Linear Regression

Lokeshwari, Sriprasath

February 9, 2020

California Housing Dataset prediction using Linear Regression

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

require(tidyverse)
California <-
read.csv("https://personal.utdallas.edu/~sxg180154/housing.csv")
```

Analysing the Dataset - California

View the california Dataset, getting summary of the Dataset (California) and dimension of the California Dataset

```
View(California)
summary(California)

##    longitude         latitude     housing_median_age  total_rooms
##  Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2
##  1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448
##  Median :-118.5   Median :34.26   Median :29.00      Median : 2127
##  Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636
##  3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148
##  Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320
##
##  total_bedrooms     population      households     median_income
##  Min.   :   1.0   Min.   :    3   Min.   :   1.0   Min.   : 0.4999
##  1st Qu.: 296.0   1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634
##  Median : 435.0   Median : 1166   Median : 409.0   Median : 3.5348
##  Mean   : 537.9   Mean   : 1425   Mean   : 499.5   Mean   : 3.8707
##  3rd Qu.: 647.0   3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432
##  Max.   :6445.0   Max.   :35682   Max.   :6082.0   Max.   :15.0001
```

```
##   NA's    :207
##   median_house_value   ocean_proximity
##   Min.   : 14999     <1H OCEAN :9136
##   1st Qu.:119600     INLAND    :6551
##   Median :179700     ISLAND    :   5
##   Mean   :206856     NEAR BAY  :2290
##   3rd Qu.:264725     NEAR OCEAN:2658
##   Max.   :500001
##

dim(California)

## [1] 20640    10
```

Finding the percentage of Null values in each column to eliminate if there are more than 50% NULL values in a column

```
for(i in 1:ncol(California)) {
  colName <- colnames(California[i])
  pctNull <- sum(is.na(California[,i]))/length(California[,i])
    print(paste("Column ", colName, " has ", round(pctNull*100, 3), "% of
nulls"))
}

## [1] "Column  longitude  has  0 % of nulls"
## [1] "Column  latitude  has  0 % of nulls"
## [1] "Column  housing_median_age  has  0 % of nulls"
## [1] "Column  total_rooms  has  0 % of nulls"
## [1] "Column  total_bedrooms  has  1.003 % of nulls"
## [1] "Column  population  has  0 % of nulls"
## [1] "Column  households  has  0 % of nulls"
## [1] "Column  median_income  has  0 % of nulls"
## [1] "Column  median_house_value  has  0 % of nulls"
## [1] "Column  ocean_proximity  has  0 % of nulls"
```

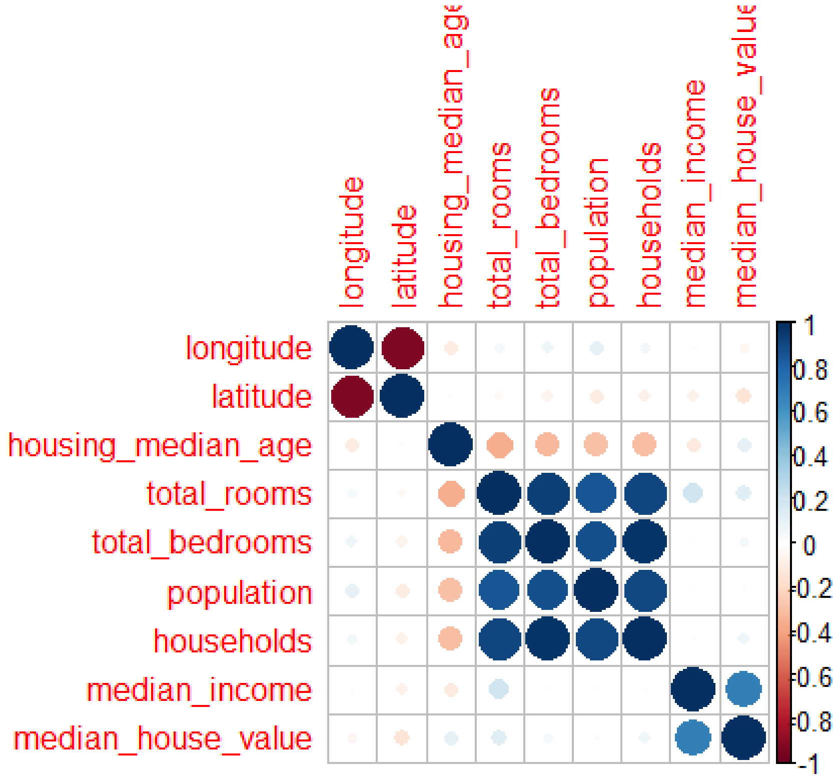
Cleaning all the NULL values in each row using exclude() function

From the dimension of the original California data and the cleaned data we can conclude that 207 rows with NULL values are eliminated

```
cali_clean <- na.exclude(California)
dim(cali_clean)

## [1] 20433    10
```

Getting a correlation plot for visuaization using corrplot()

```
library(corrplot)

## corrplot 0.84 loaded

require(corrplot)
Mat <- cor(cali_clean[sapply(cali_clean, is.numeric)])
corrplot(Mat)
```



Visual Representation and correlation of the Dataset - California

From the graph we can estimate that the correlation betweeen

Total_rooms - Households

Population - Total_rooms

Population - Households

Median_Income - Median_House_value are highly correlated

Getting the correlation values for each predictors and the output value(median_house_value)

Features having highest correlation among the predictors

```
cor(cali_clean$total_rooms, cali_clean$households)

## [1] 0.9189915

cor(cali_clean$population,cali_clean$total_rooms)

## [1] 0.8572813

cor(cali_clean$population,cali_clean$households)

## [1] 0.9071859
```

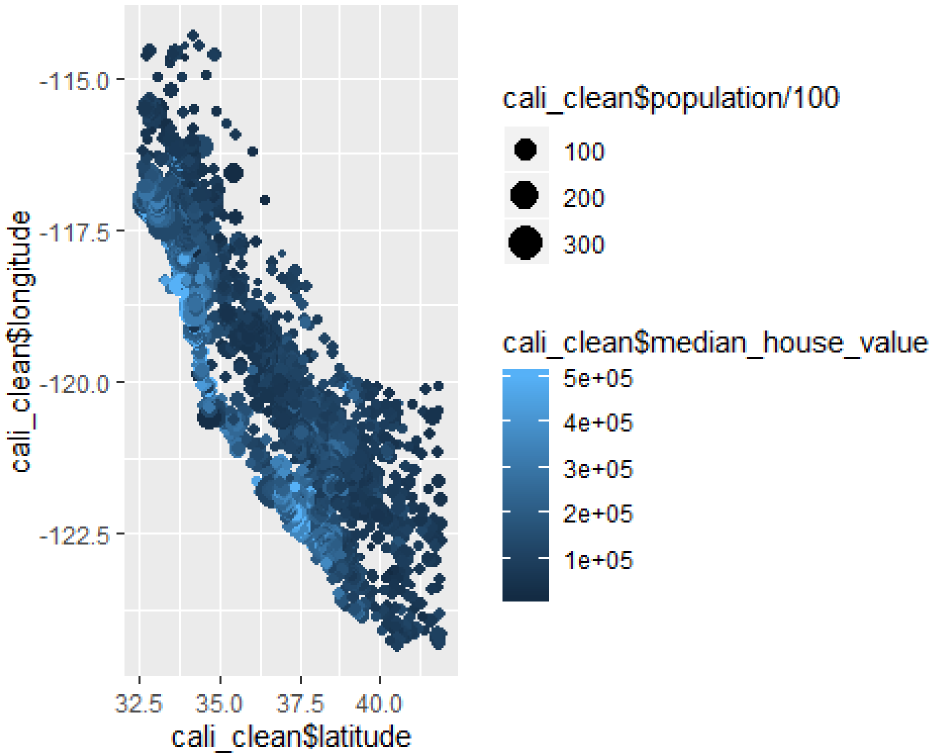
=> Correlation between the output value(median_house_value) and all other predictors

```r
cor(cali_clean$median_income,cali_clean$median_house_value)

## [1] 0.6883555

cor(cali_clean$longitude,cali_clean$median_house_value)

## [1] -0.04539822

cor(cali_clean$latitude,cali_clean$median_house_value)

## [1] -0.1446382

cor(cali_clean$housing_median_age,cali_clean$median_house_value)

## [1] 0.106432

cor(cali_clean$total_rooms,cali_clean$median_house_value)

## [1] 0.1332941

cor(cali_clean$total_bedrooms,cali_clean$median_house_value)

## [1] 0.04968618

cor(cali_clean$population,cali_clean$median_house_value)

## [1] -0.02529973

cor(cali_clean$households,cali_clean$median_house_value)

## [1] 0.06489355
```

Plotting the housing_value based on the latitude, longitude and population

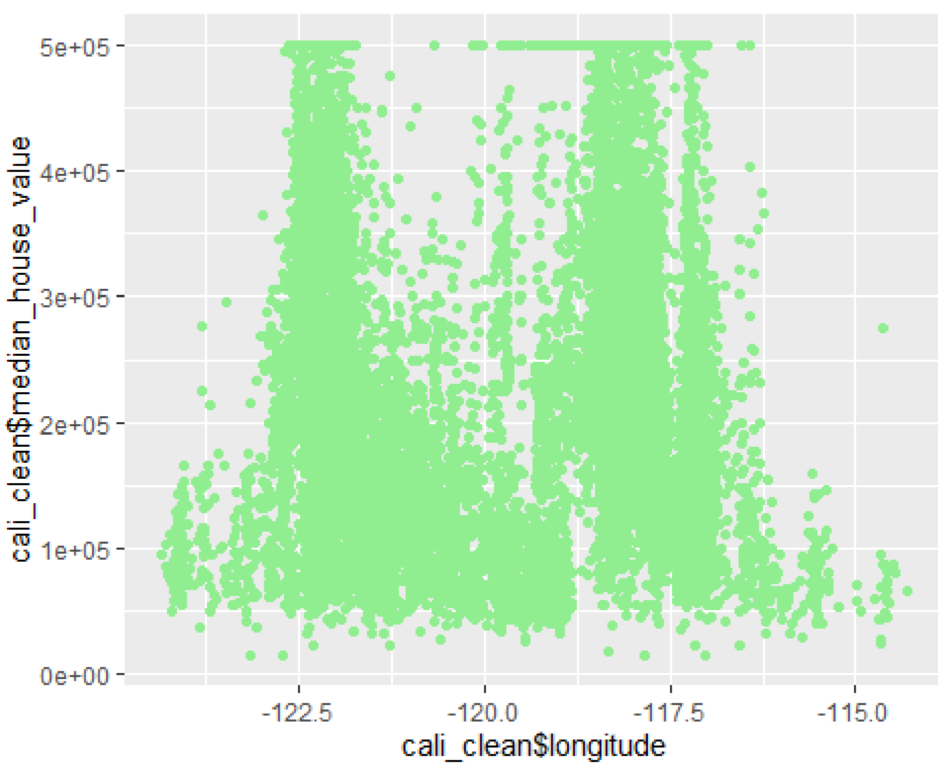
This plot shows that the region with the highest population has high housing value

```r
ggplot(data = cali_clean,mapping =
aes(cali_clean$latitude,y=cali_clean$longitude, color =
cali_clean$median_house_value, size =
cali_clean$population/100))+geom_point()
```



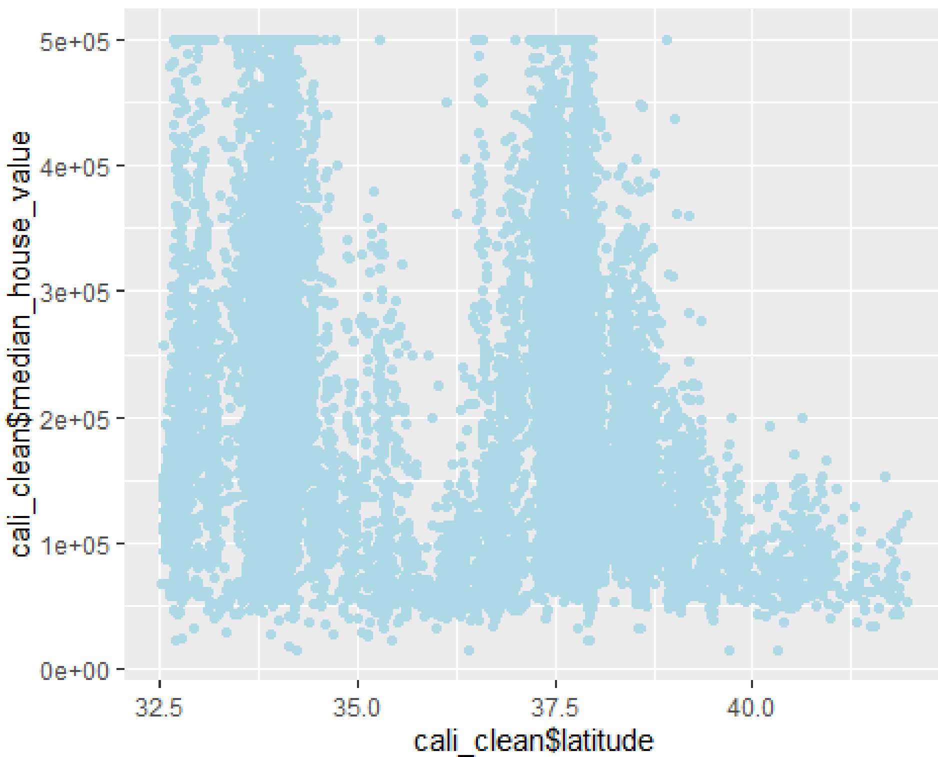
=> Plotting the point ggplot of all predictors vs the housing value

```
ggplot(data = cali_clean,mapping= aes(cali_clean$longitude,
cali_clean$median_house_value))+geom_point(color = "light green")
```



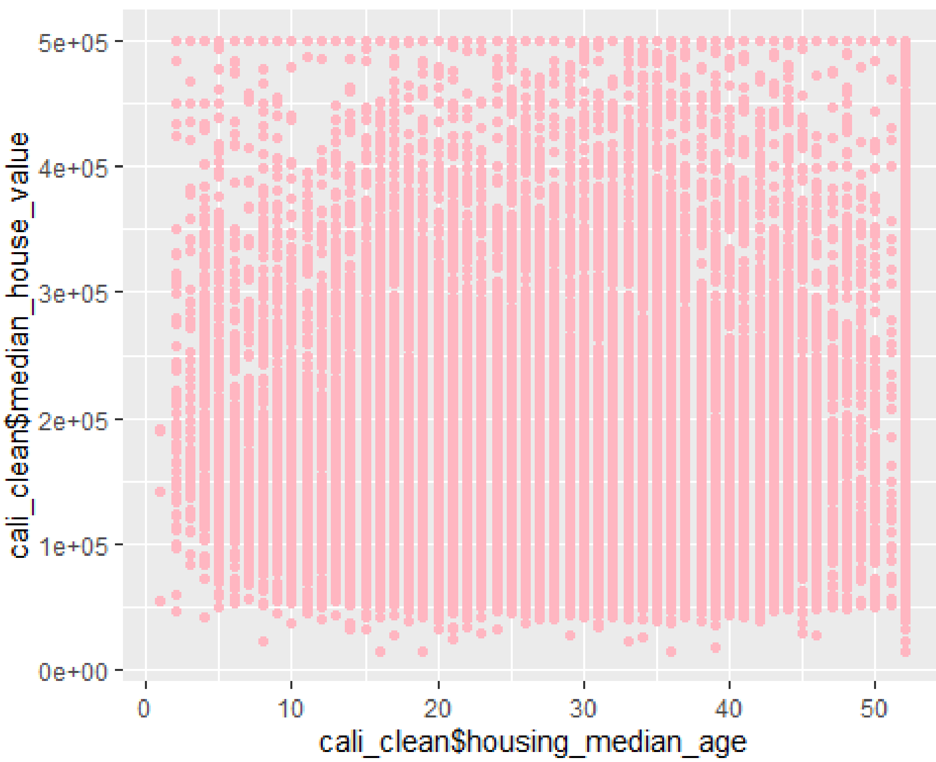
This graph concludes that the median_house_value is higher from -122.5 to -117.5 longitude. And the number of houses are more with median_house_value around 1e+05 to 3e+05 and longitude from -122.4 to -117.5

```
ggplot(data = cali_clean,mapping= aes(cali_clean$latitude,
cali_clean$median_house_value))+geom_point(color = "light blue")
```



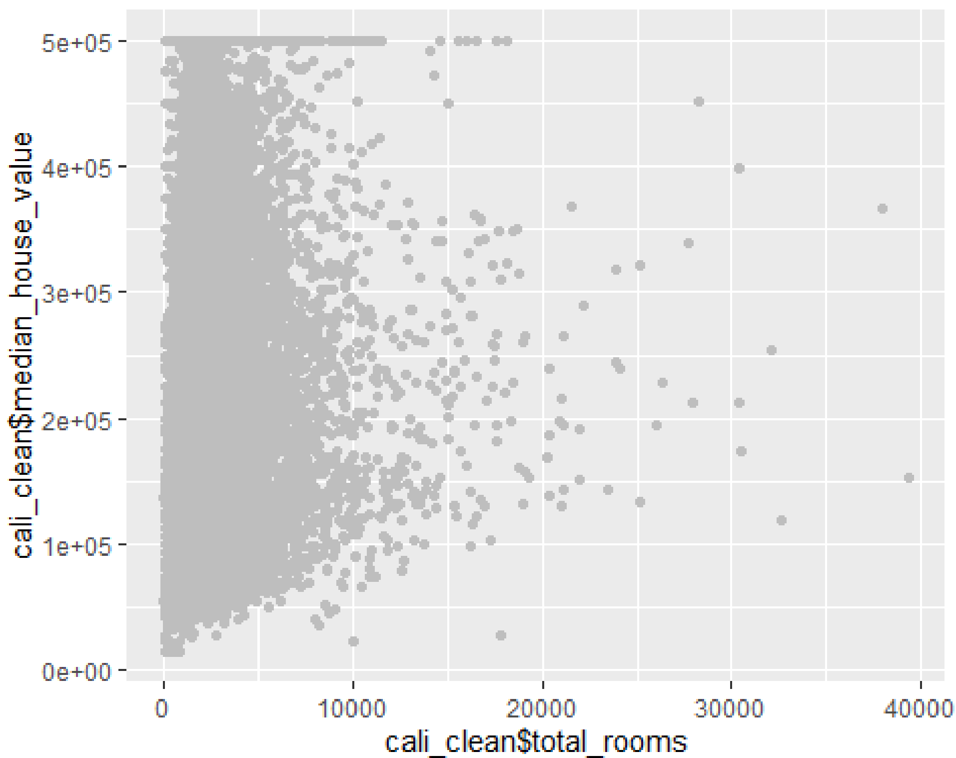
This graph concludes that the median_house_value is higher at latitude 32.5 and 37.5. The number of houses are more with the median_house_value around 1e+05 to 4e+05 and laitude from 32.5 to 40.0

```
ggplot(data = cali_clean,mapping= aes(cali_clean$housing_median_age,
cali_clean$median_house_value))+geom_point(color = "light pink")
```



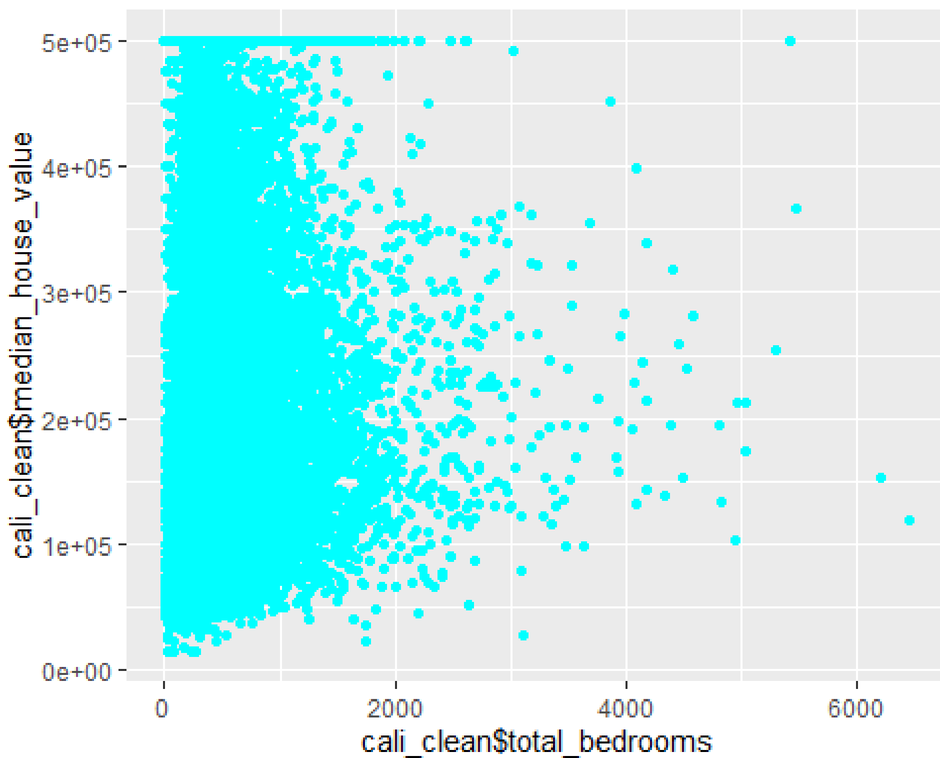
This graph shows that the median_house_value is same for all ages of houses. The houses are spread over all values of the houses and different age

```
ggplot(data = cali_clean,mapping= aes(cali_clean$total_rooms,
cali_clean$median_house_value))+geom_point(color = "grey")
```



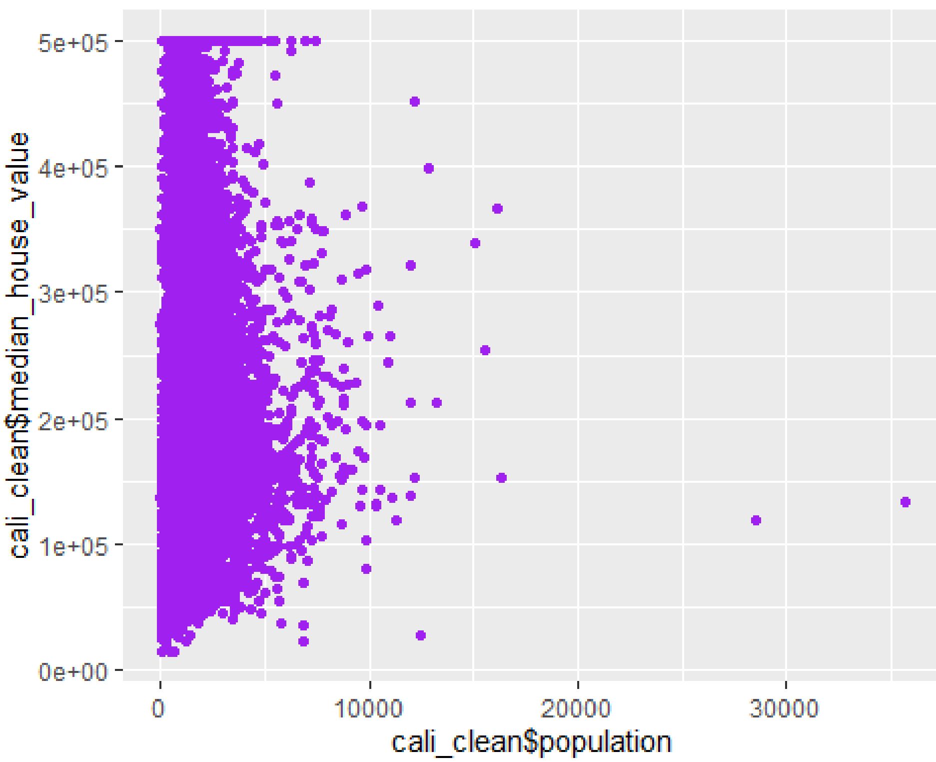
Almost 90% of the house area have 0 to 5000 rooms in a particular area. The median_house_value is evenly distributed for the area with total_rooms from 0 to 5000.

```
ggplot(data = cali_clean,mapping= aes(cali_clean$total_bedrooms,
cali_clean$median_house_value))+geom_point(color = "cyan")
```



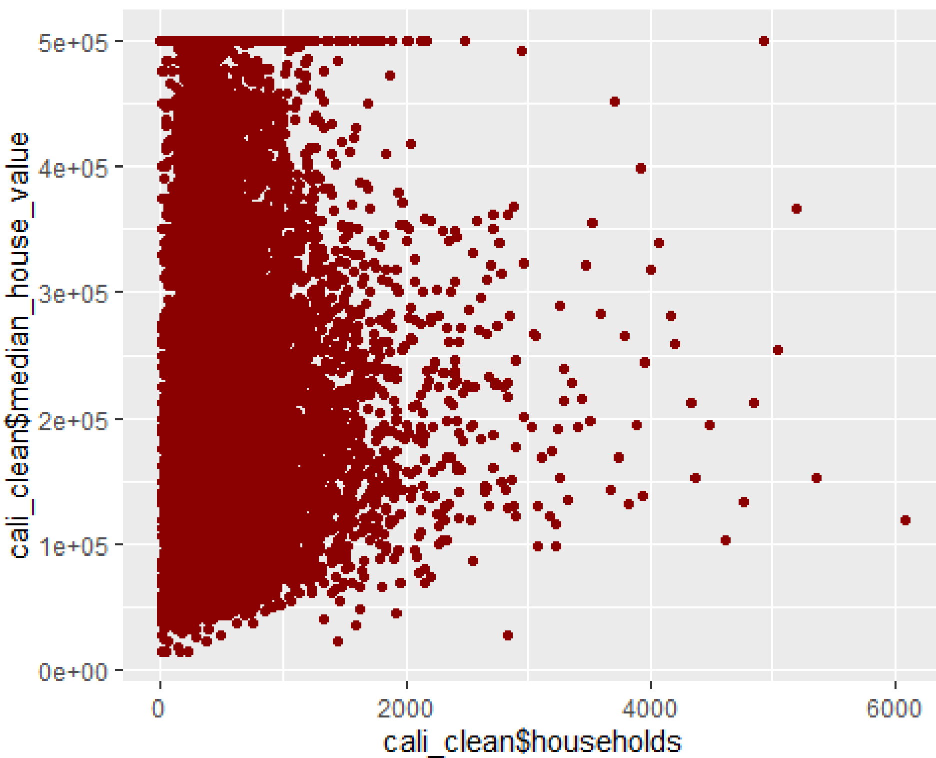
Almost 80% of the house area have 0 to 1000 bedrooms in a particular area. The median_house_value is evenly distributed for the area with total_bedrooms from 0 to 1000.

```
ggplot(data = cali_clean,mapping= aes(cali_clean$population,
cali_clean$median_house_value))+geom_point(color = "purple")
```



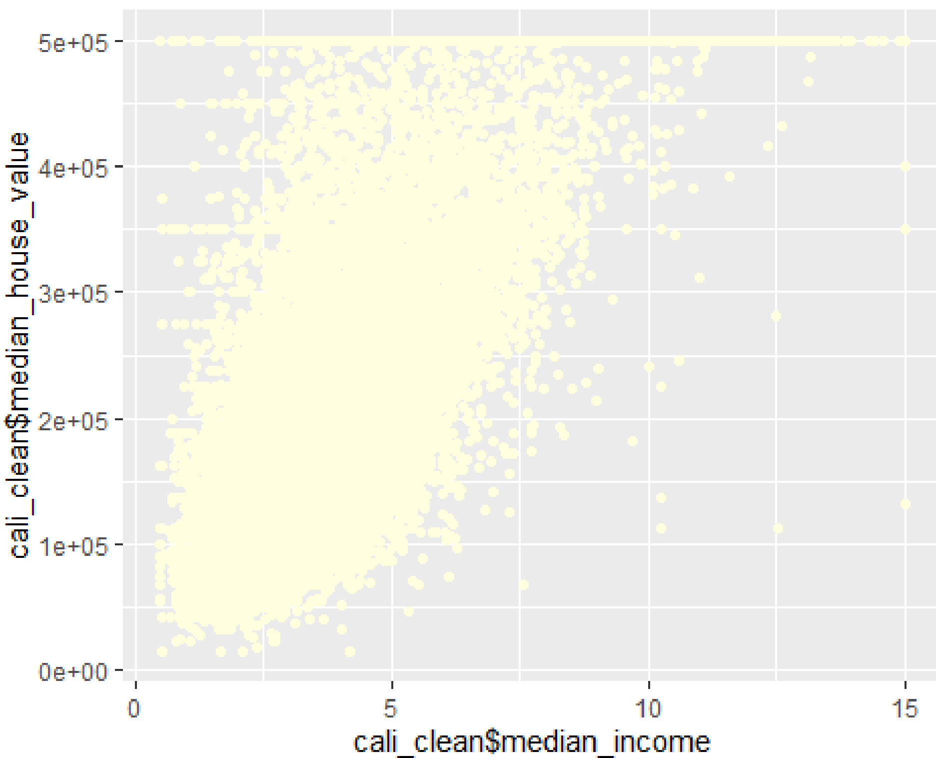
Almost 92% of the house area have 0 to 5000 population count in a particular area. The median_house_value is evenly distributed for the area with population count from 0 to 5000

```
ggplot(data = cali_clean,mapping= aes(cali_clean$households,
cali_clean$median_house_value))+geom_point(color = "dark red")
```



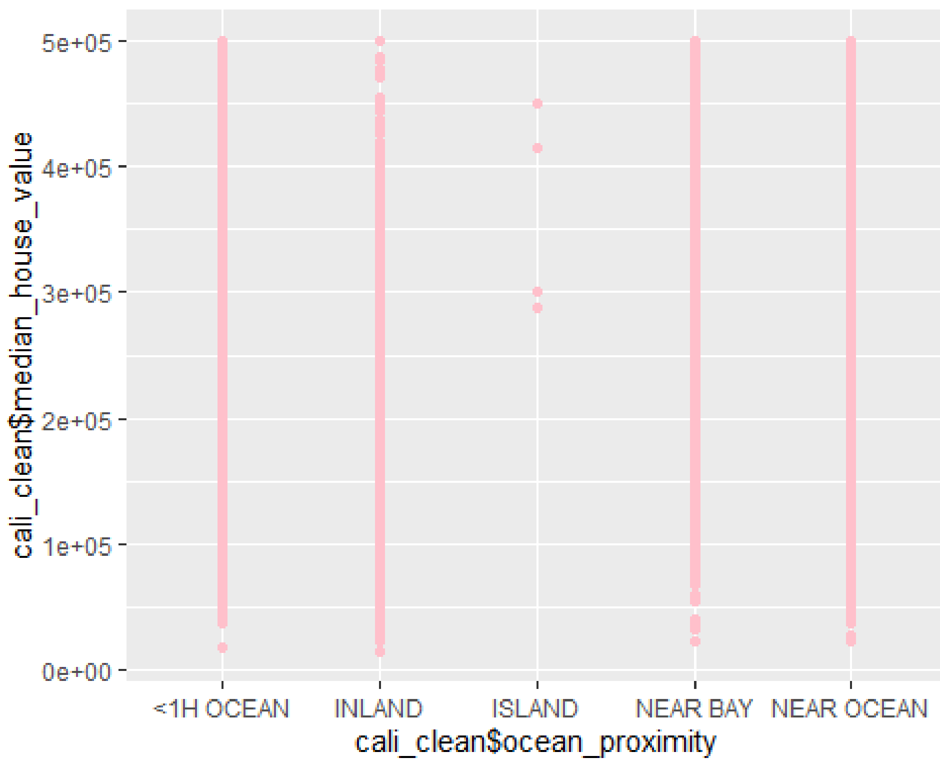
Almost 90% of the house area have 0 to 1000 households in a particular area. The median_house_value is evenly distributed for the area with households from 0 to 1000.

```
ggplot(data = cali_clean,mapping= aes(cali_clean$median_income,
cali_clean$median_house_value))+geom_point(color = "light yellow")
```



The graph is linear for the house_value and median_income. The people who gets high income have bought houses with high house_value. The median_house_value gradually increases with increase in house_value.

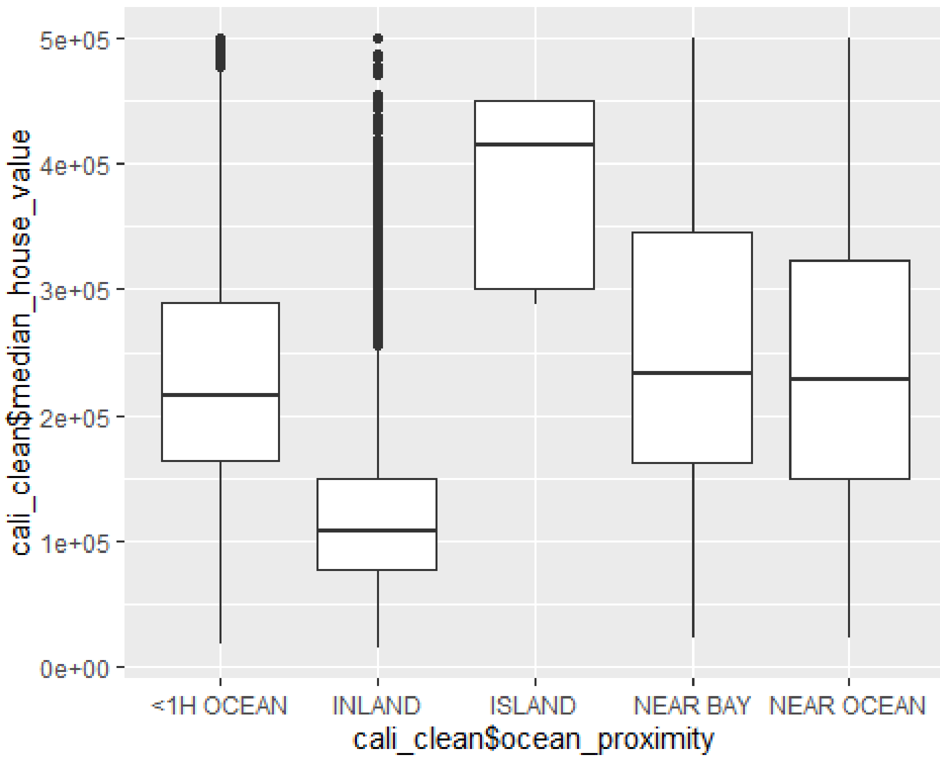
```
ggplot(data = cali_clean,mapping= aes(cali_clean$ocean_proximity,
cali_clean$median_house_value))+geom_point(color = "pink")
```



We can conclude that the there are less number of houses in near ocean and near bay and exactly 5 houses in island. The house_value is more in other area compared to Island.

=> Finding the concentration of house values based on the ocean proximity

```
ggplot(data = cali_clean,mapping = aes(x = cali_clean$ocean_proximity, y=
cali_clean$median_house_value))+geom_boxplot()
```



We can conclude that the there are less number of houses in near ocean and near bay and exactly 5 houses in island

But the Median_house_value increases for the houses near ocean and near bay than the houses in the <1HOcean and Inland

=> Number of houses in each category of ocean proximity

```
group_ocean = cali_clean %>% group_by(ocean_proximity) %>% summarise(Number =
n()) %>% arrange(desc(Number))
group_ocean

## # A tibble: 5 x 2
##   ocean_proximity Number
##   <fct>            <int>
## 1 <1H OCEAN         9034
## 2 INLAND            6496
## 3 NEAR OCEAN        2628
## 4 NEAR BAY          2270
## 5 ISLAND               5
```

We can conclude that there are less houses near the ocean and bay and exactly 5 houses in the island

=> Top 10 costliest houses in the california housing data

```
cost = cali_clean %>% filter(!is.na(median_house_value)) %>%
arrange(desc(median_house_value)) %>% head(10)
cost

##    longitude latitude housing_median_age total_rooms total_bedrooms
population
## 1    -122.27    37.80                 52         249             78
396
## 2    -122.25    37.87                 52         609            236
1349
## 3    -122.24    37.86                 52        1668            225
517
## 4    -122.24    37.85                 52        3726            474
1366
## 5    -122.23    37.83                 52        2990            379
947
## 6    -122.22    37.82                 39        2492            310
808
## 7    -122.22    37.82                 42        2991            335
1018
## 8    -122.23    37.82                 52        3242            366
1001
## 9    -122.23    37.82                 52        3494            396
1192
## 10   -122.23    37.82                 52        1611            203
556
##    households median_income median_house_value ocean_proximity
## 1          85        1.2434             500001        NEAR BAY
## 2         250        1.1696             500001        NEAR BAY
## 3         214        7.8521             500001        NEAR BAY
## 4         496        9.3959             500001        NEAR BAY
## 5         361        7.8772             500001        NEAR BAY
## 6         315       11.8603             500001        NEAR BAY
## 7         335       13.4990             500001        NEAR BAY
## 8         352       12.2138             500001        NEAR BAY
## 9         383       12.3804             500001        NEAR BAY
## 10        179        8.7477             500001        NEAR BAY
```

=> Splitting rooms_per_household, bedrooms_per_household, population_per_household

```
rooms = trunc(cali_clean$total_rooms/cali_clean$households)
bedrooms = trunc(cali_clean$total_bedrooms/cali_clean$total_rooms)
popu_per_house = trunc(cali_clean$population/cali_clean$households)
```

Adding the additional columns(rooms_per_household, bedrooms_per_room, population_per_husehold) in the cali_full dataframe

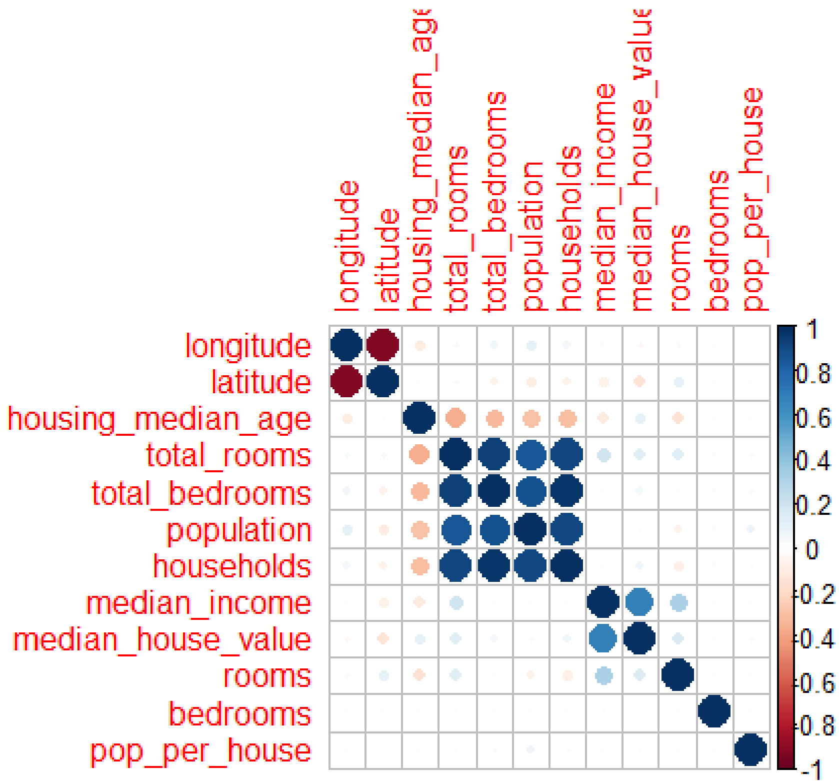
Splitting the data into the smaller dataset with respect to every household and population

```
cali_full = cali_clean %>% add_column(rooms = rooms,bedrooms = bedrooms,
pop_per_house = popu_per_house)
```

=> Again finding the correlation between the predcitors and the newly added columns

This correlation is also similar to the original data where we obtained the derived dataset i.e. rooms_per_household, bedrooms_per_room and

```
matrix <- cor(cali_full[sapply(cali_full, is.numeric)])
corrplot(matrix)
```




```
cor(matrix)

##                      longitude    latitude housing_median_age
total_rooms
## longitude          1.0000000000 -0.98144232        -0.18560762
0.14161045
## latitude          -0.9814423246  1.00000000         0.10672804 -
0.12179737
## housing_median_age -0.1856076195  0.10672804         1.00000000 -
0.75046065
## total_rooms        0.1416104506 -0.12179737        -0.75046065
1.00000000
## total_bedrooms     0.1658263736 -0.14011352        -0.69693145
```

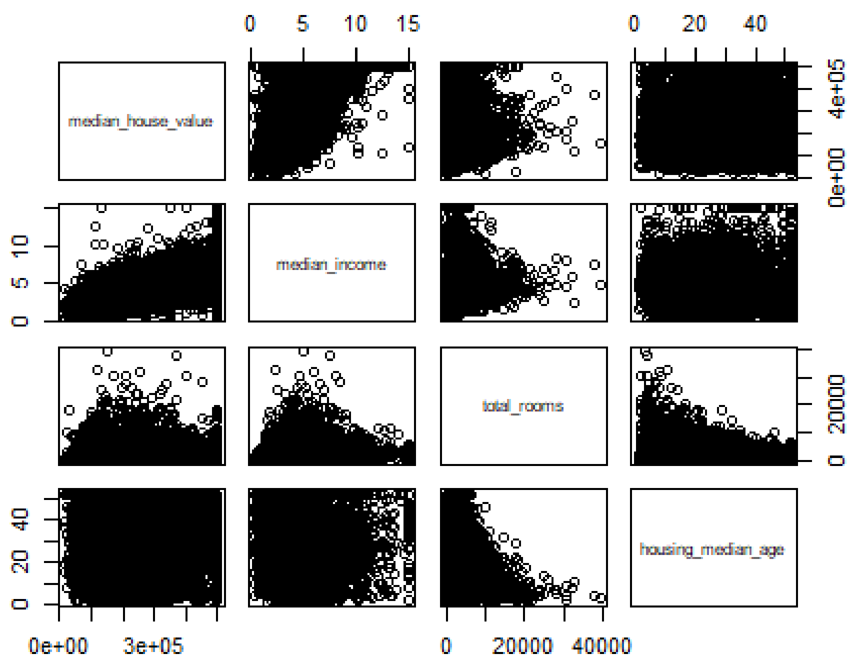
```
0.98650546
## population          0.2035423324 -0.17615988        -0.68441755
0.97013502
## households          0.1615291972 -0.14274025        -0.67889602
0.98285046
## median_income      -0.0006528726 -0.10613198        -0.14303440 -
0.02771175
## median_house_value  0.0012360285 -0.15879449         0.09101670 -
0.08072260
## rooms              -0.1234104348  0.16565900        -0.21599125 -
0.10244660
## bedrooms           -0.0065163600  0.01934191         0.04322066 -
0.23655525
## pop_per_house      -0.0062305844  0.02071009         0.06267989 -
0.23979760
##                    total_bedrooms population households median_income
## longitude               0.1658264  0.2035423  0.1615292 -0.0006528726
## latitude               -0.1401135 -0.1761599 -0.1427403 -0.1061319837
## housing_median_age     -0.6969314 -0.6844175 -0.6788960 -0.1430343952
## total_rooms             0.9865055  0.9701350  0.9828505 -0.0277117515
## total_bedrooms          1.0000000  0.9886538  0.9982912 -0.1702024466
## population              0.9886538  1.0000000  0.9911631 -0.2011689005
## households              0.9982912  0.9911631  1.0000000 -0.1653555078
## median_income          -0.1702024 -0.2011689 -0.1653555  1.0000000000
## median_house_value     -0.1803002 -0.2248021 -0.1664323  0.8852253895
## rooms                  -0.2177188 -0.2756743 -0.2615215  0.4801820119
## bedrooms               -0.2087264 -0.2040689 -0.2056358 -0.1715180939
## pop_per_house          -0.2155424 -0.1370039 -0.2101102 -0.1422231415
##                    median_house_value      rooms    bedrooms pop_per_house
## longitude                 0.001236029 -0.1234104 -0.00651636  -0.006230584
## latitude                 -0.158794486  0.1656590  0.01934191   0.020710093
## housing_median_age        0.091016699 -0.2159912  0.04322066   0.062679886
## total_rooms              -0.080722599 -0.1024466 -0.23655525  -0.239797598
## total_bedrooms           -0.180300241 -0.2177188 -0.20872643  -0.215542353
## population               -0.224802122 -0.2756743 -0.20406894  -0.137003899
## households               -0.166432318 -0.2615215 -0.20563575  -0.210110176
## median_income             0.885225389  0.4801820 -0.17151809  -0.142223141
## median_house_value        1.000000000  0.2535645 -0.16161228  -0.193169329
## rooms                     0.253564468  1.0000000 -0.14655978  -0.128032759
## bedrooms                 -0.161612276 -0.1465598  1.00000000  -0.069676508
## pop_per_house            -0.193169329 -0.1280328 -0.06967651   1.000000000
```

The Median_house_value is more correlated to the median_house_value

From the correlation matrix we get the value of correlation between the each predictor and the output value(median_house_value)

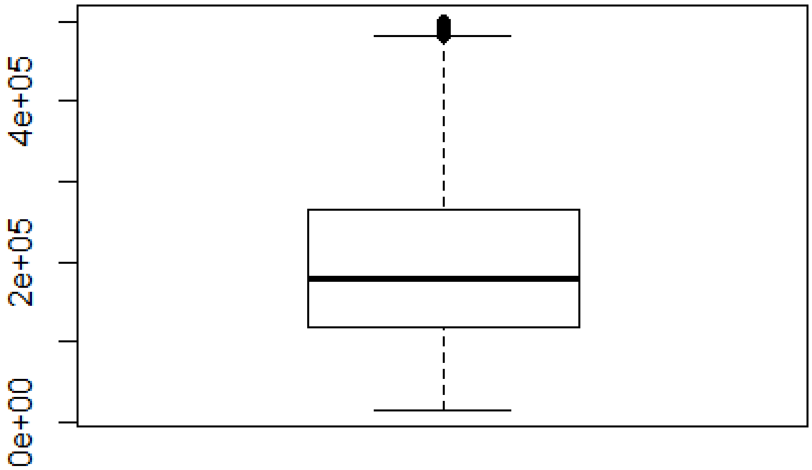
=> Plotting the graph between the median_house_value, median_income , total_rooms and housing _median_age

```
plot(cali_full[,c("median_house_value","median_income","total_rooms","housing_median_age")])
```



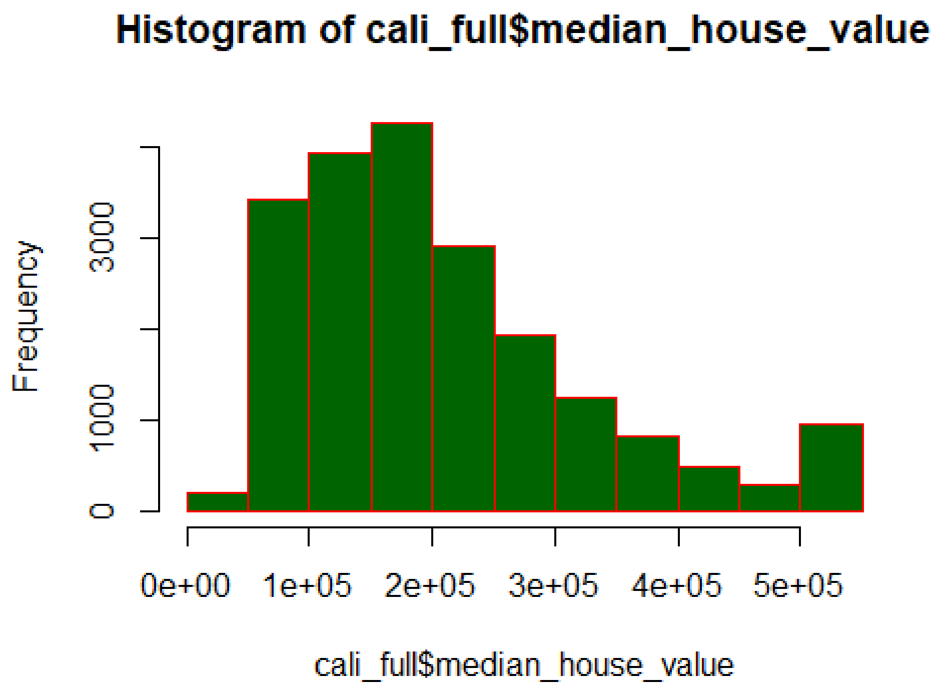
This plot shows the distribution of median_house_value , median_income , total_rooms and median_housing_age among each other predictor.

```
boxplot(cali_full$median_house_value)
```



=> The quantile range of median_house_value is plotted using box_plot. From this we can conclude that the min house_value is from 1e+05 to 5e+05 with some outliers. And the median value of the house_value is 2e+00

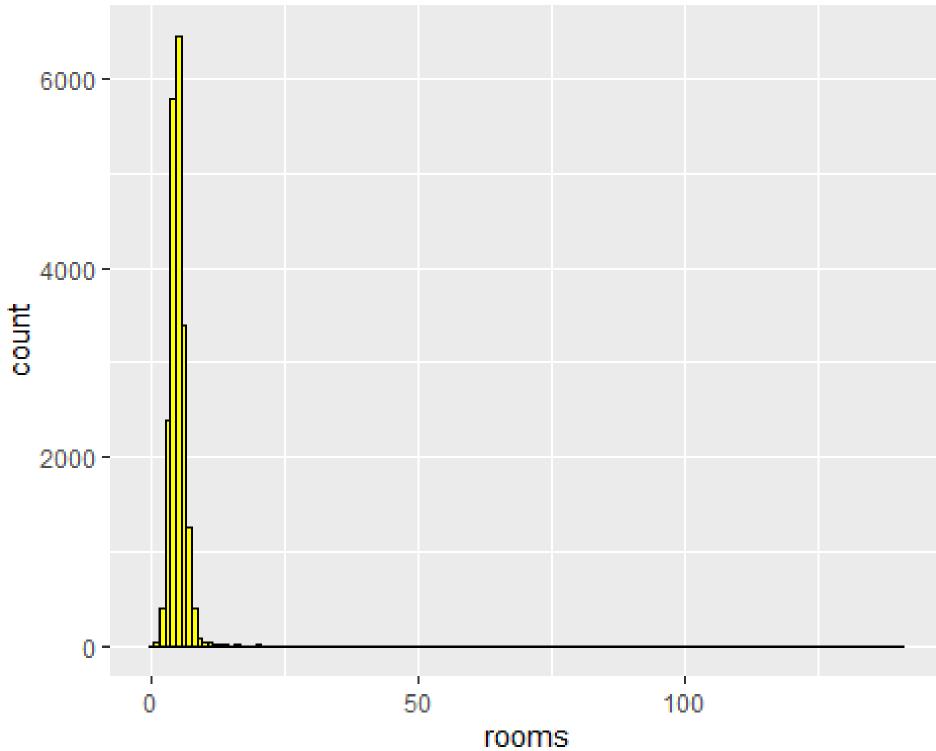
```
hist(cali_full$median_house_value,col = "dark green", border = 2)
```



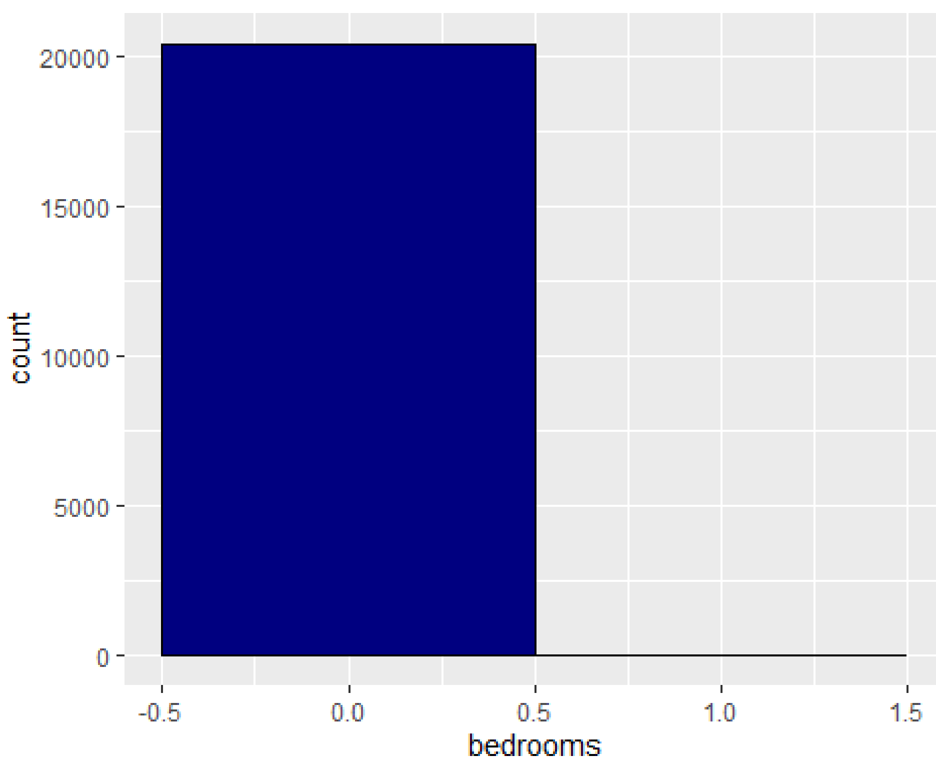
=> From the histogram we can conclude that the distribution is normal with Gaussian curve. The curve peaks at 2e+05 median_house_value and gradually reduces.

=> ggplot for the derived features from the original dataset

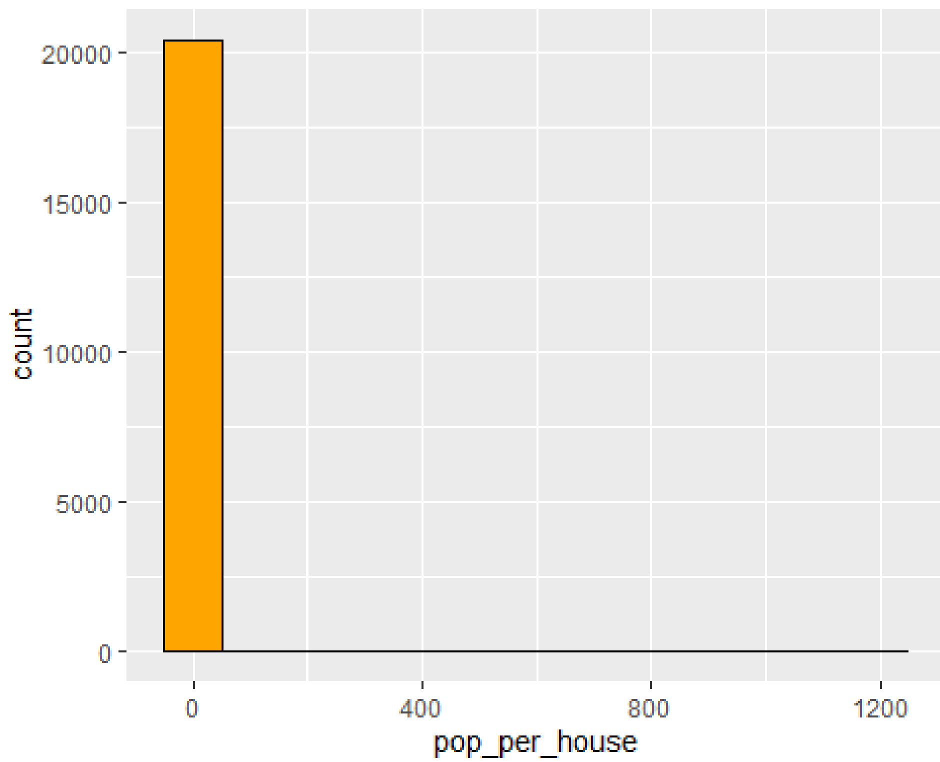
```
ggplot(cali_full, aes(x = rooms)) + geom_histogram(fill = "yellow", color =
"black", binwidth = 1)
```




```
ggplot(cali_full, aes(bedrooms)) + geom_histogram(color = "black", fill =
"navy blue", binwidth = 1)
```



```r
ggplot(cali_full, aes(x = pop_per_house)) + geom_histogram(fill = " orange",
color = "black", binwidth = 100)
```



The ggplot shows the histogram of rooms_per_household, bedroom_per_room and population_per_household. SO the rooms per house ranges from 0 to 10, bedrooms per room differs from 0 to 2 and population per household ranges from 0 to 100.

Developing a model using linear regression method

=> Creating model for each features of the dataset to find out the f stat and R^2 value

To find out the importance of each feature

```r
fit1 <- lm(median_house_value ~ longitude,cali_clean)
summary(fit1)

##
## Call:
## lm(formula = median_house_value ~ longitude, data = cali_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -201280  -86579  -26354   56598  301351
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -105885.6    48153.4  -2.199   0.0279 *
## longitude     -2615.6      402.7  -6.496 8.45e-11 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 115300 on 20431 degrees of freedom
## Multiple R-squared:  0.002061,	Adjusted R-squared:  0.002012
## F-statistic:  42.2 on 1 and 20431 DF,  p-value: 8.45e-11
```

```
fit2 <- lm(median_house_value ~ latitude , cali_clean)
summary(fit2)
```

```
##
## Call:
## lm(formula = median_house_value ~ latitude, data = cali_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -207211  -84082  -30082   57066  318746
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 485352.2    13352.6   36.35   <2e-16 ***
## latitude     -7815.4      374.1  -20.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114200 on 20431 degrees of freedom
## Multiple R-squared:  0.02092,	Adjusted R-squared:  0.02087
## F-statistic: 436.6 on 1 and 20431 DF,  p-value: < 2.2e-16
```

```
fit3 <- lm(median_house_value ~ housing_median_age, cali_clean)
summary(fit3)
```

```
##
## Call:
## lm(formula = median_house_value ~ housing_median_age, data = cali_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -214665  -85114  -25771   58290  319123
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        178926.58    1994.76    89.7   <2e-16 ***
## housing_median_age    975.72      63.77    15.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114800 on 20431 degrees of freedom
## Multiple R-squared:  0.01133,	Adjusted R-squared:  0.01128
## F-statistic: 234.1 on 1 and 20431 DF,  p-value: < 2.2e-16
```

```r
fit4 <- lm(median_house_value ~ total_rooms, cali_clean)
summary(fit4)
```

```
##
## Call:
## lm(formula = median_house_value ~ total_rooms, data = cali_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -311460  -86505  -26706   55721  311644
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.883e+05  1.254e+03  150.13   <2e-16 ***
## total_rooms 7.041e+00  3.663e-01   19.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114400 on 20431 degrees of freedom
## Multiple R-squared:  0.01777,    Adjusted R-squared:  0.01772
## F-statistic: 369.6 on 1 and 20431 DF,  p-value: < 2.2e-16
```

```r
fit5 <- lm(median_house_value ~ total_bedrooms, cali_clean)
summary(fit5)
```

```
##
## Call:
## lm(formula = median_house_value ~ total_bedrooms, data = cali_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -213629  -87479  -27730   57317  300444
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.995e+05  1.308e+03 152.568  < 2e-16 ***
## total_bedrooms 1.361e+01  1.914e+00   7.111 1.19e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 115300 on 20431 degrees of freedom
## Multiple R-squared:  0.002469,   Adjusted R-squared:  0.00242
## F-statistic: 50.56 on 1 and 20431 DF,  p-value: 1.192e-12
```

```r
fit6 <- lm(median_house_value ~ population, cali_clean)
summary(fit6)
```

```
##
## Call:
## lm(formula = median_house_value ~ population, data = cali_clean)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -195491  -86980  -26885   58117  308615
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.105e+05  1.297e+03 162.318  < 2e-16 ***
## population  -2.577e+00  7.124e-01  -3.617 0.000298 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 115400 on 20431 degrees of freedom
## Multiple R-squared:  0.0006401,  Adjusted R-squared:  0.0005912
## F-statistic: 13.09 on 1 and 20431 DF,  p-value: 0.0002983
```

```
fit7 <- lm(median_house_value ~ households, cali_clean)
summary(fit7)
```

```
##
## Call:
## lm(formula = median_house_value ~ households, data = cali_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -224153  -86962  -27933   56931  302903
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.971e+05  1.326e+03 148.644   <2e-16 ***
## households  1.959e+01  2.108e+00   9.295   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 115200 on 20431 degrees of freedom
## Multiple R-squared:  0.004211,   Adjusted R-squared:  0.004162
## F-statistic:  86.4 on 1 and 20431 DF,  p-value: < 2.2e-16
```

```
fit8 <- lm(median_house_value ~ median_income, cali_clean)
summary(fit8)
```

```
##
## Call:
## lm(formula = median_house_value ~ median_income, data = cali_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -541167  -55858  -16955   36895  434180
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    44906.4     1330.0   33.77   <2e-16 ***
```

```
## median_income  41837.1      308.4  135.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83740 on 20431 degrees of freedom
## Multiple R-squared:  0.4738, Adjusted R-squared:  0.4738
## F-statistic: 1.84e+04 on 1 and 20431 DF,  p-value: < 2.2e-16
```

From all these model we can conclude that the Latitude, Housing_median_age, Total_rooms and Median_income has high F-statistic and R^2 value. Since the correlation between median_income and median_house_value is higher we consider it for the creating the first model.

```
model1 <- lm(median_house_value ~ median_income, cali_full)
summary(model1)
```

```
##
## Call:
## lm(formula = median_house_value ~ median_income, data = cali_full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -541167  -55858  -16955   36895  434180
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    44906.4     1330.0   33.77   <2e-16 ***
## median_income  41837.1      308.4  135.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83740 on 20431 degrees of freedom
## Multiple R-squared:  0.4738, Adjusted R-squared:  0.4738
## F-statistic: 1.84e+04 on 1 and 20431 DF,  p-value: < 2.2e-16
```

Standard_error in this model is 83740 but the r^2 value and the f statistic value are higher which suits for a best model.

Consideration of Categorical data

Since there is only one categorical data and it is correlated with the median_house_value

We have to consider the ocean proximity by converting it into quantitative variable.

=> Converting the categorical variable into integer values which eases the method to find the correlation between the predictor and the output value.

```
cali_full$ocean_proximity =factor(cali_full$ocean_proximity, level = c("<1H
OCEAN","INLAND","ISLAND","NEAR BAY","NEAR OCEAN"), labels = c(1,2,3,4,5))
```

=> Model with all the features of cali_full dataset including the Categorical variable and derived features

```
model2 <- lm(median_house_value ~ ., cali_full)
summary(model2)

##
## Call:
## lm(formula = median_house_value ~ ., data = cali_full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -554800  -42684  -10402   28926  779971
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.351e+06  8.870e+04 -26.507  < 2e-16 ***
## longitude         -2.775e+04  1.027e+03 -27.007  < 2e-16 ***
## latitude          -2.653e+04  1.015e+03 -26.139  < 2e-16 ***
## housing_median_age  1.079e+03  4.386e+01  24.610  < 2e-16 ***
## total_rooms       -7.308e+00  8.120e-01  -9.000  < 2e-16 ***
## total_bedrooms     8.545e+01  7.201e+00  11.866  < 2e-16 ***
## population        -3.847e+01  1.108e+00 -34.707  < 2e-16 ***
## households         7.406e+01  8.239e+00   8.989  < 2e-16 ***
## median_income      3.862e+04  3.503e+02 110.251  < 2e-16 ***
## ocean_proximity2  -3.905e+04  1.743e+03 -22.407  < 2e-16 ***
## ocean_proximity3   1.523e+05  3.071e+04   4.959 7.15e-07 ***
## ocean_proximity4  -3.975e+03  1.911e+03  -2.080   0.0376 *
## ocean_proximity5   3.932e+03  1.569e+03   2.507   0.0122 *
## rooms              1.670e+03  2.447e+02   6.823 9.18e-12 ***
## bedrooms           7.508e+04  3.962e+04   1.895   0.0581 .
## pop_per_house      6.488e+01  4.741e+01   1.368   0.1712
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68580 on 20417 degrees of freedom
## Multiple R-squared:  0.6473, Adjusted R-squared:  0.6471
## F-statistic:  2499 on 15 and 20417 DF,  p-value: < 2.2e-16
```

This model concludes with F-statistic - 2499 and R^2 - 0.6473 values and relatively with a lesser Standard error of 68580. Hence this model is comparitively a good fit than other models.This model includes the categorical variables and the derived features.

=> The model with longitude and latitude features.

```
model3 <- lm(median_house_value ~ longitude+latitude, cali_full)
summary(model3)

##
## Call:
## lm(formula = median_house_value ~ longitude + latitude, data = cali_full)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -315933  -67617  -22892   46127  483223
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5822326.7    82532.4  -70.55   <2e-16 ***
## longitude     -71135.6      921.3  -77.21   <2e-16 ***
## latitude      -69500.8      864.0  -80.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 100500 on 20430 degrees of freedom
## Multiple R-squared:  0.2421, Adjusted R-squared:  0.242
## F-statistic:  3263 on 2 and 20430 DF,  p-value: < 2.2e-16
```

This model includes just the logitude and latitude feature of the dataset. The F-statistic increase from 2499 to 3263 but the R^2 value is reduced from 0.6473 to 0.2421.. And the Standard error rate is also increased to 100500.

=> The model with all the features except longitude and latitude and derived features.

```
model4 <- lm(median_house_value ~
housing_median_age+households+total_rooms+population+total_bedrooms,
cali_full)
summary(model4)
```

```
##
## Call:
## lm(formula = median_house_value ~ housing_median_age + households +
##     total_rooms + population + total_bedrooms, data = cali_full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -455244  -78410  -15962   54588 1177669
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.475e+05  2.462e+03   59.90   <2e-16 ***
## housing_median_age  1.603e+03  6.337e+01   25.29   <2e-16 ***
## households          2.848e+02  1.115e+01   25.54   <2e-16 ***
## total_rooms         4.592e+01  9.658e-01   47.55   <2e-16 ***
## population         -6.490e+01  1.604e+00  -40.47   <2e-16 ***
## total_bedrooms     -2.926e+02  9.700e+00  -30.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105900 on 20427 degrees of freedom
## Multiple R-squared:  0.1588, Adjusted R-squared:  0.1586
## F-statistic:   771 on 5 and 20427 DF,  p-value: < 2.2e-16
```

This model includes all other features except the logitude and latitude feature of the data set. The F-statistic has reduced to 771 and the R^2 value is also reduced to 0.1588. This model is not a good fit for linear regression. Hence some features needed to be altered.

=> The model with all features except derived features.

```
model5<- lm(median_house_value ~
longitude+latitude+housing_median_age+total_rooms+population+median_income+oc
ean_proximity, cali_full)
summary(model5)

##
## Call:
## lm(formula = median_house_value ~ longitude + latitude +
housing_median_age +
##     total_rooms + population + median_income + ocean_proximity,
##     data = cali_full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -500858  -45222  -11829   30045  506447
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.340e+06  8.965e+04 -26.105  < 2e-16 ***
## longitude          -2.812e+04  1.040e+03 -27.041  < 2e-16 ***
## latitude           -2.702e+04  1.028e+03 -26.292  < 2e-16 ***
## housing_median_age  1.004e+03  4.511e+01  22.245  < 2e-16 ***
## total_rooms         1.522e+01  4.969e-01  30.639  < 2e-16 ***
## population         -2.565e+01  9.327e-01 -27.504  < 2e-16 ***
## median_income       3.363e+04  3.011e+02 111.695  < 2e-16 ***
## ocean_proximity2   -4.688e+04  1.782e+03 -26.307  < 2e-16 ***
## ocean_proximity3    1.526e+05  3.165e+04   4.822 1.43e-06 ***
## ocean_proximity4   -1.390e+03  1.969e+03  -0.706   0.4802
## ocean_proximity5    3.935e+03  1.616e+03   2.435   0.0149 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70710 on 20422 degrees of freedom
## Multiple R-squared:  0.625,  Adjusted R-squared:  0.6248
## F-statistic:  3403 on 10 and 20422 DF,  p-value: < 2.2e-16
```

Including all the features except the features we splitted - rooms_per_house, bedrooms_per_room, population_per_house. The f-statistics is 3403 and R^2 value is 0.625. The std error is 70710. The F-statistic and R^2 values are higher which can prove this to be a best model. But the Standard error is comapritively higher at 70710.

When compared the R^2 ans F-statistic we can conclude that the both is higher if we consider all the features including the feaures we extracted from the existing predictors

i.e. model2 has comparitively higher F-statistic and R^2 value and lesser Standard error.

=> Forward and Backward selection for creating a model.

=> Extracting features for forward and backward selection

```
long <- cali_full$longitude
lat <- cali_full$latitude
age <- cali_full$housing_median_age
rooms <- cali_full$total_rooms
bed <- cali_full$total_bedrooms
pop <- cali_full$population
house <- cali_full$households
inc <- cali_full$median_income
ocean <- cali_full$ocean_proximity
```

=> Forward selection based on AIC.

```
fit.forward <- step(lm(cali_full$median_house_value ~ 1),
                    scope = list(upper = ~ long + lat + age + rooms + bed +
pop + house +inc+ocean),direction = "forward")

## Start:  AIC=476354.2
## cali_full$median_house_value ~ 1
##
##          Df  Sum of Sq        RSS    AIC
## + inc     1 1.2901e+14 1.4326e+14 463235
## + ocean   4 6.4787e+13 2.0748e+14 470810
## + lat     1 5.6958e+12 2.6657e+14 475924
## + rooms   1 4.8374e+12 2.6743e+14 475990
## + age     1 3.0842e+12 2.6918e+14 476123
## + house   1 1.1466e+12 2.7112e+14 476270
## + bed     1 6.7214e+11 2.7159e+14 476306
## + long    1 5.6114e+11 2.7170e+14 476314
## + pop     1 1.7427e+11 2.7209e+14 476343
## <none>                2.7226e+14 476354
##
## Step:  AIC=463235.5
## cali_full$median_house_value ~ inc
##
##          Df  Sum of Sq        RSS    AIC
## + ocean   4 3.1118e+13 1.1214e+14 458239
## + age     1 9.7438e+12 1.3351e+14 461798
```

```
## + lat    1 2.2109e+12 1.4105e+14 462920
## + house  1 8.4322e+11 1.4241e+14 463117
## + bed    1 8.2372e+11 1.4243e+14 463120
## + long   1 3.2780e+11 1.4293e+14 463191
## + pop    1 2.2585e+11 1.4303e+14 463205
## <none>                1.4326e+14 463235
## + rooms  1 2.4139e+09 1.4325e+14 463237
##
## Step:  AIC=458239.3
## cali_full$median_house_value ~ inc + ocean
##
##         Df  Sum of Sq        RSS    AIC
## + age    1 2.4385e+12 1.0970e+14 457792
## + bed    1 7.8257e+11 1.1136e+14 458098
## + house  1 5.6530e+11 1.1157e+14 458138
## + long   1 3.0098e+11 1.1184e+14 458186
## + pop    1 2.3972e+11 1.1190e+14 458198
## + rooms  1 1.4979e+11 1.1199e+14 458214
## <none>                1.1214e+14 458239
## + lat    1 1.0523e+09 1.1214e+14 458241
##
## Step:  AIC=457792.1
## cali_full$median_house_value ~ inc + ocean + age
##
##         Df  Sum of Sq        RSS    AIC
## + bed    1 2.2820e+12 1.0742e+14 457365
## + house  1 1.7976e+12 1.0790e+14 457456
## + rooms  1 9.6652e+11 1.0873e+14 457613
## + long   1 2.4961e+11 1.0945e+14 457748
## <none>                1.0970e+14 457792
## + lat    1 7.5363e+08 1.0970e+14 457794
## + pop    1 7.1702e+07 1.0970e+14 457794
##
## Step:  AIC=457364.5
## cali_full$median_house_value ~ inc + ocean + age + bed
##
##         Df  Sum of Sq        RSS    AIC
## + pop    1 6.9098e+12 1.0051e+14 456008
## + rooms  1 2.0006e+12 1.0542e+14 456982
## + house  1 4.2757e+11 1.0699e+14 457285
## + long   1 3.4912e+11 1.0707e+14 457300
## + lat    1 1.7758e+10 1.0740e+14 457363
## <none>                1.0742e+14 457365
##
## Step:  AIC=456008
## cali_full$median_house_value ~ inc + ocean + age + bed + pop
##
##         Df  Sum of Sq        RSS    AIC
## + house  1 5.6424e+11 9.9944e+13 455895
## + rooms  1 4.7980e+11 1.0003e+14 455912
```

```
## + long   1 2.6979e+11 1.0024e+14 455955
## <none>                1.0051e+14 456008
## + lat    1 3.2783e+08 1.0051e+14 456010
##
## Step:  AIC=455895
## cali_full$median_house_value ~ inc + ocean + age + bed + pop +
##     house
##
##          Df  Sum of Sq        RSS    AIC
## + rooms  1 4.2825e+11 9.9515e+13 455809
## + long   1 1.9564e+11 9.9748e+13 455857
## <none>                9.9944e+13 455895
## + lat    1 4.9236e+09 9.9939e+13 455896
##
## Step:  AIC=455809.2
## cali_full$median_house_value ~ inc + ocean + age + bed + pop +
##     house + rooms
##
##         Df  Sum of Sq        RSS    AIC
## + long  1 2.2779e+11 9.9288e+13 455764
## <none>               9.9515e+13 455809
## + lat   1 5.2044e+08 9.9515e+13 455811
##
## Step:  AIC=455764.4
## cali_full$median_house_value ~ inc + ocean + age + bed + pop +
##     house + rooms + long
##
##         Df  Sum of Sq        RSS    AIC
## + lat   1 3.0323e+12 9.6255e+13 455133
## <none>               9.9288e+13 455764
##
## Step:  AIC=455132.6
## cali_full$median_house_value ~ inc + ocean + age + bed + pop +
##     house + rooms + long + lat
```

```r
summary(fit.forward)
```

```
##
## Call:
## lm(formula = cali_full$median_house_value ~ inc + ocean + age +
##     bed + pop + house + rooms + long + lat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -556980  -42683  -10497   28765  779052
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.270e+06  8.801e+04 -25.791  < 2e-16 ***
## inc          3.926e+04  3.380e+02 116.151  < 2e-16 ***
```

```
## ocean2      -3.928e+04  1.744e+03 -22.522  < 2e-16 ***
## ocean3       1.529e+05  3.074e+04   4.974 6.62e-07 ***
## ocean4      -3.954e+03  1.913e+03  -2.067  0.03879 *
## ocean5       4.278e+03  1.570e+03   2.726  0.00642 **
## age          1.073e+03  4.389e+01  24.439  < 2e-16 ***
## bed          1.006e+02  6.869e+00  14.640  < 2e-16 ***
## pop         -3.797e+01  1.076e+00 -35.282  < 2e-16 ***
## house        4.962e+01  7.451e+00   6.659 2.83e-11 ***
## rooms       -6.193e+00  7.915e-01  -7.825 5.32e-15 ***
## long        -2.681e+04  1.020e+03 -26.296  < 2e-16 ***
## lat         -2.548e+04  1.005e+03 -25.363  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68660 on 20420 degrees of freedom
## Multiple R-squared:  0.6465, Adjusted R-squared:  0.6463
## F-statistic:  3112 on 12 and 20420 DF,  p-value: < 2.2e-16
```

=> Backward elimination based on AIC.

```
fit.backward <-step(lm(cali_full$median_house_value  ~ long + lat + age +
rooms + bed + pop + house +inc+ocean),
                     scope = list(lower = ~1),direction = "backward")

## Start:  AIC=455132.6
## cali_full$median_house_value ~ long + lat + age + rooms + bed +
##     pop + house + inc + ocean
##
##         Df  Sum of Sq        RSS    AIC
## <none>                 9.6255e+13 455133
## - house  1 2.0901e+11 9.6464e+13 455175
## - rooms  1 2.8863e+11 9.6544e+13 455192
## - bed    1 1.0103e+12 9.7266e+13 455344
## - ocean  4 2.6007e+12 9.8856e+13 455669
## - age    1 2.8154e+12 9.9071e+13 455720
## - lat    1 3.0323e+12 9.9288e+13 455764
## - long   1 3.2595e+12 9.9515e+13 455811
## - pop    1 5.8679e+12 1.0212e+14 456340
## - inc    1 6.3594e+13 1.5985e+14 465495

summary(fit.backward)

##
## Call:
## lm(formula = cali_full$median_house_value ~ long + lat + age +
##     rooms + bed + pop + house + inc + ocean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -556980  -42683  -10497   28765  779052
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2.270e+06  8.801e+04 -25.791  < 2e-16 ***
## long        -2.681e+04  1.020e+03 -26.296  < 2e-16 ***
## lat         -2.548e+04  1.005e+03 -25.363  < 2e-16 ***
## age          1.073e+03  4.389e+01  24.439  < 2e-16 ***
## rooms       -6.193e+00  7.915e-01  -7.825 5.32e-15 ***
## bed          1.006e+02  6.869e+00  14.640  < 2e-16 ***
## pop         -3.797e+01  1.076e+00 -35.282  < 2e-16 ***
## house        4.962e+01  7.451e+00   6.659 2.83e-11 ***
## inc          3.926e+04  3.380e+02 116.151  < 2e-16 ***
## ocean2      -3.928e+04  1.744e+03 -22.522  < 2e-16 ***
## ocean3       1.529e+05  3.074e+04   4.974 6.62e-07 ***
## ocean4      -3.954e+03  1.913e+03  -2.067  0.03879 *  
## ocean5       4.278e+03  1.570e+03   2.726  0.00642 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 68660 on 20420 degrees of freedom
## Multiple R-squared:  0.6465, Adjusted R-squared:  0.6463 
## F-statistic:  3112 on 12 and 20420 DF,  p-value: < 2.2e-16
```

=> Both forward/backward selection based on AIC.

```
fit.both <-step(lm(cali_full$median_house_value ~ 1),
                 scope = list(lower = ~1,
                              upper = ~ long + lat + age + rooms + bed + pop
+ house +inc+ocean),
                 direction = "both")
```

```
## Start:  AIC=476354.2
## cali_full$median_house_value ~ 1
## 
##           Df  Sum of Sq        RSS    AIC
## + inc      1 1.2901e+14 1.4326e+14 463235
## + ocean    4 6.4787e+13 2.0748e+14 470810
## + lat      1 5.6958e+12 2.6657e+14 475924
## + rooms    1 4.8374e+12 2.6743e+14 475990
## + age      1 3.0842e+12 2.6918e+14 476123
## + house    1 1.1466e+12 2.7112e+14 476270
## + bed      1 6.7214e+11 2.7159e+14 476306
## + long     1 5.6114e+11 2.7170e+14 476314
## + pop      1 1.7427e+11 2.7209e+14 476343
## <none>                  2.7226e+14 476354
## 
## Step:  AIC=463235.5
## cali_full$median_house_value ~ inc
## 
##           Df  Sum of Sq        RSS    AIC
## + ocean    4 3.1118e+13 1.1214e+14 458239
## + age      1 9.7438e+12 1.3351e+14 461798
```

```
## + lat    1 2.2109e+12 1.4105e+14 462920
## + house  1 8.4322e+11 1.4241e+14 463117
## + bed    1 8.2372e+11 1.4243e+14 463120
## + long   1 3.2780e+11 1.4293e+14 463191
## + pop    1 2.2585e+11 1.4303e+14 463205
## <none>                1.4326e+14 463235
## + rooms  1 2.4139e+09 1.4325e+14 463237
## - inc    1 1.2901e+14 2.7226e+14 476354
##
## Step:  AIC=458239.3
## cali_full$median_house_value ~ inc + ocean
##
##         Df  Sum of Sq        RSS    AIC
## + age    1 2.4385e+12 1.0970e+14 457792
## + bed    1 7.8257e+11 1.1136e+14 458098
## + house  1 5.6530e+11 1.1157e+14 458138
## + long   1 3.0098e+11 1.1184e+14 458186
## + pop    1 2.3972e+11 1.1190e+14 458198
## + rooms  1 1.4979e+11 1.1199e+14 458214
## <none>                1.1214e+14 458239
## + lat    1 1.0523e+09 1.1214e+14 458241
## - ocean  4 3.1118e+13 1.4326e+14 463235
## - inc    1 9.5339e+13 2.0748e+14 470810
##
## Step:  AIC=457792.1
## cali_full$median_house_value ~ inc + ocean + age
##
##         Df  Sum of Sq        RSS    AIC
## + bed    1 2.2820e+12 1.0742e+14 457365
## + house  1 1.7976e+12 1.0790e+14 457456
## + rooms  1 9.6652e+11 1.0873e+14 457613
## + long   1 2.4961e+11 1.0945e+14 457748
## <none>                1.0970e+14 457792
## + lat    1 7.5363e+08 1.0970e+14 457794
## + pop    1 7.1702e+07 1.0970e+14 457794
## - age    1 2.4385e+12 1.1214e+14 458239
## - ocean  4 2.3813e+13 1.3351e+14 461798
## - inc    1 9.7680e+13 2.0738e+14 470802
##
## Step:  AIC=457364.5
## cali_full$median_house_value ~ inc + ocean + age + bed
##
##         Df  Sum of Sq        RSS    AIC
## + pop    1 6.9098e+12 1.0051e+14 456008
## + rooms  1 2.0006e+12 1.0542e+14 456982
## + house  1 4.2757e+11 1.0699e+14 457285
## + long   1 3.4912e+11 1.0707e+14 457300
## + lat    1 1.7758e+10 1.0740e+14 457363
## <none>                1.0742e+14 457365
## - bed    1 2.2820e+12 1.0970e+14 457792
```

```
## - age    1 3.9379e+12 1.1136e+14 458098
## - ocean  4 2.1987e+13 1.2940e+14 461162
## - inc    1 9.9436e+13 2.0685e+14 470752
##
## Step:  AIC=456008
## cali_full$median_house_value ~ inc + ocean + age + bed + pop
##
##         Df  Sum of Sq        RSS    AIC
## + house  1 5.6424e+11 9.9944e+13 455895
## + rooms  1 4.7980e+11 1.0003e+14 455912
## + long   1 2.6979e+11 1.0024e+14 455955
## <none>               1.0051e+14 456008
## + lat    1 3.2783e+08 1.0051e+14 456010
## - age    1 3.7682e+12 1.0428e+14 456758
## - pop    1 6.9098e+12 1.0742e+14 457365
## - bed    1 9.1917e+12 1.0970e+14 457794
## - ocean  4 2.2385e+13 1.2289e+14 460109
## - inc    1 1.0002e+14 2.0053e+14 470119
##
## Step:  AIC=455895
## cali_full$median_house_value ~ inc + ocean + age + bed + pop +
##     house
##
##         Df  Sum of Sq        RSS    AIC
## + rooms  1 4.2825e+11 9.9515e+13 455809
## + long   1 1.9564e+11 9.9748e+13 455857
## <none>               9.9944e+13 455895
## + lat    1 4.9236e+09 9.9939e+13 455896
## - bed    1 3.2220e+11 1.0027e+14 455959
## - house  1 5.6424e+11 1.0051e+14 456008
## - age    1 3.6169e+12 1.0356e+14 456619
## - pop    1 7.0464e+12 1.0699e+14 457285
## - ocean  4 2.1209e+13 1.2115e+14 459819
## - inc    1 9.8230e+13 1.9817e+14 469880
##
## Step:  AIC=455809.2
## cali_full$median_house_value ~ inc + ocean + age + bed + pop +
##     house + rooms
##
##         Df  Sum of Sq        RSS    AIC
## + long   1 2.2779e+11 9.9288e+13 455764
## <none>               9.9515e+13 455809
## + lat    1 5.2044e+08 9.9515e+13 455811
## - rooms  1 4.2825e+11 9.9944e+13 455895
## - house  1 5.1268e+11 1.0003e+14 455912
## - bed    1 6.7328e+11 1.0019e+14 455945
## - age    1 3.4718e+12 1.0299e+14 456508
## - pop    1 5.6343e+12 1.0515e+14 456933
## - ocean  4 1.8445e+13 1.1796e+14 459276
## - inc    1 6.8970e+13 1.6849e+14 466566
```

```
##
## Step:  AIC=455764.4
## cali_full$median_house_value ~ inc + ocean + age + bed + pop +
##     house + rooms + long
##
##         Df  Sum of Sq        RSS    AIC
## + lat    1 3.0323e+12 9.6255e+13 455133
## <none>                9.9288e+13 455764
## - long   1 2.2779e+11 9.9515e+13 455809
## - house  1 4.3550e+11 9.9723e+13 455852
## - rooms  1 4.6040e+11 9.9748e+13 455857
## - bed    1 7.5484e+11 1.0004e+14 455917
## - age    1 3.4602e+12 1.0275e+14 456462
## - pop    1 5.3940e+12 1.0468e+14 456843
## - ocean  4 1.8655e+13 1.1794e+14 459274
## - inc    1 6.9023e+13 1.6831e+14 466547
##
## Step:  AIC=455132.6
## cali_full$median_house_value ~ inc + ocean + age + bed + pop +
##     house + rooms + long + lat
##
##         Df  Sum of Sq        RSS    AIC
## <none>                9.6255e+13 455133
## - house  1 2.0901e+11 9.6464e+13 455175
## - rooms  1 2.8863e+11 9.6544e+13 455192
## - bed    1 1.0103e+12 9.7266e+13 455344
## - ocean  4 2.6007e+12 9.8856e+13 455669
## - age    1 2.8154e+12 9.9071e+13 455720
## - lat    1 3.0323e+12 9.9288e+13 455764
## - long   1 3.2595e+12 9.9515e+13 455811
## - pop    1 5.8679e+12 1.0212e+14 456340
## - inc    1 6.3594e+13 1.5985e+14 465495
```

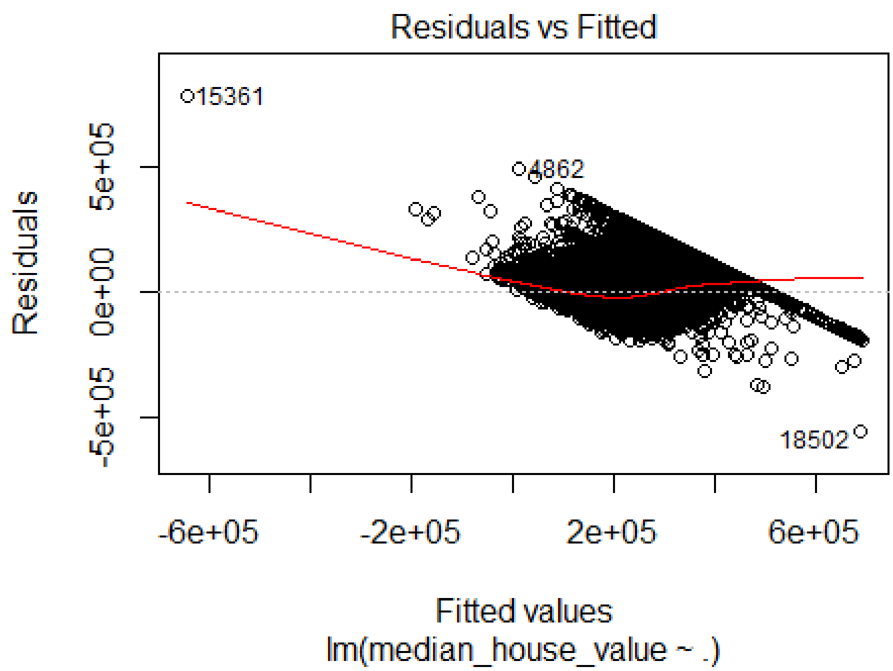
```
summary(fit.both)
```

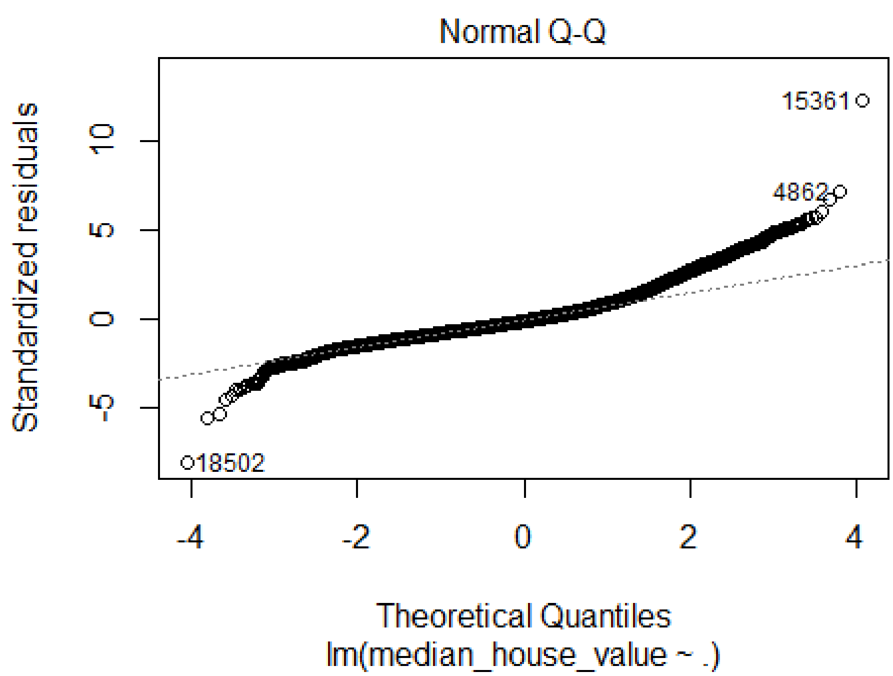
```
##
## Call:
## lm(formula = cali_full$median_house_value ~ inc + ocean + age +
##     bed + pop + house + rooms + long + lat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -556980  -42683  -10497   28765  779052
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.270e+06  8.801e+04 -25.791  < 2e-16 ***
## inc          3.926e+04  3.380e+02 116.151  < 2e-16 ***
## ocean2      -3.928e+04  1.744e+03 -22.522  < 2e-16 ***
## ocean3       1.529e+05  3.074e+04   4.974 6.62e-07 ***
```

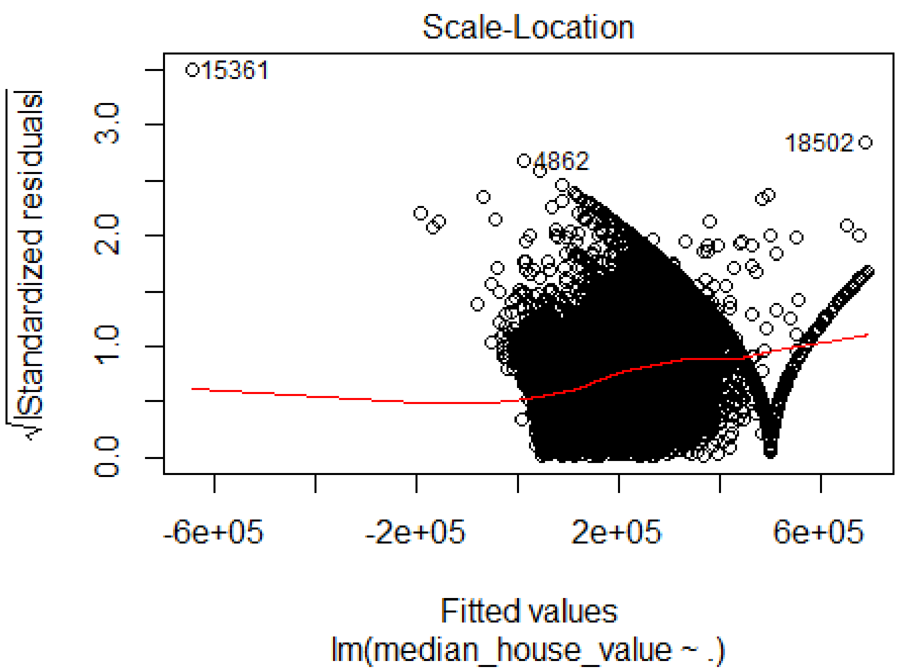
```
## ocean4      -3.954e+03  1.913e+03  -2.067  0.03879 *
## ocean5       4.278e+03  1.570e+03   2.726  0.00642 **
## age          1.073e+03  4.389e+01  24.439  < 2e-16 ***
## bed          1.006e+02  6.869e+00  14.640  < 2e-16 ***
## pop         -3.797e+01  1.076e+00 -35.282  < 2e-16 ***
## house        4.962e+01  7.451e+00   6.659 2.83e-11 ***
## rooms       -6.193e+00  7.915e-01  -7.825 5.32e-15 ***
## long        -2.681e+04  1.020e+03 -26.296  < 2e-16 ***
## lat         -2.548e+04  1.005e+03 -25.363  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68660 on 20420 degrees of freedom
## Multiple R-squared:  0.6465, Adjusted R-squared:  0.6463
## F-statistic:  3112 on 12 and 20420 DF,  p-value: < 2.2e-16
```

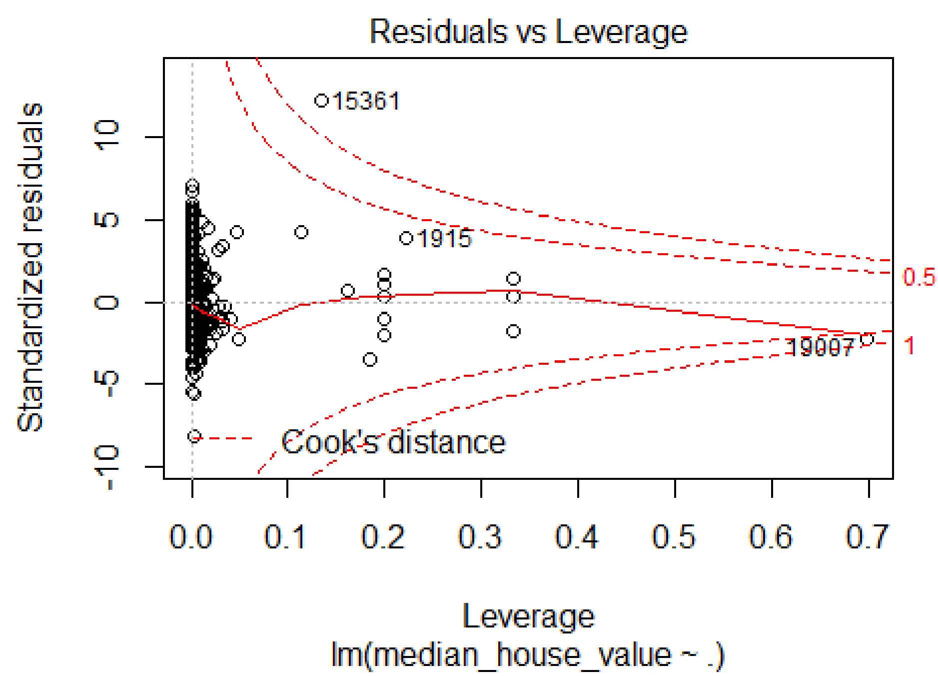
Since model2 is comparitively the best model, visualizing it using plot function which returns the following graphs - the Residual vs Fitted, Normal QQ plot, Scale-Location, Residuals vs leverage.

```
plot(model2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(median_house_value ~ .)
15361
4862
18502
5e+05
0e+00
-5e+05
-6e+05
-2e+05
2e+05
6e+05


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(median_house_value ~ .)
15361
4862
18502
10
5
0
-5
-4
-2
0
2
4


Scale-Location
15361
4862
18502
√|Standardized residuals|
0.0
1.0
2.0
3.0
-6e+05
-2e+05
2e+05
6e+05
Fitted values
lm(median_house_value ~ .)


Residuals vs Leverage
15361
1915
19007
0.5
1
Cook's distance
Standardized residuals
-10
-5
0
5
10
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Leverage
lm(median_house_value ~ .)

Predicting the outcome value or Testing the model

Obtaining the predicted value by using one of the value of the trained data.

```
df <- as.data.frame(cali_full[1,])
df

##   longitude latitude housing_median_age total_rooms total_bedrooms
population
## 1   -122.23    37.88                 41         880            129
322
##   households median_income median_house_value ocean_proximity rooms
bedrooms
## 1        126        8.3252             452600               4     6
0
##   pop_per_house
## 1             2

df[,"median_house_value"] <- NULL
df

##   longitude latitude housing_median_age total_rooms total_bedrooms
population
## 1   -122.23    37.88                 41         880            129
322
##   households median_income ocean_proximity rooms bedrooms pop_per_house
## 1        126        8.3252               4     6        0             2
```

From the df variable we can see that the actual value is 452600. Then the median_house_value is eliminated and passed the data for testing the model.

```
predict.lm(model2,df )

##        1
## 409162.3
```

The predicted value 409162.3 concludes that the error rate between the Actual - Predicted output is very small. Hence this model is a best fitted linear regression model.