*Data used: California Housing data set*

*Note: Section 1 contains Analysis and Section 2 conatins the R script.*

Section 1: Analysis

# READING AND VIEWING DATASET

1)Reading the data from csv file into data using read.csv command.

data <- read.csv(file="https://docs.google.com/spreadsheets/d/e/2PACX-1vQAJvpd8Dk\_0pgS0t\_njNRLanJcxBn5Em7OO7Ew8wckt9Lr0SqZ0gLa48c9v1svsZu1w4RE1Mf1BKgM/pub?gid=1870803675&single=true&output=csv", header=TRUE, sep=",")

// HERE WE WILL HAVE TO UPLOAD THE DATA ONLINE

//C:/Users/maman/Downloads/housing.csv

2)Invoking a spreadsheet-style data viewer using View command

View(data)

3)Retrieve names of all features of the data:-

names(data)

Output:

[1] "longitude" "latitude" "housing\_median\_age" "total\_rooms" "total\_bedrooms"

[6] "population" "households" "median\_income" "median\_house\_value" "ocean\_proximity"

DATA CLEANING :

4)Printing names of columns which contain more than 1% null values.  
for(i in 1:ncol(data)) {  
 colName <- colnames(data[i])  
 pctNull <- sum(is.na(data[,i]))/length(data[,i])  
 if (pctNull >0.01) {  
 print(paste("Column ", colName, " has ", round(pctNull\*100, 3), "% of nulls"))  
 }  
}  
Output: "Column total\_bedrooms has 1.003 % of nulls"

5)Again checking if there are any null values in total\_bedrooms.

table(is.na(data$total\_bedrooms))

Output:

FALSE TRUE

20433 207

6)Replacing null values in total\_bedrooms column with the median of that column so we won’t have any null values in our column.

data$total\_bedrooms[is.na(data$total\_bedrooms)] = median(data$total\_bedrooms,na.rm = TRUE)

7)Checking if there are any null values in total\_bedrooms now.  
table(is.na(data$total\_bedrooms))

Output:

FALSE

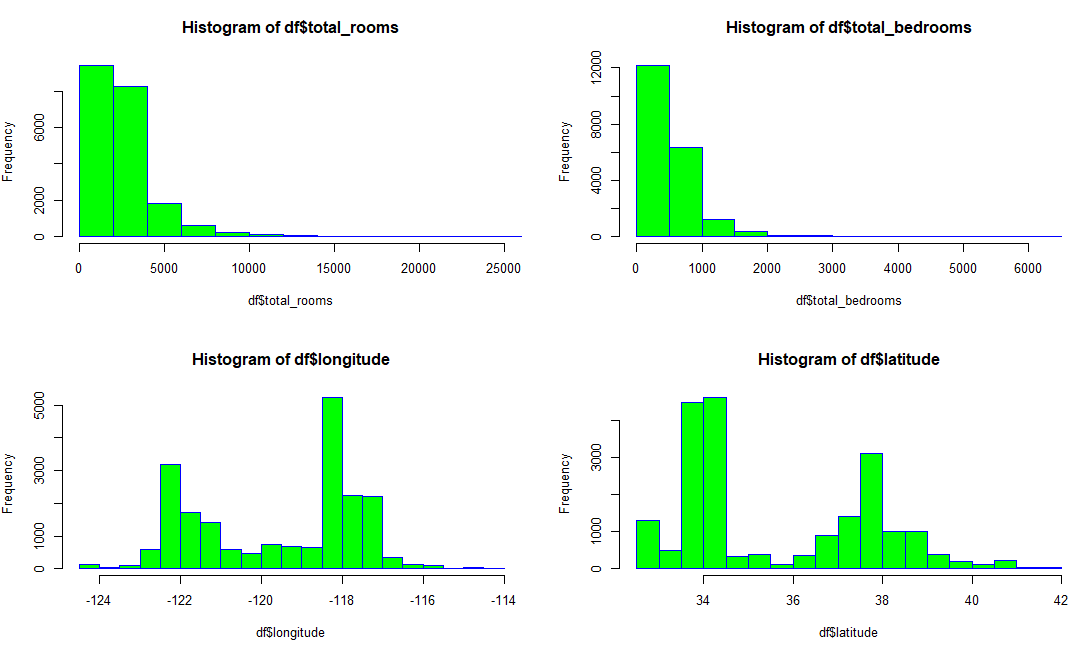
20640

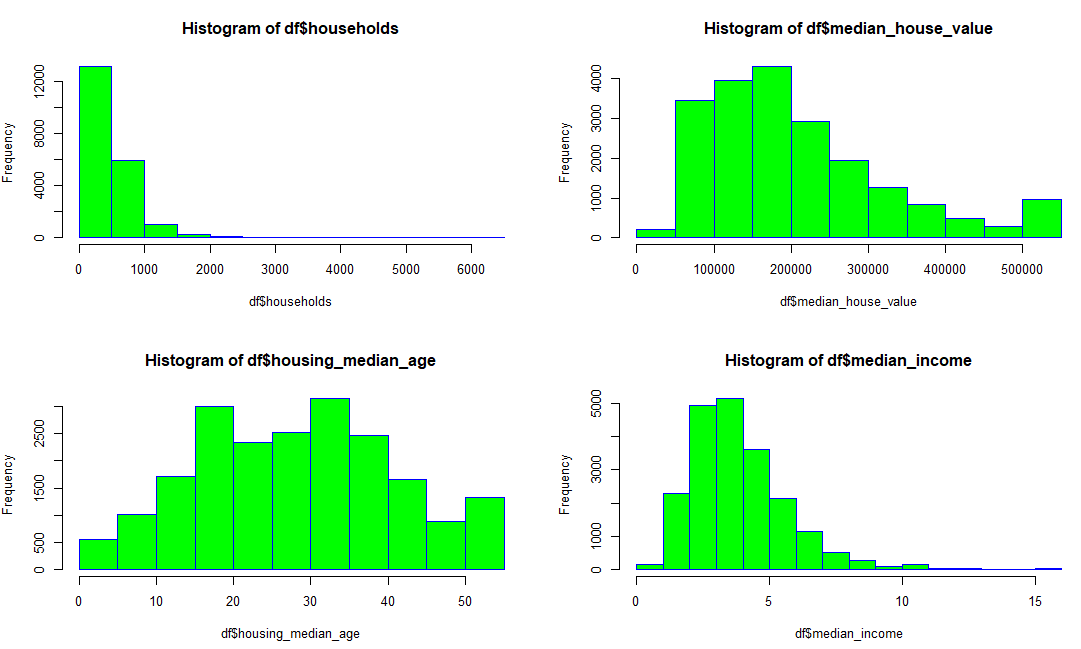
8)Categorization of values in ocean\_proximity columns into levels 1 to 5.  
data$ocean\_proximity=as.numeric(factor(data$ocean\_proximity,levels = c('<1H OCEAN','INLAND','ISLAND','NEAR BAY', 'NEAR OCEAN'),labels = c(1,2,3,4,5)))

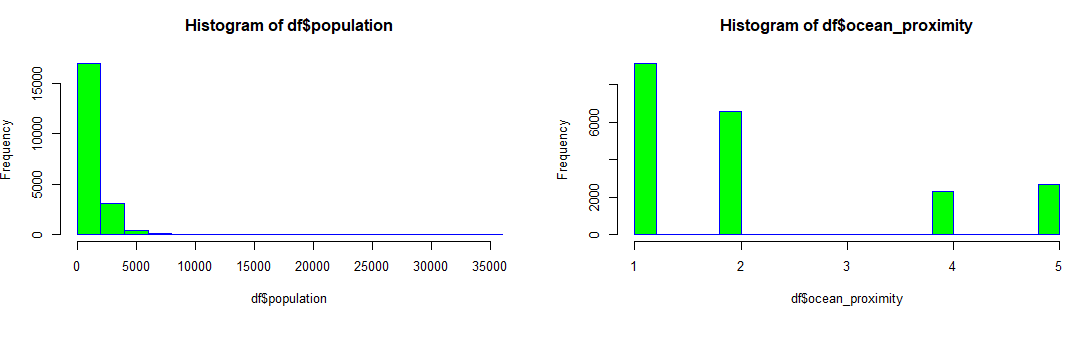
9)Viewing the data after categorization of ocean\_proximity values.  
View(data)

# EXPLORATORY ANALYSIS

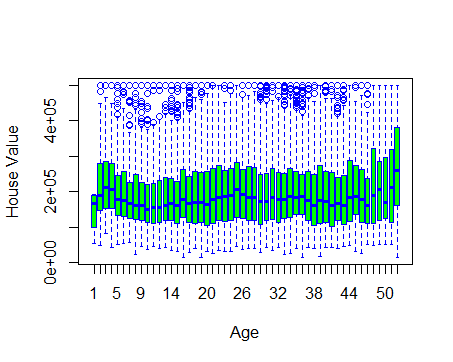
We draw the histogram plots of all the variables to check the range of values and their frequencies.



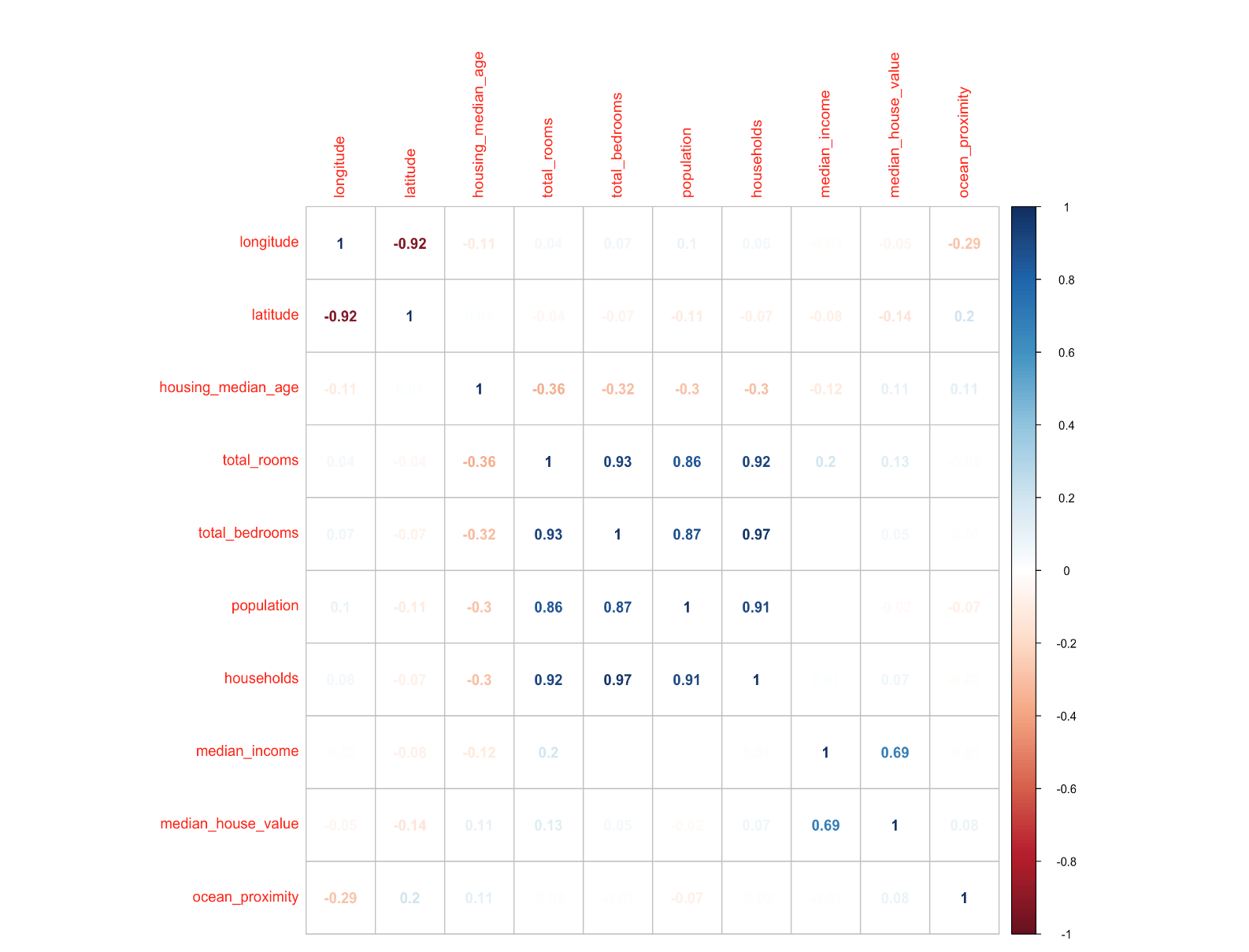




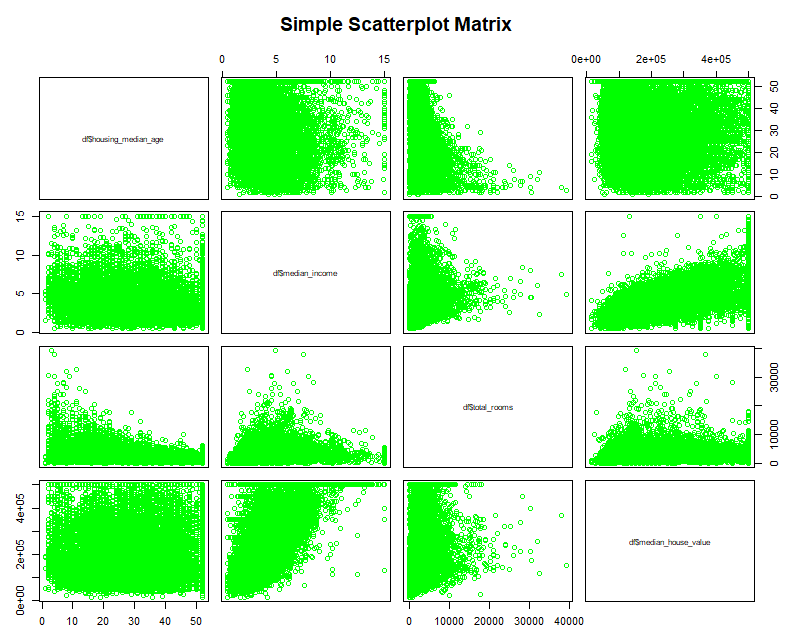
We observe the following:

1. There are close to 1500 housing blocks with homes which are old in age (>50 years).
2. There are houses worth more than 500,000... even in the 90s when this data was collected!
3. The median income has been scaled to have a value between 0 to 15.
4. 
5. From the above box plot we notice that from the age of 40, the purchasing value of the house increases.

10)Correlation matrix for visualizing all columns and their correlation.   
require(corrplot)  
M <- cor(data)  
corrplot(M, method = "number")



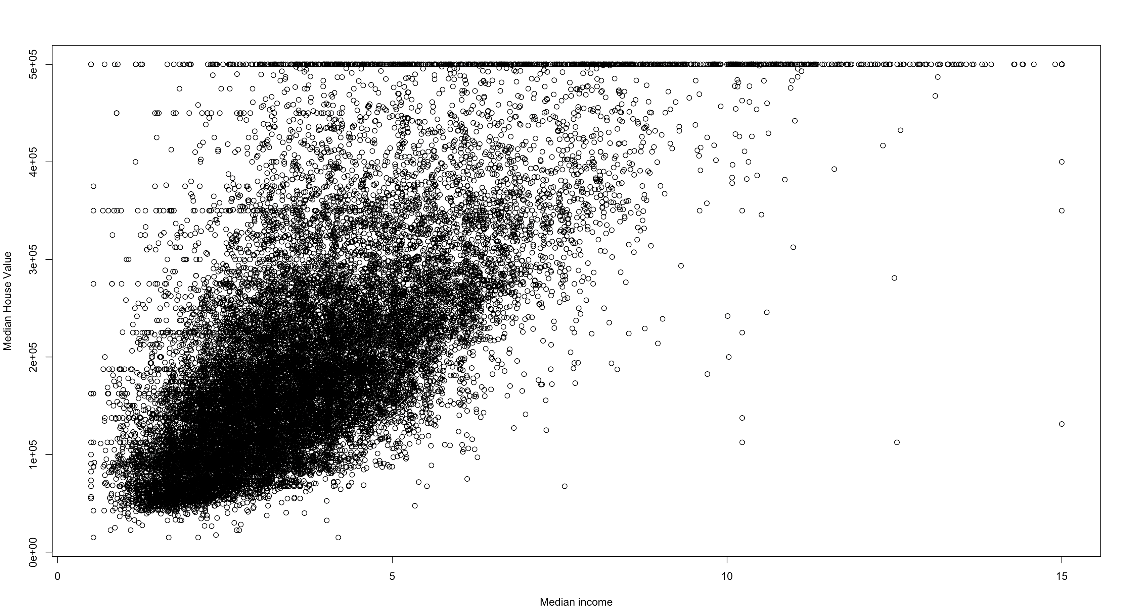
Summary- the above plot, we see that median\_house\_value is strongly correlated with median\_income (0.69) among other variables. The variables total\_rooms, latitude and house\_median\_age have a correlation of 0.2 and 0.12 and are correlated to the median\_house\_value to a lesser degree.



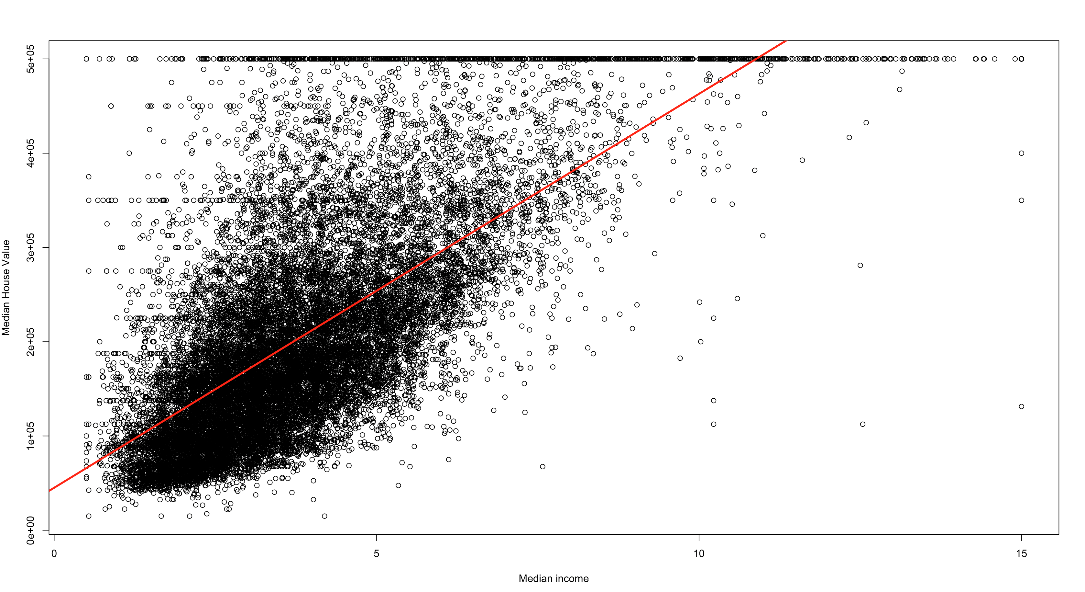
12)Plotting median\_house\_value against median\_income to visualize the correlation.

plot(data$median\_income,data$median\_house\_value, xlab="Median income", ylab="Median House Value")

Output:



abline(lm.fit,lwd=3,col="red"



The graph is showing positive linear co-relation between Median\_house\_vale and Median\_income.

# LINEAR REGRESSION

12) First we will try to use a model which has the attribute median\_house\_value with the highest correlation with median\_income.

attach(data)

lm.fit=lm(median\_house\_value ~ median\_income)  
summary(lm.fit)

Output:

Call:

lm(median\_house\_value ~ median\_income)

Residuals:

Min 1Q Median 3Q Max

-540697 -55950 -16979 36978 434023

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 45085.6 1322.9 34.08 <2e-16 \*\*\*

median\_income 41793.8 306.8 136.22 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 83740 on 20638 degrees of freedom

Multiple R-squared: 0.4734, Adjusted R-squared: 0.4734

F-statistic: 1.856e+04 on 1 and 20638 DF,p-value: < 2.2e-16

Summary- Here we are getting the Adjusted R-sqaured value as 0.4734 with one variable.

14)Predicting the median\_house\_value as a function of median\_income and total\_rooms using lm.

lm.fit1=lm(formula = data$median\_house\_value ~ data$median\_income+data$total\_rooms)  
summary(lm.fit1)

Output:

Call:

lm(formula = data$median\_house\_value ~ data$median\_income + data$total\_rooms)

Residuals:

Min 1Q Median 3Q Max

-541275 -55944 -17010 36993 433865

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 45290.4598 1406.7797 32.194 <2e-16 \*\*\*

data$median\_income 41820.3923 313.0120 133.606 <2e-16 \*\*\*

data$total\_rooms -0.1167 0.2726 -0.428 0.669

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 83740 on 20637 degrees of freedom

Multiple R-squared: 0.4735, Adjusted R-squared: 0.4734

F-statistic: 9278 on 2 and 20637 DF, p-value: < 2.2e-16

Summary- So when we add total rooms along with median\_income there is no significant change in R-square, as we can see that the p-vale of total\_rooms in this model is high which means that it is not significant in this model.

14)Predicting the median\_house\_value as a function of median\_income and hosuing\_median\_age using lm.

lm.fit2=lm(formula = data$median\_house\_value ~ data$median\_income + data$housing\_median\_age)   
summary(lm.fit2)

Output:

Call:

lm(formula = data$median\_house\_value ~ data$median\_income + data$housing\_median\_age)

Residuals:

Min 1Q Median 3Q Max

-596748 -53834 -15000 36719 446725

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -10189.03 1915.41 -5.32 1.05e-07 \*\*\*

data$median\_income 43169.19 298.36 144.69 < 2e-16 \*\*\*

data$housing\_median\_age 1744.13 45.04 38.73 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80850 on 20637 degrees of freedom

Multiple R-squared: 0.5091, Adjusted R-squared: 0.5091

F-statistic: 1.07e+04 on 2 and 20637 DF, p-value: < 2.2e-16

Summary- So when we add house\_median\_age along with median\_income there is some change in R-square, as we can see that the p-vale of house\_median\_age in this model is <0.05 which means that it is significant in this model.

15)Predicting the median\_house\_value as a function of median\_income and median\_income,housing\_median\_age and total\_rooms using lm.

lm.fit3=lm(formula = data$median\_house\_value ~ data$median\_income + data$housing\_median\_age+data$total\_rooms)   
summary(lm.fit3)  
Output:

Call:

lm(formula = data$median\_house\_value ~ data$median\_income + data$housing\_median\_age + data$total\_rooms)

Residuals:

Min 1Q Median 3Q Max

-584934 -53506 -15123 36448 450986

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.433e+04 2.160e+03 -11.26 <2e-16 \*\*\*

data$median\_income 4.247e+04 3.012e+02 140.98 <2e-16 \*\*\*

data$housing\_median\_age 1.975e+03 4.780e+01 41.32 <2e-16 \*\*\*

data$total\_rooms 3.888e+00 2.793e-01 13.92 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80480 on 20636 degrees of freedom

Multiple R-squared: 0.5137, Adjusted R-squared: 0.5136

F-statistic: 7266 on 3 and 20636 DF, p-value: < 2.2e-16

Summary- So when we add house\_median\_age and total rooms along with median\_income there is again slight change in R-square, as we can see that the p-vale of house\_median\_age and total\_rooms in this model is <0.05 which means that it is significant in this model.

17)Predicting the median\_house\_value as a function of all other columns using lm.

lm.fit5=lm(formula = data$median\_house\_value ~ ., data = data)  
  
summary(lm.fit5)  
  
Output:

Call:

lm(formula = data$median\_house\_value ~ ., data = data)

Residuals:

Min 1Q Median 3Q Max

-556662 -43911 -11448 30487 818858

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.566e+06 6.544e+04 -54.500 <2e-16 \*\*\*

longitude -4.257e+04 7.436e+02 -57.246 <2e-16 \*\*\*

latitude -4.245e+04 6.885e+02 -61.654 <2e-16 \*\*\*

housing\_median\_age 1.144e+03 4.327e+01 26.445 <2e-16 \*\*\*

total\_rooms -6.633e+00 7.766e-01 -8.541 <2e-16 \*\*\*

total\_bedrooms 8.114e+01 5.997e+00 13.531 <2e-16 \*\*\*

population -3.985e+01 1.078e+00 -36.970 <2e-16 \*\*\*

households 7.933e+01 6.747e+00 11.757 <2e-16 \*\*\*

median\_income 3.976e+04 3.333e+02 119.296 <2e-16 \*\*\*

ocean\_proximity 7.103e+01 3.675e+02 0.193 0.847

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69680 on 20630 degrees of freedom

Multiple R-squared: 0.6356, Adjusted R-squared: 0.6354

F-statistic: 3998 on 9 and 20630 DF, p-value: < 2.2e-16

Summary - Here we can see that in the full model, the Adjusted R-squared is 0.635 and we also note that for the variable ocean\_proximity, the p value is 0.847 which is greater than the threshold value of 0.05 meaning that this variabls is not significant in the model

16)Predicting the median\_house\_value as a function of all columns except ocean\_proximity using lm. We see how much difference will it make if we remove ocean\_proximity. As we can see from the value of Adjusted R-squared(0.6355), it did not impact the model.

lm.fit4=lm(formula = data$median\_house\_value ~ data$longitude + data$latitude + data$housing\_median\_age+data$total\_rooms+data$total\_bedrooms+data$population+data$households+data$median\_income, data = data)  
summary(lm.fit4)

Output:

Call:

lm(formula = data$median\_house\_value ~ data$longitude + data$latitude +

data$housing\_median\_age + data$total\_rooms + data$total\_bedrooms +

data$population + data$households + data$median\_income, data = data)

Residuals:

Min 1Q Median 3Q Max

-556726 -43914 -11463 30460 819314

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.570e+06 6.262e+04 -57.014 <2e-16 \*\*\*

data$longitude -4.261e+04 7.141e+02 -59.668 <2e-16 \*\*\*

data$latitude -4.248e+04 6.746e+02 -62.967 <2e-16 \*\*\*

data$housing\_median\_age 1.144e+03 4.324e+01 26.468 <2e-16 \*\*\*

data$total\_rooms -6.621e+00 7.741e-01 -8.553 <2e-16 \*\*\*

data$total\_bedrooms 8.116e+01 5.996e+00 13.537 <2e-16 \*\*\*

data$population -3.987e+01 1.072e+00 -37.208 <2e-16 \*\*\*

data$households 7.930e+01 6.746e+00 11.756 <2e-16 \*\*\*

data$median\_income 3.975e+04 3.318e+02 119.825 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69670 on 20631 degrees of freedom

Multiple R-squared: 0.6356, Adjusted R-squared: 0.6355

F-statistic: 4498 on 8 and 20631 DF, p-value: < 2.2e-16

# Non-linear Model

Now we see the fit of a non-linear model.

lm.fit6=lm(formula = data$median\_house\_value ~ (data$median\_income\*data$ocean\_proximity)+data$longitude + data$latitude + data$housing\_median\_age+ data$total\_rooms+data$total\_bedrooms+data$population+data$households+data$median\_income)  
summary(lm.fit6)  
Output:Call:

lm(formula = data$median\_house\_value ~ (data$median\_income \*

data$ocean\_proximity) + data$longitude + data$latitude +

data$housing\_median\_age + data$total\_rooms + data$total\_bedrooms +

data$population + data$households + data$median\_income)

Residuals:

Min 1Q Median 3Q Max

-539762 -43798 -11492 30467 825555

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.562e+06 6.536e+04 -54.500 < 2e-16 \*\*\*

data$median\_income 3.701e+04 5.099e+02 72.569 < 2e-16 \*\*\*

data$ocean\_proximity -4.903e+03 7.887e+02 -6.217 5.17e-10 \*\*\*

data$longitude -4.270e+04 7.430e+02 -57.470 < 2e-16 \*\*\*

data$latitude -4.267e+04 6.883e+02 -61.982 < 2e-16 \*\*\*

data$housing\_median\_age 1.135e+03 4.323e+01 26.263 < 2e-16 \*\*\*

data$total\_rooms -6.338e+00 7.767e-01 -8.160 3.53e-16 \*\*\*

data$total\_bedrooms 7.958e+01 5.994e+00 13.278 < 2e-16 \*\*\*

data$population -4.025e+01 1.078e+00 -37.336 < 2e-16 \*\*\*

data$households 8.035e+01 6.741e+00 11.920 < 2e-16 \*\*\*

data$median\_income:data$ocean\_proximity 1.214e+03 1.703e+02 7.126 1.07e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69590 on 20629 degrees of freedom

Multiple R-squared: 0.6365, Adjusted R-squared: 0.6363

F-statistic: 3612 on 10 and 20629 DF, p-value: < 2.2e-16

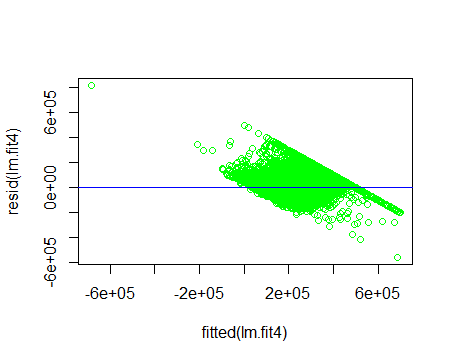
Summary- Even if we are making the model non-linear there is not much of an improvement from the previous models in terms of Adjusted R-squared.

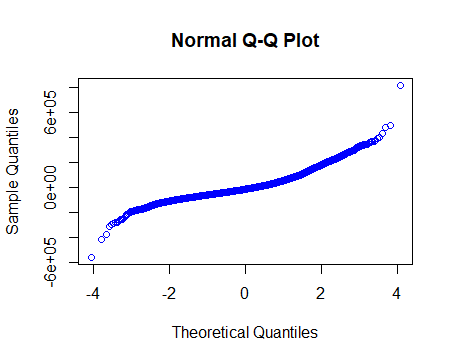
Plot(lm.fit6)

>plot(fitted(lm.fit4),resid(lm.fit4))

>abline(h=0,col="blue")

>qqnorm(resid(lm.fit4),col="blue")





We plotted the reduced fitted model against the residuals and obtained the plot below. We also

tested the normality assumption of the residuals. There do not seem to be any non-constant

vertical scatter in the plot in the scatter plot with fitted and residual values. However, is an

outlier in the plot as can be observed below. The QQ plot has a slight curvature but the

normality assumption seems to hold true.

Table comparing models-

|  |  |  |
| --- | --- | --- |
| Variables considered against Median\_House\_Value | Adjusted R-Square | F-statistics |
| Median\_income | 0.4734 | 1.856e+04 on 1 and 20638 DF |
| Median\_income+Total\_Rooms | 0.4734 | 9278 on 2 and 20637 DF |
| Median\_income+Housing\_Median\_Age | 0.5091 | 1.07e+04 on 2 and 20637 DF |
| Median\_income+Total\_Rooms+ Housing\_Median\_Age | 0.5136 | 7266 on 3 and 20636 DF |
| All variables – (Ocean\_proximity) | 0.6355 | 4498 on 8 and 20631 DF |
| All variables | 0.6354 | 3998 on 9 and 20630 DF |
| (Non-Linear) (Median\_Income\*Ocean\_proximity)+rest of the variables | 0.6363 | 3612 on 10 and 20629 DF |

# Partial F-Test

We compared this reduced model against the full model used in the beginning using ANOVA

function in R. Our Hypothesis assumption are as follows:

*Null Hypothesis*: No additional parameters in the full model that are not included in the reduced

model are useful

*Alternative Hypothesis*: At least one parameter in the full model that is not included in the

reduced model are useful.

> anova(lm.fit5,lm.fit4)

Analysis of Variance Table

Model 1: data$median\_house\_value ~ longitude + latitude + housing\_median\_age +

total\_rooms + total\_bedrooms + population + households +

median\_income + ocean\_proximity

Model 2: data$median\_house\_value ~ data$longitude + data$latitude + data$housing\_median\_age +

data$total\_rooms + data$total\_bedrooms + data$population +

data$households + data$median\_income

Res.Df RSS Df Sum of Sq F Pr(>F)

1 20630 1.0015e+14

2 20631 1.0015e+14 -1 -181313480 0.0373 0.8468

The p-value is 0.8468 which greater than our assumed level of significance of 0.05. We accept

null hypothesis. Based on our null hypothesis that additional parameter(ocan proximity) included in the full

model are not useful. We accept the null hypothesis that additional parameters are not

necessary to use in the model.

# Using the predict function

predict(lm.fit,data.frame(median\_income=(c(5,10,15))),interval = "confidence")

Output:

fit lwr upr

1 254054.8 252725.8 255383.9

2 463024.1 459165.1 466883.0

3 671993.3 665203.7 678782.9

predict(lm.fit,data.frame(median\_income=(c(5,10,15))),interval = "prediction")

Output:

fit lwr upr

1 254054.8 89917.07 418192.6

2 463024.1 298846.34 627201.8

3 671993.3 507720.58 836266.1

# Conclusion

Based on the forward selection , we found that the model with attributes longitude , latitude , housing\_median\_age, total\_rooms , total\_bedrooms , population , households and median\_income predictor seems to have the highest Adjusted R-Squared and the lowest residual standard error.

Section2 :R Code

data <- read.csv(file="https://docs.google.com/spreadsheets/d/e/2PACX-1vQAJvpd8Dk\_0pgS0t\_njNRLanJcxBn5Em7OO7Ew8wckt9Lr0SqZ0gLa48c9v1svsZu1w4RE1Mf1BKgM/pub?gid=1870803675&single=true&output=csv", header=TRUE, sep=",")

View(data)

names(data)

for(i in 1:ncol(data)) {

colName <- colnames(data[i])

pctNull <- sum(is.na(data[,i]))/length(data[,i])

if (pctNull >0.01) {

print(paste("Column ", colName, " has ", round(pctNull\*100, 3), "% of nulls"))

}

}

data$total\_bedrooms[is.na(data$total\_bedrooms)]=median(data$total\_bedrooms,na.rm = TRUE)

table(is.na(data$total\_bedrooms))

data$ocean\_proximity=as.numeric(factor(data$ocean\_proximity,levels = c('<1H OCEAN','INLAND','ISLAND','NEAR BAY', 'NEAR OCEAN'),labels = c(1,2,3,4,5)))

View(data)

require(corrplot)

M <- cor(data)

corrplot(M, method = "number")

attach(data)

lm.fit=lm(median\_house\_value ~ median\_income)

summary(lm.fit)

plot(data$median\_income,data$median\_house\_value, xlab="Median income", ylab="Median House Value")

boxplot(data$median\_house\_value ~ data$housing\_median\_age, xlab="Age", ylab="House Value")

abline(lm.fit,lwd=3,col="red")

predict(lm.fit,data.frame(median\_income=(c(5,10,15))),interval = "confidence")

predict(lm.fit,data.frame(xhefv=(c(5,10,15))), interval="predicition")

plot(data$median\_income,data$median\_house\_value, xlab="Median income", ylab="Median House Value")

abline(lm.fit,lwd=3,col="red")

lm.fit1=lm(formula = data$median\_house\_value ~ data$median\_income+data$total\_rooms)

summary(lm.fit1)

lm.fit2=lm(formula = data$median\_house\_value ~ data$median\_income + data$housing\_median\_age)

summary(lm.fit2)

lm.fit3=lm(formula = data$median\_house\_value ~ data$median\_income + data$housing\_median\_age+data$total\_rooms)

summary(lm.fit3)

lm.fit4=lm(formula = data$median\_house\_value ~ data$longitude + data$latitude + data$housing\_median\_age+ data$total\_rooms+data$total\_bedrooms+data$population+data$households+data$median\_income, data = data)

summary(lm.fit4)

lm.fit5=lm(formula = data$median\_house\_value ~ ., data = data)

summary(lm.fit5)

#anova(lm.fit5,lm.fit4)

lm.fit6=lm(formula = data$median\_house\_value ~ (data$median\_income\*data$ocean\_proximity)+data$longitude + data$latitude + data$housing\_median\_age+ data$total\_rooms+data$total\_bedrooms+data$population+data$households+data$median\_income)

summary(lm.fit6)

# Residual plot and checking for normality assumption for the reduced model

plot(fitted(lm.fit4),resid(lm.fit4))

abline(h=0,col="blue")

qqnorm(resid(lm.fit4),col="blue")