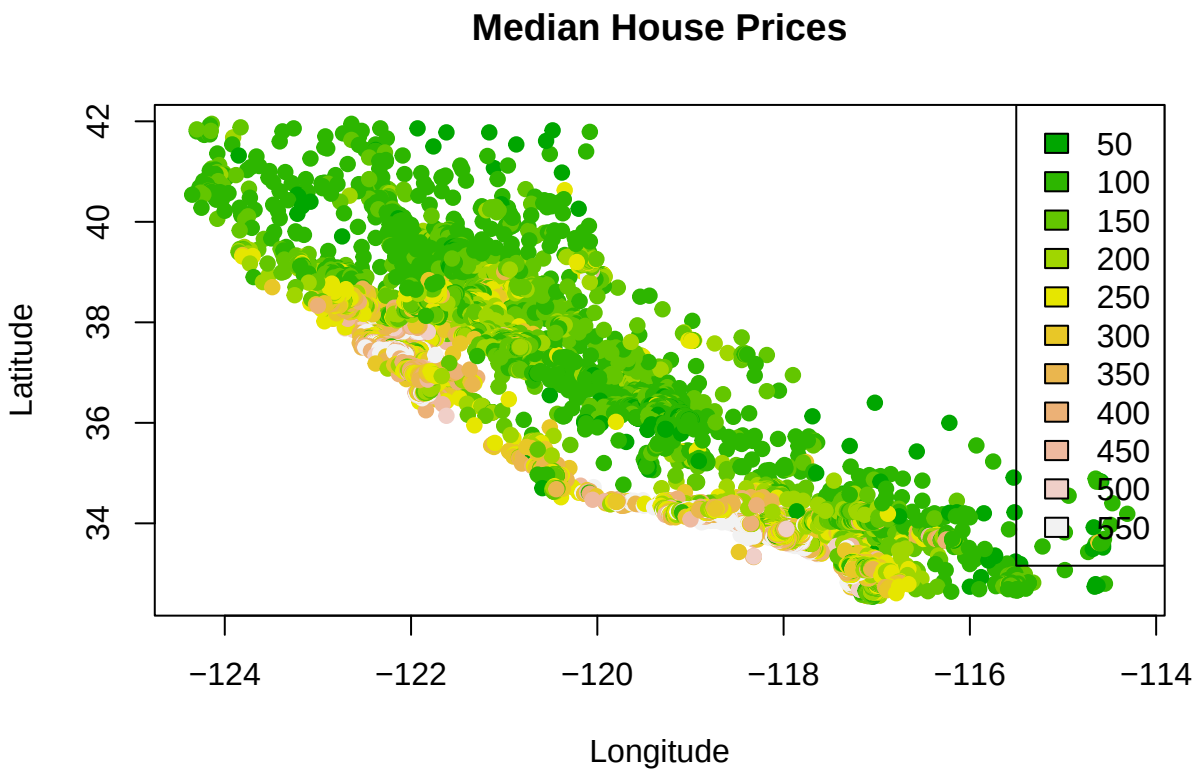
Lab3

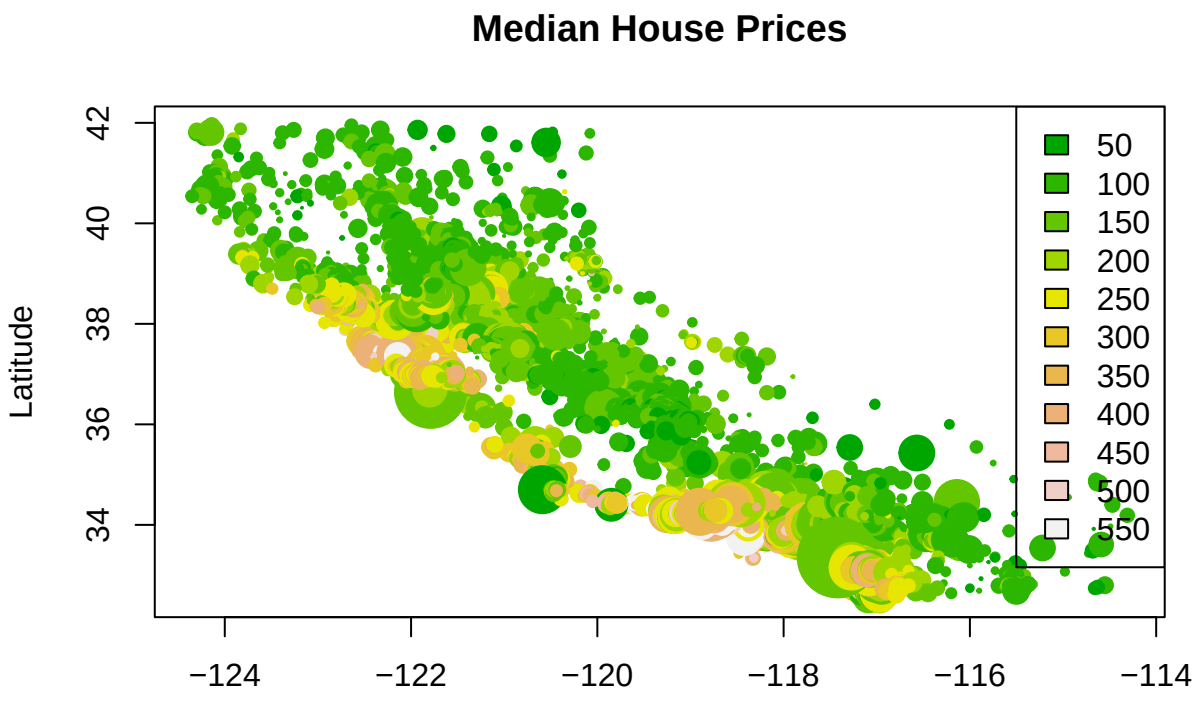
Joshua Jansen-Montoya

2022-10-11

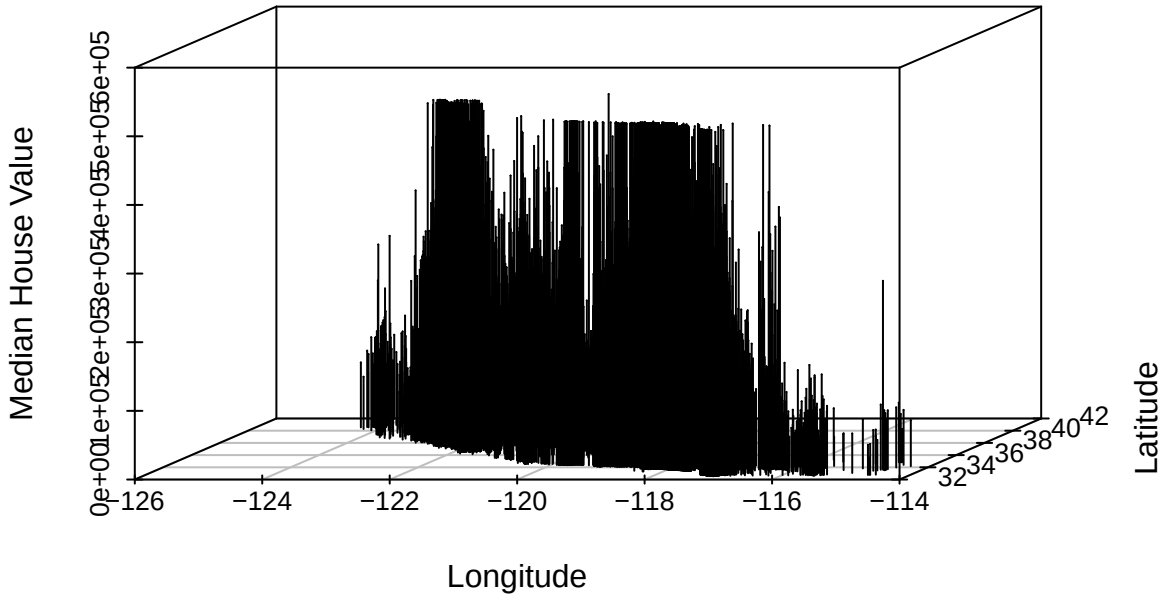
```
calif = read.table("http://www.stat.cmu.edu/~cshalizi/uADA/12/hw/01/cadata.dat", header=TRUE)

library(scatterplot3d)
plot(calif$Longitude,calif$Latitude,pch=21, col=terrain.colors(11)[1+floor(calif$MedianHouseValue/50e3)]
       bg=terrain.colors(11)[1+floor(calif$MedianHouseValue/50e3)],
       xlab="Longitude",ylab="Latitude",main="Median House Prices")
legend(x="topright",legend=(50*(1:11)),fill=terrain.colors(11))
```




```
plot(calif$Longitude,calif$Latitude,pch=21,
     col=terrain.colors(11)[1+floor(calif$MedianHouseValue/50e3)], bg=terrain.colors(11)[1+floor(calif$
     cex=sqrt(calif$Population/median(calif$Population)),
     xlab="Longitude",ylab="Latitude",main="Median House Prices",
     sub="Circle area proportional to population")
legend(x="topright",legend=(50*(1:11)),fill=terrain.colors(11))
```




```
scatterplot3d(calif$Longitude,calif$Latitude,calif$MedianHouseValue,
              pch=20,scale.y=0.4,type="h",cex.symbol=0.01,
              xlab="Longitude",ylab="Latitude",zlab="Median House Value")
```




```
fit = lm(MedianHouseValue ~ ., data=calif)
summary(fit)
```

```
##
## Call:
```

```
## lm(formula = MedianHouseValue ~ ., data = calif)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -563013  -43592  -11327   30307  803996
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.594e+06  6.254e+04 -57.468  < 2e-16 ***
## MedianIncome    4.025e+04  3.351e+02 120.123  < 2e-16 ***
## MedianHouseAge  1.156e+03  4.317e+01  26.787  < 2e-16 ***
## TotalRooms     -8.182e+00  7.881e-01 -10.381  < 2e-16 ***
## TotalBedrooms   1.134e+02  6.902e+00  16.432  < 2e-16 ***
## Population     -3.854e+01  1.079e+00 -35.716  < 2e-16 ***
## Households      4.831e+01  7.515e+00   6.429 1.32e-10 ***
## Latitude       -4.258e+04  6.733e+02 -63.240  < 2e-16 ***
## Longitude      -4.282e+04  7.130e+02 -60.061  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69530 on 20631 degrees of freedom
## Multiple R-squared:  0.6371, Adjusted R-squared:  0.637
## F-statistic:  4528 on 8 and 20631 DF,  p-value: < 2.2e-16
```

```
signif(coefficients(fit),3)
```

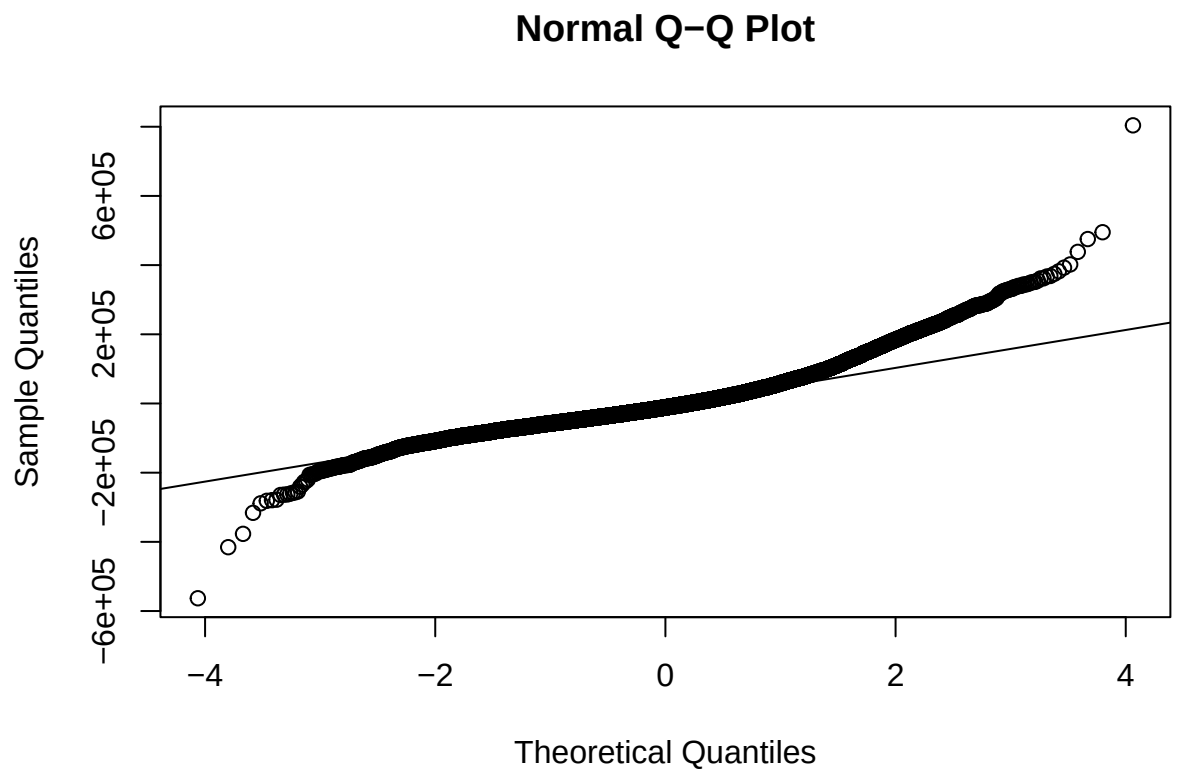
```
##    (Intercept)   MedianIncome MedianHouseAge     TotalRooms  TotalBedrooms
##      -3.59e+06       4.02e+04       1.16e+03      -8.18e+00       1.13e+02
##     Population     Households       Latitude      Longitude
##      -3.85e+01       4.83e+01      -4.26e+04      -4.28e+04
```

```
signif(summary(fit)$r.squared,digits=3)
```

```
## [1] 0.637
```

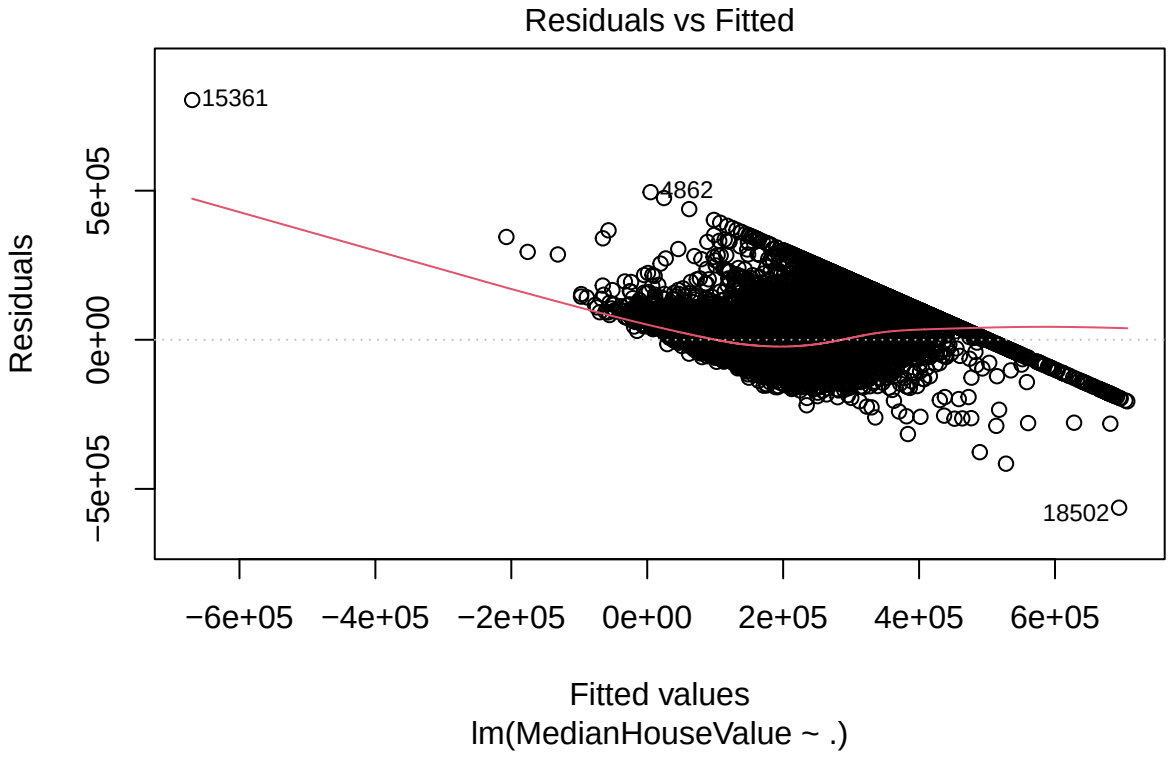
The variables that seem the most important seem to be longitude latitude, and median income which makes sense given the fact that there is a massive Difference in the price of a house in California depending on whre you are from such as with Bakersfield versus San Francisco. Similarly, if you make a lot more money than other people, then it would make sense that you would be able to afford a more expensive house which is why Median income afffecting the output makes a lot of sense.

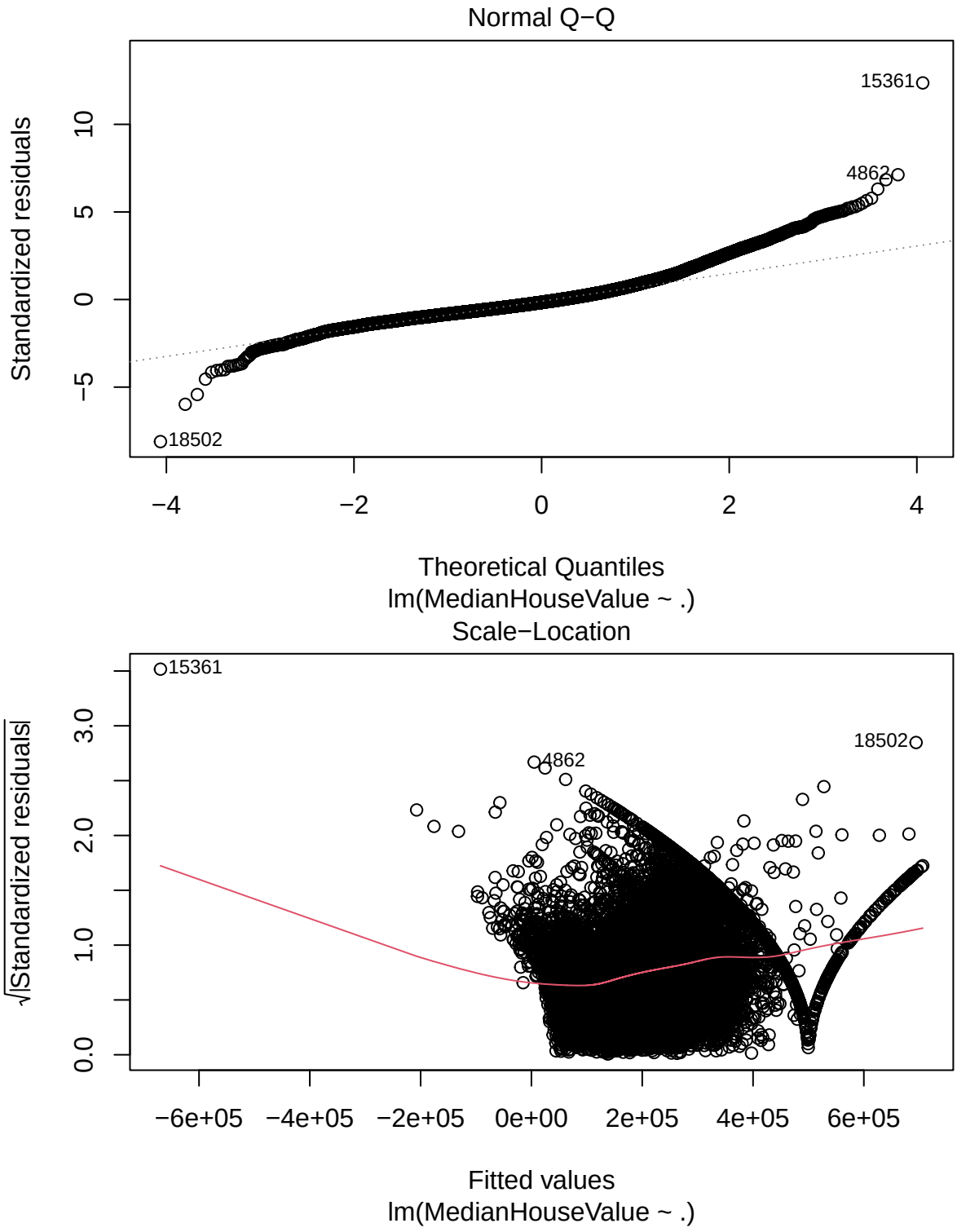
```
  qqnorm(resid(fit))
  qqline(resid(fit))
```

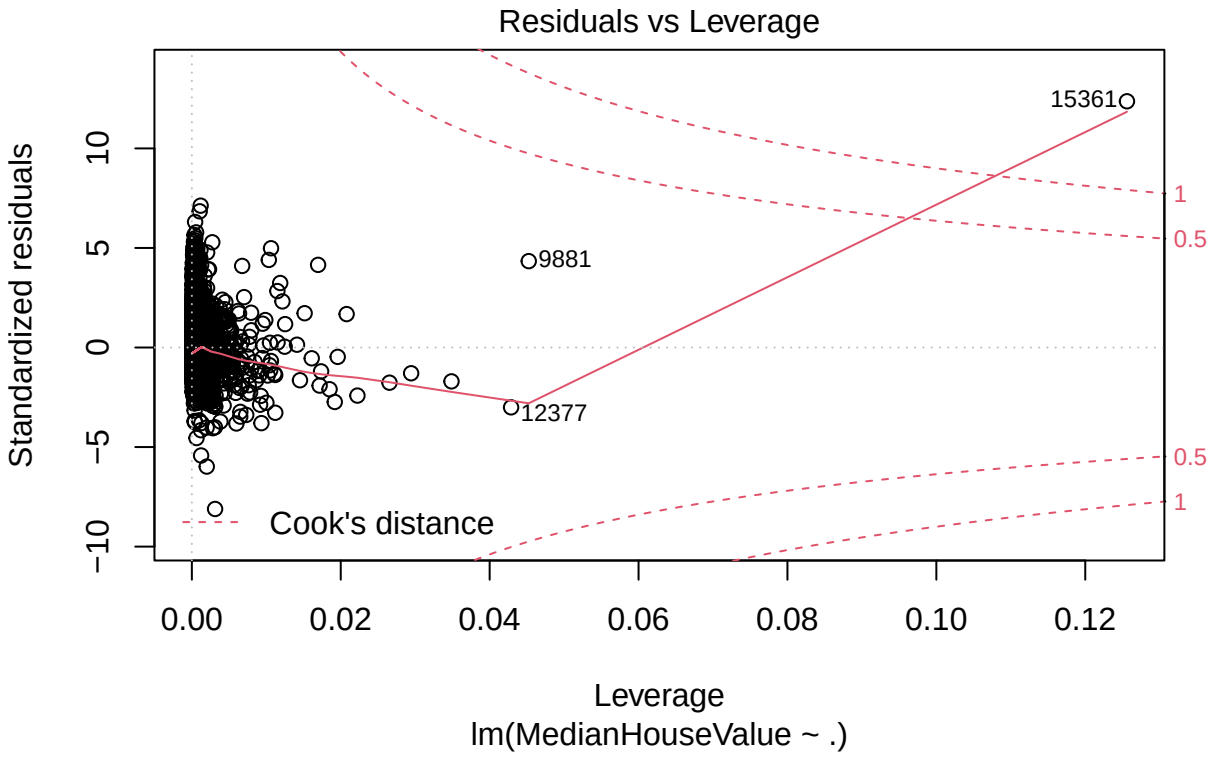


Looking at our residuals, we can see that we do not quite have our normalacy.

```
plot(fit)
```




Normal Q−Q
Standardized residuals
15361
4862
18502
−4
−2
0
2
4
10
5
0
−5
Theoretical Quantiles
lm(MedianHouseValue ~ .)
Scale−Location
√|Standardized residuals|
15361
4862
18502
3.0
2.0
1.0
0.0
−6e+05
−4e+05
−2e+05
0e+00
2e+05
4e+05
6e+05
Fitted values
lm(MedianHouseValue ~ .)




```
calif_sample = calif[sample(nrow(calif), 5000),]
fit_sample = lm(MedianHouseValue ~ ., data=calif_sample)
shapiro.test(residuals(fit_sample))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit_sample)
## W = 0.93498, p-value < 2.2e-16
```

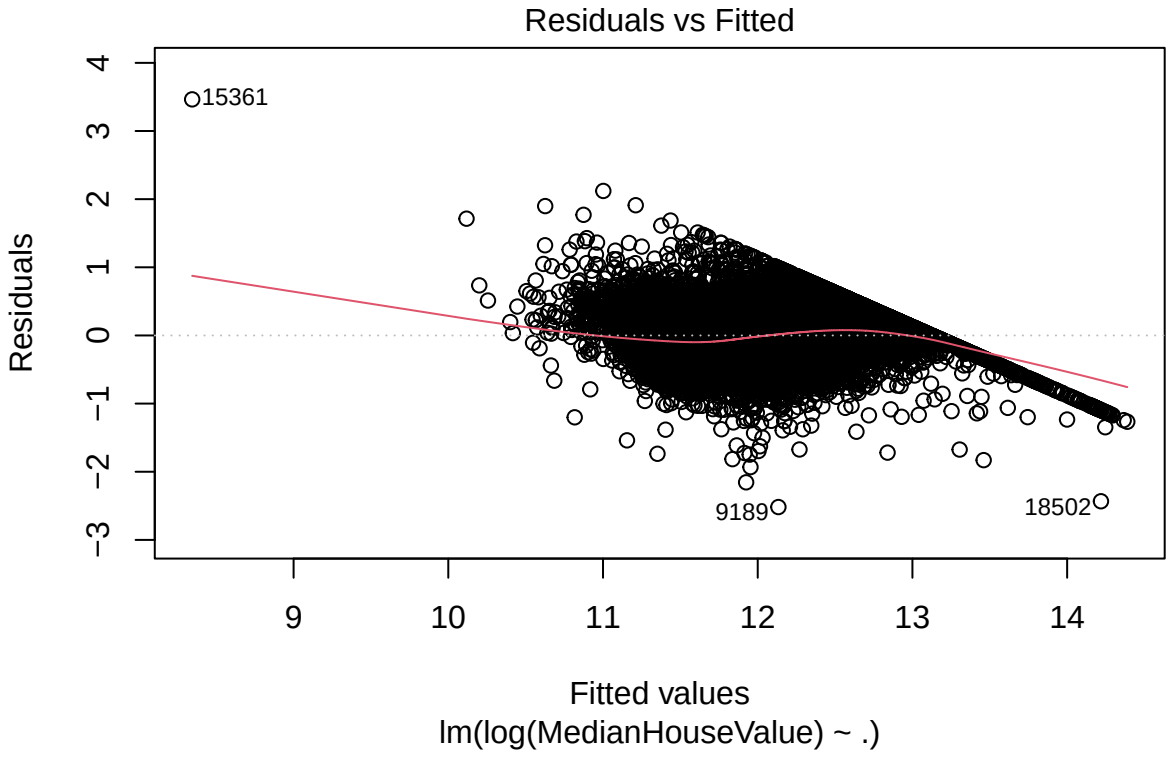
Looking at the plot of our residuals versus our fitted, we can see that there is a clustering of points together and that the red line is not approximately fit nor close to zero, and thus, we can determine that our residuals are not normal. Our Shapiro-Wilks test using a random sample of our data agrees with this with high probability that our data is not normal.

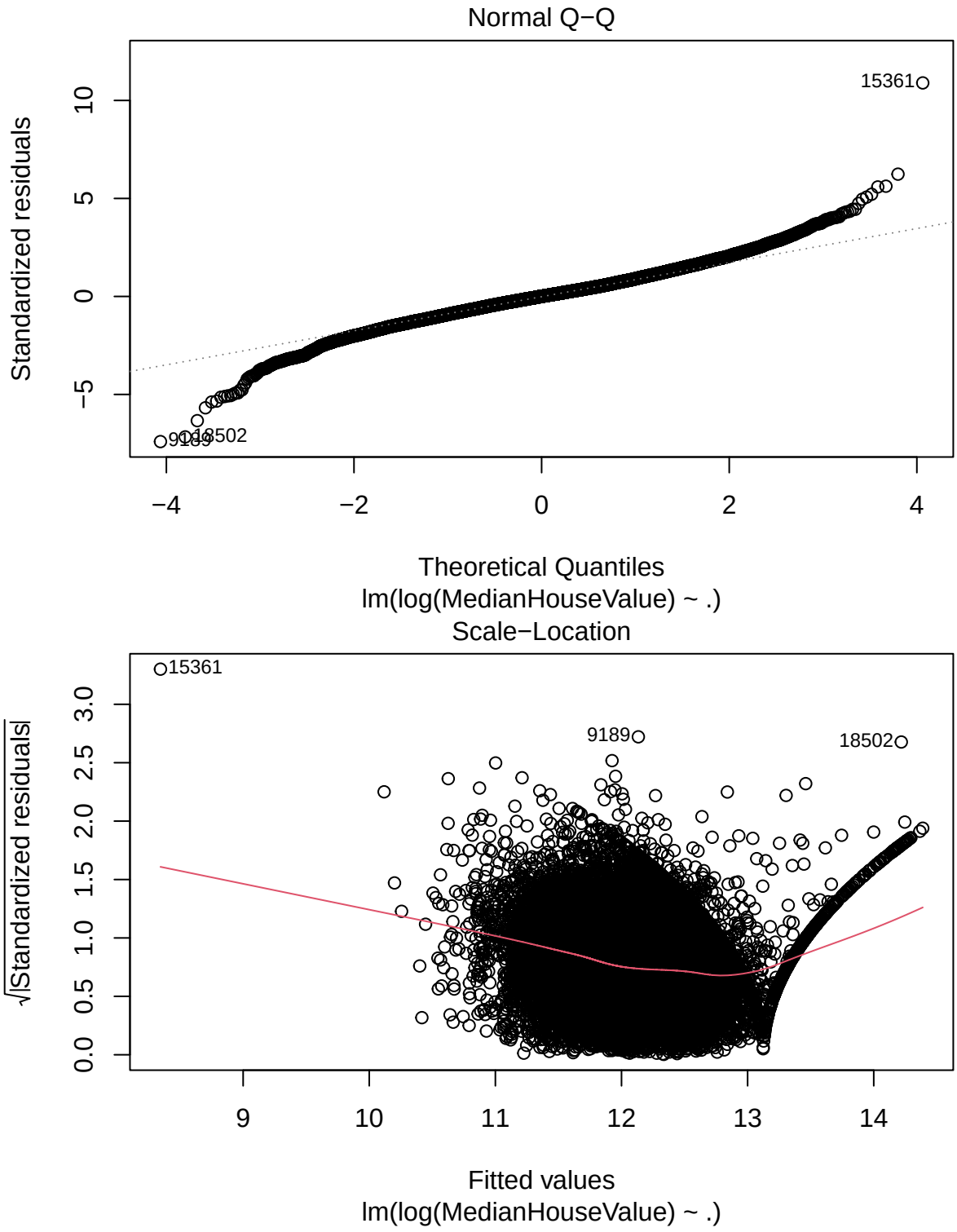
```
fitLog = lm(log(MedianHouseValue) ~ ., data=calif)
summary(fitLog)
```

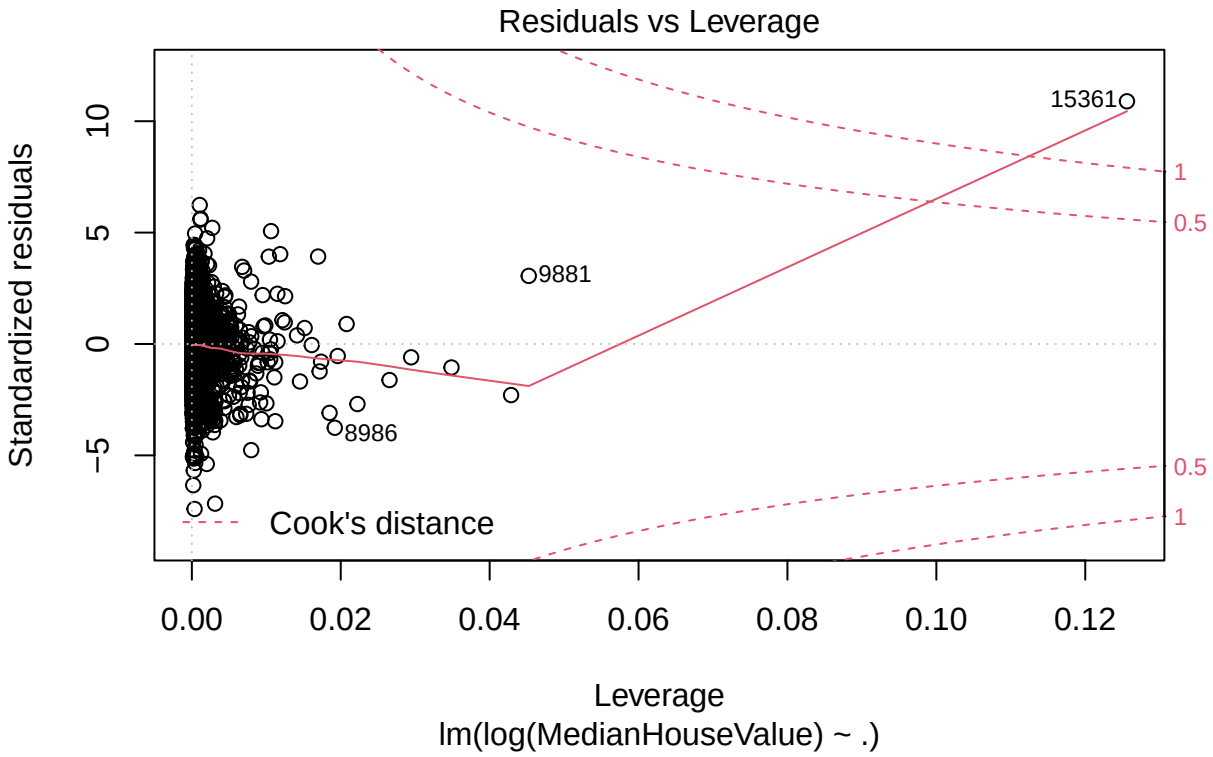
```
##
## Call:
## lm(formula = log(MedianHouseValue) ~ ., data = calif)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5180 -0.2038  0.0016  0.1949  3.4641
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.180e+01  3.059e-01 -38.570  < 2e-16 ***
## MedianIncome    1.782e-01  1.639e-03 108.753  < 2e-16 ***
## MedianHouseAge  3.261e-03  2.111e-04  15.446  < 2e-16 ***
```

```
## TotalRooms     -3.186e-05  3.855e-06  -8.265  < 2e-16 ***
## TotalBedrooms   4.798e-04  3.375e-05  14.215  < 2e-16 ***
## Population     -1.725e-04  5.277e-06 -32.687  < 2e-16 ***
## Households      2.493e-04  3.675e-05   6.783 1.21e-11 ***
## Latitude       -2.801e-01  3.293e-03 -85.078  < 2e-16 ***
## Longitude      -2.762e-01  3.487e-03 -79.212  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.34 on 20631 degrees of freedom
## Multiple R-squared:  0.6432, Adjusted R-squared:  0.643
## F-statistic:  4648 on 8 and 20631 DF,  p-value: < 2.2e-16
```

```
plot(fitLog)
```




Normal Q−Q
Standardized residuals
15361
9189
18502
Theoretical Quantiles
lm(log(MedianHouseValue) ~ .)
Scale−Location
√|Standardized residuals|
15361
9189
18502
Fitted values
lm(log(MedianHouseValue) ~ .)



We prefer the original model because even it still is not perfectly normal, the logged values are not normal either. We would instead prefer the more natural representation of the points in this instance which leads us to defer to the first model.