Introduction

This California housing dataset is from the StatLib repository and the information was collected from all block groups in California from the 1990 Census. In this sample, a block group on average includes 1425 individuals living in a geographically compact area. This dataset provides insight on over 19,000 block groups spread across the state. Our quantitative response variable is the median house value. We are interested in determining which of our six predictor variables best predict median house value. The variables are:

- housing median age
- total rooms
- population
- geographical coordinates (latitude and longitude)
- proximity to the ocean (1=Near Ocean, 2=Near Bay, 3=<1 Hour to Ocean, 4=Inland)
- median income of the residents.

```
# Read in the data
housing_data = read.csv(file="C:/Users/Lara/OneDrive/Desktop/r_projects/MATH327/housing_data.csv",
                        header=TRUE,
                        col.names=c("longitude", "latitude", "housing_median_age", "total_rooms",
                                    "population", "households", "median_income", "median_house_valu
e",
                                    "ocean_proximity", "ocean_proximity_label"))
```

```
#clean data
dim(housing_data)
```

```
## [1] 19643    10
```

```
housing_data = housing_data[complete.cases (housing_data), ]

# removing near bay because it is very similar to near ocean
housing_data <- housing_data[housing_data$ocean_proximity != "NEAR BAY",]
dim(housing_data)
```

```
## [1] 17555    10
```

###Exploratory analysis

```
library(MASS)
library(tidyr)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

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

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library (corrplot)
```

```
## corrplot 0.92 loaded
```

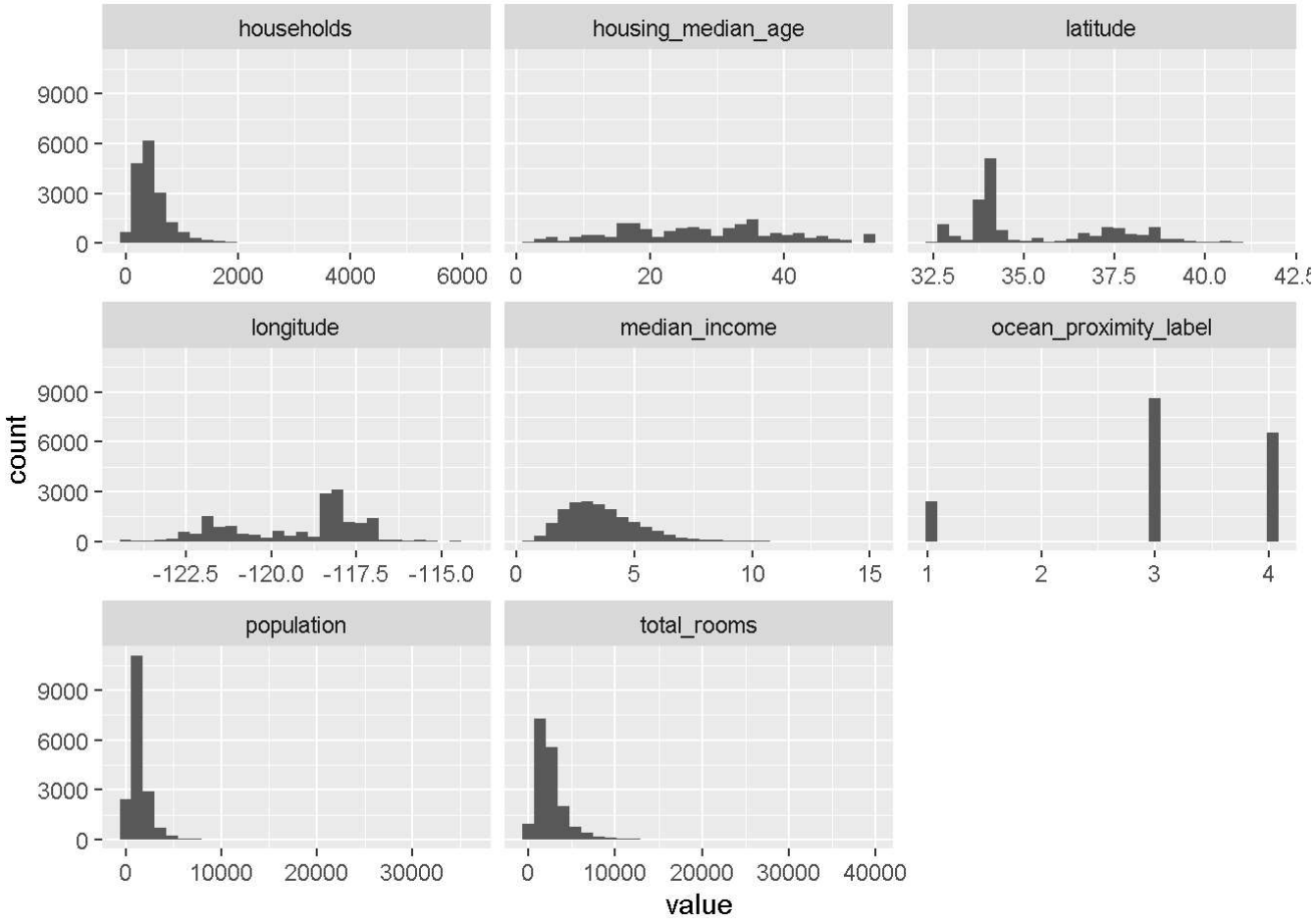
```
library(car)
```

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library (corrplot)



# exclude 8th and 9th columns
housing_data_long <- housing_data %>%
  dplyr::select(-c(8,9)) %>%
  gather(key="variable", value="value")

# Plot
ggplot(housing_data_long, aes(x=value)) +
  geom_histogram(bins=30) +
  facet_wrap(~variable, scales='free_x')
```



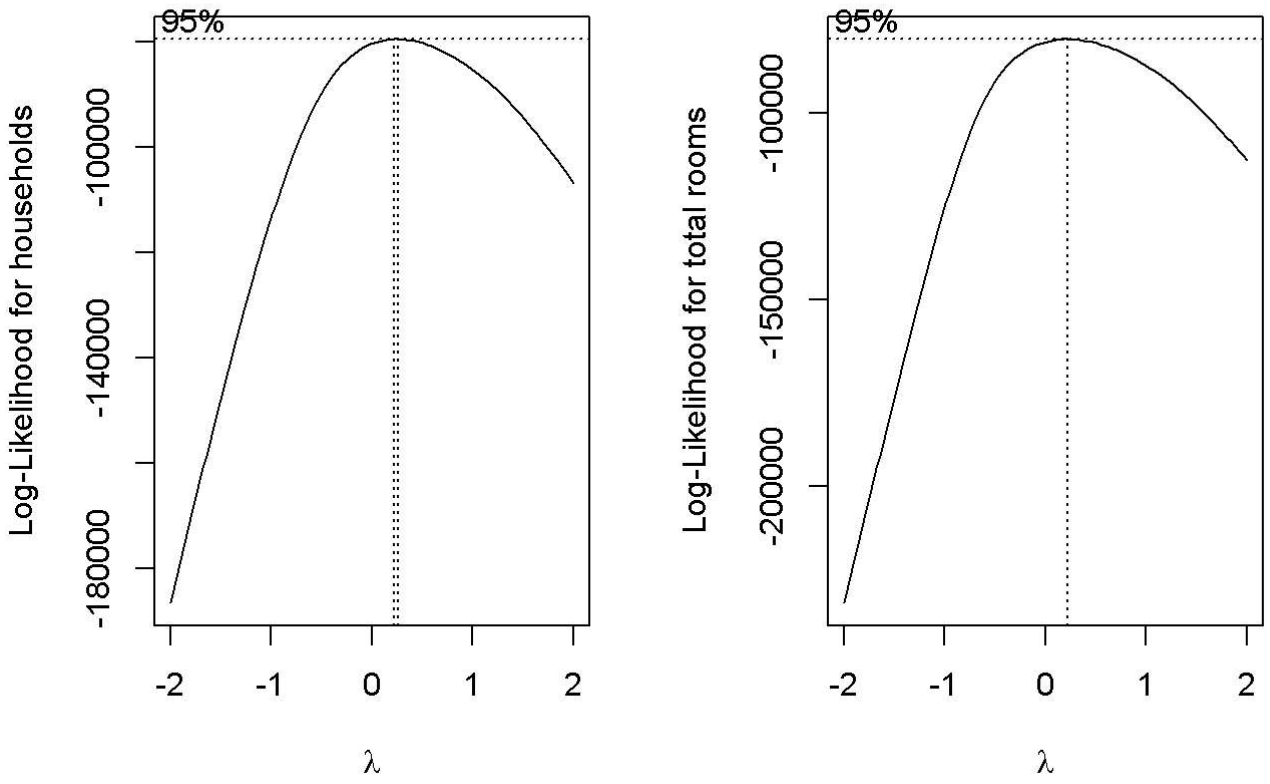

```
summary(housing_data)
```

```
##    longitude         latitude     housing_median_age  total_rooms
##  Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2
##  1st Qu.:-121.2   1st Qu.:33.90   1st Qu.:17.00      1st Qu.: 1442
##  Median :-118.4   Median :34.16   Median :27.00      Median : 2120
##  Mean   :-119.2   Mean   :35.40   Mean   :27.31      Mean   : 2637
##  3rd Qu.:-117.9   3rd Qu.:37.35   3rd Qu.:36.00      3rd Qu.: 3140
##  Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320
##    population      households     median_income     median_house_value
##  Min.   :    3   Min.   :   2.0   Min.   : 0.4999   Min.   : 14999
##  1st Qu.:  807   1st Qu.: 282.0   1st Qu.: 2.5000   1st Qu.:112900
##  Median : 1198   Median : 412.0   Median : 3.4107   Median :168300
##  Mean   : 1464   Mean   : 502.5   Mean   : 3.6508   Mean   :186785
##  3rd Qu.: 1774   3rd Qu.: 606.5   3rd Qu.: 4.5572   3rd Qu.:240300
##  Max.   :35682   Max.   :6082.0   Max.   :15.0001   Max.   :499100
##  ocean_proximity    ocean_proximity_label
##  Length:17555       Min.   :1.000
##  Class :character   1st Qu.:3.000
##  Mode  :character   Median :3.000
##                     Mean   :3.094
##                     3rd Qu.:4.000
##                     Max.   :4.000
```

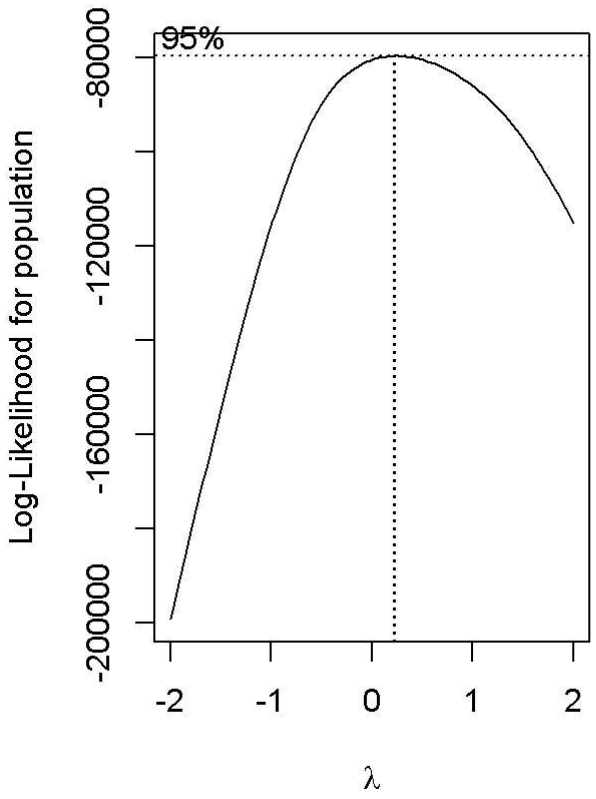
- The minimum household per block is 2 and maximum is 6082. Most of the households are below around 1500 per block.
- Median house age ranges from 1 to 52 years with average of 27 years.
- Latitude ranges from about 32 to 42 degrees. Most of the houses are within 33 to 35 degrees.
- The minimum longitude is -124.3 degrees to maximum of -114.3 degrees. Most of the houses are between -120 to -117 degrees.
- Minimum median house value $149990 to maximum of $4991000. There could have been houses above the maximum value as well but we removed those values because all of them were capped at 5 million dollars.
- Median household income from $4999 to 150 thousands dollars. Most of the income are between $5000 to $50000.
- ocean proximity measures how close a house is from the ocean. 1 being closeand 4 beaing farther away. Most of the houses are farther away from
- There are 3 categories for ocean proximity. Most of the houses are less than 1 hours from the ocean (1=Near Ocean, 2=Near Bay, 3=<1 Hour to Ocean, 4=Inland). Most of the house are less than an hour away from the ocean.
- The minimum population in a block is 3 where maximum is 35682 people with median of 1198 people per block.
- Total rooms in a block is 2 with maximum room being 39320. Most of the room are between 1000 to 10000 in a block.

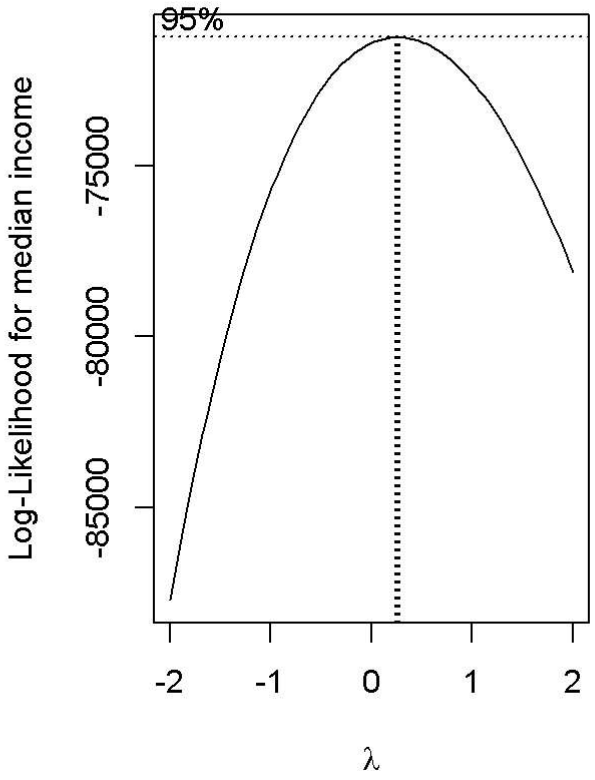
We can use the Box-Cox method with an “empty” model to help decide how to transform individual predictor variables.

```
#only includes postitve variables
par (mfrow=c(1,2))
MASS::boxcox (lm (housing_data$households ~ 1), ylab = "Log-Likelihood for households")
MASS::boxcox (lm (housing_data$total_rooms ~ 1), ylab = "Log-Likelihood for total rooms")
```




```
MASS::boxcox(lm(housing_data$population ~ 1), ylab = "Log-Likelihood for population")
MASS::boxcox(lm(housing_data$median_income ~ 1), ylab = "Log-Likelihood for median income")
```






```
#  λ that maximizes the log-likelihood for households
bc_result_households <- MASS::boxcox(lm(housing_data$households ~ 1), plotit = FALSE)
lambda_best_households <- bc_result_households$x[which.max(bc_result_households$y)]
print(paste("Best lambda for households:", lambda_best_households))
```

```
## [1] "Best lambda for households: 0.2"
```

```
# λ that maximizes the log-likelihood for total_rooms
bc_result_total_rooms <- MASS::boxcox(lm(housing_data$total_rooms ~ 1), plotit = FALSE)
lambda_best_total_rooms <- bc_result_total_rooms$x[which.max(bc_result_total_rooms$y)]
print(paste("Best lambda for total rooms:", lambda_best_total_rooms))
```

```
## [1] "Best lambda for total rooms: 0.2"
```

```
# λ that maximizes the log-likelihood for population
bc_result_population <- MASS::boxcox(lm(housing_data$population ~ 1), plotit = FALSE)
lambda_best_population <- bc_result_population$x[which.max(bc_result_population$y)]
print(paste("Best lambda for population:", lambda_best_population))
```

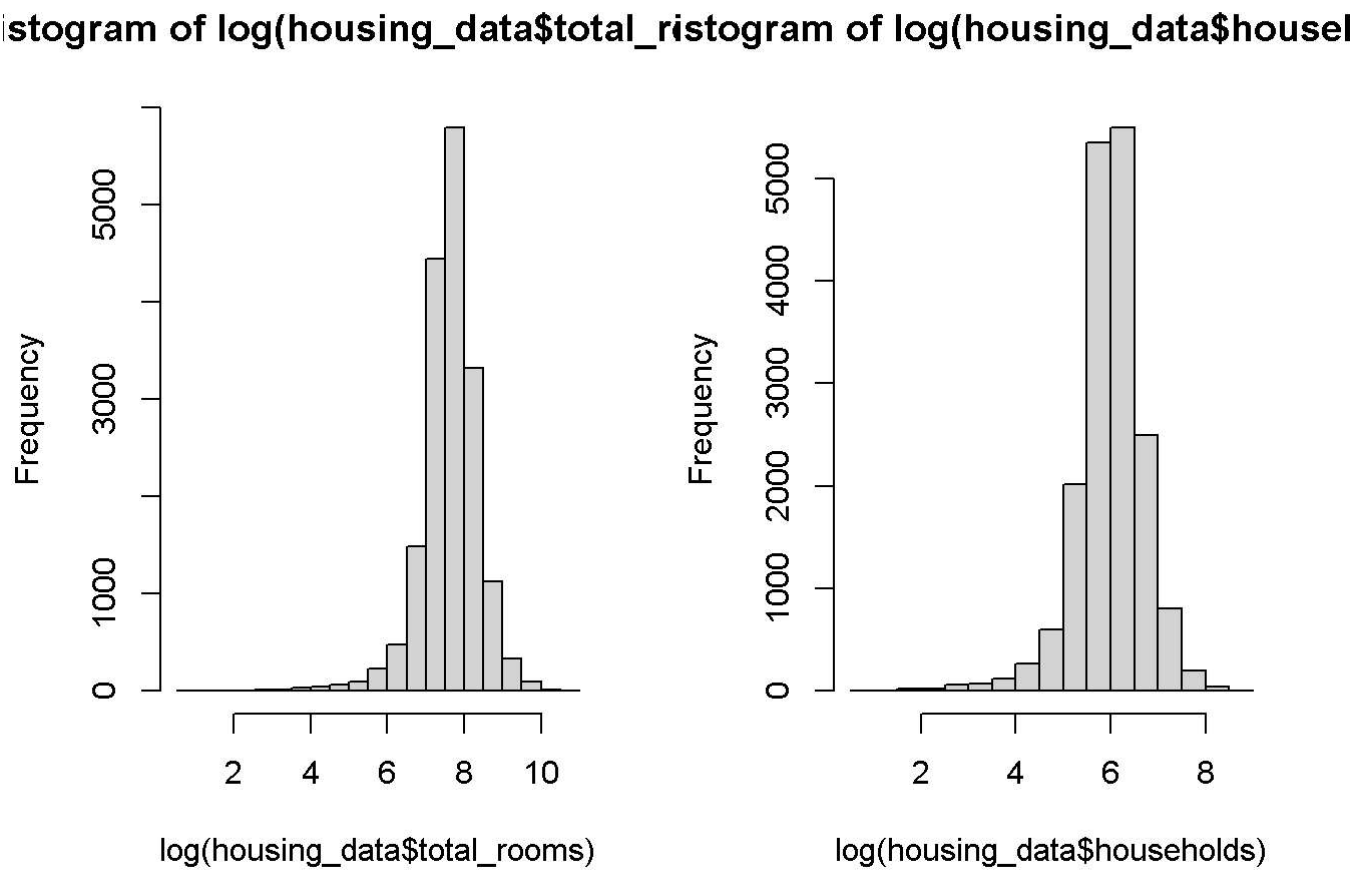
```
## [1] "Best lambda for population: 0.2"
```

```
# λ that maximizes the log-likelihood for median_income
bc_result_median_income <- MASS::boxcox(lm(housing_data$median_income ~ 1), plotit = FALSE)
lambda_best_median_income <- bc_result_median_income$x[which.max(bc_result_median_income$y)]
print(paste("Best lambda for median income:", lambda_best_median_income))
```

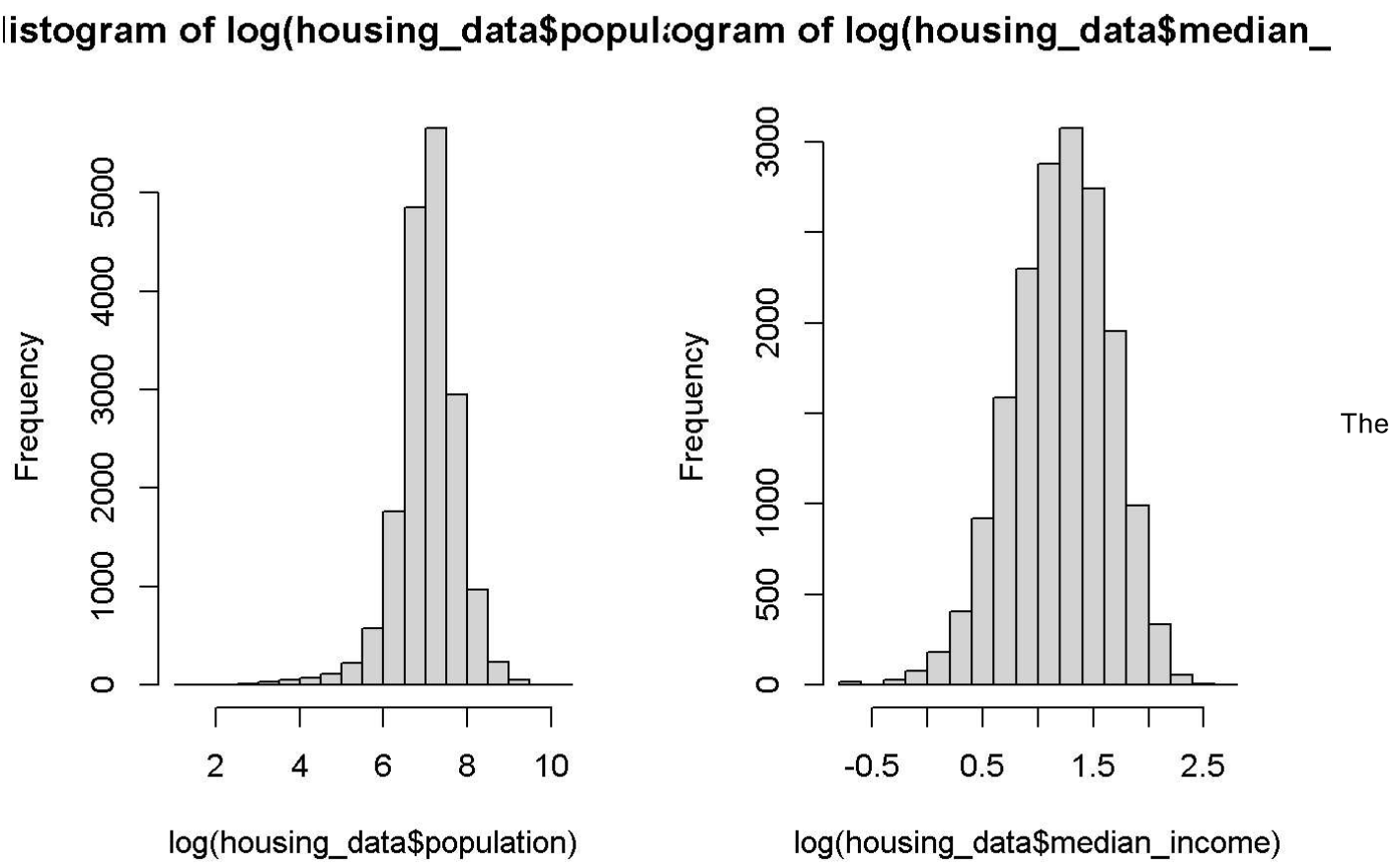
```
## [1] "Best lambda for median income: 0.3"
```

The Box-Cox analysis shows that the λ value is 0.2 for households, total rooms, and population and 0.3 for median income which is close enough to zero to indicate that a logarithmic transformation is practical.

```
housing_data$log.total_rooms = log (housing_data$total_rooms)
housing_data$log.households = log (housing_data$households)
housing_data$log.population = log(housing_data$population)
housing_data$log.median_income = log(housing_data$median_income)

par (mfrow=c(1,2))
hist (log (housing_data$total_rooms))
hist (log (housing_data$households))
```



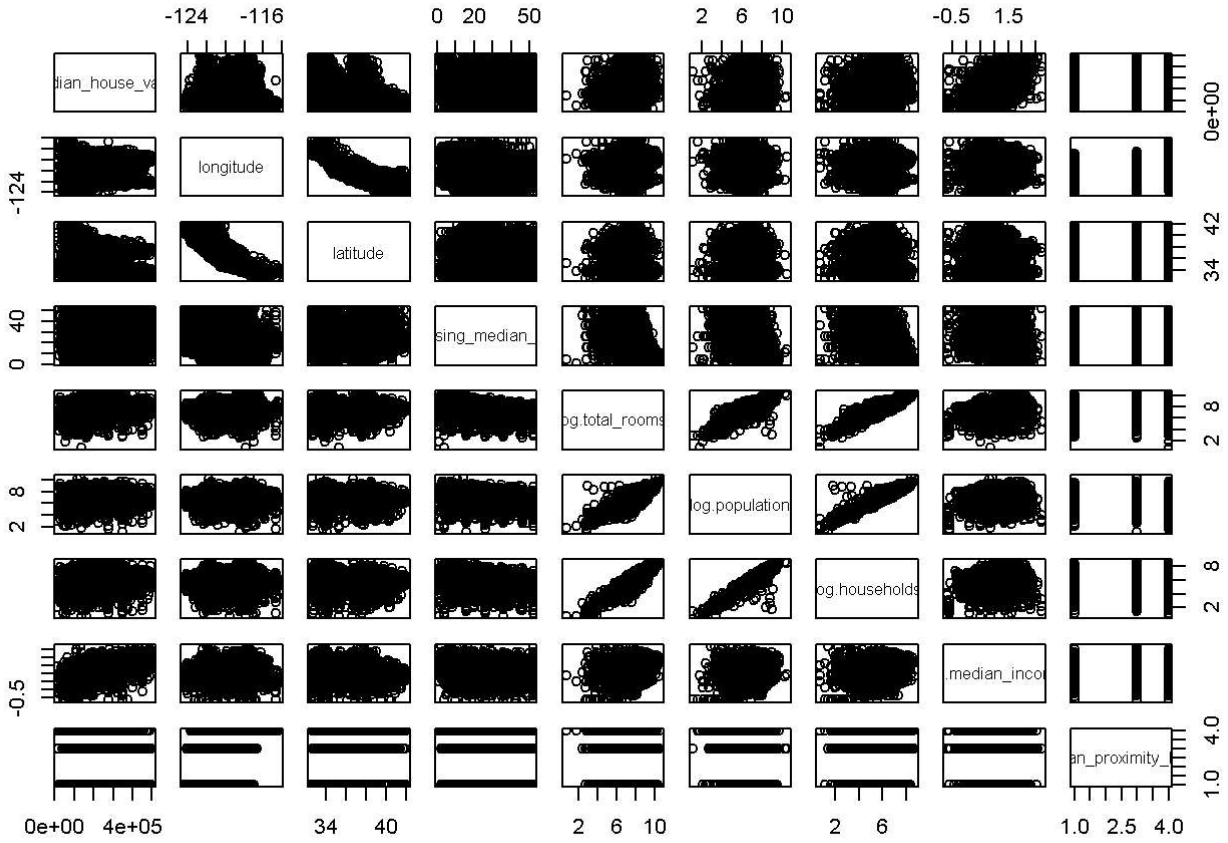

```
hist(log(housing_data$population))
hist(log(housing_data$median_income))
```



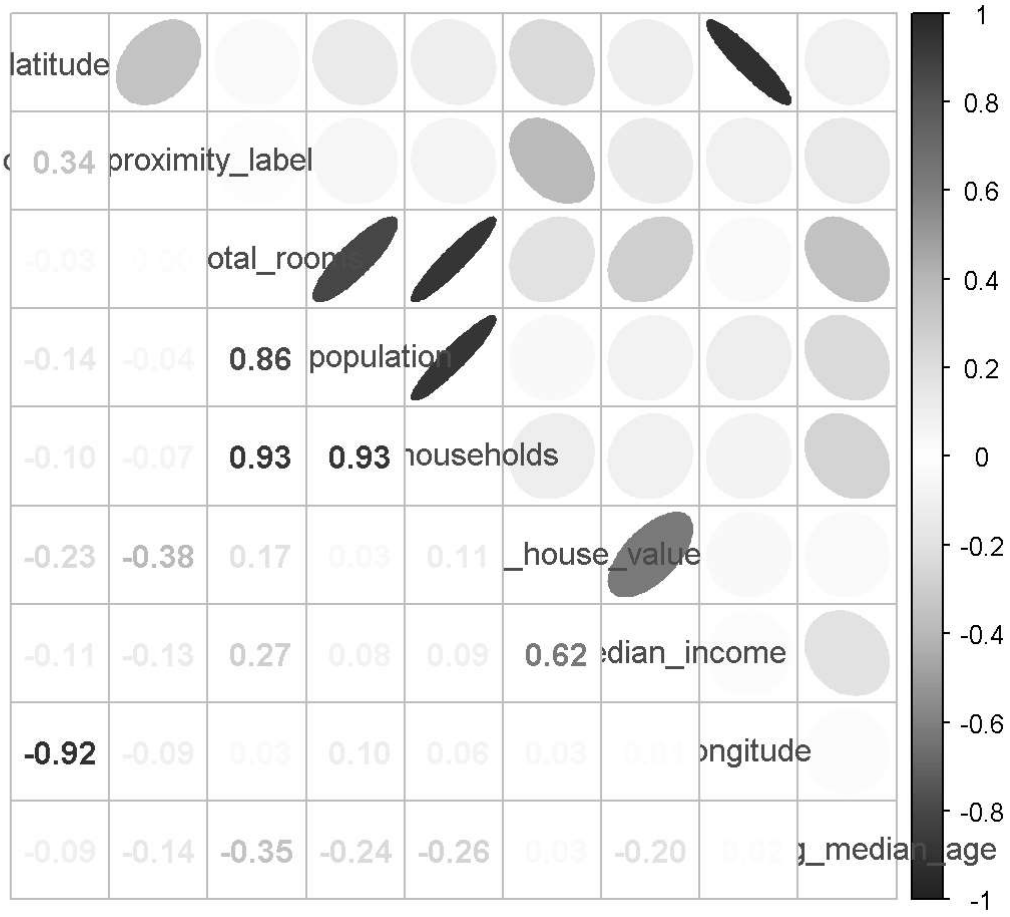
The distributions of all log transformations are much closer to symmetric compared to their un-logged versions.

We are going to examine the pairwise correlations.

```
housing.corr = housing_data [, c("median_house_value", "longitude", "latitude",
                                 "housing_median_age", "log.total_rooms", "log.population","log.ho
useholds", "log.median_income", "ocean_proximity_label")]
plot (housing.corr)
```




```
cormat = cor (housing.corr)
## corrplot 0.92 loaded
corrplot.mixed (cormat, lower='number', upper='ellipse',
                order = 'hclust')
```



Model 1

The first model will include all individual predictors.

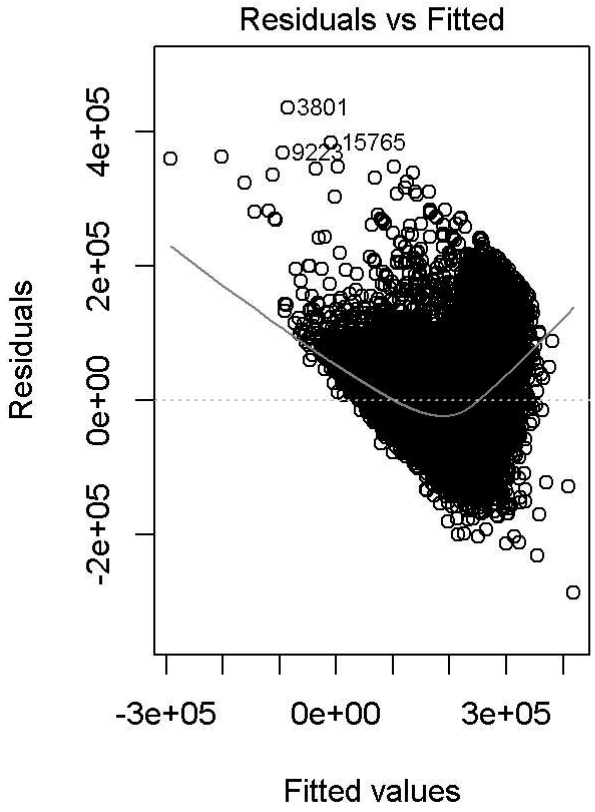
```
### First-order model with all predictors
fit1 = lm (median_house_value ~ longitude + latitude + housing_median_age + log.total_rooms + log.
population +
             log.households + log.median_income + ocean_proximity_label, data=housing_data)
summary (fit1)
```

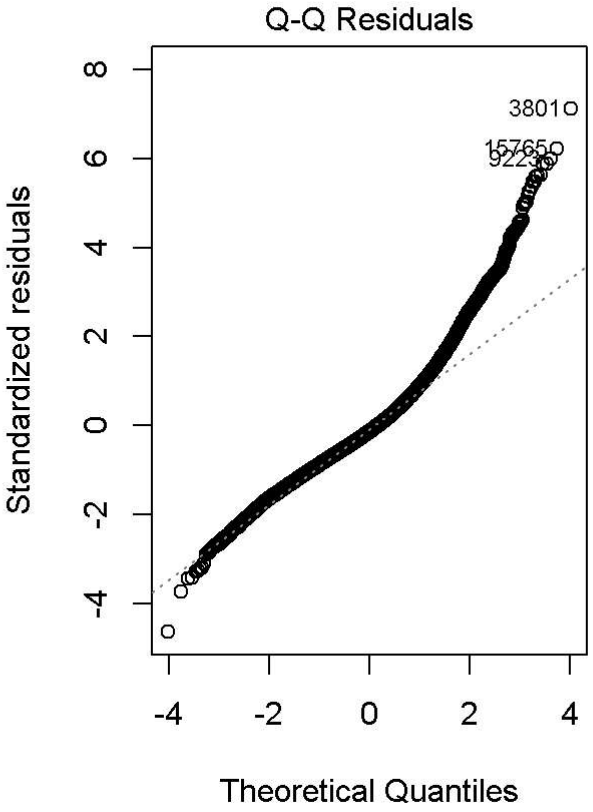
```
##
## Call:
## lm(formula = median_house_value ~ longitude + latitude + housing_median_age +
##     log.total_rooms + log.population + log.households + log.median_income +
##     ocean_proximity_label, data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -286803  -39961   -8486   30484  436232
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.823e+06  7.625e+04 -37.027   <2e-16 ***
## longitude             -3.540e+04  8.563e+02 -41.338   <2e-16 ***
## latitude              -3.600e+04  8.203e+02 -43.886   <2e-16 ***
## housing_median_age     5.913e+02  4.319e+01  13.692   <2e-16 ***
## log.total_rooms        3.912e+03  2.317e+03   1.688   0.0914 .
## log.population        -7.432e+04  1.792e+03 -41.469   <2e-16 ***
## log.households         7.126e+04  2.880e+03  24.748   <2e-16 ***
## log.median_income      1.159e+05  1.425e+03  81.298   <2e-16 ***
## ocean_proximity_label -8.014e+03  6.466e+02 -12.393   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61930 on 17546 degrees of freedom
## Multiple R-squared:  0.5765, Adjusted R-squared:  0.5763
## F-statistic:  2985 on 8 and 17546 DF,  p-value: < 2.2e-16
```

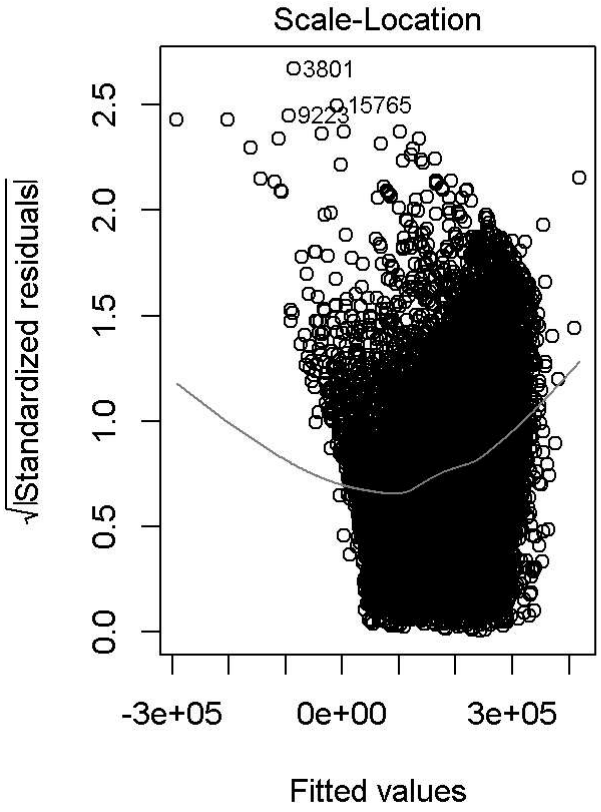
Using all predictor variables, the model explains 57.63% of the variation in sales price (adjusted R2). The residual standard error is $61,930 (sales price). Based on the P value the most significant predictors are longitude, latitude, housing median age, population, households,median income, and ocean proximity

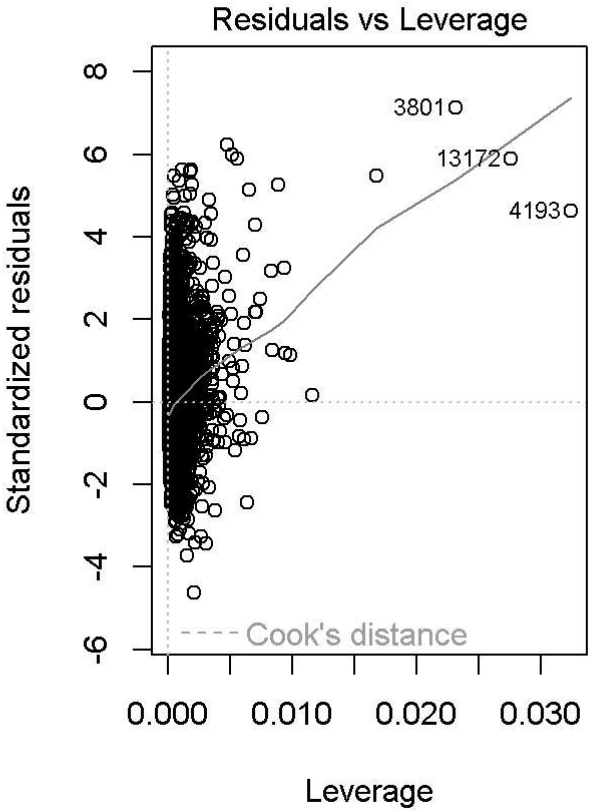
###Residual Analysis - First-order model

```
par (mfrow=c(1,2))
plot (fit1)
```


Residuals vs Fitted
Residuals
Fitted values
3801
15765
9223
-3e+05
0e+00
3e+05
-2e+05
0e+00
2e+05
4e+05


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
3801
15765
-4
-2
0
2
4
6
8


Scale-Location
√|Standardized residuals|
Fitted values
3801
15765
9223
-3e+05
0e+00
3e+05
0.0
0.5
1.0
1.5
2.0
2.5


Residuals vs Leverage
Standardized residuals
Leverage
3801
13172
4193
Cook's distance
0.000
0.010
0.020
0.030
-6
-4
-2
0
2
4
6
8

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

Residual analysis of the first-order model indicates some concerns:

1. The residuals vs fitted plot has curvature

2.The Normal Q-Q plot shows that the residuals do not follow a normal distribution. The top tail is more spread out then we would expect from a normal distribution.

3. The scale-location plot is curved

Residual analysis of the first-order model indicates some positive aspects:

4. In the residuals vs leverage plot there are no points outside or on the cook's distance line.

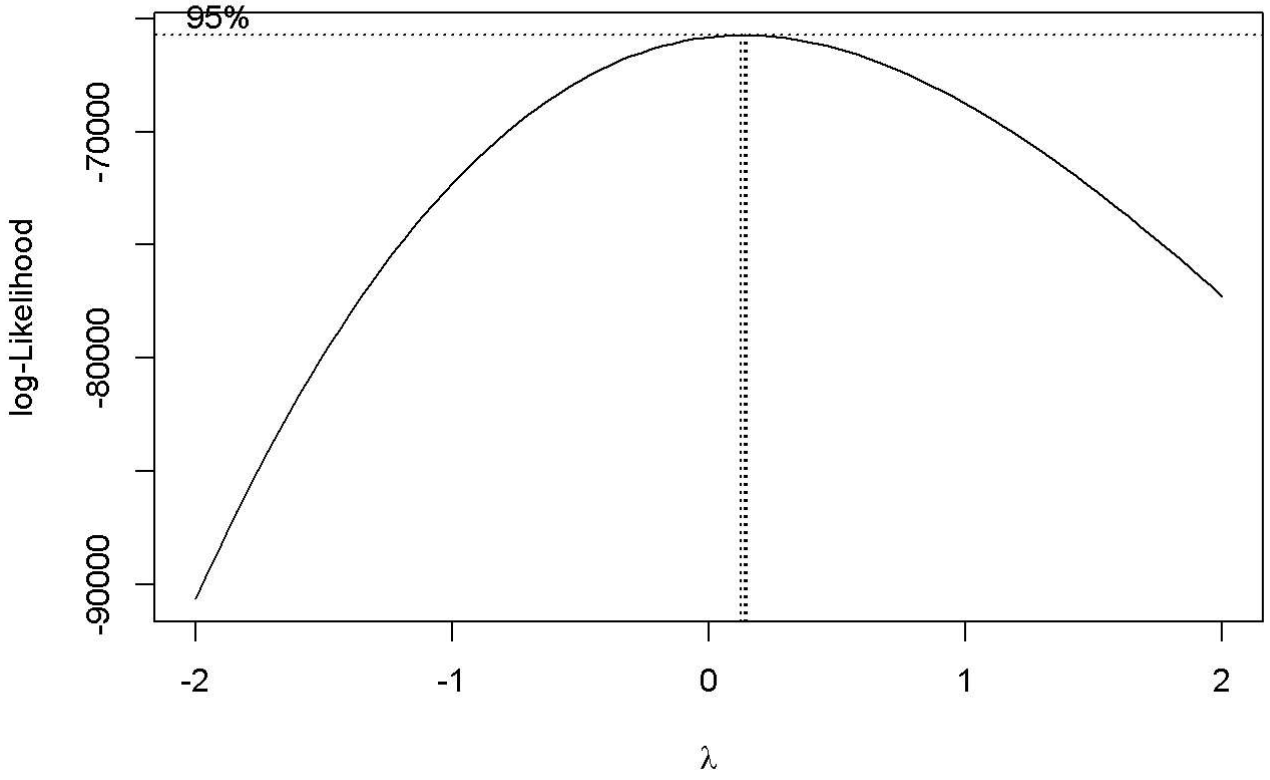
After viewing the highest cooks distances we have determined since non are above 0.5 no data points should be removed

```
 #calculate residuals and cooks distance

influence_df <-(data.frame(resid = residuals(fit1), cooks = cooks.distance(fit1)))

# Sort based on Cook's distance in descending order
sorted_influence_df <- influence_df[order(-influence_df$cooks),]

head(sorted_influence_df)
```

```
##          resid      cooks
## 3801  436231.6 0.13369306
## 13172 359682.8 0.10921798
## 4193  281809.8 0.07965299
## 11953 336596.9 0.05709618
## 13622 324603.7 0.02750807
## 6939  363775.6 0.02166068
```

Box-cox analysis on the first model indicates a log-transformation of the outcome variable is desired since the highest lambda value is closest to 0.

```
# Do Box Cox analysis
MASS::boxcox (fit1)
```



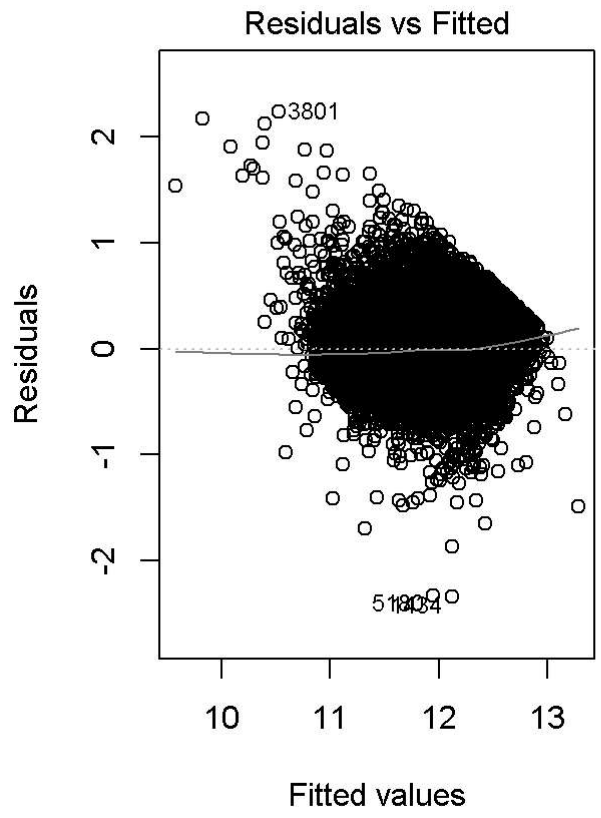
Model 2 - log scale median_house_value vs. predictor variables

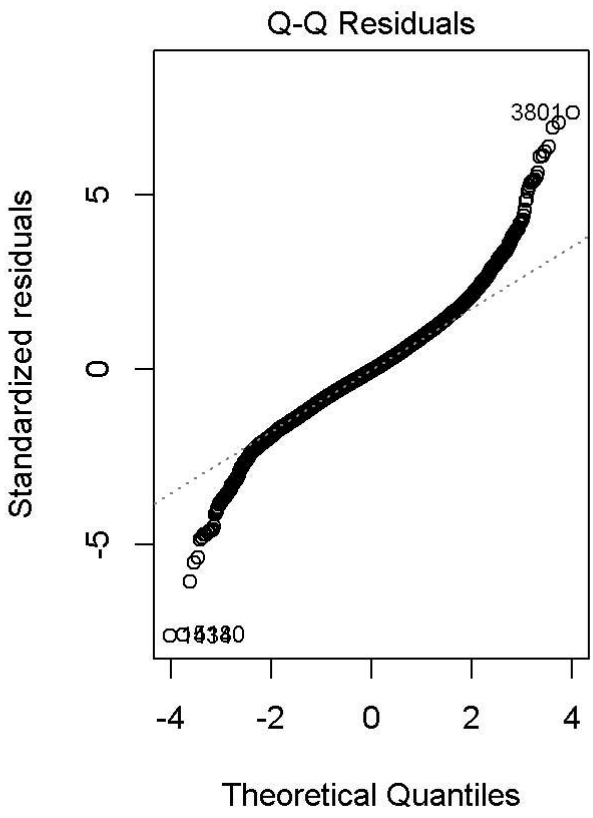
```
fit2 = lm (log(median_house_value) ~ longitude + latitude + housing_median_age + log.total_rooms +
log.population +
             log.households + log.median_income + as.factor (ocean_proximity_label),
              data=housing_data)
model2_summary = summary (fit2)

model2_summary
```

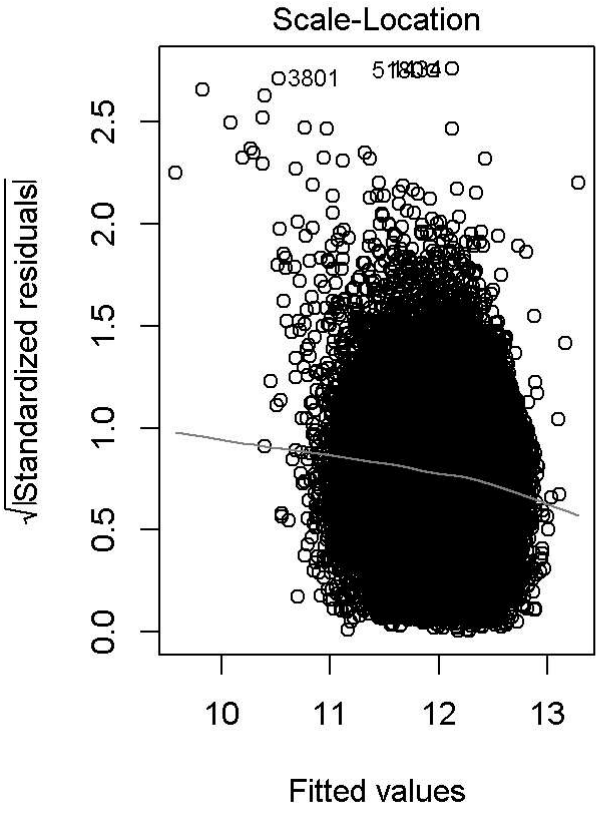
```
##
## Call:
## lm(formula = log(median_house_value) ~ longitude + latitude +
##     housing_median_age + log.total_rooms + log.population + log.households +
##     log.median_income + as.factor(ocean_proximity_label), data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34643 -0.18718 -0.01224  0.17853  2.23902
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -0.2638149  0.4091660  -0.645   0.5191
## longitude                       -0.1408634  0.0047015 -29.961  < 2e-16 ***
## latitude                        -0.1399435  0.0046568 -30.052  < 2e-16 ***
## housing_median_age               0.0011800  0.0002158   5.468 4.60e-08 ***
## log.total_rooms                  0.0197217  0.0116344   1.695   0.0901 .
## log.population                  -0.3717623  0.0089145 -41.703  < 2e-16 ***
## log.households                   0.3647769  0.0143980  25.335  < 2e-16 ***
## log.median_income                0.6294471  0.0072178  87.207  < 2e-16 ***
## as.factor(ocean_proximity_label)3  0.0488084  0.0073523   6.639 3.26e-11 ***
## as.factor(ocean_proximity_label)4 -0.2735894  0.0103153 -26.523  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.308 on 17545 degrees of freedom
## Multiple R-squared:  0.6654, Adjusted R-squared:  0.6653
## F-statistic:  3877 on 9 and 17545 DF,  p-value: < 2.2e-16
```

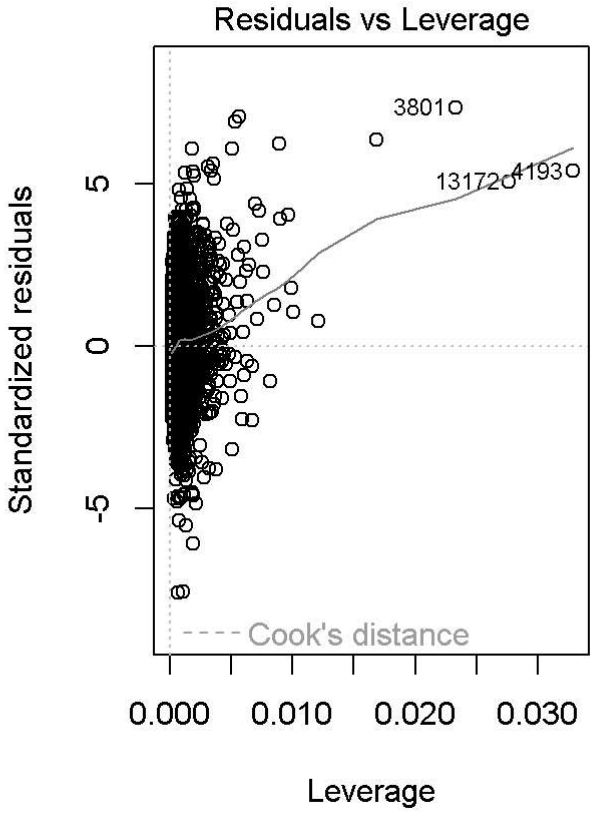
The model explains 66.54 percent variation in log median house value. The residual standard error is 0.31 log dollars.

```
par (mfrow=c(1,2))
plot (fit2)
```


Residuals vs Fitted
Residuals
Fitted values
3801
10
11
12
13
-2
-1
0
1
2


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
3801
-4
-2
0
2
4
-5
0
5


Scale-Location
√|Standardized residuals|
Fitted values
3801
10
11
12
13
0.0
0.5
1.0
1.5
2.0
2.5


Residuals vs Leverage
Standardized residuals
Leverage
3801
13172
41930
Cook's distance
0.000
0.010
0.020
0.030
-5
0
5

Although there are still some skewness in both tails it is better than the first model which has more skewness. The residuals in the residuals vs fitted plot seems to be spread out evenly both above and below zero and the plot has minimal curvature. The Scale location is also more straight than in the model 1. Overall everything looks better than the model 1.

Let's look into the parameters of the model 2 in more details.

```
confint(fit2)
```

```
##                                         2.5 %       97.5 %
## (Intercept)                       -1.0658208579  0.538191031
## longitude                         -0.1500788570 -0.131647887
## latitude                          -0.1490712342 -0.130815729
## housing_median_age                 0.0007570506  0.001602966
## log.total_rooms                   -0.0030827809  0.042526209
## log.population                    -0.3892356323 -0.354289036
## log.households                     0.3365555076  0.392998371
## log.median_income                  0.6152994675  0.643594776
## as.factor(ocean_proximity_label)3  0.0343971192  0.063219616
## as.factor(ocean_proximity_label)4 -0.2938084542 -0.253370335
```

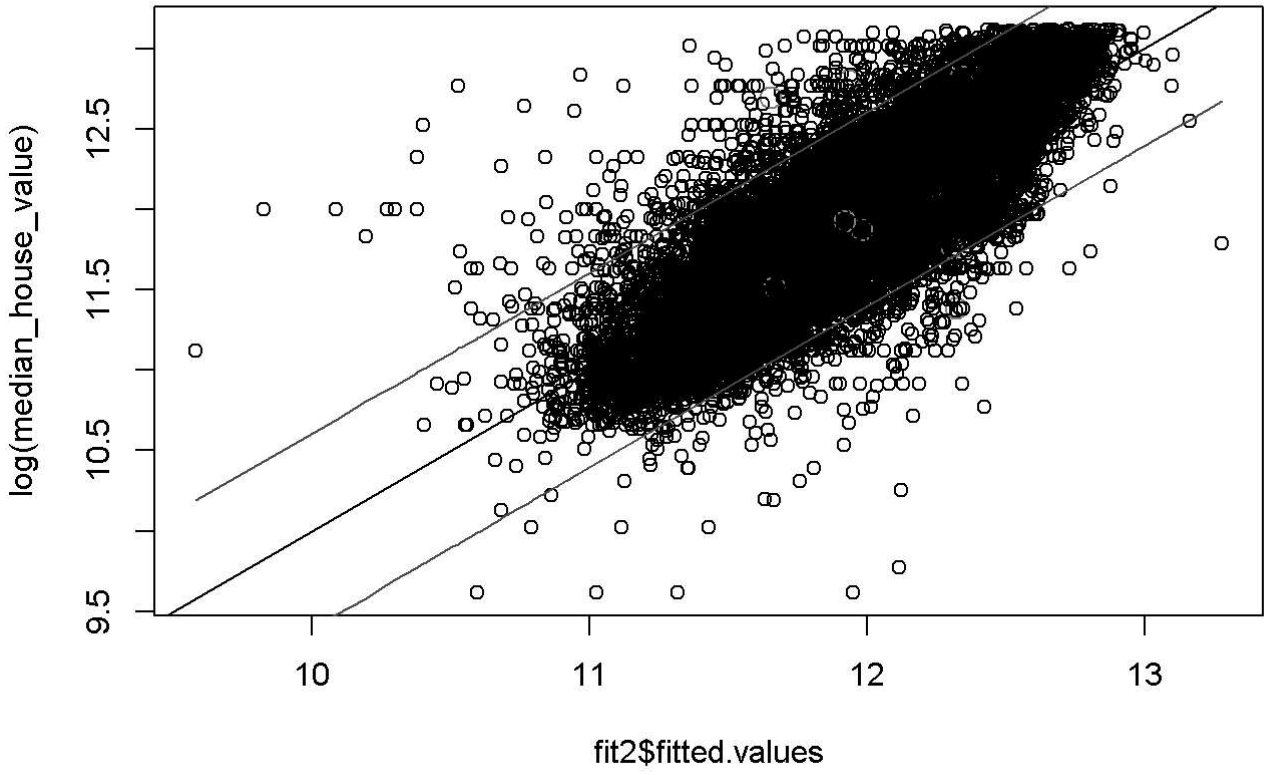
Qualitative interpretations of estimates, each one accounting for the other predictors in the model (i.e., preface each statement with, "All other things being the same, …"):

- House located at the lower longitude tend to cost more on average.
- House with lower latitude value tend to cost more on average.
- Old homes tend to have higher prices on average. This sounds counter intuitive. But the increase in median house age doesn't increase the house price by much.
- Homes with higher total rooms tend to cost more on average although the significance is near boarder line (0.0425).
- Homes located in more dense places are tend to cost less than in places with less population on average.
- House located inland are tend to be cheaper on average compared to houses that are less than an hour away from the ocean.
- Houses located in higher number of household blocks tend to cost more on average.
- Having higher income is associated with higher median house value on average.

```
plot (log(median_house_value) ~ fit2$fitted.values, data=housing_data,
      col=ifelse (1:522==104,2,1), cex=ifelse(1:522==104,1.5,1))
abline (0, 1)

# Add prediction limits to previous plot

fit2.pred = as.data.frame (predict (fit2, interval="prediction"))
```

```
## Warning in predict.lm(fit2, interval = "prediction"): predictions on current data refer to _fut
ure_ responses
```

```
## Warning in predict.lm(fit2, interval = "prediction"): predictions on current data refer to _fut
ure_ responses
order2 = order (fit2.pred$fit)
lines (fit2.pred$fit [order2], fit2.pred$lwr[order2], col='red')
lines (fit2.pred$fit [order2], fit2.pred$upr[order2], col='red')
```

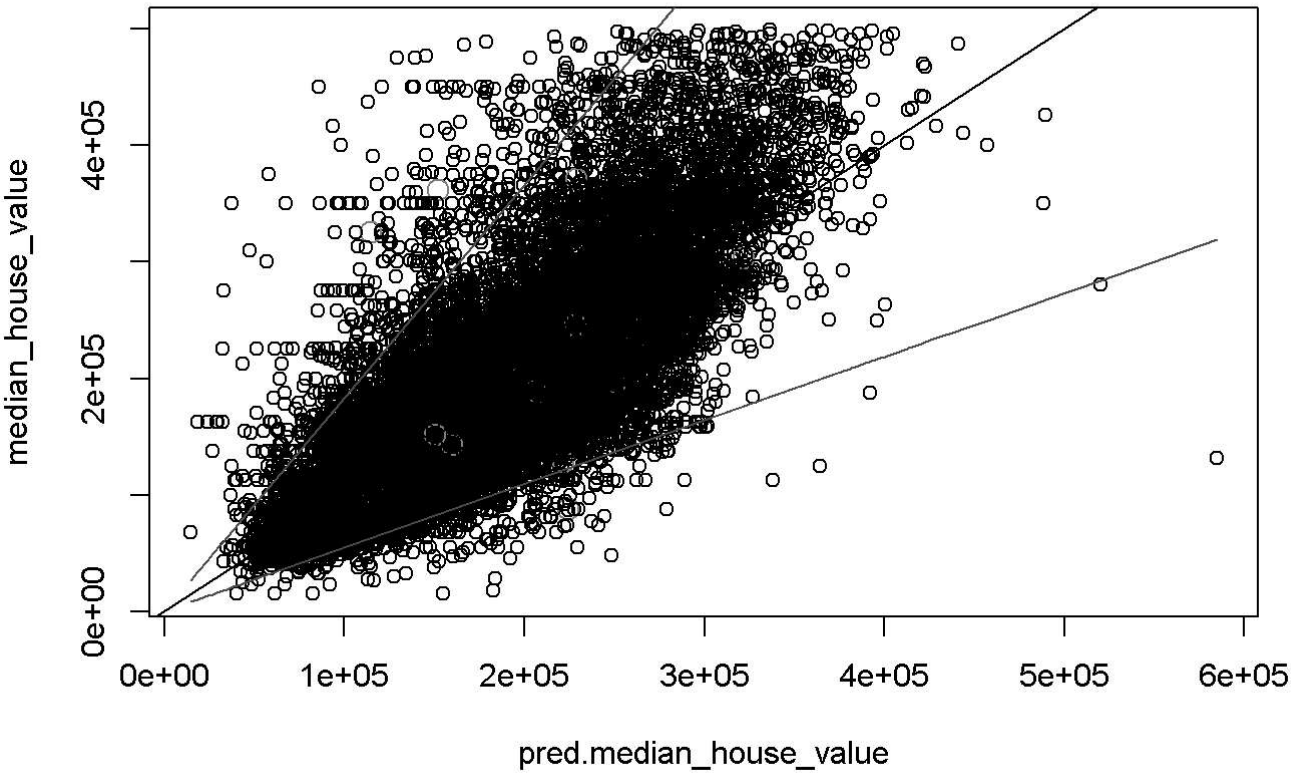


Most of the observed and predicted values of log median house value are very close. Some observed values or much high or low than the predicted values but it is consistent with the r value of 66.5%.

```
# Back-transform fitted sales prices values and plot
fit2.pred = as.data.frame (predict (fit2, interval="prediction"))
```

```
## Warning in predict.lm(fit2, interval = "prediction"): predictions on current data refer to _fut
ure_ responses
```

```
order2 = order (fit2.pred$fit)
pred.median_house_value = exp (fit2$fitted.values)
plot (median_house_value ~ pred.median_house_value, data=housing_data,
      col=ifelse (1:522==104,2,1), cex=ifelse(1:522==104,1.5,1))
abline (0, 1)

# Add prediction limits to previous plot

lines (exp (fit2.pred$fit [order2]), exp (fit2.pred$lwr[order2]), col='red')
lines (exp (fit2.pred$fit [order2]), exp (fit2.pred$upr[order2]), col='red')
```



The plot above shows the observed median house value vs. fitted median house value price, back-transformed from Model 2 to the original scale (dollars). On original sales price scale, predicted median house value are more variable (i.e., less precise) than the observed home values.

Stepwise regression

ext, we apply step-wise regression to the full log(sales price) model using the AIC criterion.

```
fit2aic = step (fit2, direction='both')
```

```
## Start:  AIC=-41341.16
## log(median_house_value) ~ longitude + latitude + housing_median_age +
##     log.total_rooms + log.population + log.households + log.median_income +
##     as.factor(ocean_proximity_label)
##
##                                   Df Sum of Sq    RSS    AIC
## <none>                                         1664.0 -41341
## - log.total_rooms                  1      0.27 1664.3 -41340
## - housing_median_age               1      2.84 1666.9 -41313
## - log.households                   1     60.88 1724.9 -40712
## - longitude                        1     85.14 1749.2 -40467
## - latitude                         1     85.65 1749.7 -40462
## - as.factor(ocean_proximity_label)  2    158.86 1822.9 -39744
## - log.population                   1    164.95 1829.0 -39684
## - log.median_income                1    721.30 2385.3 -35022
```

```
summary(fit2aic)
```

```
##
## Call:
## lm(formula = log(median_house_value) ~ longitude + latitude +
##     housing_median_age + log.total_rooms + log.population + log.households +
##     log.median_income + as.factor(ocean_proximity_label), data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34643 -0.18718 -0.01224  0.17853  2.23902
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -0.2638149  0.4091660  -0.645   0.5191
## longitude                         -0.1408634  0.0047015 -29.961  < 2e-16 ***
## latitude                          -0.1399435  0.0046568 -30.052  < 2e-16 ***
## housing_median_age                 0.0011800  0.0002158   5.468 4.60e-08 ***
## log.total_rooms                    0.0197217  0.0116344   1.695   0.0901 .
## log.population                    -0.3717623  0.0089145 -41.703  < 2e-16 ***
## log.households                     0.3647769  0.0143980  25.335  < 2e-16 ***
## log.median_income                  0.6294471  0.0072178  87.207  < 2e-16 ***
## as.factor(ocean_proximity_label)3  0.0488084  0.0073523   6.639 3.26e-11 ***
## as.factor(ocean_proximity_label)4 -0.2735894  0.0103153 -26.523  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.308 on 17545 degrees of freedom
## Multiple R-squared:  0.6654, Adjusted R-squared:  0.6653
## F-statistic:  3877 on 9 and 17545 DF,  p-value: < 2.2e-16
```

Nothing was removed.we will apply step-wise regression using the SBC criterion, which is more restrictive about retaining predictor variables.

```
n = dim (housing_data)[1]
fit2sbc = step (fit2, direction='both', k=log(n))
```

```
## Start:  AIC=-41263.43
## log(median_house_value) ~ longitude + latitude + housing_median_age +
##     log.total_rooms + log.population + log.households + log.median_income +
##     as.factor(ocean_proximity_label)
##
##                                   Df Sum of Sq    RSS    AIC
## - log.total_rooms                  1      0.27 1664.3 -41270
## <none>                                       1664.0 -41263
## - housing_median_age               1      2.84 1666.9 -41243
## - log.households                   1     60.88 1724.9 -40642
## - longitude                        1     85.14 1749.2 -40397
## - latitude                         1     85.65 1749.7 -40392
## - as.factor(ocean_proximity_label)  2    158.86 1822.9 -39682
## - log.population                   1    164.95 1829.0 -39614
## - log.median_income                1    721.30 2385.3 -34952
##
## Step:  AIC=-41270.32
## log(median_house_value) ~ longitude + latitude + housing_median_age +
##     log.population + log.households + log.median_income + as.factor(ocean_proximity_label)
##
##                                   Df Sum of Sq    RSS    AIC
## <none>                                       1664.3 -41270
## + log.total_rooms                  1      0.27 1664.0 -41263
## - housing_median_age               1      2.68 1667.0 -41252
## - longitude                        1     87.20 1751.5 -40384
## - latitude                         1     88.25 1752.6 -40373
## - as.factor(ocean_proximity_label)  2    159.65 1824.0 -39682
## - log.population                   1    165.48 1829.8 -39616
## - log.households                   1    173.13 1837.4 -39543
## - log.median_income                1   1188.66 2853.0 -31819
```

SBC criterion removed log of total rooms from the model.

```
summary (fit2sbc)
```

```
##
## Call:
## lm(formula = log(median_house_value) ~ longitude + latitude +
##     housing_median_age + log.population + log.households + log.median_income +
##     as.factor(ocean_proximity_label), data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.35052 -0.18720 -0.01238  0.17864  2.24990
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -0.0902698  0.3961719  -0.228     0.82
## longitude                       -0.1391252  0.0045886 -30.320  < 2e-16 ***
## latitude                        -0.1380933  0.0045273 -30.502  < 2e-16 ***
## housing_median_age               0.0011396  0.0002145   5.313 1.09e-07 ***
## log.population                  -0.3722026  0.0089112 -41.768  < 2e-16 ***
## log.households                   0.3838487  0.0089848  42.722  < 2e-16 ***
## log.median_income                0.6369753  0.0056901 111.944  < 2e-16 ***
## as.factor(ocean_proximity_label)3  0.0476967  0.0073234   6.513 7.57e-11 ***
## as.factor(ocean_proximity_label)4 -0.2731371  0.0103124 -26.486  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.308 on 17546 degrees of freedom
## Multiple R-squared:  0.6654, Adjusted R-squared:  0.6652
## F-statistic:  4361 on 8 and 17546 DF,  p-value: < 2.2e-16
```

Both AIC and BIC stepwise has veru similar r squared of 66.5% and residual standard error of 0.308. We will proceed with the BIC model.

Model 2 with Interactions

Next, we well add all possible two-way interaction effects to the BIC step-wise regression model obtained above.

```
mycenter = function (x) x - mean (x)

housing_data$latitiude.c = mycenter (housing_data$latitude)
housing_data$longitude.c = mycenter (housing_data$longitude)
housing_data$housing_median_age.c = mycenter (housing_data$housing_median_age)
housing_data$log.population.c = mycenter (housing_data$log.population)
housing_data$log.households.c = mycenter (housing_data$log.households)
housing_data$log.median_income.c = mycenter (housing_data$log.median_income)


fit2sbc.int = lm(log(median_house_value) ~
                    (latitiude.c + longitude.c + housing_median_age.c + log.population.c + log.hou
seholds.c + log.median_income.c +  as.factor(ocean_proximity ))^2, data=housing_data)



summary(fit2sbc.int)
```

```
##
## Call:
## lm(formula = log(median_house_value) ~ (latitiude.c + longitude.c +
##     housing_median_age.c + log.population.c + log.households.c +
##     log.median_income.c + as.factor(ocean_proximity))^2, data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4839 -0.1735 -0.0099  0.1708  2.7553
##
## Coefficients:
##                                                      Estimate Std. Error
## (Intercept)                                         1.217e+01  9.096e-03
## latitiude.c                                        -1.147e-01  1.368e-02
## longitude.c                                        -1.119e-01  1.226e-02
## housing_median_age.c                                3.446e-05  3.647e-04
## log.population.c                                   -3.994e-01  1.458e-02
## log.households.c                                    4.319e-01  1.475e-02
## log.median_income.c                                 5.138e-01  9.595e-03
## as.factor(ocean_proximity)INLAND                   -3.440e-01  9.657e-03
## as.factor(ocean_proximity)NEAR OCEAN               -1.012e-01  1.475e-02
## latitiude.c:longitude.c                             1.043e-02  1.006e-03
## latitiude.c:housing_median_age.c                   -5.974e-03  4.087e-04
## latitiude.c:log.population.c                        1.823e-02  1.371e-02
## latitiude.c:log.households.c                       -3.117e-02  1.403e-02
## latitiude.c:log.median_income.c                    -1.224e-01  1.149e-02
## latitiude.c:as.factor(ocean_proximity)INLAND       -2.404e-02  1.454e-02
## latitiude.c:as.factor(ocean_proximity)NEAR OCEAN   -9.118e-02  1.840e-02
## longitude.c:housing_median_age.c                   -5.363e-03  4.164e-04
## longitude.c:log.population.c                        6.853e-03  1.392e-02
## longitude.c:log.households.c                       -1.426e-02  1.427e-02
## longitude.c:log.median_income.c                    -1.343e-01  1.144e-02
## longitude.c:as.factor(ocean_proximity)INLAND       -4.698e-02  1.396e-02
## longitude.c:as.factor(ocean_proximity)NEAR OCEAN   -1.157e-01  1.771e-02
## housing_median_age.c:log.population.c              -6.002e-03  6.941e-04
## housing_median_age.c:log.households.c               7.138e-03  6.902e-04
## housing_median_age.c:log.median_income.c           -2.734e-04  4.478e-04
## housing_median_age.c:as.factor(ocean_proximity)INLAND      2.960e-03  6.769e-04
## housing_median_age.c:as.factor(ocean_proximity)NEAR OCEAN  5.460e-04  6.356e-04
## log.population.c:log.households.c                   1.307e-02  2.007e-03
## log.population.c:log.median_income.c                7.996e-02  1.733e-02
## log.population.c:as.factor(ocean_proximity)INLAND   1.193e-01  2.754e-02
## log.population.c:as.factor(ocean_proximity)NEAR OCEAN      4.739e-02  2.499e-02
## log.households.c:log.median_income.c               -8.351e-03  1.775e-02
## log.households.c:as.factor(ocean_proximity)INLAND  -1.239e-01  2.788e-02
## log.households.c:as.factor(ocean_proximity)NEAR OCEAN      -3.313e-02  2.527e-02
## log.median_income.c:as.factor(ocean_proximity)INLAND       2.986e-01  1.857e-02
## log.median_income.c:as.factor(ocean_proximity)NEAR OCEAN   5.541e-02  1.721e-02
##                                                     t value Pr(>|t|)
## (Intercept)                                        1337.668  < 2e-16 ***
## latitiude.c                                          -8.388  < 2e-16 ***
## longitude.c                                          -9.124  < 2e-16 ***
```

```
## housing_median_age.c                                        0.094 0.924736
## log.population.c                                          -27.403  < 2e-16 ***
## log.households.c                                           29.287  < 2e-16 ***
## log.median_income.c                                        53.552  < 2e-16 ***
## as.factor(ocean_proximity)INLAND                          -35.622  < 2e-16 ***
## as.factor(ocean_proximity)NEAR OCEAN                       -6.861 7.05e-12 ***
## latitiude.c:longitude.c                                    10.365  < 2e-16 ***
## latitiude.c:housing_median_age.c                          -14.618  < 2e-16 ***
## latitiude.c:log.population.c                                1.330 0.183624
## latitiude.c:log.households.c                               -2.221 0.026370 *
## latitiude.c:log.median_income.c                           -10.657  < 2e-16 ***
## latitiude.c:as.factor(ocean_proximity)INLAND               -1.653 0.098253 .
## latitiude.c:as.factor(ocean_proximity)NEAR OCEAN           -4.956 7.25e-07 ***
## longitude.c:housing_median_age.c                          -12.881  < 2e-16 ***
## longitude.c:log.population.c                                0.492 0.622578
## longitude.c:log.households.c                               -0.999 0.317887
## longitude.c:log.median_income.c                           -11.732  < 2e-16 ***
## longitude.c:as.factor(ocean_proximity)INLAND               -3.366 0.000764 ***
## longitude.c:as.factor(ocean_proximity)NEAR OCEAN           -6.534 6.60e-11 ***
## housing_median_age.c:log.population.c                      -8.647  < 2e-16 ***
## housing_median_age.c:log.households.c                      10.342  < 2e-16 ***
## housing_median_age.c:log.median_income.c                   -0.611 0.541447
## housing_median_age.c:as.factor(ocean_proximity)INLAND       4.373 1.23e-05 ***
## housing_median_age.c:as.factor(ocean_proximity)NEAR OCEAN   0.859 0.390397
## log.population.c:log.households.c                           6.512 7.61e-11 ***
## log.population.c:log.median_income.c                        4.614 3.98e-06 ***
## log.population.c:as.factor(ocean_proximity)INLAND           4.331 1.49e-05 ***
## log.population.c:as.factor(ocean_proximity)NEAR OCEAN       1.896 0.057986 .
## log.households.c:log.median_income.c                       -0.471 0.637948
## log.households.c:as.factor(ocean_proximity)INLAND          -4.445 8.86e-06 ***
## log.households.c:as.factor(ocean_proximity)NEAR OCEAN      -1.311 0.189846
## log.median_income.c:as.factor(ocean_proximity)INLAND       16.078  < 2e-16 ***
## log.median_income.c:as.factor(ocean_proximity)NEAR OCEAN    3.219 0.001289 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.295 on 17519 degrees of freedom
## Multiple R-squared:  0.6935, Adjusted R-squared:  0.6929
## F-statistic:  1133 on 35 and 17519 DF,  p-value: < 2.2e-16
```

This model has a higher adjusted r squared of 69.29% with a lower residual standard error of 0.295.

Interaction model step-wise regression

A step-wise regression will be used to get a more streamlined model.

```
nrows = fit2sbc.int$rank + fit2sbc.int$df.residual
fit2.int.sbc = step (fit2sbc.int, direction='both', k=log(nrows))
```

```
summary(fit2.int.sbc)
```

```
##
## Call:
## lm(formula = log(median_house_value) ~ latitiude.c + longitude.c +
##     housing_median_age.c + log.population.c + log.households.c +
##     log.median_income.c + as.factor(ocean_proximity) + latitiude.c:longitude.c +
##     latitiude.c:housing_median_age.c + latitiude.c:log.households.c +
##     latitiude.c:log.median_income.c + latitiude.c:as.factor(ocean_proximity) +
##     longitude.c:housing_median_age.c + longitude.c:log.median_income.c +
##     longitude.c:as.factor(ocean_proximity) + housing_median_age.c:log.population.c +
##     housing_median_age.c:log.households.c + log.population.c:log.households.c +
##     log.population.c:log.median_income.c + log.population.c:as.factor(ocean_proximity) +
##     log.households.c:as.factor(ocean_proximity) + log.median_income.c:as.factor(ocean_proximit
y),
##     data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48006 -0.17452 -0.00988  0.17105  2.76558
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                      12.1674282  0.0090386
## latitiude.c                                      -0.1122906  0.0136478
## longitude.c                                      -0.1097186  0.0122422
## housing_median_age.c                              0.0012794  0.0002206
## log.population.c                                 -0.4156247  0.0127169
## log.households.c                                  0.4545539  0.0128503
## log.median_income.c                               0.5246487  0.0093149
## as.factor(ocean_proximity)INLAND                 -0.3490958  0.0095162
## as.factor(ocean_proximity)NEAR OCEAN             -0.0982425  0.0146336
## latitiude.c:longitude.c                           0.0105642  0.0009941
## latitiude.c:housing_median_age.c                 -0.0046759  0.0002611
## latitiude.c:log.households.c                     -0.0059887  0.0016664
## latitiude.c:log.median_income.c                  -0.1141712  0.0111219
## latitiude.c:as.factor(ocean_proximity)INLAND     -0.0206935  0.0144423
## latitiude.c:as.factor(ocean_proximity)NEAR OCEAN -0.0904996  0.0183133
## longitude.c:housing_median_age.c                 -0.0041658  0.0002927
## longitude.c:log.median_income.c                  -0.1276773  0.0111091
## longitude.c:as.factor(ocean_proximity)INLAND     -0.0435329  0.0138157
## longitude.c:as.factor(ocean_proximity)NEAR OCEAN -0.1140377  0.0175861
## housing_median_age.c:log.population.c            -0.0055582  0.0006799
## housing_median_age.c:log.households.c             0.0066538  0.0006773
## log.population.c:log.households.c                 0.0127066  0.0019984
## log.population.c:log.median_income.c              0.0739667  0.0058623
## log.population.c:as.factor(ocean_proximity)INLAND 0.1526014  0.0192817
## log.population.c:as.factor(ocean_proximity)NEAR OCEAN 0.0427637  0.0242241
## log.households.c:as.factor(ocean_proximity)INLAND -0.1732484  0.0195886
## log.households.c:as.factor(ocean_proximity)NEAR OCEAN -0.0308952  0.0244217
## log.median_income.c:as.factor(ocean_proximity)INLAND 0.2797732  0.0178389
## log.median_income.c:as.factor(ocean_proximity)NEAR OCEAN 0.0518295  0.0168703
##                                                    t value Pr(>|t|)
## (Intercept)                                       1346.168  < 2e-16 ***
```

```
## latitiude.c                                          -8.228  < 2e-16 ***
## longitude.c                                          -8.962  < 2e-16 ***
## housing_median_age.c                                  5.801 6.72e-09 ***
## log.population.c                                    -32.683  < 2e-16 ***
## log.households.c                                     35.373  < 2e-16 ***
## log.median_income.c                                  56.323  < 2e-16 ***
## as.factor(ocean_proximity)INLAND                    -36.684  < 2e-16 ***
## as.factor(ocean_proximity)NEAR OCEAN                 -6.713 1.96e-11 ***
## latitiude.c:longitude.c                              10.627  < 2e-16 ***
## latitiude.c:housing_median_age.c                    -17.910  < 2e-16 ***
## latitiude.c:log.households.c                         -3.594 0.000327 ***
## latitiude.c:log.median_income.c                     -10.265  < 2e-16 ***
## latitiude.c:as.factor(ocean_proximity)INLAND         -1.433 0.151920
## latitiude.c:as.factor(ocean_proximity)NEAR OCEAN     -4.942 7.81e-07 ***
## longitude.c:housing_median_age.c                    -14.231  < 2e-16 ***
## longitude.c:log.median_income.c                     -11.493  < 2e-16 ***
## longitude.c:as.factor(ocean_proximity)INLAND         -3.151 0.001630 **
## longitude.c:as.factor(ocean_proximity)NEAR OCEAN     -6.485 9.14e-11 ***
## housing_median_age.c:log.population.c                -8.175 3.16e-16 ***
## housing_median_age.c:log.households.c                 9.824  < 2e-16 ***
## log.population.c:log.households.c                     6.359 2.09e-10 ***
## log.population.c:log.median_income.c                 12.617  < 2e-16 ***
## log.population.c:as.factor(ocean_proximity)INLAND     7.914 2.63e-15 ***
## log.population.c:as.factor(ocean_proximity)NEAR OCEAN 1.765 0.077524 .
## log.households.c:as.factor(ocean_proximity)INLAND    -8.844  < 2e-16 ***
## log.households.c:as.factor(ocean_proximity)NEAR OCEAN -1.265 0.205861
## log.median_income.c:as.factor(ocean_proximity)INLAND 15.683  < 2e-16 ***
## log.median_income.c:as.factor(ocean_proximity)NEAR OCEAN  3.072 0.002128 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2952 on 17526 degrees of freedom
## Multiple R-squared:  0.693,  Adjusted R-squared:  0.6925
## F-statistic:  1413 on 28 and 17526 DF,  p-value: < 2.2e-16
```

The interaction between housing median as and ocean proximity, latitude and log population, longitude and log population, log household and median income, and housing median age and log median income was removed under the SBC criterion.

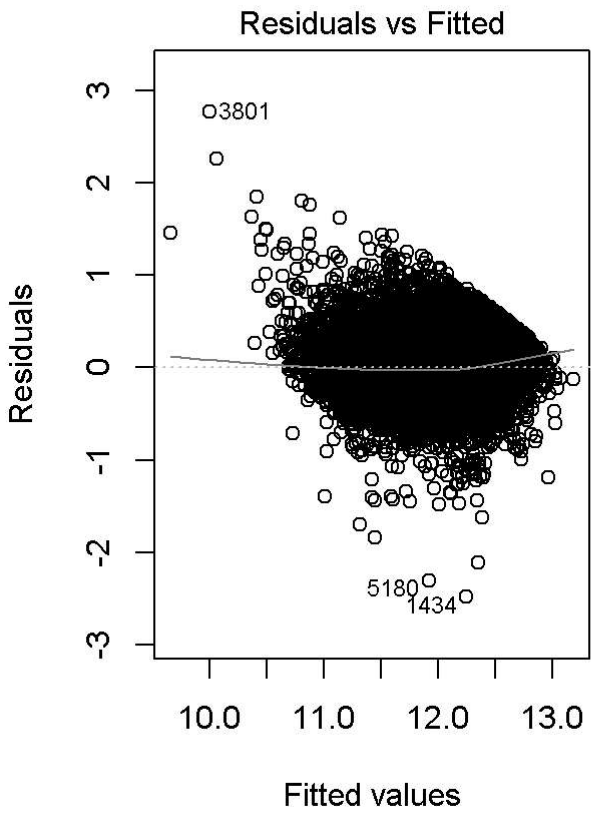
The r squared is 69.25% and RSE is 0.2952

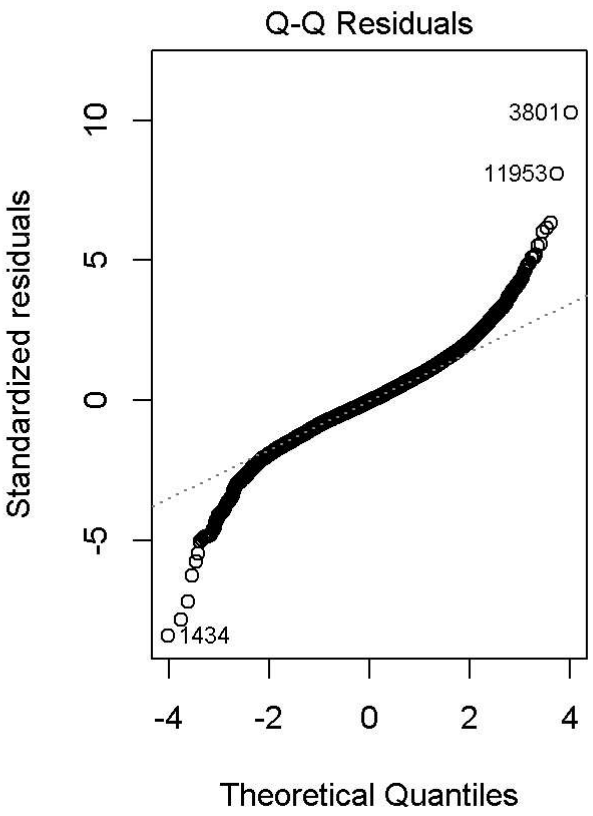
Final Model

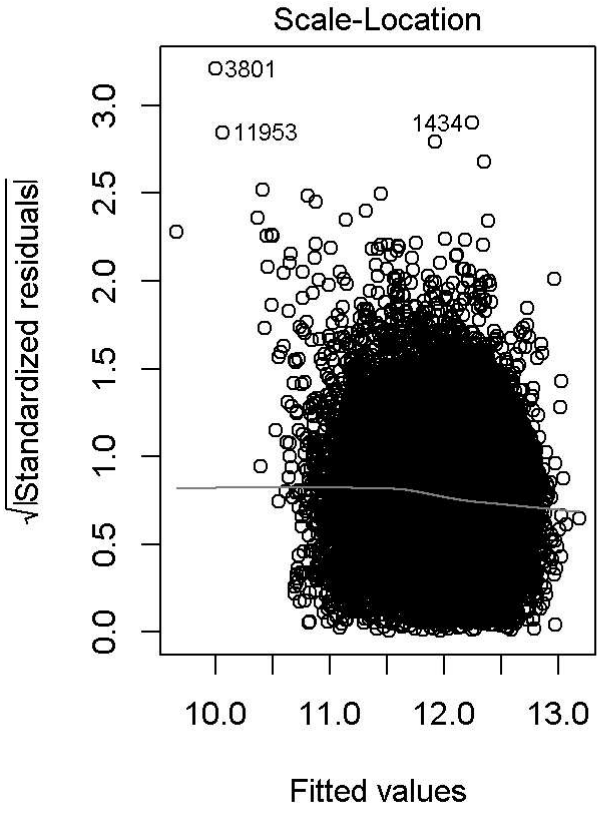
Residual vs. Influence Analysis

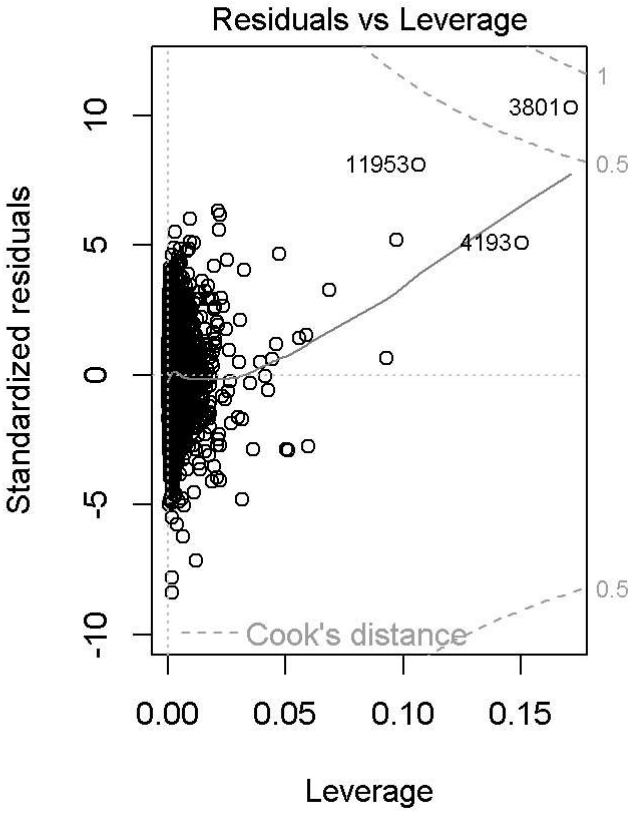
Next we will look at the residual and influence analysis for this model.

```
par (mfrow = c(1,2))
plot (fit2.int.sbc)
```


Residuals vs Fitted
Residuals
Fitted values
3801
5180
1434
10.0
11.0
12.0
13.0
-3
-2
-1
0
1
2
3

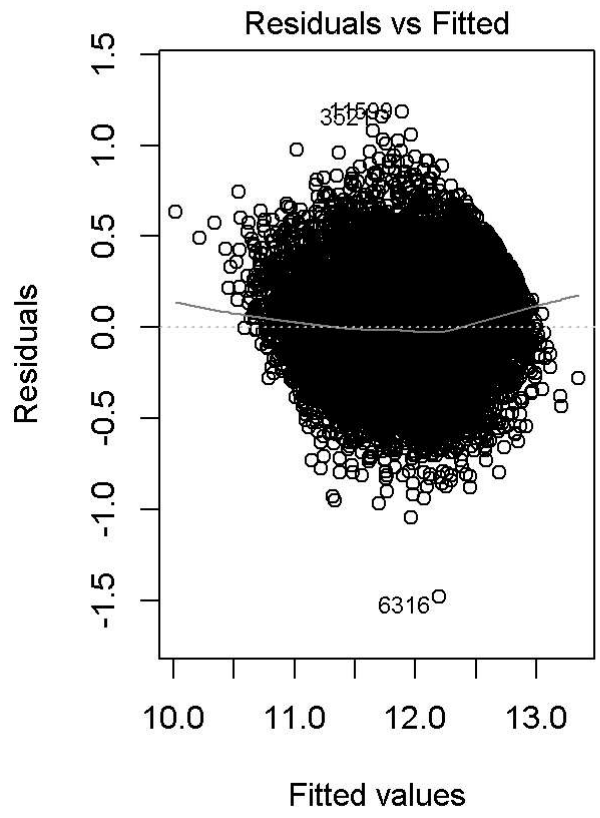

Q-Q Residuals
Standardized residuals
Theoretical Quantiles
3801
11953
1434
-4
-2
0
2
4
-5
0
5
10

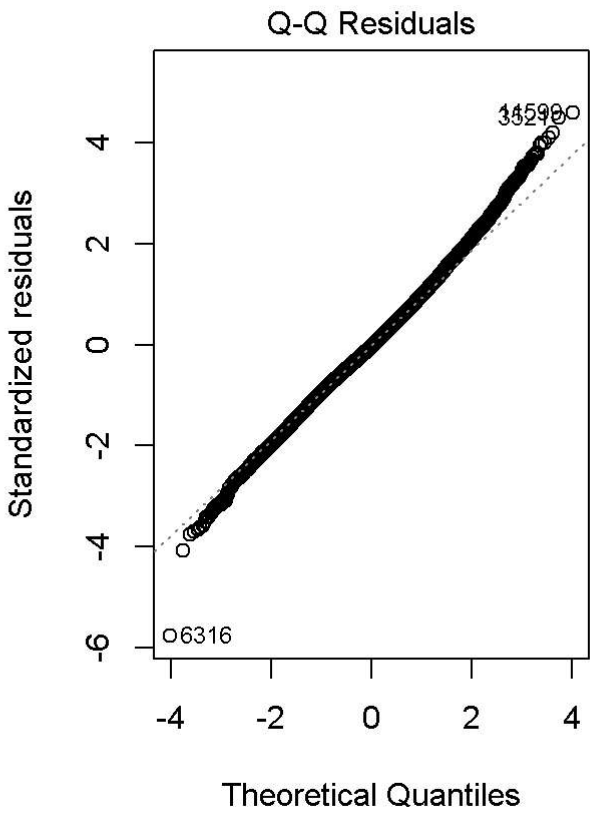

Scale-Location
√|Standardized residuals|
Fitted values
3801
11953
1434
10.0
11.0
12.0
13.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

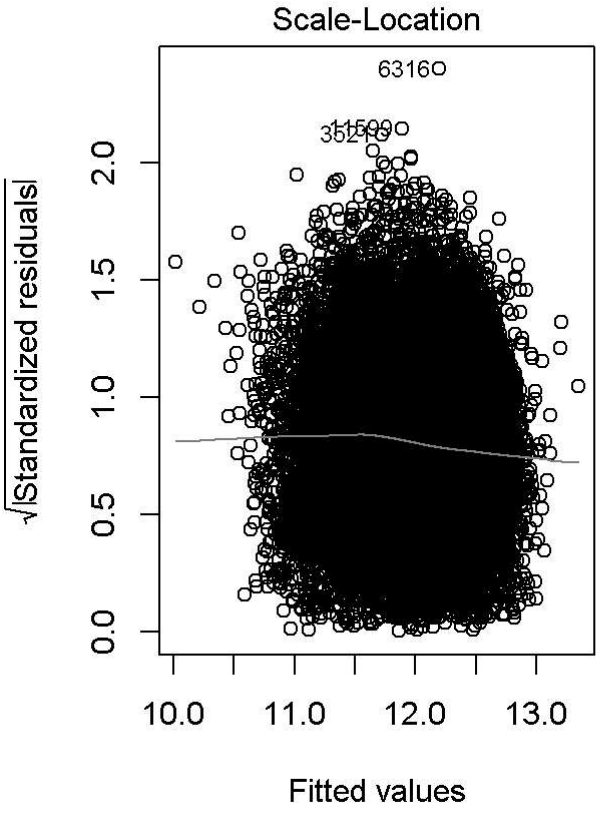

Residuals vs Leverage
Standardized residuals
Leverage
3801
11953
4193
Cook's distance
1
0.5
0.5
0.00
0.05
0.10
0.15
-10
-5
0
5
10

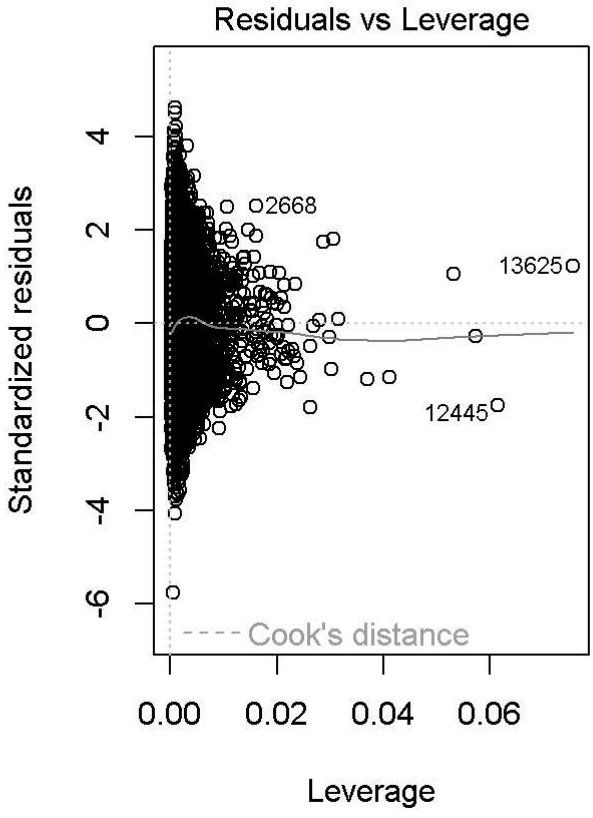
The cooks distance shows that there are data points that are influential observations within the model. Due to the presence of influential observations we will consider removing data points that has cooks' distance values greater than 3 standard deviations above and below the mean cooks distance value. Here is the residual analysis of the edited model.

```
# Compute Cook's distance
cooksD <- cooks.distance(fit2.int.sbc)

# Identify influential points based on a threshold of 3 times the mean Cook's distance
threshold <- 3 * mean(cooksD, na.rm = TRUE)
influential_points <- which(cooksD > threshold)

# Remove influential points
housing_data_v2 <- housing_data[-influential_points, ]

# Refit the model using the cleaned data
fit2.int.sbc_cleaned <- lm(log(median_house_value) ~ latitiude.c + longitude.c + housing_median_ag
e.c +
    log.population.c + log.households.c + log.median_income.c +
    as.factor(ocean_proximity) + latitiude.c:longitude.c + latitiude.c:housing_median_age.c +
    latitiude.c:log.households.c + latitiude.c:log.median_income.c +
    latitiude.c:as.factor(ocean_proximity) + longitude.c:housing_median_age.c +
    longitude.c:log.median_income.c + longitude.c:as.factor(ocean_proximity) +
    housing_median_age.c:log.population.c + housing_median_age.c:log.households.c +
    log.population.c:log.households.c + log.population.c:log.median_income.c +
    log.population.c:as.factor(ocean_proximity) + log.households.c:as.factor(ocean_proximity) +
    log.median_income.c:as.factor(ocean_proximity), data = housing_data_v2 )

# plots for the cleaned data
par(mfrow = c(1,2))
plot(fit2.int.sbc_cleaned)
```









All

Residual plots without influential points have better linearity, homoscedasticity, and normality.

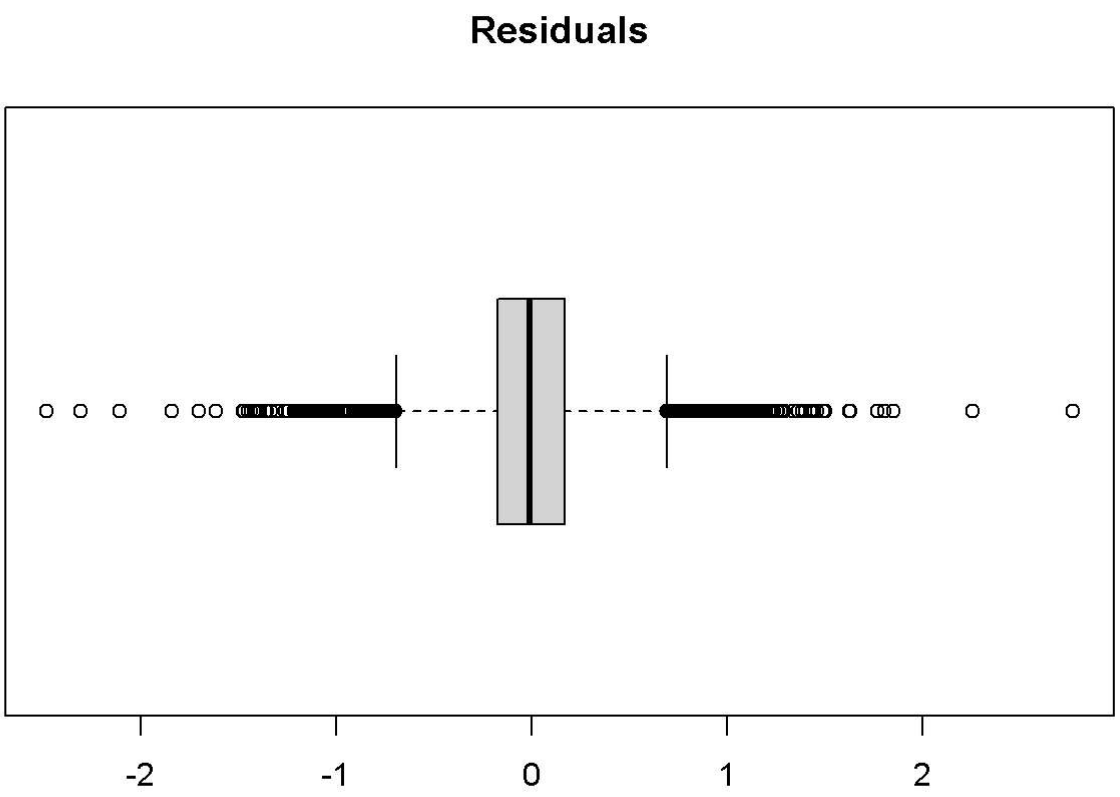
When comparing with the previous residual plots where influential points aren't excluded there are some considerations.

Pros of Removing Influential Points: * Improved Normality: The improvement in the Q-Q plot suggests better adherence to the assumption of normality of residuals. * Stable Model: The more symmetrical leverage plot suggests that the model may be more stable without these points.

Cons and Cautions: * Similar Residual Plots: Since the scale-location and fitted vs. residuals plots are similar we can argue that the removal of points doesn't significantly enhance these aspects. * Although we will not test these models on external datasets to confirm there is the possibility that removal could lead to overfitting * Since the data points are valid and come from a reliable source(1990 census), they may represent important variability in the data that could be important for accurate predictions and interpretations.

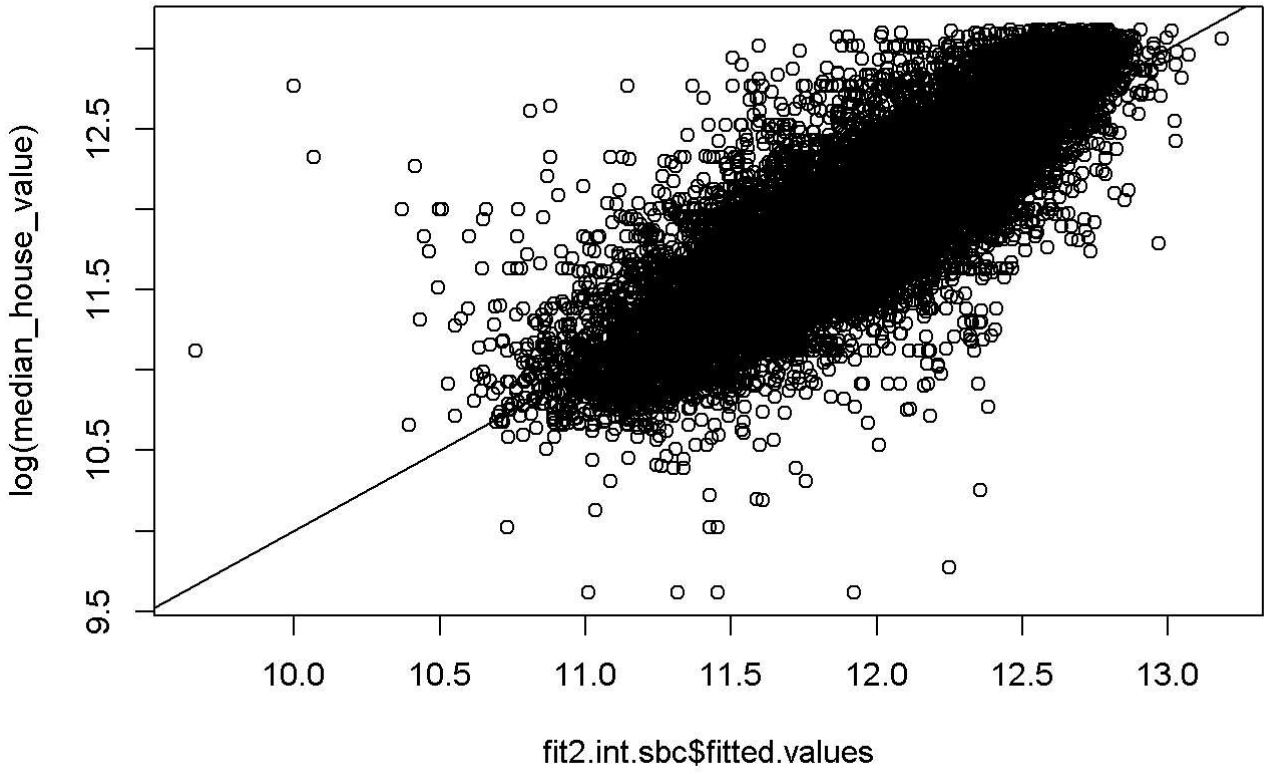
After considering these pros and Cons we have decided that there is no clear benefit to removing influential points and due to the validity of the data it will likely do more harm than good.

```
boxplot(residuals(fit2.int.sbc), horizontal = TRUE, main = "Residuals")
```



The box plot of residuals shows some outliers but has good symmetry

```
plot (log (median_house_value) ~ fit2.int.sbc$fitted.values, data=housing_data)
abline (0, 1)
```



The plot of log median house value vs. fitted log median house value shows a good fit of the model to the data.

```
car::vif (fit2.int.sbc, type='predictor')
```

```
## GVIFs computed for predictors
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## latitiude.c          192701.60170 12        1.660365
## longitude.c            1260.42251  9        1.486794
## housing_median_age.c   1598.97923 12        1.359858
## log.population.c         29.31018  8        1.235062
## log.households.c      16116.43825  9        1.712920
## log.median_income.c   17519.84832  8        1.841707
## ocean_proximity        9106.06260 11        1.513455
##
Interacts With
## latitiude.c          longitude.c, housing_median_age.c, log.households.c, log.median_income.c,
as.factor(ocean_proximity)
## longitude.c                             latitiude.c, housing_median_age.c, log.median_income.c,
as.factor(ocean_proximity)
## housing_median_age.c                                          latitiude.c, longitude.c, log.popu
lation.c, log.households.c
## log.population.c                  housing_median_age.c, log.households.c, log.median_income.c,
as.factor(ocean_proximity)
## log.households.c                        latitiude.c, housing_median_age.c, log.population.c,
as.factor(ocean_proximity)
## log.median_income.c                                  latitiude.c, longitude.c, log.population.c,
as.factor(ocean_proximity)
## ocean_proximity                         latitiude.c, longitude.c, log.population.c, log.househo
lds.c, log.median_income.c
##                                                                 Other Predictors
## latitiude.c                                  log.population.c, ocean_proximity
## longitude.c              log.population.c, log.households.c, ocean_proximity
## housing_median_age.c                     log.median_income.c, ocean_proximity
## log.population.c                   latitiude.c, longitude.c, ocean_proximity
## log.households.c           longitude.c, log.median_income.c, ocean_proximity
## log.median_income.c  housing_median_age.c, log.households.c, ocean_proximity
## ocean_proximity                        housing_median_age.c, ocean_proximity
```

All of the VIF values are less than 5, which is good for the collinearity assessment.

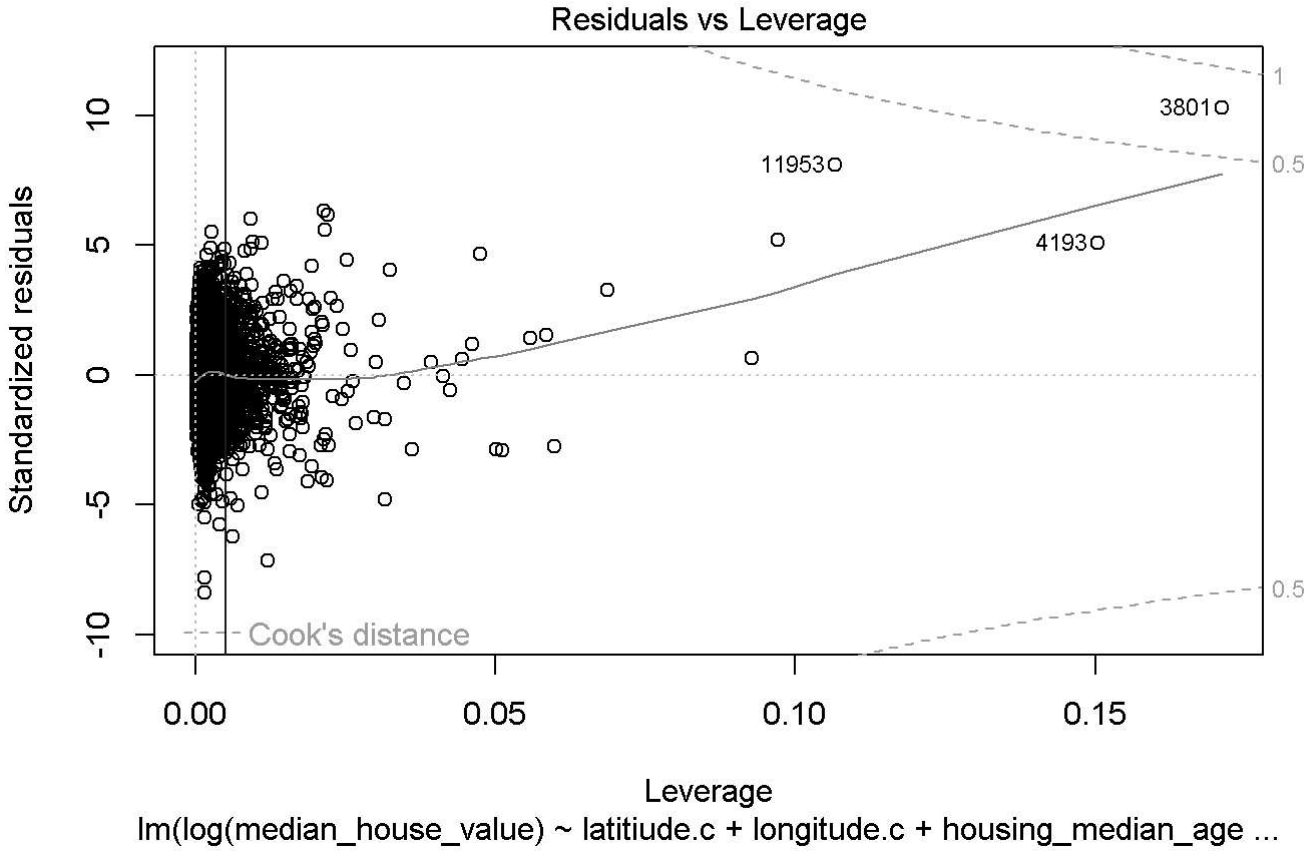
###Leverage cutoff

```
(levg.cutoff = 3*fit2.int.sbc$rank / nrows)
```

```
## [1] 0.004955853
```

The leverage cutoff is 0.005

```
plot (fit2.int.sbc, which=5)
abline (v=levg.cutoff, col='blue')
```




```
hatvals = hatvalues (fit2.int.sbc) > levg.cutoff
rows = sum (hatvals)
rows
```

```
## [1] 657
```

There are 668 (0.037 % of the total sample size) above the cutoff. We expect about 0.5 %..

```
summary (cooks.distance(fit2.int.sbc))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000000 0.0000023 0.0000119 0.0001688 0.0000438 0.7538831
```

The maximum Cook's distance in our dataset is 0.754. This is above the 0.5 threshold but we have decided to keep these influential points.

Interaction Plot

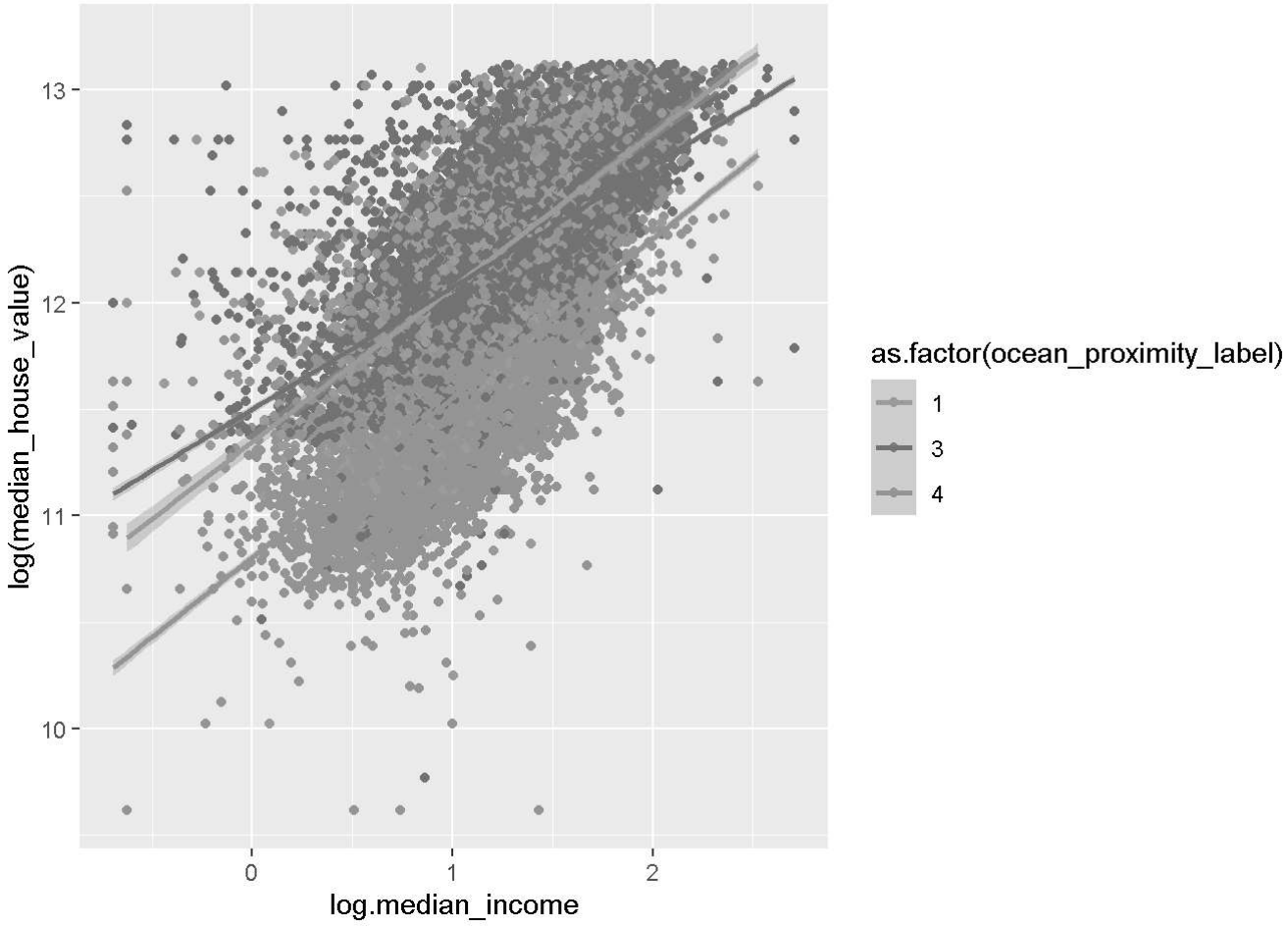
Next, we plot the interaction effects.

```
par (mfrow=c(1,1))
# Function to categorize a continuous variable into its quartiles
categorize = function (x) {
  quartiles = summary (x) [c(2, 3, 5)]
  result = rep ("Q1", length (x))
  result [(quartiles[1] < x) & (x <= quartiles [2])] = "Q2"
  result [(quartiles[2] < x) & (x <= quartiles [3])] = "Q3"
  result [quartiles[3] < x] = "Q4"
  return (result)
}
```

```
with (housing_data,
      qplot (x=log.median_income, y=log(median_house_value), color=as.factor (ocean_proximity_labe
l)) +
        geom_smooth (method="lm"))
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

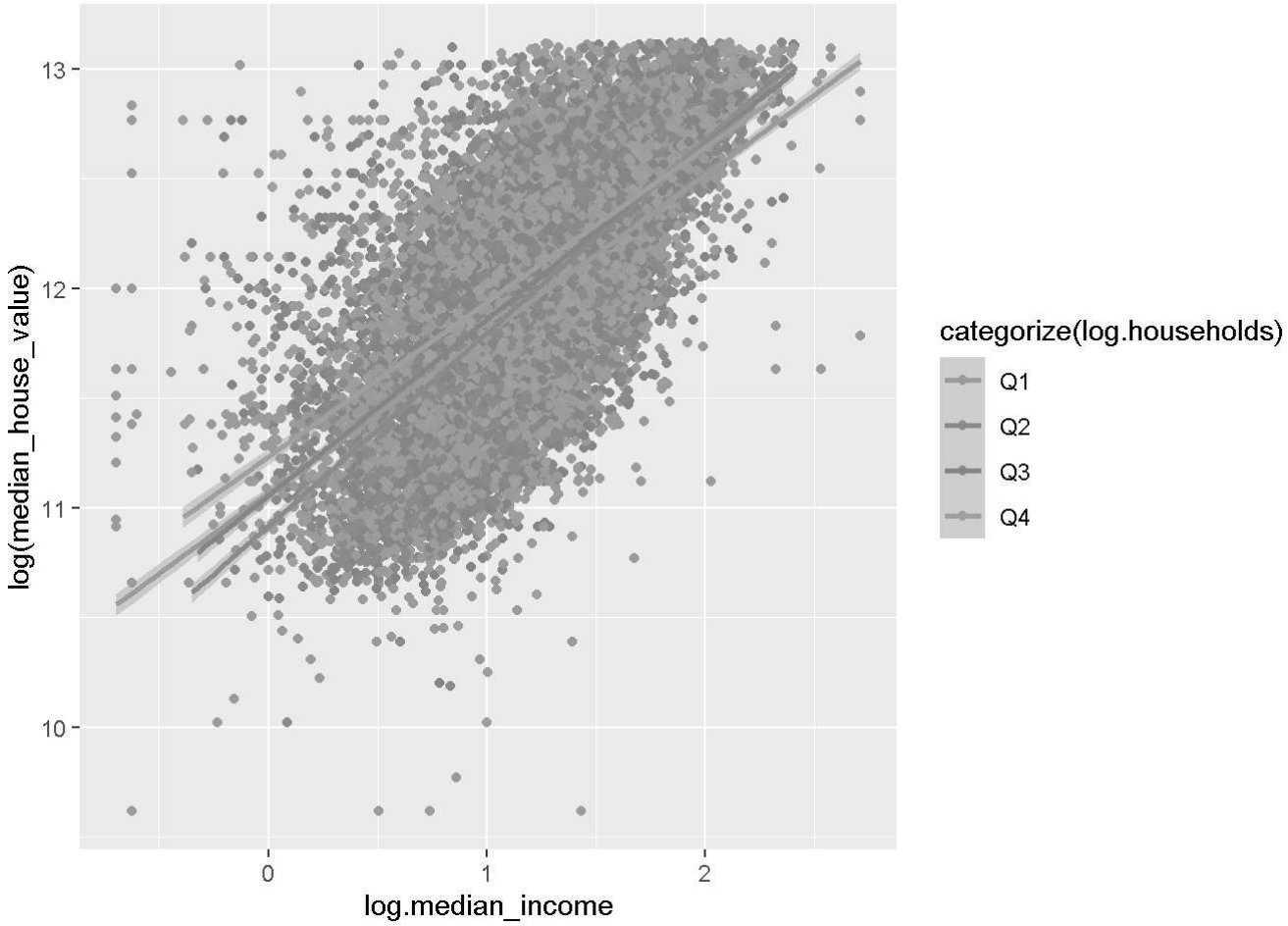
```
## `geom_smooth()` using formula = 'y ~ x'
```



This plot shows that the relationship between log median house value and log median income is stronger for houses that are in land (4) that the houses in other areas.

```
with (housing_data,
      qplot (x=log.median_income, y=log(median_house_value), color=categorize(log.households)) +
        geom_smooth (method="lm"))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This plot shows that relationship between log of median house value and log of median income is stronger with the number of household not having too much (Q4) or too less (Q1) per block.

```
preds = predict (fit2.int.sbc, interval='prediction')
```

```
## Warning in predict.lm(fit2.int.sbc, interval = "prediction"): predictions on current data refer
to _future_ responses
```

```
## In predict.lm(fit2.int.sbc, interval = "prediction"): predictions on current data refer to _fut
ure_ responses
housing_data$pred.median_house_value = exp (preds [,1])
housing_data$pred.lower = exp (preds [,2])
housing_data$pred.upper = exp (preds [,3])

# Print the results for 6 selected homes, 2 each with
# low, middle, and high prices
housing_data [c(100, 500, 5000, 357, 15000, 4000), c("households", "median_income", "longitude",
"latitude", "median_house_value", "pred.median_house_value",  "housing_median_age", "population",
"pred.lower", "pred.upper")]
```

```
##       households median_income longitude latitude median_house_value
## 209          642        2.8438   -121.98    38.37             111500
## 662          219        1.9018   -119.98    38.94             109700
## 5687        4072        6.6288   -117.87    34.04             339700
## 497          428        3.8438   -122.41    37.66             250700
## 16808        851        3.9722   -122.31    37.55             354100
## 4558         335        4.0000   -118.28    33.80             165500
##       pred.median_house_value housing_median_age population pred.lower
## 209                 110336.01                 21       2018   61854.41
## 662                  79730.32                 25        503   44672.59
## 5687                311507.95                  7      15037  174409.61
## 497                 238973.09                 37       1255  133946.58
## 16808               272743.39                 27       1877  152848.78
## 4558                213075.68                 38       1207  119461.38
##       pred.upper
## 209     196817.6
## 662     142300.3
## 5687    556375.3
## 497     426350.1
## 16808   486683.4
## 4558    380049.5
```

```
housing_data$in.interval = ifelse (housing_data$pred.lower <= housing_data$median_house_value &
                                   housing_data$median_house_value <= housing_data$pred.upper,
                                   1, 0)
mean (housing_data$in.interval)
```

```
## [1] 0.9505554
```

All the houses values fall under the predicted interval. Among all of the houses in the data set, 95% have prediction intervals that contain the observed median house value. Which is just you would expect.

Conclusions

The final model has adjusted R squared of about 69%, which means that 69% of the variation in log sales median house value is explained by the model. The residual standard error is 0.29 log dollars. Based on our residual analysis, this model is consistent with the conditions for doing linear regression.

Some the features with interaction we have plotted them above to see the relationship. We found out that relationship between log median house value and log median income is stronger when house is in land than when house are near ocean. Relationship between log median house value and log median income is strong with median number of households compared to blocks with high too high or too low number of households.

The other predictor variables can be interpreted qualitatively as follows: Homes in higher median income blocks tend to cost more Homes with higher total rooms tend to cost more Homes that are in highly populated blocks tend to cost less.