Calfornia Housing dataset - Linear regression

Team

Harichandana Epuri (hxe170000), Akhila Kancharana (axk180025)

loading the California Housing Dataset

```
california_data <-
read.csv("https://utdallas.edu/~axk180025/CaliforniaHousingPrice/housing.csv"
)
```

Dimensions of the dataset

```
dim(california_data)

## [1] 20640    10
```

Columns in the dataset

```
colnames(california_data)

##  [1] "longitude"          "latitude"           "housing_median_age"
##  [4] "total_rooms"        "total_bedrooms"     "population"
##  [7] "households"         "median_income"      "median_house_value"
## [10] "ocean_proximity"
```

Structure of the dataset

```
str(california_data)

## 'data.frame':    20640 obs. of  10 variables:
##  $ longitude         : num  -122 -122 -122 -122 -122 ...
##  $ latitude          : num  37.9 37.9 37.9 37.9 37.9 ...
##  $ housing_median_age: num  41 21 52 52 52 52 52 52 42 52 ...
##  $ total_rooms       : num  880 7099 1467 1274 1627 ...
##  $ total_bedrooms    : num  129 1106 190 235 280 ...
##  $ population        : num  322 2401 496 558 565 ...
##  $ households        : num  126 1138 177 219 259 ...
##  $ median_income     : num  8.33 8.3 7.26 5.64 3.85 ...
##  $ median_house_value: num  452600 358500 352100 341300 342200 ...
##  $ ocean_proximity   : Factor w/ 5 levels "<1H OCEAN","INLAND",..: 4 4 4 4
4 4 4 4 4 4 ...
```

Summary of the dataset

```
summary(california_data)

##    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
##
```

Data preprocessing

Finding the NA values in the dataset in each column

```
NA_count_of_each_col<-sapply(california_data,function(x) sum(is.na(x)==TRUE))
NA_count_of_each_col
```

```
##          longitude           latitude housing_median_age
total_rooms
##                  0                  0                  0
0
##     total_bedrooms         population         households
median_income
##                207                  0                  0
0
## median_house_value    ocean_proximity
##                  0                  0
```

Removing the rows with the total_bedrooms column as NA

```
california_data_clean <- na.omit(california_data)
```

Some statistics and plots found from the dataset

Finding the number of houses in each proximity

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

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

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

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

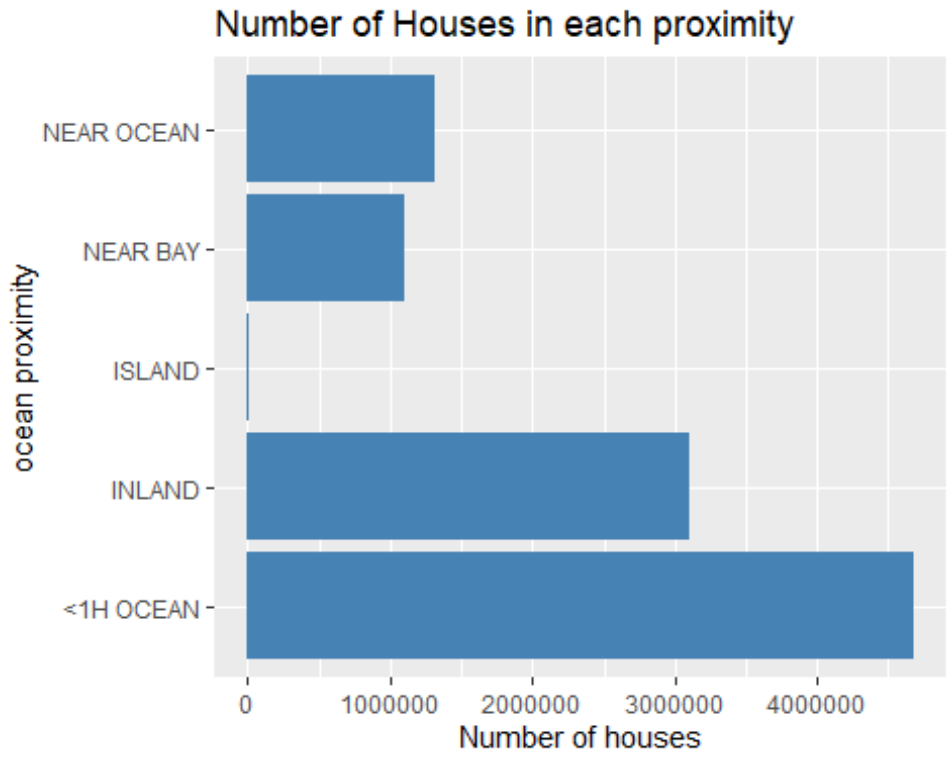
```
options(scipen=10000)
houses_each_proximity <- california_data_clean %>% group_by(ocean_proximity)
%>% summarise(number_of_houses = sum(households))
houses_each_proximity
```

```
## # A tibble: 5 x 2
##   ocean_proximity number_of_houses
##   <fct>                      <dbl>
## 1 <1H OCEAN                4674364
## 2 INLAND                   3105133
## 3 ISLAND                      1383
## 4 NEAR BAY                 1106026
## 5 NEAR OCEAN               1318018
```

```
format(houses_each_proximity, scientific = FALSE)
```

```
## [1] "# A tibble: 5 x 2"                  "  ocean_proximity
number_of_houses"
## [3] "  <fct>                      <dbl>" "1 <1H OCEAN
4674364"
## [5] "2 INLAND                   3105133" "3 ISLAND
1383"
## [7] "4 NEAR BAY                 1106026" "5 NEAR OCEAN
1318018"
```

```
ggplot(houses_each_proximity, aes(ocean_proximity, number_of_houses),fill =
ocean_proximity)+   geom_bar(stat = "identity" , fill="steelblue")+
theme(legend.position = "none")+   labs(x = "ocean proximity", y = "Number of
houses",        title = "Number of Houses in each proximity")+ coord_flip()
```

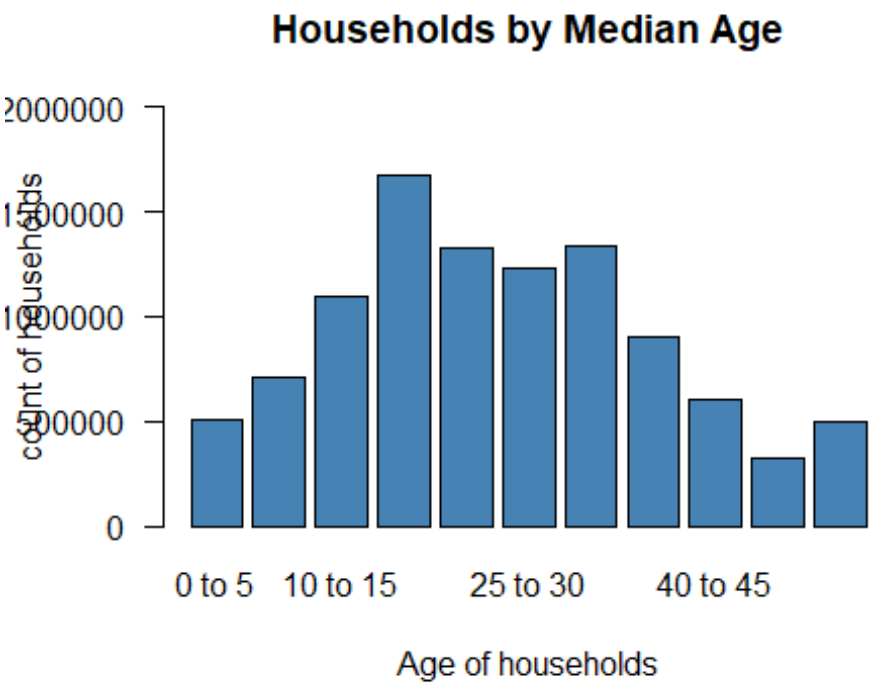


Sumamry

It is inferred from the above graph that there are more households (4,730,118 households) situated 1 hour away from ocean (<1H OCEAN) as compared to other areas. It is followed by households (3,127,759) that are situated inland (INLAND) and there are least number of households in island.

Households by median_age

```
age <- c(0,5,10,15,20,25,30,35,40,45,50,55)
header <- c("0 to 5","5 to 10", "10 to 15", "15 to 20", "20 to 25","25 to
30",
            "30 to 35","35 to 40","40 to 45","45 to 50", "50 to 55")
partition_data <- transform(california_data_clean, age_cut =
cut(housing_median_age, age, labels = header))

barplot( height =
rowsum(partition_data$households,partition_data$age_cut)[,1], las = 1, col =
"steelblue", main = "Households by Median Age", ylab = "count of households",
xlab = "Age of households"
         , ylim = c(0,2000000))
```

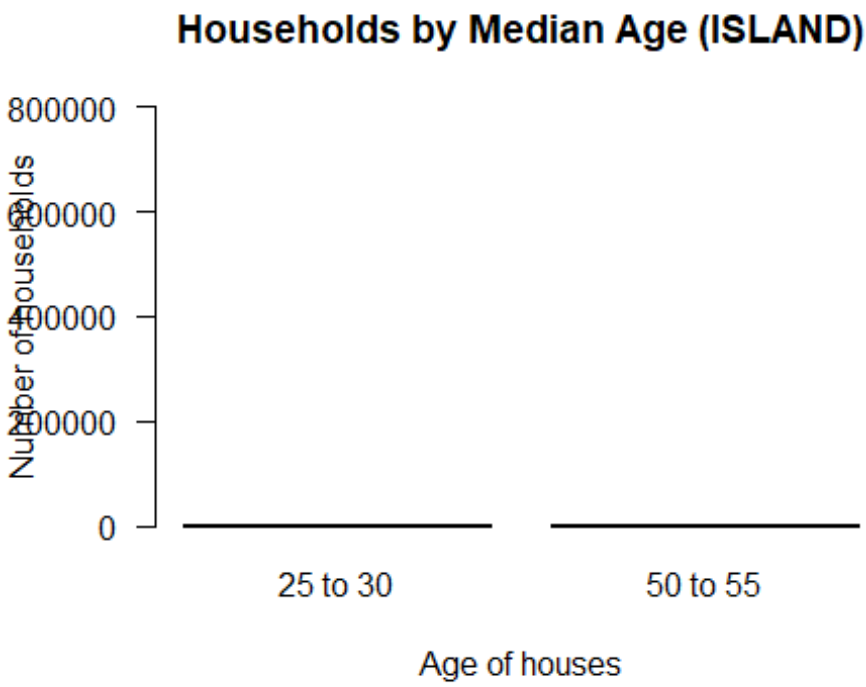


Summary

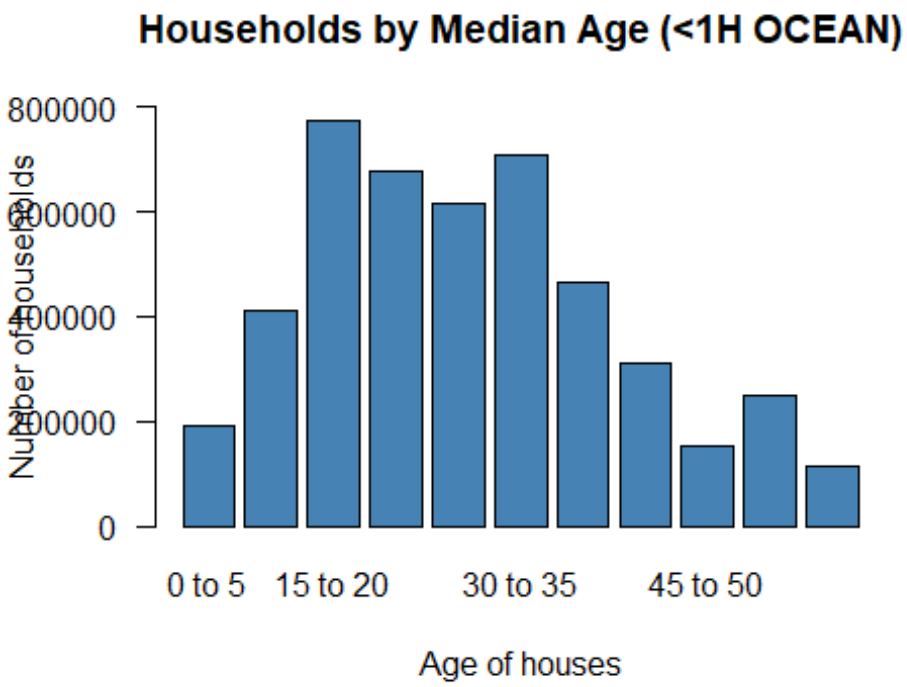
The above grapoh looks like a normal distribution and most of the households are 15-35 years old.

Households by median age in all the proximities

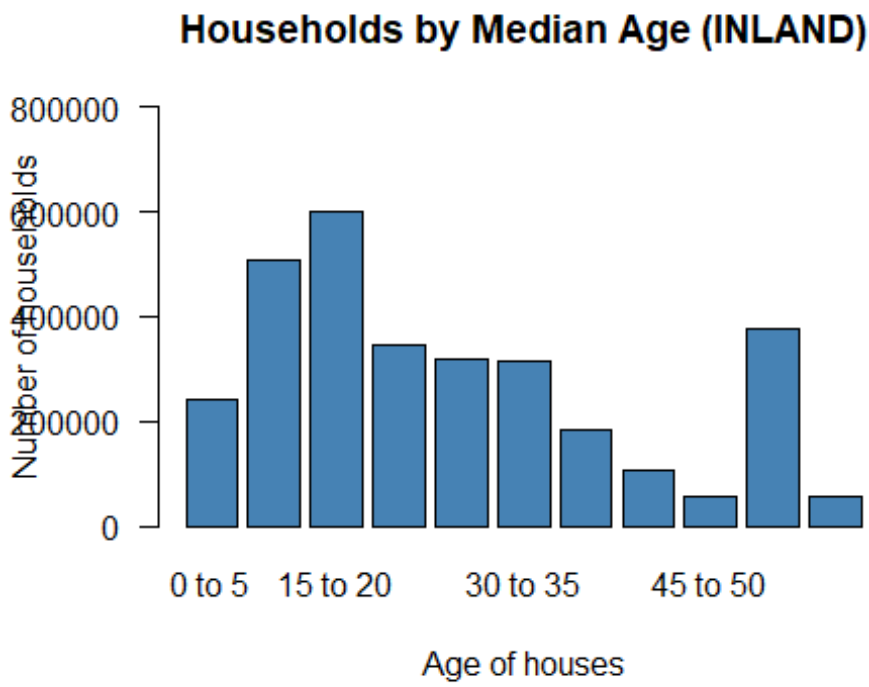
```
options(scipen=10000)
island_houses <- data.frame(matrix(nrow=5,ncol = 2))
colnames(island_houses) <- c("households","agecut")
island_houses[,1]<-
c(partition_data$households[partition_data$ocean_proximity=="ISLAND"])
island_houses[,2]<-
c(header[partition_data$age_cut[partition_data$ocean_proximity=="ISLAND"]])


barplot( height = rowsum(island_houses$households,island_houses$agecut)[,1],
las = 1, col = "#03B0F6", main = "Households by Median Age (ISLAND)", ylab =
"Number of households"
         , xlab = "Age of houses", ylim = c(0,800000))
```



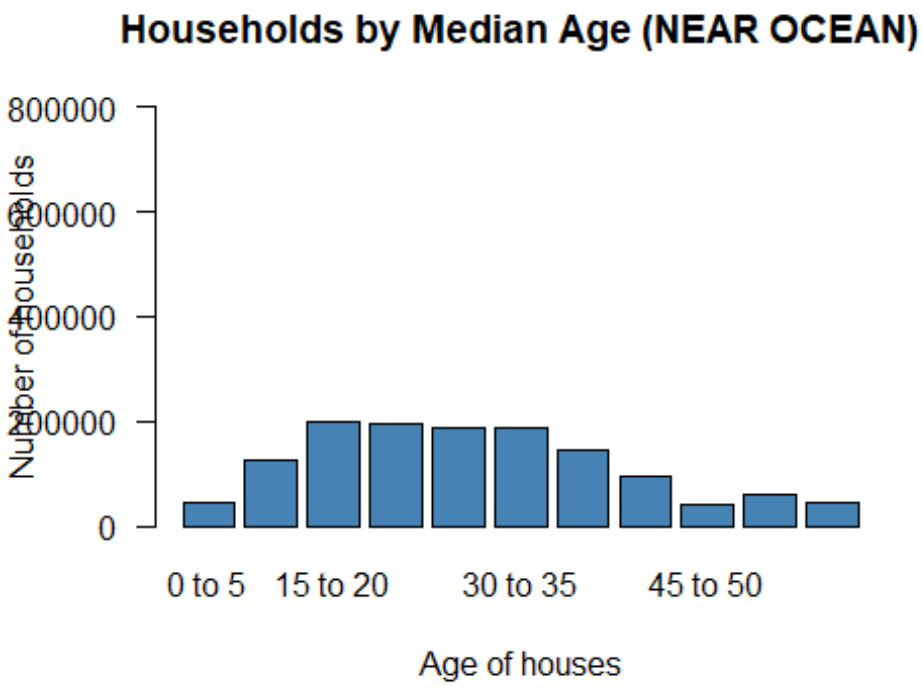

```
H1 <- data.frame(matrix(nrow=9034,ncol = 2))
colnames(H1) <- c("households","agecut")
H1[,1]<-c(partition_data$households[partition_data$ocean_proximity=="<1H
OCEAN"])
H1[,2]<-c(header[partition_data$age_cut[partition_data$ocean_proximity=="<1H
OCEAN"]])


barplot( height = rowsum(H1$households,H1$agecut)[,1], las = 1, col =
"steelblue", main = "Households by Median Age (<1H OCEAN)", ylab = "Number of
households"
         , xlab = "Age of houses", ylim = c(0,800000))
```



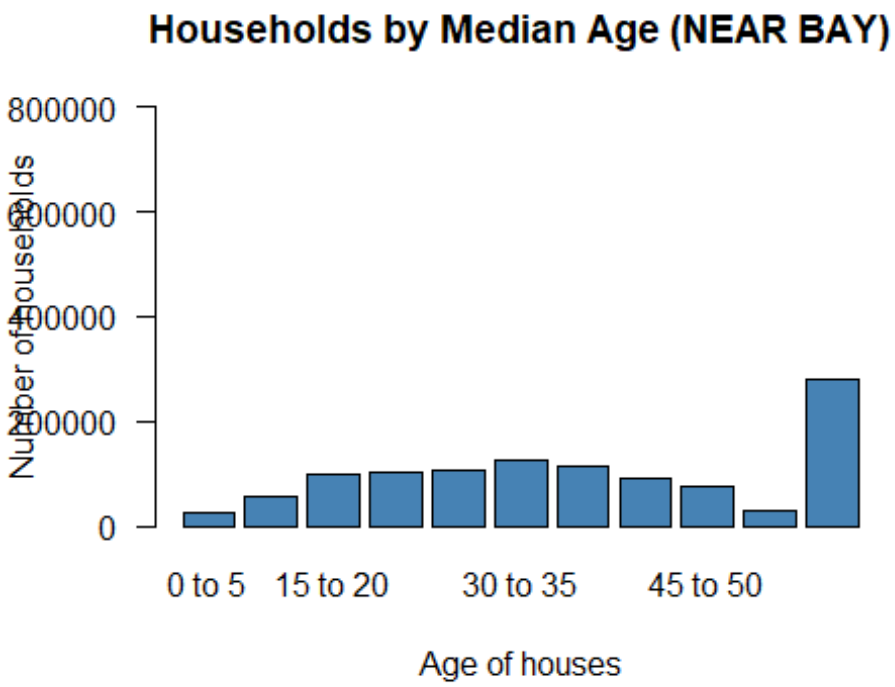

```
INLAND <- data.frame(matrix(nrow=6496,ncol = 2))
colnames(INLAND) <- c("households","agecut")
INLAND[,1]<-
c(partition_data$households[partition_data$ocean_proximity=="INLAND"])
INLAND[,2]<-
c(header[partition_data$age_cut[partition_data$ocean_proximity=="INLAND"]])

barplot( height = rowsum(INLAND$households,INLAND$agecut)[,1], las = 1, col =
"steelblue", main = "Households by Median Age (INLAND)", ylab = "Number of
households"
         , xlab = "Age of houses", ylim = c(0,800000))
```




```
ocean <- data.frame(matrix(nrow=2628,ncol = 2))
colnames(ocean) <- c("households","agecut")
ocean[,1]<-c(partition_data$households[partition_data$ocean_proximity=="NEAR
OCEAN"])
ocean[,2]<-
c(header[partition_data$age_cut[partition_data$ocean_proximity=="NEAR
OCEAN"]])

barplot( height = rowsum(ocean$households,ocean$agecut)[,1], las = 1, col =
"steelblue", main = "Households by Median Age (NEAR OCEAN)", ylab = "Number
of households"
         , xlab = "Age of houses", ylim = c(0,800000))
```




```
bay <- data.frame(matrix(nrow=2270,ncol = 2))
colnames(bay) <- c("households","agecut")
bay[,1]<-c(partition_data$households[partition_data$ocean_proximity=="NEAR
BAY"])
bay[,2]<-
c(header[partition_data$age_cut[partition_data$ocean_proximity=="NEAR BAY"]])



barplot( height = rowsum(bay$households,bay$agecut)[,1], las = 1, col =
"steelblue", main = "Households by Median Age (NEAR BAY)", ylab = "Number of
households"
         , xlab = "Age of houses", ylim = c(0,800000))
```

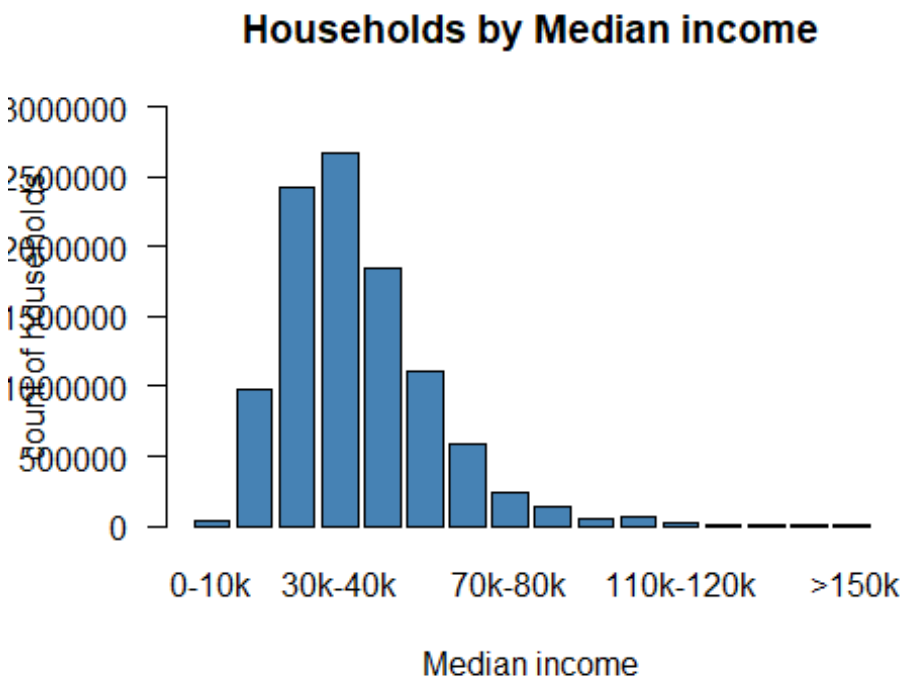


Summary

From the above graphs, it can be inferred that expect in the Near Bay location, most of the households are in the age group of 10-40 years. This means that the constructiog of new households may be common these areas.But in the Near bay area, most of the households are more than 50 years old. The median age of houses near bay and near near ocean is very less compared to inland and <1H ocean

Households by median income

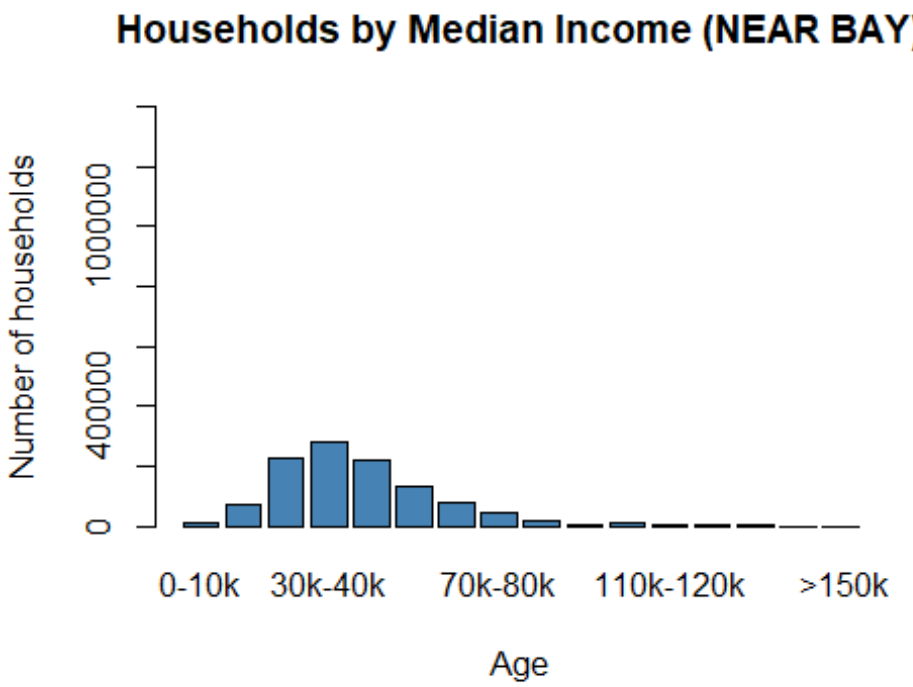
```
income_level <- c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,Inf)
header1 <- c("0-10k", "10k-20k", "20k-30k","30k-40k", "40k-50k", "50k-
60k","60k-70k","70k-80k","80k-90k","90k-100k"
                  ,"100k-110k","110k-120k","120k-130k","130k-140k","140k-
150k"," >150k")
partition_data1 <- transform(california_data_clean, income_cut =
cut(median_income, income_level, labels = header1))
barplot( height =
rowsum(partition_data1$households,partition_data1$income_cut)[,1], las = 1,
col = "steelblue", main = "Households by Median income", ylab = "count of
households", xlab = "Median income", ylim = c(0,3000000))
```



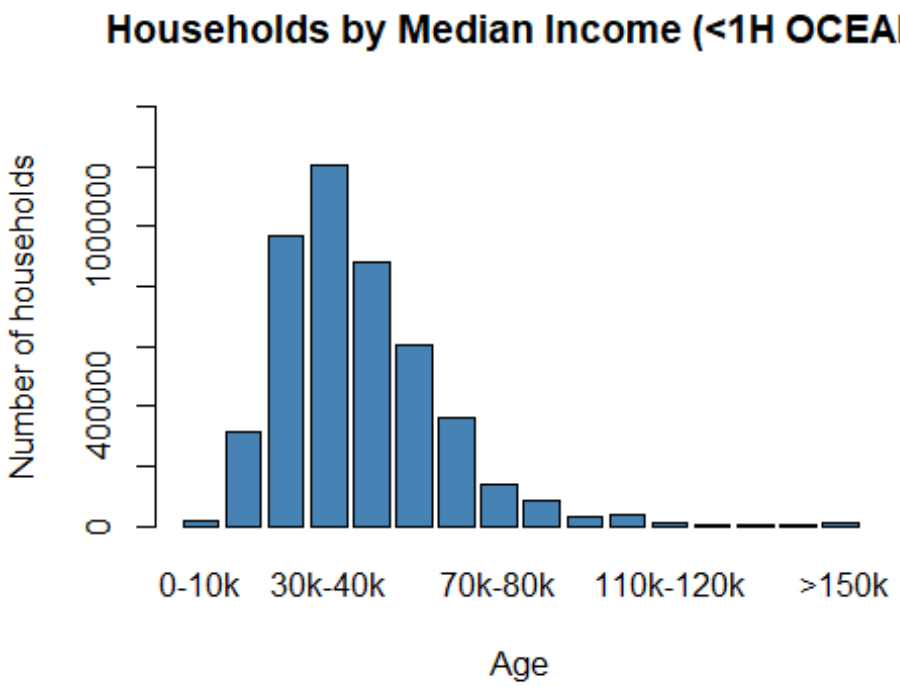
Summary

From the above graph, it can be inferred that the median income of most of the households is in the range of 20k to 50k and there are very few households with the median income of above 90k.

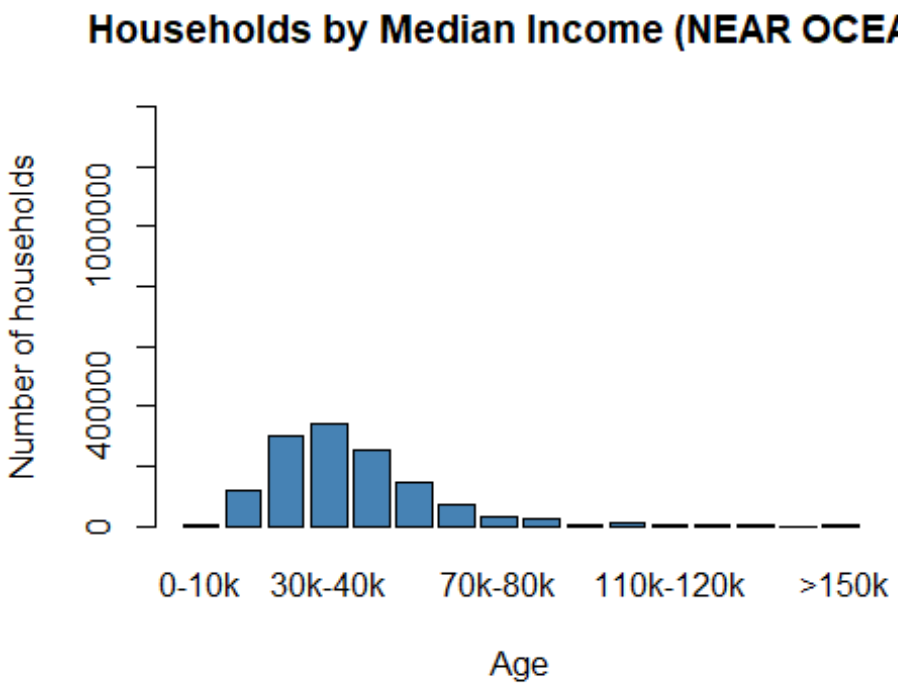
```
bay_income <- data.frame(matrix(nrow=2270,ncol = 2))
colnames(bay_income) <- c("households","incomecut")
bay_income[,1]<-
c(partition_data1$households[partition_data1$ocean_proximity=="NEAR BAY"])
bay_income[,2]<-
c(header1[partition_data1$income_cut[partition_data1$ocean_proximity=="NEAR
BAY"]])

barplot( height =
rowsum(bay_income$households,bay_income$incomecut)[,1][header1], col =
"steelblue", main = "Households by Median Income (NEAR BAY)", ylab = "Number
of households", xlab = "Age", ylim = c(0,1400000))
```



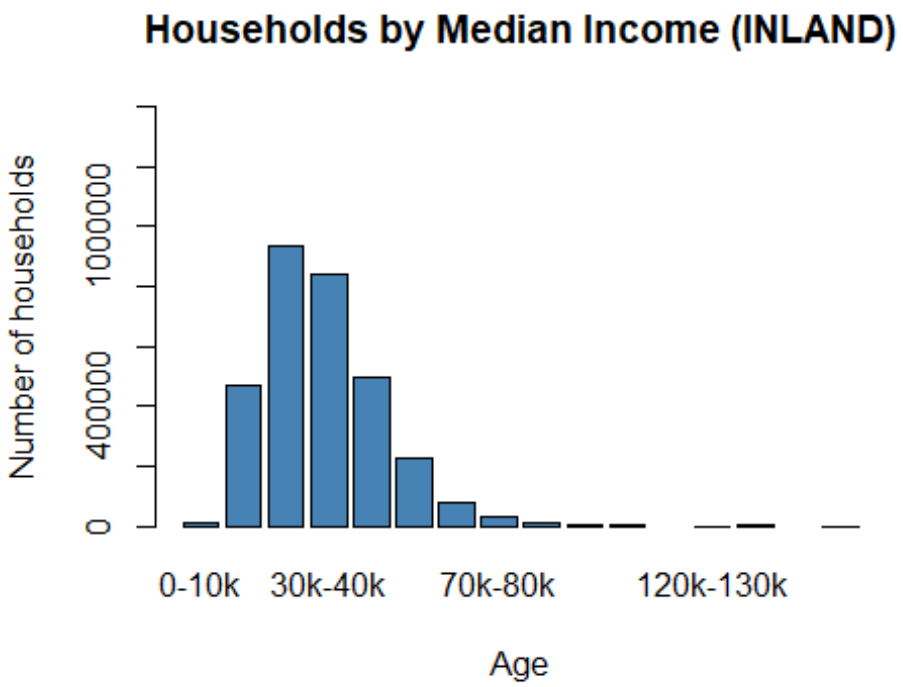

```
h1_income <- data.frame(matrix(nrow=9034,ncol = 2))
colnames(h1_income) <- c("households","incomecut")
h1_income[,1]<-
c(partition_data1$households[partition_data1$ocean_proximity=="<1H OCEAN"])
h1_income[,2]<-
c(header1[partition_data1$income_cut[partition_data1$ocean_proximity=="<1H
OCEAN"]])

barplot( height =
rowsum(h1_income$households,h1_income$incomecut)[,1][header1], col =
"steelblue", main = "Households by Median Income (<1H OCEAN)", ylab = "Number
of households", xlab = "Age", ylim = c(0,1400000))
```



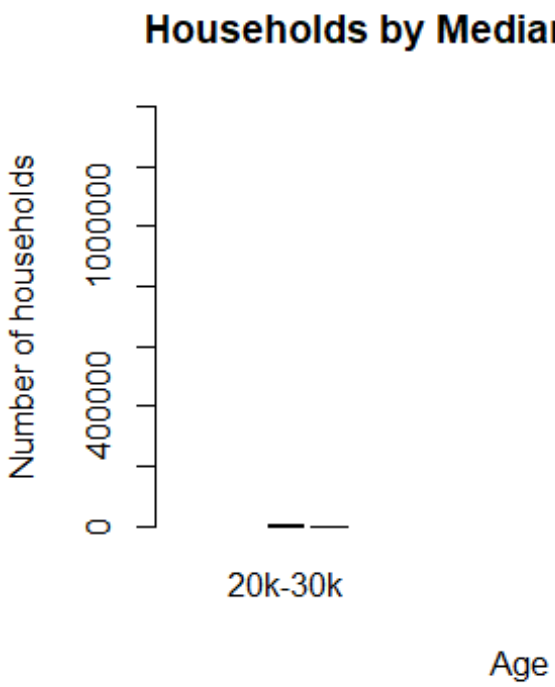

```
ocean_income <- data.frame(matrix(nrow=2628,ncol = 2))
colnames(ocean_income) <- c("households","incomecut")
ocean_income[,1]<-
c(partition_data1$households[partition_data1$ocean_proximity=="NEAR OCEAN"])
ocean_income[,2]<-
c(header1[partition_data1$income_cut[partition_data1$ocean_proximity=="NEAR
OCEAN"]])

barplot( height =
rowsum(ocean_income$households,ocean_income$incomecut)[,1][header1], col =
"steelblue", main = "Households by Median Income (NEAR OCEAN)", ylab =
"Number of households", xlab = "Age", ylim = c(0,1400000))
```




```
inland_income <- data.frame(matrix(nrow=6496,ncol = 2))
colnames(inland_income) <- c("households","incomecut")
inland_income[,1]<-
c(partition_data1$households[partition_data1$ocean_proximity=="INLAND"])
inland_income[,2]<-
c(header1[partition_data1$income_cut[partition_data1$ocean_proximity=="INLAND
"]])

barplot( height =
rowsum(inland_income$households,inland_income$incomecut)[,1][header1], col =
"steelblue", main = "Households by Median Income (INLAND)", ylab = "Number of
households", xlab = "Age", ylim = c(0,1400000))
```




```
island_income <- data.frame(matrix(nrow=5,ncol = 2))
colnames(island_income) <- c("households","incomecut")
island_income[,1]<-
c(partition_data1$households[partition_data1$ocean_proximity=="ISLAND"])
island_income[,2]<-
c(header1[partition_data1$income_cut[partition_data1$ocean_proximity=="ISLAND
"]])

barplot( height =
rowsum(island_income$households,island_income$incomecut)[,1][header1], col =
"steelblue", main = "Households by Median Income (ISLAND)", ylab = "Number of
households", xlab = "Age", ylim = c(0,1400000))
```

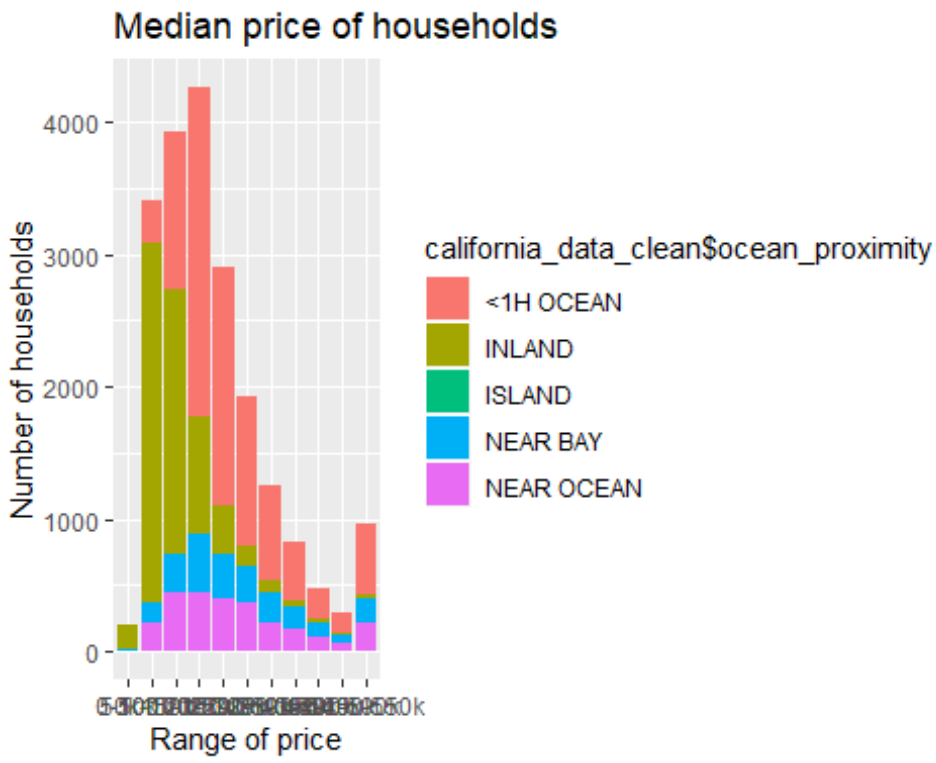


Summary

All of them follow similar trend. They have an median income of in the range of $30K - $40K except in inland. IN inland, the median income of households are in the 20k - 30k range, which is quiet obvious as it is away from the ocean and expense of living is less and hence income is less.Also, if considering the previous graphs regarding the number of households for median age, it is similar to the median income, it may also indicate that since more people live in <1H ocean or inland and hence the house prices are high compared to near ocean and near bay.

Count of households in various price ranges in all the proximities

```
value_cut <- c(price <-
c(0,50000,100000,150000,200000,250000,300000,350000,400000,450000,500000,5500
00))
header2 <- c("0-50k", "50k-100k", "100k-150k","150k-200k", "200k-250k",
"250k-300k","300k-350k","350k-400k","400k-450k","450k-500k","500k-550k")
partition_data2 <- transform(california_data_clean, cut_value =
cut(median_house_value, value_cut, labels = header2))
graph <- ggplot(data = california_data_clean) + geom_bar(map = aes(x =
partition_data2$cut_value, fill = california_data_clean$ocean_proximity))
graph + labs(x="Range of price",y="Number of households",title="Median price
of households") + labs(colour = "ocean proximity")
```

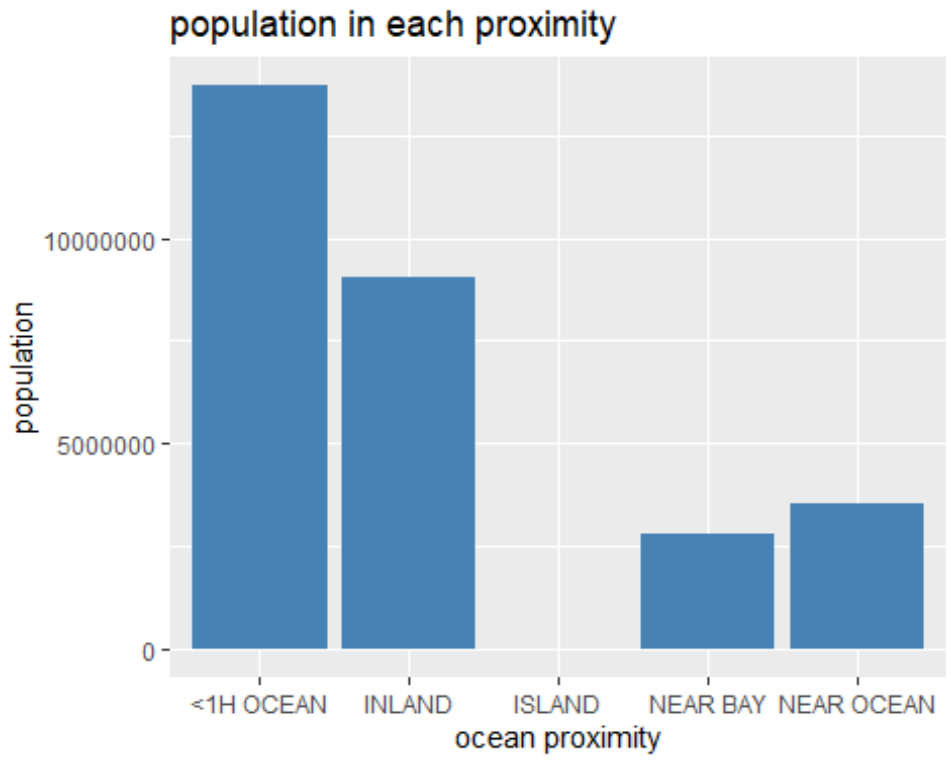


Summary

From the above plot, it can be inferred that most of the households in inland are cheaper and there like very few or no households in the price range of 0-50k.

Population in each proximity

```
options(scipen=10000)
pop_each_proximity <- california_data_clean %>% group_by(ocean_proximity) %>%
summarise(popu = sum(population))

ggplot(pop_each_proximity, aes(ocean_proximity, popu))+   geom_bar(stat =
"identity" , fill="steelblue")+   theme(legend.position = "none")+   labs(x =
"ocean proximity", y = "population",        title = "population in each
proximity")
```

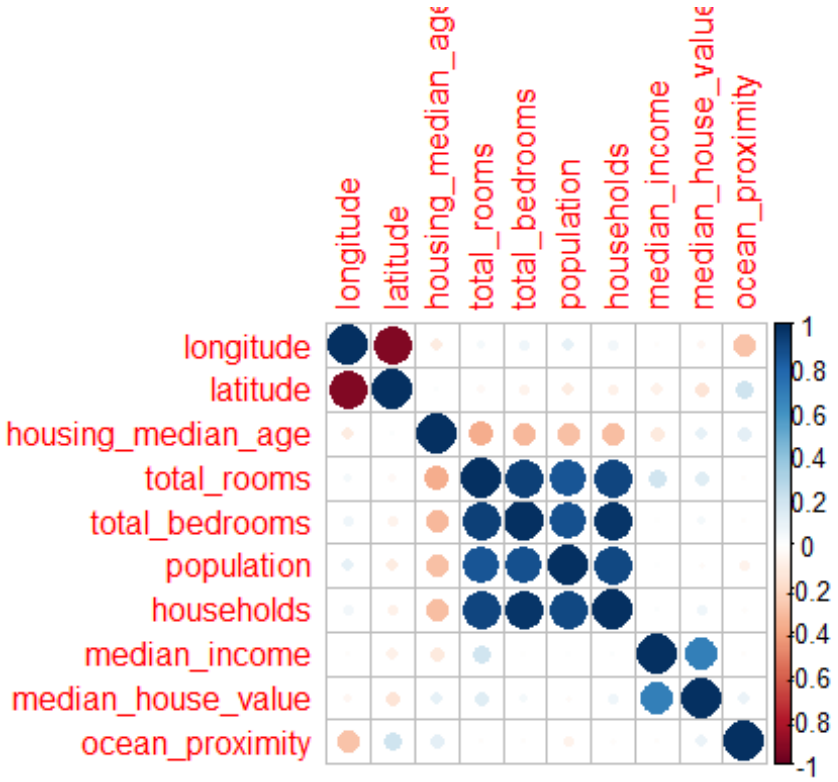


Sumamry

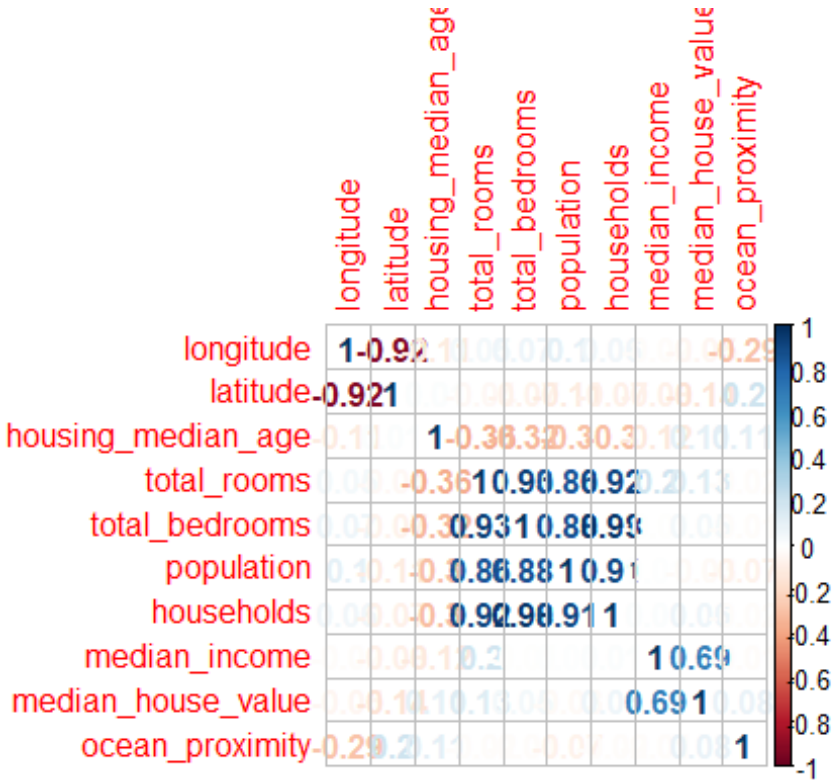
From the above graph, it can be inferred that most of the people live in the areas situated 1 hour away from ocean (<1H OCEAN), followed by inlands and the least in the islands.

Correlation plot

```
require(corrplot)

## Loading required package: corrplot

## Warning: package 'corrplot' was built under R version 3.6.2

## corrplot 0.84 loaded

california_data_clean$ocean_proximity <-
as.numeric(california_data_clean$ocean_proximity)
correlation  <- cor(california_data_clean)
corrplot(correlation, method = "circle")
```




```
corMat <- as.data.frame(corrplot(correlation,method = "number"))
```



Finding the corelation of median_house value with every other attribute

```
print(paste(row.names(corMat), corMat$median_house_value))
```

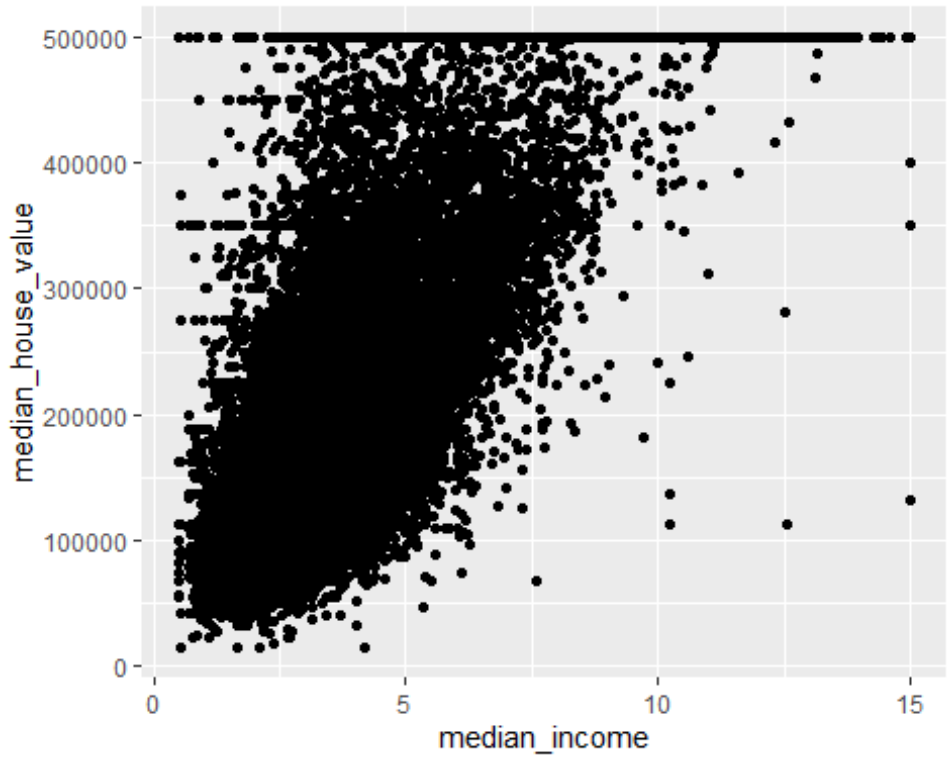
```
##  [1] "longitude -0.0453982193344448"
##  [2] "latitude -0.144638211576211"
##  [3] "housing_median_age 0.106432046876155"
##  [4] "total_rooms 0.133294134808323"
##  [5] "total_bedrooms 0.0496861802473459"
##  [6] "population -0.025299732287442"
##  [7] "households 0.0648935494881417"
##  [8] "median_income 0.688355475316112"
##  [9] "median_house_value 1"
## [10] "ocean_proximity 0.0804878600204853"
```

Observation:

We see that median_income has higher corelation with the median_house_value. So while constructing our model to predict the median_house_value, we consider median_income as the first variable and then we will consider other variables in the decreasing order of their corelation with the median_house_value

Plotting median_income vs median_house_value

```
options(scipen=10000)
ggplot(aes(x=median_income,y=median_house_value),data=california_data_clean)+
  geom_point()
```

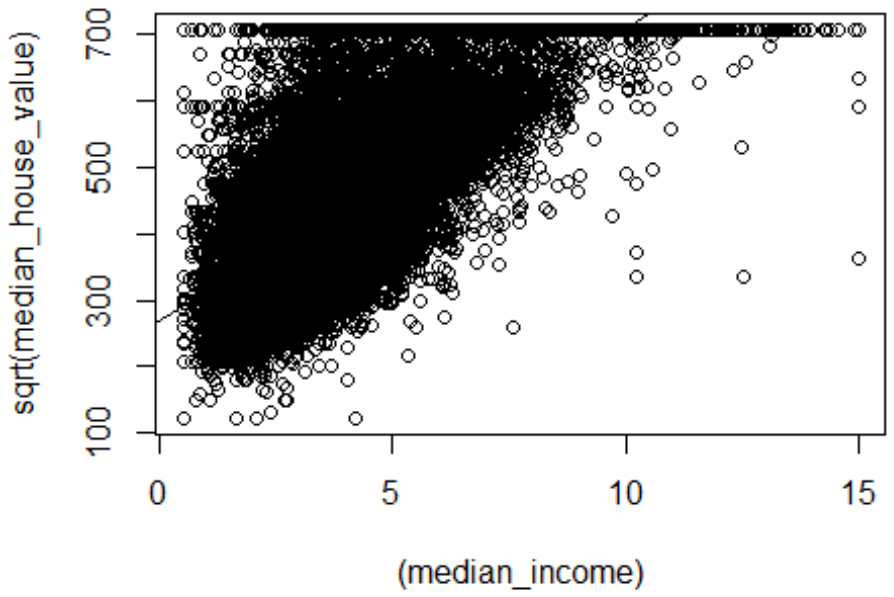


Constructing a linear regression model to predict household price

```
attach(california_data_clean)
fit1 <- lm(sqrt(california_data_clean$median_house_value) ~ median_income)
summary(fit1)
```

```
##
## Call:
## lm(formula = sqrt(california_data_clean$median_house_value) ~
##     median_income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -565.99  -62.39  -11.05   49.28  417.68
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   267.3973     1.4244   187.7 <0.0000000000000002 ***
## median_income  44.0629     0.3303   133.4 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 89.68 on 20431 degrees of freedom
## Multiple R-squared:  0.4655, Adjusted R-squared:  0.4655
## F-statistic: 1.779e+04 on 1 and 20431 DF,  p-value: < 0.00000000000000022
```

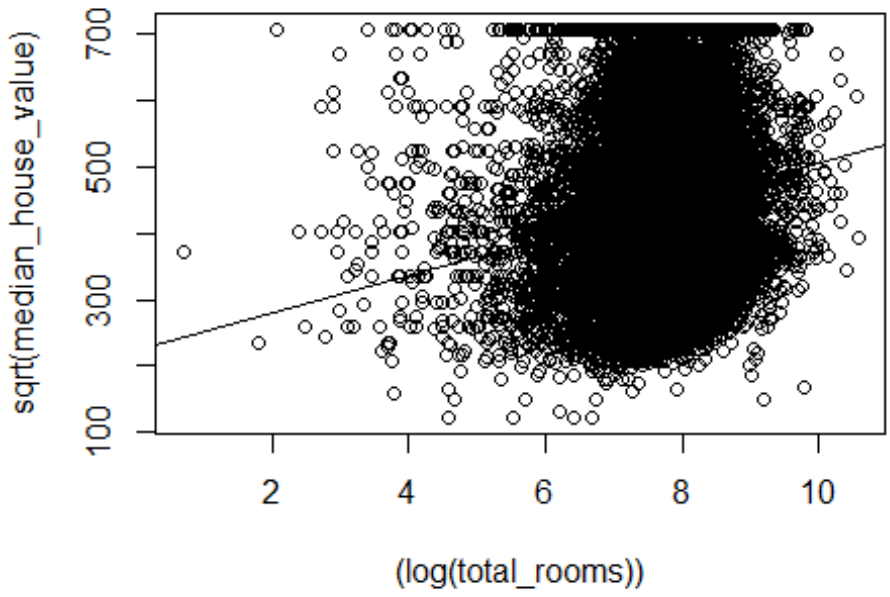
```r
plot(x= (median_income),y=sqrt(median_house_value))
abline(fit1)
```



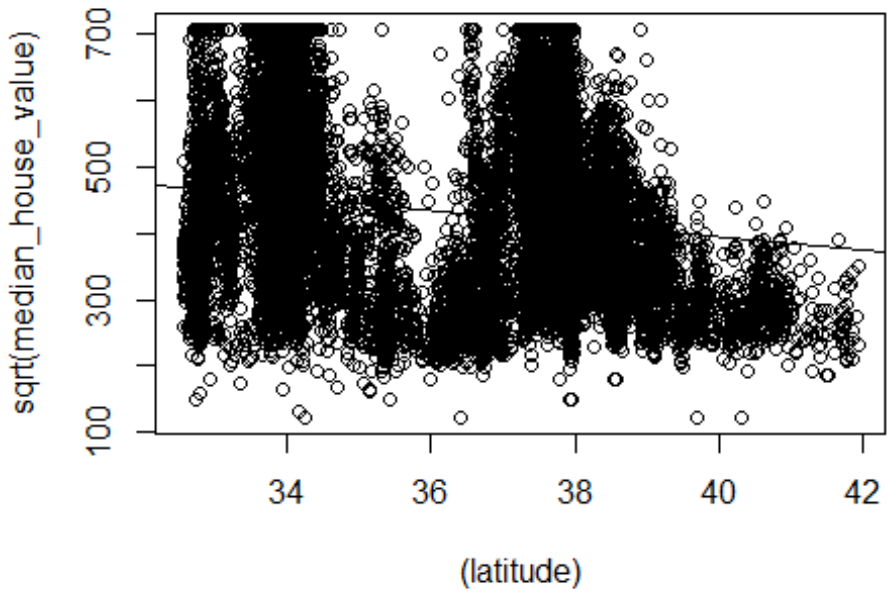

```r
fit2 <- lm(sqrt(median_house_value) ~  log(total_rooms))
summary(fit2)
```

```
##
## Call:
## lm(formula = sqrt(median_house_value) ~ log(total_rooms))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -332.61  -91.51  -13.57   72.51  426.01
##
## Coefficients:
##                  Estimate Std. Error t value            Pr(>|t|)
## (Intercept)       222.315      8.618   25.80 <0.0000000000000002 ***
## log(total_rooms)   28.269      1.124   25.15 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 120.8 on 20431 degrees of freedom
## Multiple R-squared:  0.03002,    Adjusted R-squared:  0.02997
## F-statistic: 632.3 on 1 and 20431 DF,  p-value: < 0.00000000000000022
```

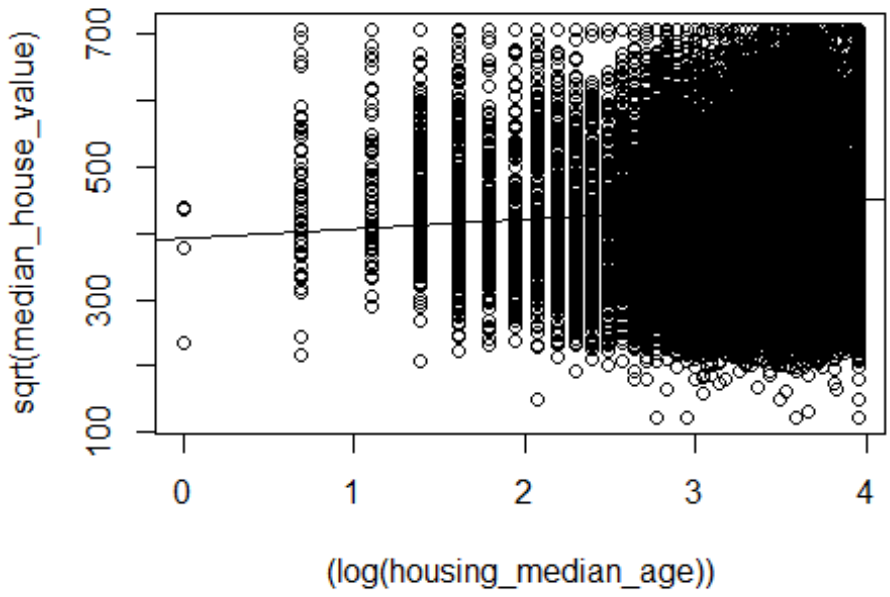
```r
plot(x= (log(total_rooms)),y=sqrt(median_house_value))
abline(fit2)
```



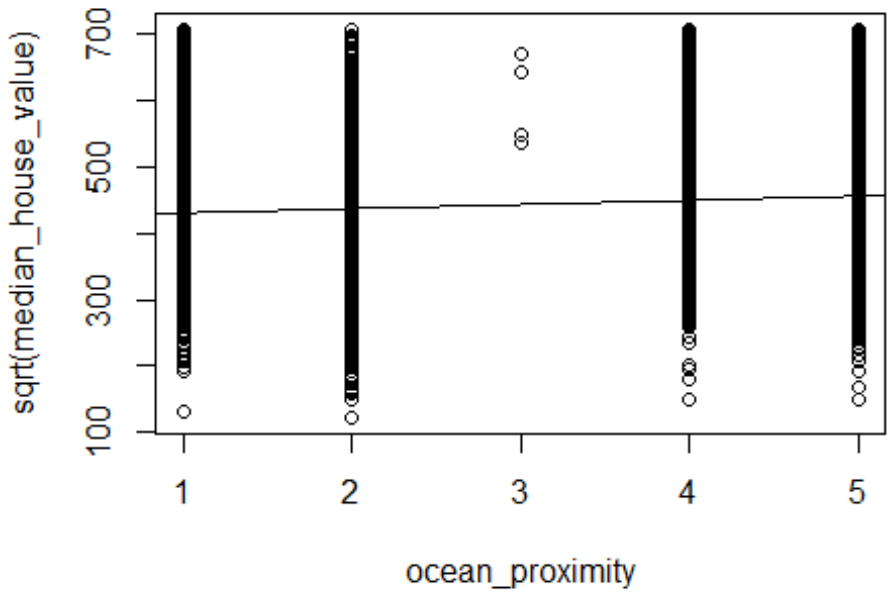

```
fit3 <- lm(sqrt(median_house_value) ~  (latitude))
summary(fit3)

##
## Call:
## lm(formula = sqrt(median_house_value) ~ (latitude))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -329.06  -89.35  -18.17   75.74  301.03
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 784.8026    14.1317   55.53 <0.0000000000000002 ***
## latitude     -9.7334     0.3959  -24.59 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 120.9 on 20431 degrees of freedom
## Multiple R-squared:  0.02874,    Adjusted R-squared:  0.02869
## F-statistic: 604.5 on 1 and 20431 DF,  p-value: < 0.00000000000000022

plot(x= (latitude),y=sqrt(median_house_value))
abline(fit3)
```



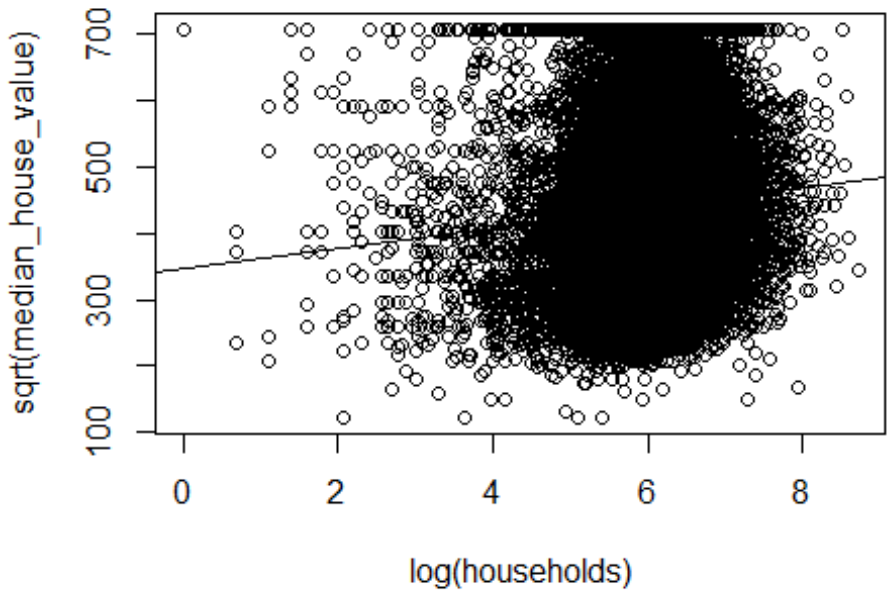

```
fit4 <- lm(sqrt(median_house_value) ~  log(housing_median_age))
summary(fit4)

##
## Call:
## lm(formula = sqrt(median_house_value) ~ log(housing_median_age))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -325.65  -90.42  -13.57   76.48  304.51
##
## Coefficients:
##                          Estimate Std. Error t value            Pr(>|t|)
## (Intercept)               392.918      4.935   79.62 <0.0000000000000002
***
## log(housing_median_age)   13.971      1.507    9.27 <0.0000000000000002
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.4 on 20431 degrees of freedom
## Multiple R-squared:  0.004189,   Adjusted R-squared:  0.00414
## F-statistic: 85.94 on 1 and 20431 DF,  p-value: < 0.00000000000000022

plot(x= (log(housing_median_age)),y=sqrt(median_house_value))
abline(fit4)
```



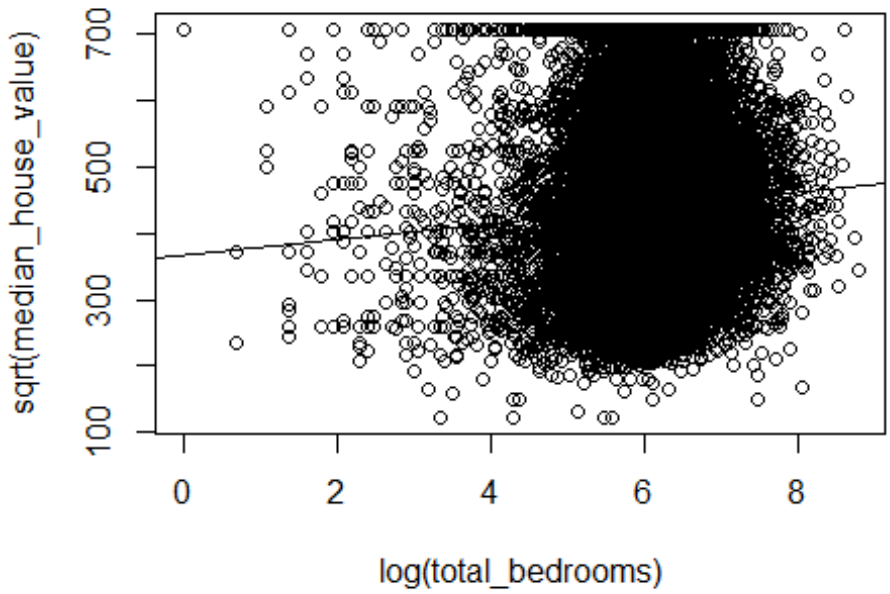

```
fit5 <- lm(sqrt(median_house_value) ~  ocean_proximity)
summary(fit5)

##
## Call:
## lm(formula = sqrt(median_house_value) ~ ocean_proximity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -314.50  -92.58  -12.60   75.29  276.18
##
## Coefficients:
##                 Estimate Std. Error t value            Pr(>|t|)
## (Intercept)     424.8913     1.5613  272.13 <0.0000000000000002 ***
## ocean_proximity   6.0386     0.6028   10.02 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.4 on 20431 degrees of freedom
## Multiple R-squared:  0.004888,   Adjusted R-squared:  0.004839
## F-statistic: 100.4 on 1 and 20431 DF,  p-value: < 0.00000000000000022

plot(x= ocean_proximity,y=sqrt(median_house_value))
abline(fit5)
```



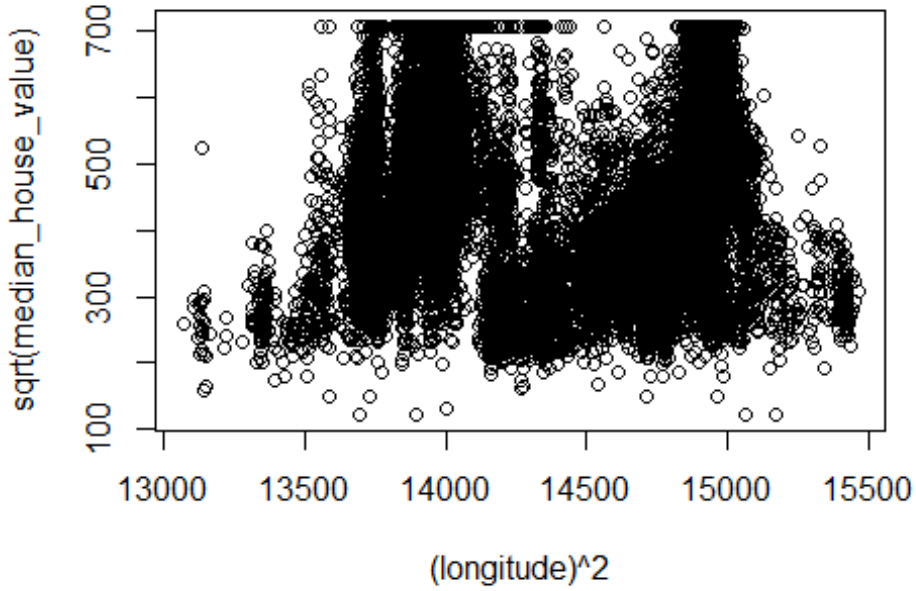

```
fit6 <- lm(sqrt(median_house_value) ~  log(households))
summary(fit6)

##
## Call:
## lm(formula = sqrt(median_house_value) ~ log(households))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -307.97  -91.44  -15.03   74.55  361.76
##
## Coefficients:
##                 Estimate Std. Error t value            Pr(>|t|)
## (Intercept)      345.349      7.022   49.18 <0.0000000000000002 ***
## log(households)   15.487      1.165   13.29 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.1 on 20431 degrees of freedom
## Multiple R-squared:  0.00857,    Adjusted R-squared:  0.008522
## F-statistic: 176.6 on 1 and 20431 DF,  p-value: < 0.00000000000000022

plot(x=log(households),y=sqrt(median_house_value))
abline(fit6)
```




```
fit7 <- lm(sqrt(median_house_value) ~  log(total_bedrooms))
summary(fit7)

##
## Call:
## lm(formula = sqrt(median_house_value) ~ log(total_bedrooms))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -310.06  -92.04  -14.75   75.34  340.18
##
## Coefficients:
##                     Estimate Std. Error t value            Pr(>|t|)
## (Intercept)          366.927      7.120   51.53 <0.0000000000000002 ***
## log(total_bedrooms)   11.741      1.168   10.05 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.4 on 20431 degrees of freedom
## Multiple R-squared:  0.00492,    Adjusted R-squared:  0.004871
## F-statistic:   101 on 1 and 20431 DF,  p-value: < 0.00000000000000022

plot(x=log(total_bedrooms),y=sqrt(median_house_value))
abline(fit7)
```

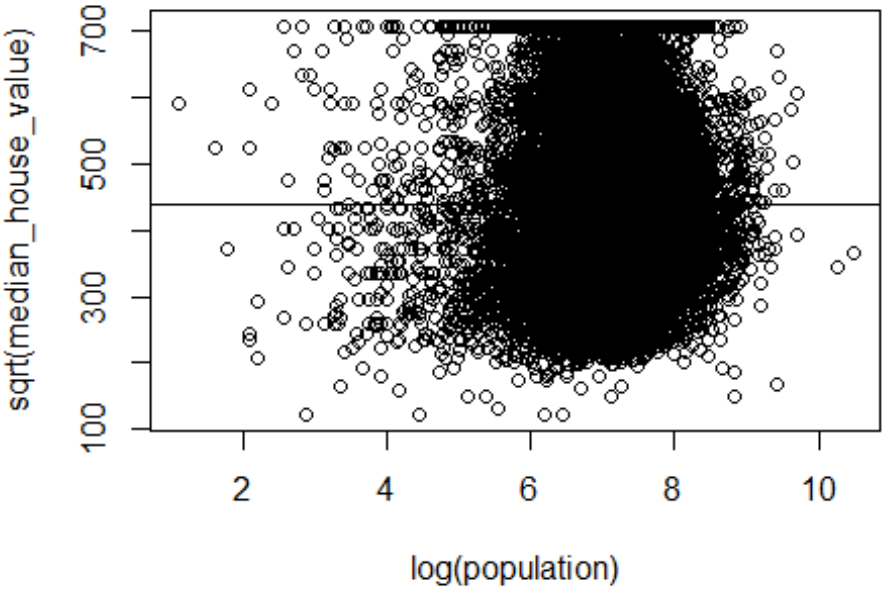



```
fit8 <- lm(sqrt(median_house_value) ~  (longitude)^2)
summary(fit8)

##
## Call:
## lm(formula = sqrt(median_house_value) ~ (longitude)^2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -323.40  -91.48  -13.41   75.66  276.02
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)  175.738     51.188   3.433    0.000598 ***
## longitude     -2.193      0.428  -5.124 0.000000302 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.6 on 20431 degrees of freedom
## Multiple R-squared:  0.001283,   Adjusted R-squared:  0.001234
## F-statistic: 26.25 on 1 and 20431 DF,  p-value: 0.0000003023

plot(x=(longitude)^2,y=sqrt(median_house_value))
abline(fit8)
```



Longitude doesn't have any linear relationship with the median_house_value. So we are not considering that in our model.

```
fit9 <- lm(sqrt(median_house_value) ~  log(population))
summary(fit9)

##
## Call:
## lm(formula = sqrt(median_house_value) ~ log(population))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -315.44  -92.29  -14.05   76.53  269.57
##
## Coefficients:
##                  Estimate Std. Error t value            Pr(>|t|)
## (Intercept)     437.28734    8.20525  53.294 <0.0000000000000002 ***
## log(population)   0.09745    1.16181   0.084               0.933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.7 on 20431 degrees of freedom
## Multiple R-squared:  3.443e-07,  Adjusted R-squared:  -4.86e-05
## F-statistic: 0.007035 on 1 and 20431 DF,  p-value: 0.9332

plot(x=log(population),y=sqrt(median_house_value))
abline(fit9)
```



From the summary of the model, we can see that population variable has high p-value. So, it is not significant in predicting the price of the household. So, we are not including population in building our model.

Incrementally adding the attributes to our model to predict the price of the household

Linear model when median_income is the only predictor variable.

```
summary(fit1)

##
## Call:
## lm(formula = sqrt(california_data_clean$median_house_value) ~
##     median_income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -565.99  -62.39  -11.05   49.28  417.68
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   267.3973     1.4244   187.7 <0.0000000000000002 ***
## median_income  44.0629     0.3303   133.4 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 89.68 on 20431 degrees of freedom
## Multiple R-squared:  0.4655, Adjusted R-squared:  0.4655
## F-statistic: 1.779e+04 on 1 and 20431 DF,  p-value: < 0.00000000000000022
```

Adding the total_rooms attribute and comparing with the previous model.

```
m1 <- lm(sqrt(median_house_value) ~  median_income + log(total_rooms))
summary(m1)
```

```
##
## Call:
## lm(formula = sqrt(median_house_value) ~ median_income + log(total_rooms))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -548.32  -62.56  -11.00   49.23  428.28
##
## Coefficients:
##                  Estimate Std. Error t value             Pr(>|t|)
## (Intercept)      226.2346     6.3905  35.402 < 0.0000000000000002 ***
## median_income     43.6050     0.3372 129.320 < 0.0000000000000002 ***
## log(total_rooms)   5.6281     0.8518   6.607      0.0000000000402 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 89.59 on 20430 degrees of freedom
## Multiple R-squared:  0.4666, Adjusted R-squared:  0.4666
## F-statistic:  8937 on 2 and 20430 DF,  p-value: < 0.00000000000000022
```

We observe that the F-statistic value has increased and the p-value of the model is low. So, total_rooms is a good predictor.

```
m2 <- lm(sqrt(median_house_value) ~  median_income + log(total_rooms) +
log(latitude))
summary(m2)
```

```
##
## Call:
## lm(formula = sqrt(median_house_value) ~ median_income + log(total_rooms) +
##     log(latitude))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -532.61  -61.23  -12.45   48.67  416.22
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)      1068.1425    38.1631  27.989 < 0.0000000000000002 ***
## median_income      43.0670     0.3340 128.941 < 0.0000000000000002 ***
## log(total_rooms)    5.3065     0.8417   6.304       0.000000000295 ***
```

```
## log(latitude)    -234.4588    10.4814 -22.369 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.51 on 20429 degrees of freedom
## Multiple R-squared:  0.4794, Adjusted R-squared:  0.4793
## F-statistic:  6270 on 3 and 20429 DF,  p-value: < 0.00000000000000022
```

We observe that the F-statistic value has decreased but the standard error value has reduced. So, we are considering latitude.

Similarly, we are adding other attributes and evaluating the model.

```
m3 <- lm(sqrt(median_house_value) ~  median_income + log(total_rooms) +
log(latitude) + housing_median_age)
summary(m3)
```

```
##
## Call:
## lm(formula = sqrt(median_house_value) ~ median_income + log(total_rooms) +
##     log(latitude) + housing_median_age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -564.97  -58.24  -10.32   47.73  441.75
##
## Coefficients:
##                      Estimate Std. Error t value            Pr(>|t|)
## (Intercept)         931.75658   36.91227   25.24 <0.0000000000000002 ***
## median_income        43.81098    0.32221  135.97 <0.0000000000000002 ***
## log(total_rooms)     15.40052    0.84908   18.14 <0.0000000000000002 ***
## log(latitude)      -234.64564   10.09444  -23.25 <0.0000000000000002 ***
## housing_median_age    1.99658    0.04996   39.97 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.24 on 20428 degrees of freedom
## Multiple R-squared:  0.5171, Adjusted R-squared:  0.517
## F-statistic:  5469 on 4 and 20428 DF,  p-value: < 0.00000000000000022
```

So far, these parameters provided the best f-statistic value. Let's add the ocean_proximity.

```
m4 <- lm(sqrt(median_house_value) ~  median_income + log(total_rooms) +
log(latitude) + housing_median_age + ocean_proximity)
summary(m4)
```

```
##
## Call:
## lm(formula = sqrt(median_house_value) ~ median_income + log(total_rooms) +
##     log(latitude) + housing_median_age + ocean_proximity)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -552.94  -57.10   -9.57   47.24  440.45
##
## Coefficients:
##                      Estimate Std. Error t value            Pr(>|t|)
## (Intercept)        1046.14187   37.25488   28.08 <0.0000000000000002 ***
## median_income        43.76345    0.31995  136.78 <0.0000000000000002 ***
## log(total_rooms)     14.99180    0.84341   17.77 <0.0000000000000002 ***
## log(latitude)      -269.40211   10.22648  -26.34 <0.0000000000000002 ***
## housing_median_age    1.89757    0.04994   38.00 <0.0000000000000002 ***
## ocean_proximity       7.33318    0.42826   17.12 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.64 on 20427 degrees of freedom
## Multiple R-squared:  0.524,  Adjusted R-squared:  0.5239
## F-statistic:  4497 on 5 and 20427 DF,  p-value: < 0.00000000000000022
```

```
m5 <- lm(sqrt(median_house_value) ~  median_income + log(total_rooms) +
housing_median_age + ocean_proximity)
summary(m5)
```

```
##
## Call:
## lm(formula = sqrt(median_house_value) ~ median_income + log(total_rooms) +
##     housing_median_age + ocean_proximity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -573.94  -58.33   -8.77   47.38  454.15
##
## Coefficients:
##                     Estimate Std. Error t value            Pr(>|t|)
## (Intercept)         81.98024    7.07431   11.59 <0.0000000000000002 ***
## median_income       44.37152    0.32448  136.75 <0.0000000000000002 ***
## log(total_rooms)    15.46863    0.85740   18.04 <0.0000000000000002 ***
## housing_median_age   1.92721    0.05077   37.96 <0.0000000000000002 ***
## ocean_proximity      5.09387    0.42680   11.94 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 86.06 on 20428 degrees of freedom
## Multiple R-squared:  0.5078, Adjusted R-squared:  0.5077
## F-statistic:  5269 on 4 and 20428 DF,  p-value: < 0.00000000000000022
```

We observe that either adding or removing the latitude attribute is not affecting the model. So, we are ignoring the latitude attribute and try with log(latitude) parameter like in m3.

```r
m6 <- lm(sqrt(median_house_value) ~  median_income + log(total_rooms) +
housing_median_age + ocean_proximity + households)
summary(m6)
```

```
##
## Call:
## lm(formula = sqrt(median_house_value) ~ median_income + log(total_rooms) +
##     housing_median_age + ocean_proximity + households)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -613.64  -57.06   -7.90   47.15  440.96
##
## Coefficients:
##                      Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        195.454536   8.991009   21.74 < 0.0000000000000002 ***
## median_income       45.910108   0.330276  139.00 < 0.0000000000000002 ***
## log(total_rooms)    -3.915429   1.283651   -3.05              0.00229 **
## housing_median_age   2.049702   0.050639   40.48 < 0.0000000000000002 ***
## ocean_proximity      5.116922   0.422641   12.11 < 0.0000000000000002 ***
## households           0.049833   0.002475   20.13 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.22 on 20427 degrees of freedom
## Multiple R-squared:  0.5174, Adjusted R-squared:  0.5173
## F-statistic:  4380 on 5 and 20427 DF,  p-value: < 0.00000000000000022
```

```r
m7 <- lm(sqrt(median_house_value) ~  median_income + log(total_rooms) +
housing_median_age + ocean_proximity + households + total_bedrooms)
summary(m7)
```

```
##
## Call:
## lm(formula = sqrt(median_house_value) ~ median_income + log(total_rooms) +
##     housing_median_age + ocean_proximity + households + total_bedrooms)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -618.51  -57.23   -7.99   47.30  439.77
##
## Coefficients:
##                      Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        199.783256   9.036293  22.109 < 0.0000000000000002 ***
## median_income       46.154169   0.334397 138.022 < 0.0000000000000002 ***
## log(total_rooms)    -4.800709   1.297533  -3.700             0.000216 ***
## housing_median_age   2.079289   0.051026  40.750 < 0.0000000000000002 ***
## ocean_proximity      5.063506   0.422597  11.982 < 0.0000000000000002 ***
## households           0.015598   0.007881   1.979             0.047801 *
## total_bedrooms       0.033180   0.007252   4.575           0.00000479 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.18 on 20426 degrees of freedom
## Multiple R-squared:  0.5179, Adjusted R-squared:  0.5177
## F-statistic:  3657 on 6 and 20426 DF,  p-value: < 0.00000000000000022
```

```
m8 <- lm(sqrt(median_house_value) ~  median_income + log(total_rooms)+
housing_median_age + ocean_proximity + households + total_bedrooms)
summary(m8)
```

```
##
## Call:
## lm(formula = sqrt(median_house_value) ~ median_income + log(total_rooms) +
##     housing_median_age + ocean_proximity + households + total_bedrooms)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -618.51  -57.23   -7.99   47.30  439.77
##
## Coefficients:
##                      Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        199.783256   9.036293  22.109 < 0.0000000000000002 ***
## median_income       46.154169   0.334397 138.022 < 0.0000000000000002 ***
## log(total_rooms)    -4.800709   1.297533  -3.700             0.000216 ***
## housing_median_age   2.079289   0.051026  40.750 < 0.0000000000000002 ***
## ocean_proximity      5.063506   0.422597  11.982 < 0.0000000000000002 ***
## households           0.015598   0.007881   1.979             0.047801 *
## total_bedrooms       0.033180   0.007252   4.575           0.00000479 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.18 on 20426 degrees of freedom
## Multiple R-squared:  0.5179, Adjusted R-squared:  0.5177
## F-statistic:  3657 on 6 and 20426 DF,  p-value: < 0.00000000000000022
```

We observe that removing the households attribute increases the F-statistic and R value is not changed. So, we are ignoring households attribute.

```
m9 <- lm(sqrt(median_house_value) ~  median_income +   housing_median_age +
ocean_proximity + log(total_rooms) + total_bedrooms + longitude)
summary(m9)
```

```
##
## Call:
## lm(formula = sqrt(median_house_value) ~ median_income + housing_median_age
+
##     ocean_proximity + log(total_rooms) + total_bedrooms + longitude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -618.50  -57.25   -7.99   47.27  439.25
##
## Coefficients:
##                     Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        239.91237   37.75355   6.355       0.000000000213 ***
## median_income       46.21264    0.33339 138.614 < 0.0000000000000002 ***
## housing_median_age   2.09163    0.05097  41.035 < 0.0000000000000002 ***
## ocean_proximity      5.18242    0.44030  11.770 < 0.0000000000000002 ***
## log(total_rooms)    -4.51013    1.29315  -3.488              0.000488 ***
## total_bedrooms       0.04666    0.00228  20.461 < 0.0000000000000002 ***
## longitude            0.35663    0.31245   1.141              0.253716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.19 on 20426 degrees of freedom
## Multiple R-squared:  0.5178, Adjusted R-squared:  0.5177
## F-statistic:  3656 on 6 and 20426 DF,  p-value: < 0.00000000000000022
```

After adding the longitude attribue, the F-statistic value is reducing and also the p-value is not significant. So, we are ignoring the longitude attribute and we also saw in the graphs that longitude doesn't have any linear relationship with the price of household attribute.

```
m10 <- lm(sqrt(median_house_value) ~  median_income +   housing_median_age +
ocean_proximity + log(total_rooms) + total_bedrooms + population)
summary(m10)

##
## Call:
## lm(formula = sqrt(median_house_value) ~ median_income + housing_median_age
+
##     ocean_proximity + log(total_rooms) + total_bedrooms + population)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -623.76  -55.29   -5.53   47.17  592.40
##
## Coefficients:
##                      Estimate Std. Error t value            Pr(>|t|)
## (Intercept)        193.980673   8.820861  21.991 <0.0000000000000002 ***
## median_income       46.247802   0.326864 141.489 <0.0000000000000002 ***
## housing_median_age   2.070359   0.049867  41.518 <0.0000000000000002 ***
## ocean_proximity      3.652785   0.417065   8.758 <0.0000000000000002 ***
## log(total_rooms)    -2.772075   1.268588  -2.185               0.0289 *
## total_bedrooms       0.117564   0.003329  35.315 <0.0000000000000002 ***
## population          -0.031104   0.001086 -28.650 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83.53 on 20426 degrees of freedom
```

```
## Multiple R-squared:  0.5364, Adjusted R-squared:  0.5363
## F-statistic:  3939 on 6 and 20426 DF,  p-value: < 0.00000000000000022
```

Addition of population to the model, reduces the F-statistic by few numbers and the p-value of the model has increased on addition of population to model, so it is not a good predictor.

```
summary(m3)

##
## Call:
## lm(formula = sqrt(median_house_value) ~ median_income + log(total_rooms) +
##     log(latitude) + housing_median_age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -564.97  -58.24  -10.32   47.73  441.75
##
## Coefficients:
##                      Estimate Std. Error t value            Pr(>|t|)
## (Intercept)         931.75658   36.91227   25.24 <0.0000000000000002 ***
## median_income        43.81098    0.32221  135.97 <0.0000000000000002 ***
## log(total_rooms)     15.40052    0.84908   18.14 <0.0000000000000002 ***
## log(latitude)      -234.64564   10.09444  -23.25 <0.0000000000000002 ***
## housing_median_age    1.99658    0.04996   39.97 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.24 on 20428 degrees of freedom
## Multiple R-squared:  0.5171, Adjusted R-squared:  0.517
## F-statistic:  5469 on 4 and 20428 DF,  p-value: < 0.00000000000000022
```

Of all the linear models, the model m3 with median_income, log(total_rooms), log(latitude), housing_median_age has the highest F-statistic value and with a compartively less residual standard error(RSS). Though the model m5 has lesser RSS, it's F-statistic is less compared to m3. All other models except these have very low F-statistic and also the p-values are high. And also, some models like m8,m7m10 show some insignificant values as well. The latitude is found insignificant by itself, but the log(latitude) has added to a significant value. Therefore, sqrt(median_house_value) ~ median_income + log(total_rooms) + log(latitude) + housing_median_age is the best model.