-02  1.830e-02  -1.318 0.187474
## zipcode98033   6.536e-01  2.796e-02  23.377  < 2e-16 ***
## zipcode98034   3.963e-01  2.991e-02  13.252  < 2e-16 ***
## zipcode98038   1.708e-01  1.735e-02   9.842  < 2e-16 ***
## zipcode98039   1.273e+00  3.702e-02  34.385  < 2e-16 ***
## zipcode98040   8.205e-01  2.172e-02  37.771  < 2e-16 ***
## zipcode98042   6.631e-02  1.449e-02   4.576 4.78e-06 ***
## zipcode98045   3.702e-01  3.319e-02  11.153  < 2e-16 ***
## zipcode98052   5.136e-01  2.879e-02  17.843  < 2e-16 ***
## zipcode98053   4.898e-01  3.125e-02  15.673  < 2e-16 ***
## zipcode98055   1.014e-01  1.722e-02   5.890 3.92e-09 ***
## zipcode98056   2.668e-01  1.874e-02  14.242  < 2e-16 ***
## zipcode98058   1.248e-01  1.640e-02   7.606 2.97e-14 ***
## zipcode98059   2.898e-01  1.852e-02  15.645  < 2e-16 ***
## zipcode98065   3.824e-01  2.995e-02  12.765  < 2e-16 ***
## zipcode98070   2.390e-01  2.357e-02  10.139  < 2e-16 ***
## zipcode98072   3.392e-01  3.350e-02  10.126  < 2e-16 ***
## zipcode98074   4.708e-01  2.761e-02  17.053  < 2e-16 ***
## zipcode98075   4.892e-01  2.663e-02  18.368  < 2e-16 ***
## zipcode98077   3.250e-01  3.505e-02   9.271  < 2e-16 ***
## zipcode98092   2.147e-02  1.328e-02   1.617 0.105960    
## zipcode98102   8.733e-01  2.915e-02  29.954  < 2e-16 ***
## zipcode98103   6.964e-01  2.679e-02  25.992  < 2e-16 ***
## zipcode98105   8.189e-01  2.754e-02  29.734  < 2e-16 ***
## zipcode98106   3.190e-01  1.991e-02  16.026  < 2e-16 ***
## zipcode98107   7.191e-01  2.758e-02  26.067  < 2e-16 ***
## zipcode98108   3.088e-01  2.212e-02  13.961  < 2e-16 ***
## zipcode98109   8.527e-01  2.920e-02  29.205  < 2e-16 ***
## zipcode98112   9.307e-01  2.539e-02  36.661  < 2e-16 ***
## zipcode98115   6.882e-01  2.723e-02  25.271  < 2e-16 ***
## zipcode98116   6.715e-01  2.211e-02  30.373  < 2e-16 ***
## zipcode98117   6.809e-01  2.755e-02  24.713  < 2e-16 ***
## zipcode98118   4.086e-01  1.939e-02  21.075  < 2e-16 ***
## zipcode98119   8.518e-01  2.716e-02  31.359  < 2e-16 ***
## zipcode98122   6.993e-01  2.395e-02  29.202  < 2e-16 ***
## zipcode98125   4.234e-01  2.942e-02  14.389  < 2e-16 ***
## zipcode98126   4.892e-01  2.030e-02  24.095  < 2e-16 ***
## zipcode98133   3.188e-01  3.034e-02  10.507  < 2e-16 ***
## zipcode98136   6.093e-01  2.085e-02  29.219  < 2e-16 ***
## zipcode98144   5.846e-01  2.227e-02  26.251  < 2e-16 ***
## zipcode98146   2.177e-01  1.875e-02  11.608  < 2e-16 ***
## zipcode98148   1.484e-01  2.790e-02   5.319 1.05e-07 ***
## zipcode98155   2.845e-01  3.164e-02   8.993  < 2e-16 ***
## zipcode98166   2.649e-01  1.722e-02  15.383  < 2e-16 ***
## zipcode98168   6.312e-02  1.798e-02   3.510 0.000449 ***
## zipcode98177   3.883e-01  3.188e-02  12.179  < 2e-16 ***
## zipcode98178   1.093e-01  1.871e-02   5.840 5.32e-09 ***
## zipcode98188   7.227e-02  1.944e-02   3.717 0.000202 ***
## zipcode98198   3.757e-02  1.464e-02   2.566 0.010302 *  
## zipcode98199   7.278e-01  2.619e-02  27.793  < 2e-16 ***
## lat            3.748e-01  6.633e-02   5.650 1.62e-08 ***
## long          -1.669e-01  5.277e-02  -3.163 0.001564 ** 
## sqft_living15  6.636e-05  3.094e-06  21.452  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1528 on 18251 degrees of freedom
## Multiple R-squared:  0.8348, Adjusted R-squared:  0.8341
## F-statistic:  1085 on 85 and 18251 DF,  p-value: < 2.2e-16
```

```
pred.step <- exp(predict(step.p, test))
step.mse <- mean((test$price_sqft - pred.step)^2)

step.mse
```

```
## [1] 2060.988
```

```
train2 <- train[,-1]
test2 <- test[,-1]

train2$waterfront <- as.numeric(train2$waterfront)
train2$zipcode <- as.numeric(train2$zipcode)

test2$waterfront <- as.numeric(test2$waterfront)
test2$zipcode <- as.numeric(test2$zipcode)


opt.lambda <- cv.glmnet(as.matrix(train2[,-18]),
                        as.matrix(log(train2$price_sqft)),
                        alpha = 1)$lambda.min
lasso <- glmnet(as.matrix(train2[,-18]),
                as.matrix(log(train2$price_sqft)), alpha = 1, lambda = opt.lambda)

coef(lasso)
```

```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept)   -7.364280e+01
## bedrooms      -5.062452e-02
## bathrooms      4.298592e-02
## sqft_lot       5.561622e-07
## floors         5.433222e-02
## waterfront     4.257176e-01
## view           7.086899e-02
## condition      4.807474e-02
## grade          1.544870e-01
## sqft_above    -2.505722e-04
## sqft_basement -2.800686e-04
## yr_built      -3.849553e-03
## yr_renovated   2.889652e-05
## zipcode       -1.416472e-03
## lat            1.403804e+00
## long          -1.521154e-01
## sqft_living15  7.431047e-05
## sqft_lot15    -2.119418e-07
```

```
p.lasso <- exp(predict(lasso, s = opt.lambda,
                   newx = as.matrix(test2[,-18])))
lasso.mse <- mean((test2$price_sqft - p.lasso)^2)

lasso.mse
```

```
## [1] 4931.009
```

```
opt.lambda2 <- cv.glmnet(as.matrix(train2[,-18]),
                        as.matrix(log(train2$price_sqft)),
                        alpha = 0)$lambda.min
ridge <- glmnet(as.matrix(train2[,-18]),
                as.matrix(log(train2$price_sqft)), alpha = 0, lambda = opt.lambda2)

coef(ridge)
```

```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                           s0
## (Intercept)   -7.503671e+01
## bedrooms      -6.057232e-02
## bathrooms      1.810302e-02
## sqft_lot       4.389692e-07
## floors         4.839973e-02
## waterfront     4.125402e-01
## view           6.781407e-02
## condition      4.816806e-02
## grade          1.228191e-01
## sqft_above    -1.727866e-04
## sqft_basement -2.142110e-04
## yr_built      -3.226353e-03
## yr_renovated   3.363895e-05
## zipcode       -1.119356e-03
## lat            1.343662e+00
## long          -1.787057e-01
## sqft_living15  5.454289e-05
## sqft_lot15    -2.356043e-07
```

```
p.ridge <- exp(predict(ridge, alpha = 0,
                   newx = as.matrix(test2[,-18])))
ridge.mse <- mean((test2$price_sqft - p.ridge)^2)

ridge.mse
```

```
## [1] 5068.003
```

```
PCR <- pcr(log(price_sqft)~., data = train, scale = TRUE, validation = "CV")

summary(PCR)
```

```
## Data:    X dimension: 18337 86
##  Y dimension: 18337 1
## Fit method: svdpc
## Number of components considered: 86
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV          0.3752   0.3716   0.3263    0.326   0.2939    0.286   0.2807
## adjCV       0.3752   0.3716   0.3262    0.326   0.2939    0.286   0.2807
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV      0.2627   0.2556   0.2556    0.2545    0.2475    0.2475    0.2449
## adjCV   0.2626   0.2554   0.2555    0.2541    0.2476    0.2477    0.2451
##        14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
## CV       0.2446    0.2444    0.2435    0.2426    0.2419    0.2418    0.2416
## adjCV    0.2447    0.2447    0.2437    0.2420    0.2419    0.2420    0.2419
##        21 comps  22 comps  23 comps  24 comps  25 comps  26 comps  27 comps
## CV       0.2411    0.2409    0.2404    0.2403    0.2400    0.2396    0.2395
## adjCV    0.2412    0.2411    0.2405    0.2406    0.2401    0.2397    0.2398
##        28 comps  29 comps  30 comps  31 comps  32 comps  33 comps  34 comps
## CV       0.2379    0.2367    0.2365    0.2362    0.2359    0.2356    0.2354
## adjCV    0.2377    0.2366    0.2367    0.2363    0.2359    0.2354    0.2354
##        35 comps  36 comps  37 comps  38 comps  39 comps  40 comps  41 comps
## CV       0.2353    0.2353    0.2349    0.2344    0.2337    0.2330    0.2323
## adjCV    0.2354    0.2354    0.2347    0.2344    0.2336    0.2329    0.2324
##        42 comps  43 comps  44 comps  45 comps  46 comps  47 comps  48 comps
## CV       0.2319    0.2315    0.2314    0.2311    0.2308    0.2306    0.2305
## adjCV    0.2316    0.2314    0.2316    0.2310    0.2308    0.2306    0.2306
##        49 comps  50 comps  51 comps  52 comps  53 comps  54 comps  55 comps
## CV       0.2301    0.2299    0.2299    0.2295    0.2292    0.2290    0.2287
## adjCV    0.2303    0.2301    0.2300    0.2296    0.2294    0.2291    0.2288
##        56 comps  57 comps  58 comps  59 comps  60 comps  61 comps  62 comps
## CV       0.2287    0.2285    0.2283    0.2280    0.2280    0.2279    0.2276
## adjCV    0.2288    0.2286    0.2284    0.2281    0.2281    0.2281    0.2277
##        63 comps  64 comps  65 comps  66 comps  67 comps  68 comps  69 comps
## CV       0.2276    0.2276    0.2275    0.2274    0.2275    0.2272    0.2262
## adjCV    0.2277    0.2277    0.2276    0.2276    0.2277    0.2273    0.2264
##        70 comps  71 comps  72 comps  73 comps  74 comps  75 comps  76 comps
## CV       0.2014    0.1923    0.1917    0.1904    0.1851    0.1841    0.1825
## adjCV    0.2009    0.1922    0.1916    0.1904    0.1850    0.1841    0.1824
##        77 comps  78 comps  79 comps  80 comps  81 comps  82 comps  83 comps
## CV       0.1824    0.1809    0.1733    0.1732    0.1720    0.1685    0.1582
## adjCV    0.1824    0.1808    0.1732    0.1732    0.1721    0.1685    0.1582
##        84 comps  85 comps  86 comps
## CV       0.1543    0.1543    0.1532
## adjCV    0.1543    0.1543    0.1532
##
## TRAINING: % variance explained
##                  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X                  5.510    8.368    10.85    12.98    14.73    16.26    17.66
## log(price_sqft)    1.908   24.383    24.47    38.69    41.98    44.12    51.05
##                  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## X                 18.96    20.25     21.50     22.74     23.95     25.15
## log(price_sqft)    53.67    53.67     54.19     56.62     56.62     57.59
##                  14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
## X                   26.34     27.54     28.73     29.92     31.12     32.31
## log(price_sqft)     57.71     57.72     58.08     58.64     58.69     58.70
##                  20 comps  21 comps  22 comps  23 comps  24 comps  25 comps
## X                   33.50     34.69     35.88     37.07     38.25     39.44
## log(price_sqft)     58.73     58.94     59.01     59.21     59.23     59.36
##                  26 comps  27 comps  28 comps  29 comps  30 comps  31 comps
## X                   40.62     41.81     42.99     44.18     45.36     46.54
## log(price_sqft)     59.52     59.52     60.21     60.54     60.56     60.73
##                  32 comps  33 comps  34 comps  35 comps  36 comps  37 comps
## X                   47.73     48.91     50.09     51.27     52.45     53.63
## log(price_sqft)     60.84     61.03     61.04     61.04     61.10     61.30
##                  38 comps  39 comps  40 comps  41 comps  42 comps  43 comps
## X                   54.81     55.99     57.17     58.35     59.52     60.70
## log(price_sqft)     61.37     61.61     61.81     62.00     62.25     62.32
##                  44 comps  45 comps  46 comps  47 comps  48 comps  49 comps
## X                   61.88     63.06     64.23     65.41     66.59     67.77
## log(price_sqft)     62.32     62.52     62.61     62.70     62.70     62.79
##                  50 comps  51 comps  52 comps  53 comps  54 comps  55 comps
## X                   68.94     70.12     71.29     72.47     73.64     74.82
## log(price_sqft)     62.85     62.89     62.99     63.06     63.13     63.23
##                  56 comps  57 comps  58 comps  59 comps  60 comps  61 comps
## X                   75.99     77.16     78.34     79.51     80.68     81.85
## log(price_sqft)     63.23     63.30     63.38     63.48     63.48     63.52
##                  62 comps  63 comps  64 comps  65 comps  66 comps  67 comps
## X                   83.02     84.19     85.36     86.52     87.69     88.86
## log(price_sqft)     63.62     63.63     63.63     63.66     63.67     63.67
##                  68 comps  69 comps  70 comps  71 comps  72 comps  73 comps
## X                   90.02     91.13     92.21     93.26     94.29     95.27
## log(price_sqft)     63.77     64.02     71.64     74.02     74.13     74.47
##                  74 comps  75 comps  76 comps  77 comps  78 comps  79 comps
## X                   96.12     96.94     97.64     98.26     98.75     99.04
## log(price_sqft)     75.89     76.13     76.55     76.56     76.97     78.86
##                  80 comps  81 comps  82 comps  83 comps  84 comps  85 comps
## X                   99.32     99.57     99.78     99.92     99.98    100.00
## log(price_sqft)     78.88     79.17     80.01     82.37     83.23     83.23
##                  86 comps
## X                  100.00
## log(price_sqft)     83.48
```

```
PCR.p <- exp(predict(PCR, test, ncomp = 86))

PCR.mse <- mean((test$price_sqft - PCR.p)^2)

PCR.mse
```

```
## [1] 2061.569
```

```
# Random Forest regression CV

# Creating a Mode function
Mode <- function(x) {
  uni <- unique(x)
  uni[which.max(tabulate(match(x, uni)))]
}

# Creating a subsample of the data to CV due to computational strain
set.seed(10)

n2 <- NULL
m2 <- NULL

for (cv in 1:10) {
  rf.sub <- train[sample(nrow(train), nrow(train)/20),]
  cv.flag <- sample(nrow(rf.sub), nrow(rf.sub)/10)
  cv.train <- rf.sub[-cv.flag,]
  cv.test <- rf.sub[cv.flag,]

  n <- NULL
  m <- NULL

  for (i in c(100,200,300,400,500,600,700,800,900,1000)) {

    ms <- NULL

    for (k in 1:10) {
      rf.cv <- randomForest(log(price_sqft)~., data = cv.train[,c(-6,-14)]
                            + as.numeric(cv.train$zipcode)
                            + as.numeric(cv.train$waterfront),
                            ntree = i, mtry = k)
      p.rf <- exp(predict(rf.cv, cv.test))
      cv.mse <- mean((cv.test$price_sqft - p.rf)^2)
      ms <- c(ms, cv.mse)
    }

    min.mse <- min(ms)
    m <- c(m, which.min(ms))
    n <- c(n, min.mse)
  }

  m2 <- c(m2, Mode(m))
  n2 <- c(n2, which.min(n)*100)

}

opt.mtry <- Mode(m2)
opt.ntree <- Mode(n2)

cat("Opt mtry:", opt.mtry, "\n")
```

```
## Opt mtry: 1
```

```
cat("Opt ntrees:", opt.ntree)
```

```
## Opt ntrees: 100
```

```
rf.p <- randomForest(log(price_sqft) ~ ., data = train[,c(-6,-14)]
                     + as.numeric(train$zipcode)
                     + as.numeric(train$waterfront), importance = TRUE,
                     ntree = opt.ntree, mtry = opt.mtry)

rf.p
```

```
## 
## Call:
##  randomForest(formula = log(price_sqft) ~ ., data = train[, c(-6,      -14)] + as.numeric(tra
in$zipcode) + as.numeric(train$waterfront),      importance = TRUE, ntree = opt.ntree, mtry = op
t.mtry) 
##                Type of random forest: regression
##                      Number of trees: 100
## No. of variables tried at each split: 1
## 
##           Mean of squared residuals: 0.02431407
##                     % Var explained: 80.52
```

```
rf.p$importance
```

```
##                   %IncMSE IncNodePurity
## date          0.001437709      46.51502
## bedrooms      0.037876210     174.57475
## bathrooms     0.040465857     170.93269
## sqft_lot      0.010973615      84.11589
## floors        0.049263666     168.57854
## view          0.053970721     154.71353
## condition     0.040382865     143.72134
## grade         0.038540517     171.81303
## sqft_above    0.009847719      75.63696
## sqft_basement 0.025284156     144.15302
## yr_built      0.013835901      77.53221
## yr_renovated  0.042285958     153.57581
## lat           0.053178687     175.07680
## long          0.058847797     199.30080
## sqft_living15 0.010841314      78.75308
## sqft_lot15    0.014116382     109.24761
```

```
pred.rf <- exp(predict(rf.p, test[,c(-6,-14)]
                     + as.numeric(test$zipcode)
                     + as.numeric(test$waterfront)))
rf.mse <- mean((test$price_sqft - pred.rf)^2)

rf.mse
```

```
## [1] 4274.357
```

```
# Generalized additive model

f <- c(4,10,11,12,13,17,18)
sm <- colnames(train[,f])
nm <- colnames(train[,c(-f, -19)])

v <- paste0("s(", sm, ")", collapse = "+")
n2 <- paste0(nm, collapse = "+")
vn <- paste0(v, "+", n2)

form <- as.formula(paste0("log(price_sqft)~", vn, collapse = ""))
gm <- gam(formula = form, data = train)

summary(gm)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(price_sqft) ~ s(sqft_lot) + s(sqft_above) + s(sqft_basement) +
##     s(yr_built) + s(yr_renovated) + s(sqft_living15) + s(sqft_lot15) +
##     date + bedrooms + bathrooms + floors + waterfront + view +
##     condition + grade + zipcode + lat + long
##
## Parametric coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.327e-02  1.079e-02  -4.012 6.05e-05 ***
## date          2.026e-04  9.585e-06  21.132  < 2e-16 ***
## bedrooms     -1.905e-02  1.695e-03 -11.237  < 2e-16 ***
## bathrooms     2.563e-02  2.794e-03   9.175  < 2e-16 ***
## floors       -5.708e-02  3.524e-03 -16.200  < 2e-16 ***
## waterfront1   5.242e-01  1.725e-02  30.379  < 2e-16 ***
## view          6.675e-02  1.814e-03  36.790  < 2e-16 ***
## condition     5.254e-02  1.984e-03  26.489  < 2e-16 ***
## grade         8.781e-02  1.893e-03  46.387  < 2e-16 ***
## zipcode98002 -8.943e-03  1.394e-02  -0.642  0.52105
## zipcode98003  3.861e-02  1.253e-02   3.080  0.00207 **
## zipcode98004  1.058e+00  2.322e-02  45.576  < 2e-16 ***
## zipcode98005  6.792e-01  2.502e-02  27.152  < 2e-16 ***
## zipcode98006  5.878e-01  2.076e-02  28.308  < 2e-16 ***
## zipcode98007  5.997e-01  2.606e-02  23.013  < 2e-16 ***
## zipcode98008  5.830e-01  2.499e-02  23.329  < 2e-16 ***
## zipcode98010  1.627e-01  2.110e-02   7.711 1.31e-14 ***
## zipcode98011  3.426e-01  3.123e-02  10.971  < 2e-16 ***
## zipcode98014  1.717e-01  3.601e-02   4.769 1.87e-06 ***
## zipcode98019  1.796e-01  3.533e-02   5.084 3.74e-07 ***
## zipcode98022  2.305e-02  1.327e-02   1.737  0.08240 .
## zipcode98023  1.325e-04  1.109e-02   0.012  0.99047
## zipcode98024  2.838e-01  3.279e-02   8.655  < 2e-16 ***
## zipcode98027  4.691e-01  2.114e-02  22.190  < 2e-16 ***
## zipcode98028  3.073e-01  2.963e-02  10.370  < 2e-16 ***
## zipcode98029  5.229e-01  2.419e-02  21.620  < 2e-16 ***
## zipcode98030  3.484e-02  1.386e-02   2.513  0.01197 *
## zipcode98031  5.513e-02  1.460e-02   3.776  0.00016 ***
## zipcode98032 -1.754e-03  1.742e-02  -0.101  0.91982
## zipcode98033  6.790e-01  2.637e-02  25.749  < 2e-16 ***
## zipcode98034  4.527e-01  2.789e-02  16.230  < 2e-16 ***
## zipcode98038  1.221e-01  1.381e-02   8.840  < 2e-16 ***
## zipcode98039  1.284e+00  3.501e-02  36.689  < 2e-16 ***
## zipcode98040  8.457e-01  2.052e-02  41.209  < 2e-16 ***
## zipcode98042  3.642e-02  1.268e-02   2.872  0.00408 **
## zipcode98045  2.472e-01  2.543e-02   9.719  < 2e-16 ***
## zipcode98052  5.344e-01  2.755e-02  19.397  < 2e-16 ***
## zipcode98053  4.659e-01  2.982e-02  15.622  < 2e-16 ***
## zipcode98055  1.116e-01  1.649e-02   6.767 1.35e-11 ***
## zipcode98056  2.654e-01  1.797e-02  14.774  < 2e-16 ***
## zipcode98058  1.277e-01  1.554e-02   8.218  < 2e-16 ***
## zipcode98059  2.750e-01  1.764e-02  15.593  < 2e-16 ***
## zipcode98065  2.981e-01  2.549e-02  11.697  < 2e-16 ***
## zipcode98070  2.917e-01  1.964e-02  14.851  < 2e-16 ***
## zipcode98072  3.679e-01  3.185e-02  11.551  < 2e-16 ***
## zipcode98074  4.500e-01  2.633e-02  17.091  < 2e-16 ***
## zipcode98075  4.524e-01  2.508e-02  18.034  < 2e-16 ***
## zipcode98077  3.252e-01  3.367e-02   9.660  < 2e-16 ***
## zipcode98092  3.883e-03  1.187e-02   0.327  0.74355    
## zipcode98102  9.246e-01  2.640e-02  35.018  < 2e-16 ***
## zipcode98103  7.518e-01  2.305e-02  32.613  < 2e-16 ***
## zipcode98105  8.639e-01  2.494e-02  34.638  < 2e-16 ***
## zipcode98106  3.338e-01  1.697e-02  19.674  < 2e-16 ***
## zipcode98107  7.754e-01  2.324e-02  33.363  < 2e-16 ***
## zipcode98108  3.336e-01  2.010e-02  16.596  < 2e-16 ***
## zipcode98109  9.126e-01  2.595e-02  35.169  < 2e-16 ***
## zipcode98112  9.823e-01  2.299e-02  42.723  < 2e-16 ***
## zipcode98115  7.350e-01  2.426e-02  30.293  < 2e-16 ***
## zipcode98116  7.239e-01  1.799e-02  40.248  < 2e-16 ***
## zipcode98117  7.380e-01  2.289e-02  32.240  < 2e-16 ***
## zipcode98118  4.239e-01  1.787e-02  23.722  < 2e-16 ***
## zipcode98119  9.172e-01  2.343e-02  39.153  < 2e-16 ***
## zipcode98122  7.437e-01  2.174e-02  34.205  < 2e-16 ***
## zipcode98125  4.790e-01  2.613e-02  18.333  < 2e-16 ***
## zipcode98126  5.146e-01  1.690e-02  30.445  < 2e-16 ***
## zipcode98133  3.833e-01  2.601e-02  14.736  < 2e-16 ***
## zipcode98136  6.561e-01  1.717e-02  38.213  < 2e-16 ***
## zipcode98144  6.112e-01  2.019e-02  30.271  < 2e-16 ***
## zipcode98146  2.470e-01  1.598e-02  15.456  < 2e-16 ***
## zipcode98148  1.896e-01  2.613e-02   7.254 4.20e-13 ***
## zipcode98155  3.373e-01  2.798e-02  12.052  < 2e-16 ***
## zipcode98166  2.979e-01  1.504e-02  19.815  < 2e-16 ***
## zipcode98168  6.890e-02  1.626e-02   4.238 2.27e-05 ***
## zipcode98177  4.810e-01  2.700e-02  17.819  < 2e-16 ***
## zipcode98178  1.139e-01  1.772e-02   6.428 1.32e-10 ***
## zipcode98188  9.144e-02  1.826e-02   5.006 5.60e-07 ***
## zipcode98198  7.612e-02  1.349e-02   5.642 1.71e-08 ***
## zipcode98199  8.054e-01  2.145e-02  37.539  < 2e-16 ***
## lat           2.692e-01  5.915e-02   4.551 5.36e-06 ***
## long          9.632e-02  2.290e-02   4.206 2.61e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Approximate significance of smooth terms:
##                     edf Ref.df        F p-value    
## s(sqft_lot)       8.704  8.968   35.934  <2e-16 ***
## s(sqft_above)     6.415  7.467  612.898  <2e-16 ***
## s(sqft_basement) 4.021  4.948 1520.126  <2e-16 ***
## s(yr_built)       8.326  8.870   69.876  <2e-16 ***
## s(yr_renovated)  3.162  3.473   76.895  <2e-16 ***
## s(sqft_living15) 5.055  6.196   65.277  <2e-16 ***
## s(sqft_lot15)    8.747  8.980    6.598  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 138/143
## R-sq.(adj) =  0.849   Deviance explained =   85%
## GCV = 0.021388  Scale est. = 0.021244  n = 18337
```

```
pred.gm <- exp(predict(gm, test))
gm.mse <- mean((test$price_sqft - pred.gm)^2)

gm.mse
```

```
## [1] 1887.063
```

```
# Averaging predictions of all models together
combined <- (pred.lm6 + pred.step +  pred.rf + pred.gm + p.lasso[,1] +
             p.ridge[,1] + as.vector(PCR.p)) / 7

combined.mse <- mean((test$price_sqft - combined)^2)

combined.mse
```

```
## [1] 2245.721
```

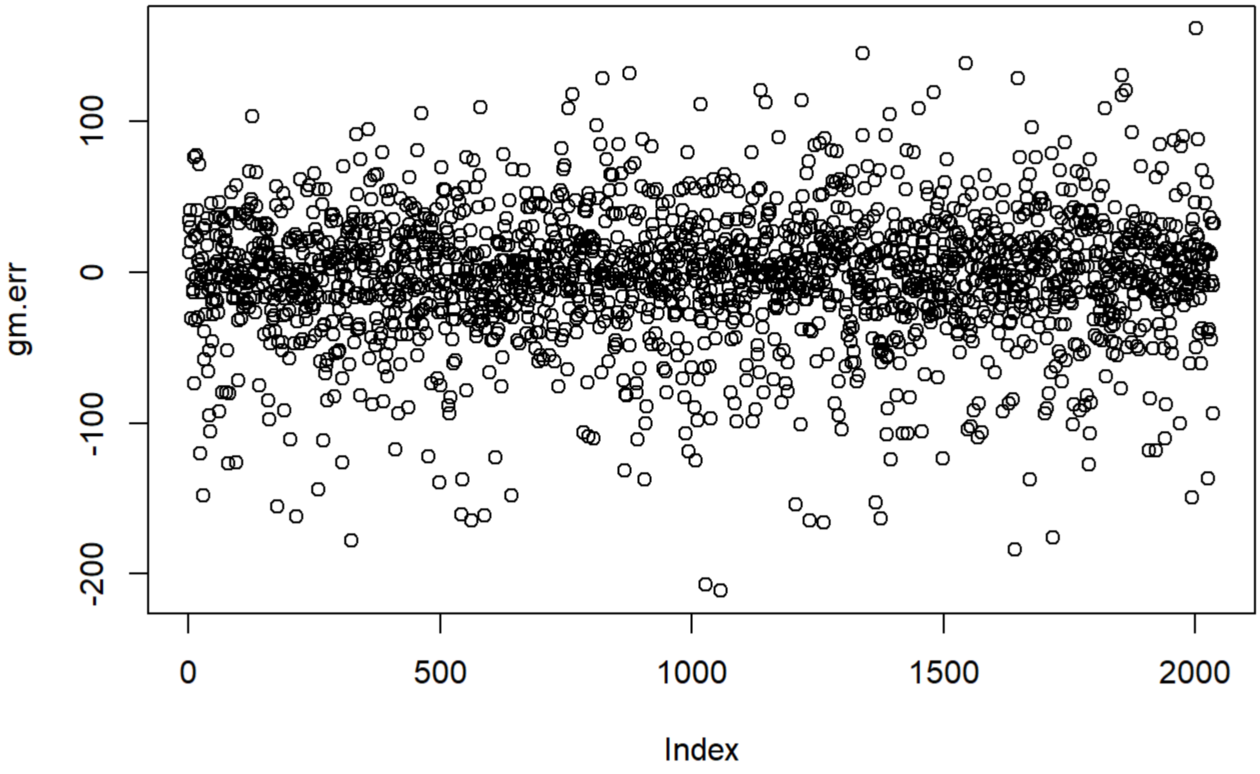
```
mses <- c(lm6.mse, step.mse, lasso.mse, ridge.mse, PCR.mse, rf.mse, gm.mse,
          combined.mse)

MSEs <- data.frame(Model = c("Linear Model", "Stepwise", "LASSO", "Ridge",
                     "PCR", "Random Forest", "GAM", "Combination"),
                   MSE = mses)

MSEs
```

```
##           Model      MSE
## 1  Linear Model 2061.513
## 2      Stepwise 2060.988
## 3         LASSO 4931.009
## 4         Ridge 5068.003
## 5           PCR 2061.569
## 6 Random Forest 4274.357
## 7           GAM 1887.063
## 8   Combination 2245.721
```

```
gm.err <- pred.gm - test$price_sqft

summary(gm.err)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## -211.3528  -22.2442   -0.6779   -2.9227   21.0632  161.7432
```

```
plot(gm.err)
```



```
test[which.max(abs(gm.err)),]
```

```
##             date bedrooms bathrooms sqft_lot floors waterfront view condition
## 10695 2015-02-25        3         1     6120    1.5          0    0         3
##       grade sqft_above sqft_basement yr_built yr_renovated zipcode     lat
## 10695     7       1140             0     1926            0   98115 47.6822
##           long sqft_living15 sqft_lot15 price_sqft
## 10695 -122.309          1800       4080   617.5439
```

```
test[which.max(gm.err),]
```

```
##            date bedrooms bathrooms sqft_lot floors waterfront view condition
## 3988 2014-07-01        1         1      833      1          0    0         4
##      grade sqft_above sqft_basement yr_built yr_renovated zipcode     lat
## 3988     7        590             0     1926            0   98122 47.6082
##          long sqft_living15 sqft_lot15 price_sqft
## 3988 -122.299           780       1617   342.3729
```