Kritagya Nepal

Number of members in Group: 1

NetId: kxn190007

Linear Regression

California Housing Dataset prediction using Linear Regression.

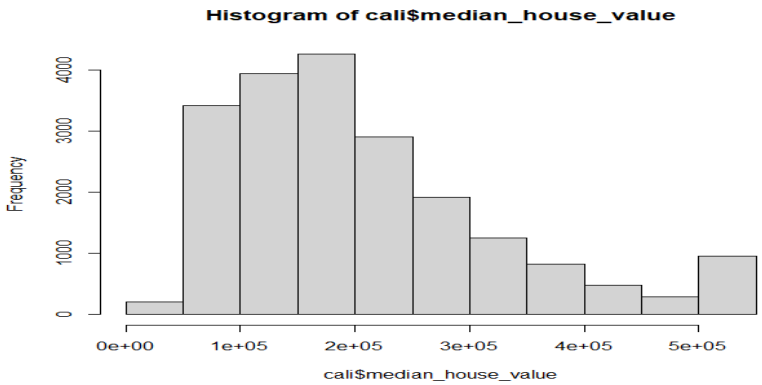
First, we look at the data as it is and see that there are 20640 rows and 10 columns. The summary of the dataset is shown below:

```
> dim(california)
[1] 20640    10
> summary(california)
   longitude          latitude      housing_median_age  total_rooms     total_bedrooms
 Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2   Min.   :   1.0
 1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448   1st Qu.: 296.0
 Median :-118.5   Median :34.26   Median :29.00      Median : 2127   Median : 435.0
 Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636   Mean   : 537.9
 3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148   3rd Qu.: 647.0
 Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320   Max.   :6445.0
                                                                    NA's   :207
   population      households     median_income     median_house_value ocean_proximity
 Min.   :    3   Min.   :   1.0   Min.   : 0.4999   Min.   : 14999     Length:20640
 1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634   1st Qu.:119600     Class :character
 Median : 1166   Median : 409.0   Median : 3.5348   Median :179700     Mode  :character
 Mean   : 1425   Mean   : 499.5   Mean   : 3.8707   Mean   :206856
 3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432   3rd Qu.:264725
 Max.   :35682   Max.   :6082.0   Max.   :15.0001   Max.   :500001
```

The null values from the data are removed by using na.exclude() command, after which we have we are left with 20433 rows and 10 column. We lose 207 rows while doing so.

Visualization:

We first plot the median_house prices using histogram and see that the house prices form a right skewed bell curve.



From the corrplot in Figure:1 we see that the following features are strongly corelated:

i. Total_rooms and Population

ii. Population and Households

iii. Total_rooms and household

iv. Median Income and Median House Value

Getting the correlation between output value (Median House Value) and all other features we see that median house value is strongly corelated with median income, since it has the highest value among all other features. This result is shown in Figure2.

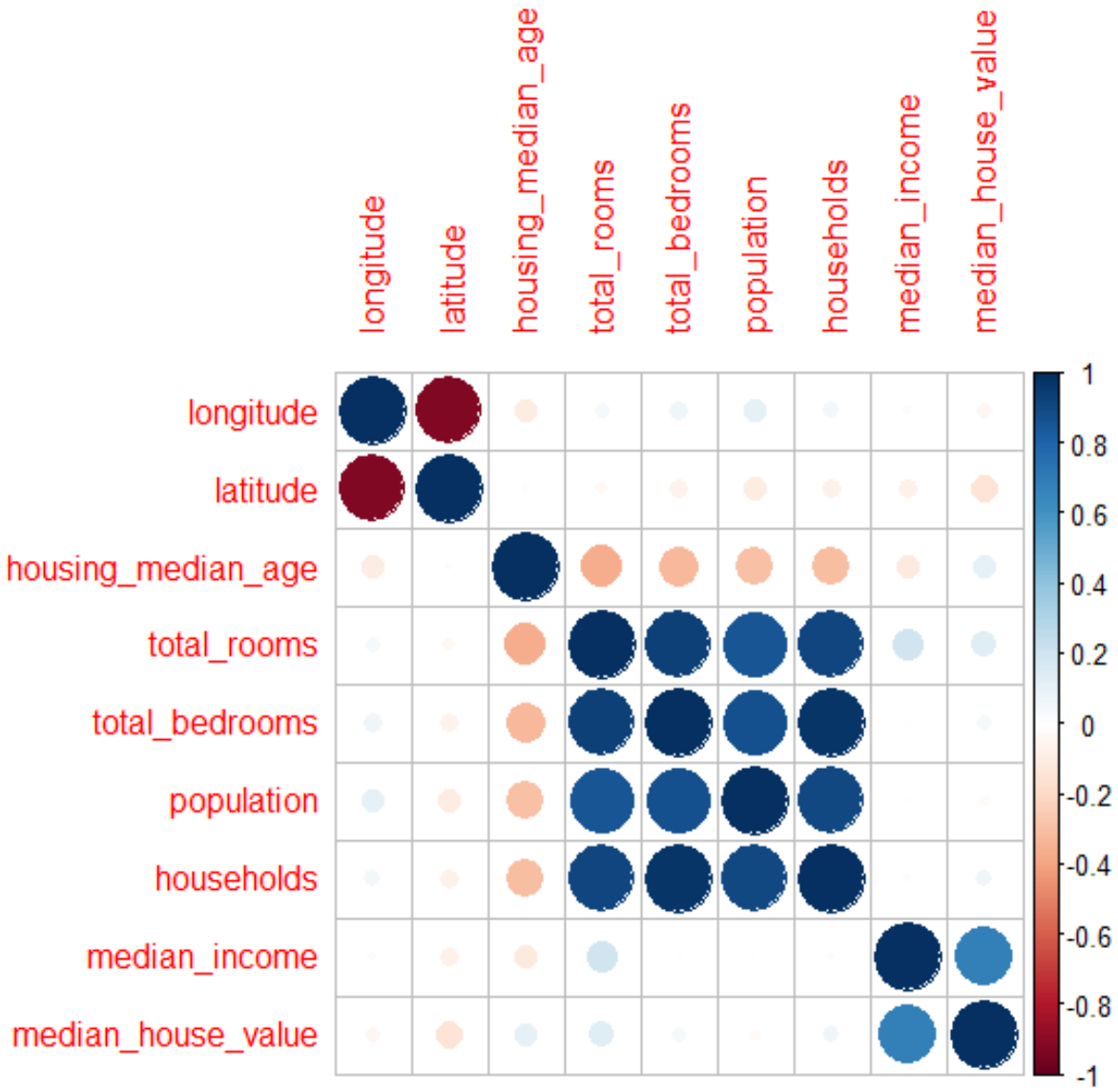


Figure 1

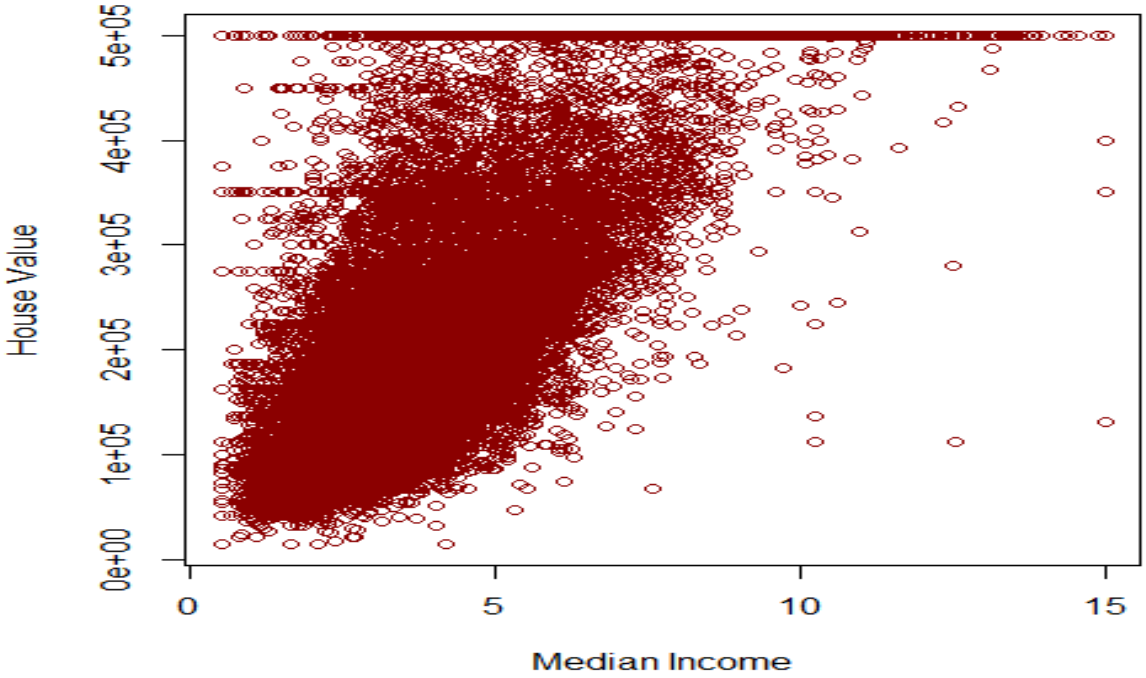
List of all features corelated with median house value:

```
> cor(cali$median_house_value,cali$longitude)
[1] -0.04539822
> cor(cali$median_house_value,cali$latitude)
[1] -0.1446382
> cor(cali$median_house_value,cali$housing_median_age)
[1] 0.106432
> cor(cali$median_house_value,cali$total_rooms)
[1] 0.1332941
> cor(cali$median_house_value,cali$total_bedrooms)
[1] 0.04968618
> cor(cali$median_house_value,cali$population)
[1] -0.02529973
> cor(cali$median_house_value,cali$households)
[1] 0.06489355
> cor(cali$median_house_value,cali$median_income)
[1] 0.6883555
```

Here we plot the housing price with comparison to some other important features.

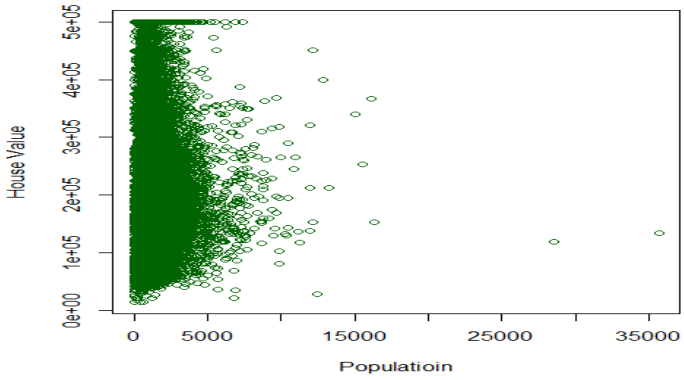
i. House Value Vs Median Income:

The house values increases with increase in median income. The data shows positive correlation between these two variables. This implies that people who have higher median income have expensive house expect some outliers.



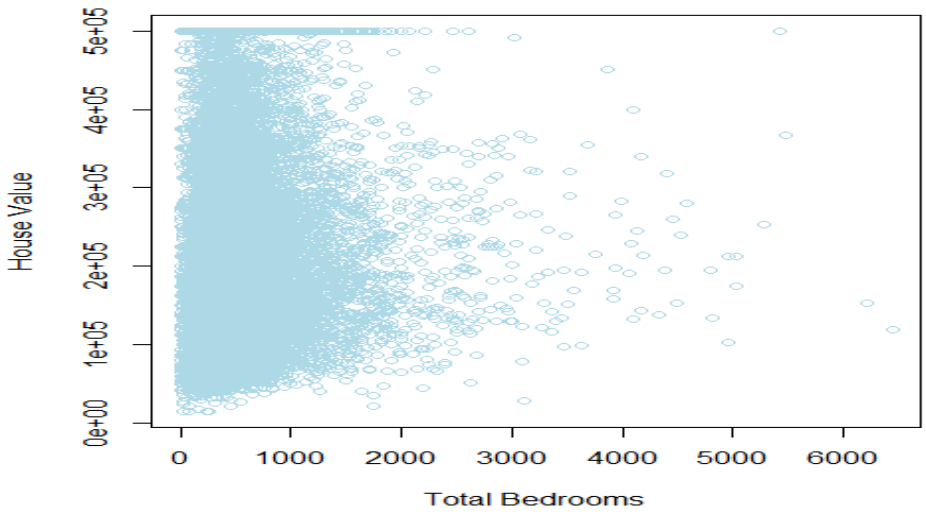
ii. House Value Vs Population:

There seems to be a big range in house price with population between 0 and 5K. This implies that the most people but house where there is population between 0 and 5K. The price is normally distributed in this region.



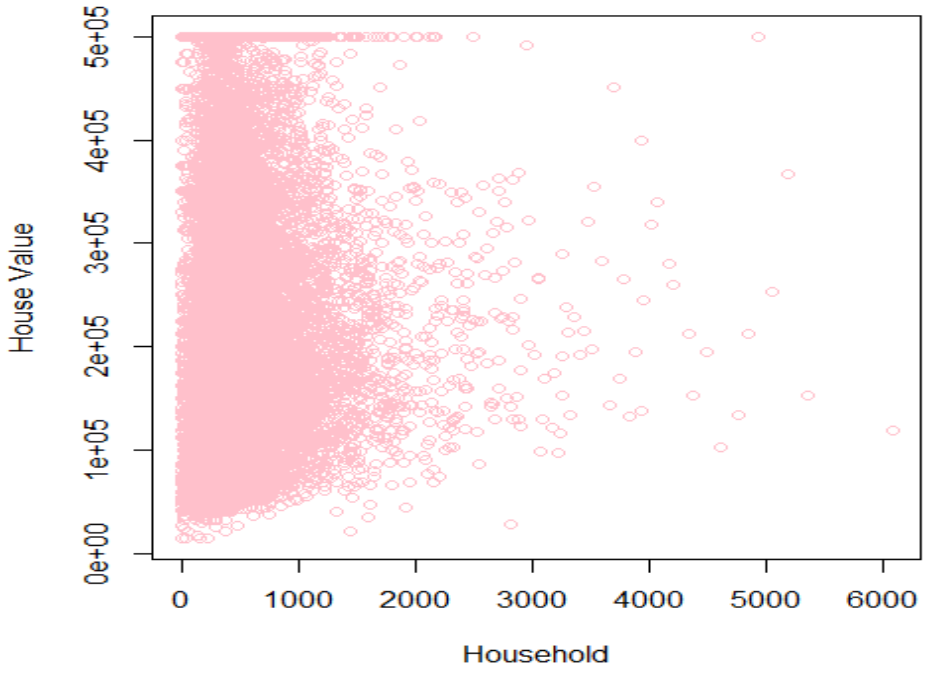
iii. House Value Vs Total Bedroom:

Most of the house have 0-1000 bedrooms, the house prices are evenly distributed around them.



iv. House Value Vs Household:

Most of the house prices are between 0-1K Households and is evenly distributed.



Linear Model:

We first create a linear model using each of the features and see the ones that most fits the given data.

```
#linear model
model1 <- lm(cali$median_house_value~cali$longitude,cali)
model2 <- lm(cali$median_house_value~cali$latitude,cali)
model3 <- lm(cali$median_house_value~cali$housing_median_age,cali)
model4 <- lm(cali$median_house_value~cali$total_rooms)
model5 <- lm(cali$median_house_value~cali$population,cali)
model6 <- lm(cali$median_house_value~cali$households,cali)
model7 <- lm(cali$median_house_value~cali$median_income,cali)
model8 <- lm(cali$median_house_value~cali$total_bedrooms)
```

After creating the model, we see the summary of each:

```
#Summary od each individual model.
summary(model1)
summary(model2)
summary(model3)
summary(model4)
summary(model5)
summary(model6)
summary(model7)
summary(model8)
```

The summary for each model are shown:

```
> summary(model1)

Call:
lm(formula = cali$median_house_value ~ cali$longitude, data = cali)

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 *
cali$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

> summary(model2)

Call:
lm(formula = cali$median_house_value ~ cali$latitude, data = cali)

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 ***
cali$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
```

```
> summary(model3)

Call:
lm(formula = cali$median_house_value ~ cali$housing_median_age,
    data = cali)

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 ***
cali$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

> summary(model4)

Call:
lm(formula = cali$median_house_value ~ cali$total_rooms)

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 ***
cali$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
```

```
> summary(model5)

Call:
lm(formula = cali$median_house_value ~ cali$population, data = cali)

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 ***
cali$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

> summary(model6)

Call:
lm(formula = cali$median_house_value ~ cali$households, data = cali)

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 ***
cali$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
```

```
> summary(model7)

Call:
lm(formula = cali$median_house_value ~ cali$median_income, data = cali)

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 ***
cali$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

> summary(model8)

Call:
lm(formula = cali$median_house_value ~ cali$total_bedrooms)

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 ***
cali$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
```

From the summary of each of them we can see that Latitude, Housing Median Age, Median Income and Total rooms has high F-Statistics and R^2 values. Since the correlation between median income and housing price is higher, we analyze their summary.

Model7 has a standard error of 83740 with 20431 degrees of freedom but the R^2 and the F-statistics are optimal hence, this model is better than others.

Now we look at a model with all features excluding the ocean_proximity.

```
> summary(model9)

Call:
lm(formula = cali$median_house_value ~ ., data = cali)

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 ***
longitude                 -2.681e+04  1.020e+03 -26.296  < 2e-16 ***
latitude                  -2.548e+04  1.005e+03 -25.363  < 2e-16 ***
housing_median_age         1.073e+03  4.389e+01  24.439  < 2e-16 ***
total_rooms               -6.193e+00  7.915e-01  -7.825 5.32e-15 ***
total_bedrooms             1.006e+02  6.869e+00  14.640  < 2e-16 ***
population                -3.797e+01  1.076e+00 -35.282  < 2e-16 ***
households                 4.962e+01  7.451e+00   6.659 2.83e-11 ***
median_income              3.926e+04  3.380e+02 116.151  < 2e-16 ***
ocean_proximityINLAND     -3.928e+04  1.744e+03 -22.522  < 2e-16 ***
ocean_proximityISLAND      1.529e+05  3.074e+04   4.974 6.62e-07 ***
ocean_proximityNEAR BAY   -3.954e+03  1.913e+03  -2.067  0.03879 *
ocean_proximityNEAR OCEAN  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
```

Here, the F-statistic is 3112 and R^2 value is 0.6463. Thus, this model is relatively a good fit. This model includes all the features including the derived features. To see the model without the derived features we create another model10, which will take into consideration, longitude, latitude, total rooms, population, median income, and house median age.

```
> summary(model10)

Call:
lm(formula = cali$median_house_value ~ cali$longitude + cali$latitude +
    cali$total_rooms + cali$population + cali$median_income +
    cali$housing_median_age, data = cali)

Residuals:
    Min      1Q  Median      3Q     Max
-503984  -47000  -13174   32448  514087

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -3.968e+06  6.326e+04  -62.72   <2e-16 ***
cali$longitude          -4.774e+04  7.196e+02  -66.34   <2e-16 ***
cali$latitude           -4.777e+04  6.798e+02  -70.27   <2e-16 ***
cali$total_rooms         1.504e+01  5.009e-01   30.02   <2e-16 ***
cali$population         -2.541e+01  9.405e-01  -27.01   <2e-16 ***
cali$median_income       3.431e+04  3.033e+02  113.10   <2e-16 ***
cali$housing_median_age  1.118e+03  4.485e+01   24.93   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71980 on 20426 degrees of freedom
Multiple R-squared:  0.6113,    Adjusted R-squared:  0.6112
F-statistic:  5353 on 6 and 20426 DF,  p-value: < 2.2e-16
```

Here, F-statistic is 5353 and R^2 is 0.6112, which is optimal considering all the important features. Thus this model best explains the data.