Data Exploration and Median House Value prediction of California housing price in 1990

Sasima Deesriwong
Instructor: Daniel D. Gutierrez

Introduction to Data Science
Final class project report UCLA Extension – Winter 2020

Contents

1.Project description and Dataset

Housing pricing in California is increasing every year due to high demanding especially when we live in the state where is the center of top leading technology company. Let's throw back time to 1990, 30 years ago and find out which variables had effect housing price value at that time.

Data set

There are 20,640 observations and 10 variables in this data set. Here are the title and definition details of each variables;

1. **longitude:** A measure of how far west a house is; a higher value is farther west

2. **latitude:** A measure of how far north a house is; a higher value is farther north

3. **housing_median_age:** Median age of a house within a block; a lower number is a newer building

4. **total_rooms:** Total number of rooms within a block

5. **total_bedrooms:** Total number of bedrooms within a block

6. **population:** Total number of people residing within a block

7. **households:** Total number of households, a group of people residing within a home unit, for a block

8. **median_income:** Median income for households within a block of houses (measured in tens of thousands of US Dollars)

9. **median_house_value:** Median house value for households within a block (measured in US Dollars)

10. **ocean_proximity:** Location of the house w.r.t ocean/sea

Please note that from all variables as mentioned above there are 9 numeric variables and 1 categorical variables.

Summary of data set

This is ‘cahousing’ data from ‘housing.csv’ file. There are 207 NA data in total bedrooms variables which needed to do some manipulation and I will walk you through in later stage.

```
> summary(cahousing)
   longitude         latitude     housing_median_age  total_rooms    total_bedrooms     population
 Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2   Min.   :   1.0   Min.   :    3
 1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448   1st Qu.: 296.0   1st Qu.:  787
 Median :-118.5   Median :34.26   Median :29.00      Median : 2127   Median : 435.0   Median : 1166
 Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636   Mean   : 537.9   Mean   : 1425
 3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148   3rd Qu.: 647.0   3rd Qu.: 1725
 Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320   Max.   :6445.0   Max.   :35682
                                                                    NA's   :207
   households     median_income     median_house_value   ocean_proximity
 Min.   :   1.0   Min.   : 0.4999   Min.   : 14999     <1H OCEAN :9136
 1st Qu.: 280.0   1st Qu.: 2.5634   1st Qu.:119600     INLAND    :6551
 Median : 409.0   Median : 3.5348   Median :179700     ISLAND    :   5
 Mean   : 499.5   Mean   : 3.8707   Mean   :206856     NEAR BAY  :2290
 3rd Qu.: 605.0   3rd Qu.: 4.7432   3rd Qu.:264725     NEAR OCEAN:2658
 Max.   :6082.0   Max.   :15.0001   Max.   :500001
```

```
> str(cahousing)
'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 ...
```

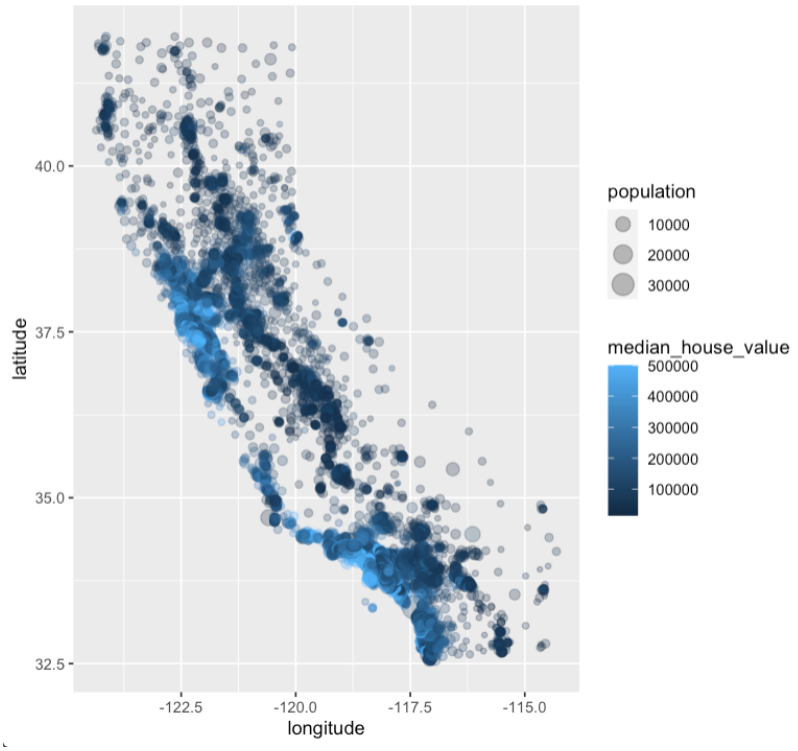
Here are the first 6 actual data

```
> head(cahousing)
  longitude latitude housing_median_age total_rooms total_bedrooms population households median_income median_house_value ocean_proximity
1   -122.23    37.88                41         880            129        322        126        8.3252            452600        NEAR BAY
2   -122.22    37.86                21        7099           1106       2401       1138        8.3014            358500        NEAR BAY
3   -122.24    37.85                52        1467            190        496        177        7.2574            352100        NEAR BAY
4   -122.25    37.85                52        1274            235        558        219        5.6431            341300        NEAR BAY
5   -122.25    37.85                52        1627            280        565        259        3.8462            342200        NEAR BAY
6   -122.25    37.85                52         919            213        413        193        4.0368            269700        NEAR BAY
```

2. Exploratory data analysis

2.1 Data visualization

Before jump into data manupulation, let see initial data visualization to understand data more clearly.



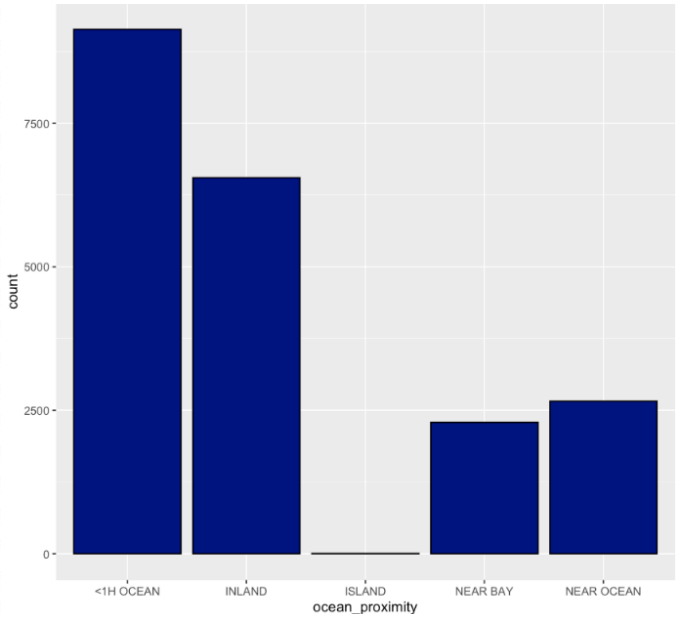
Here is plot data of latitude and longitude which results showing California map. Left side is closed to North Pacific Ocean.

The initial results of this visualization shown that at the edge of California coast , median house value are crowed with higher price when compare to others area.

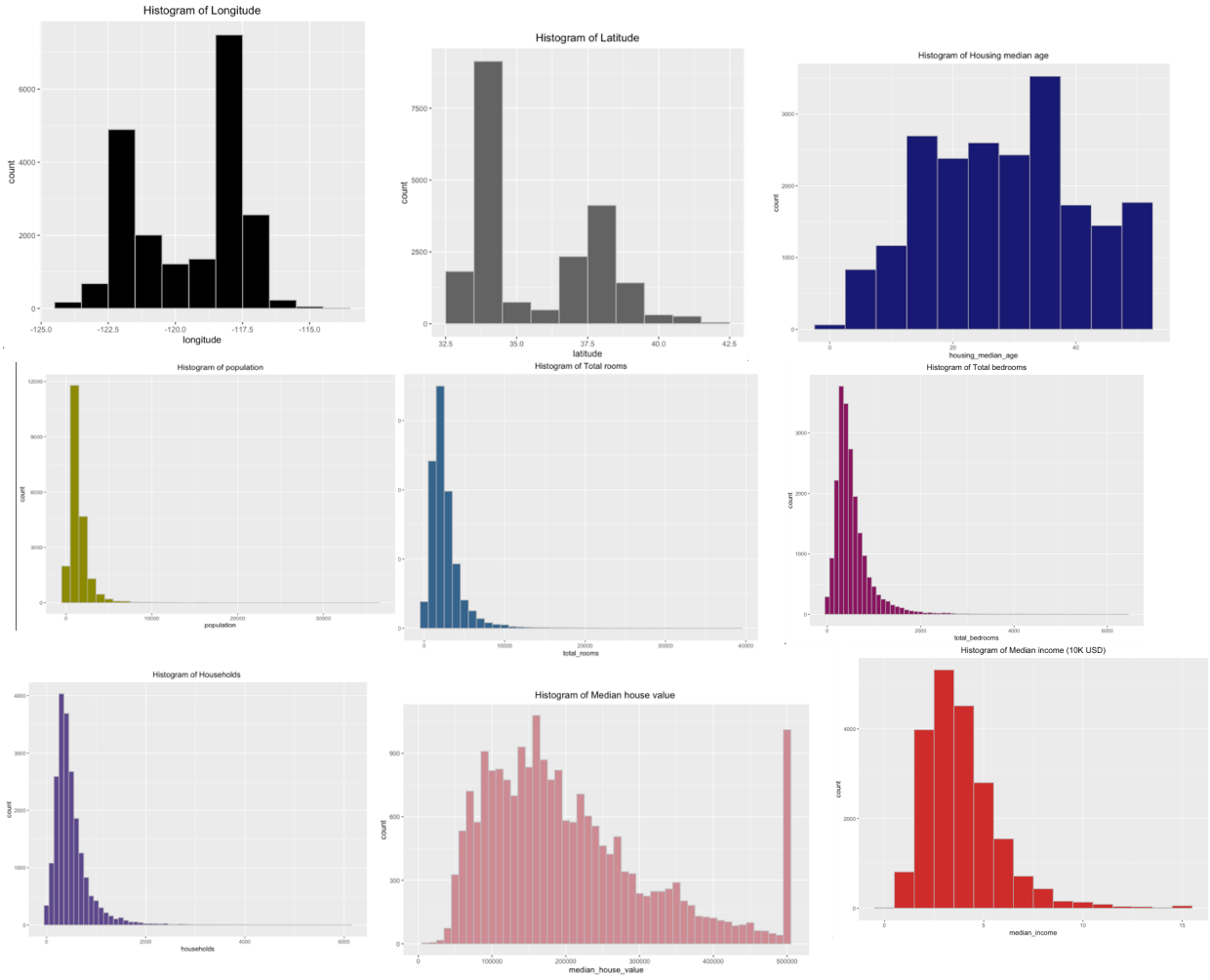
And population are crowded in coast side and in the south of California.

Bar chart on the right side is showing counts number of each housing based on ocean proximity category.

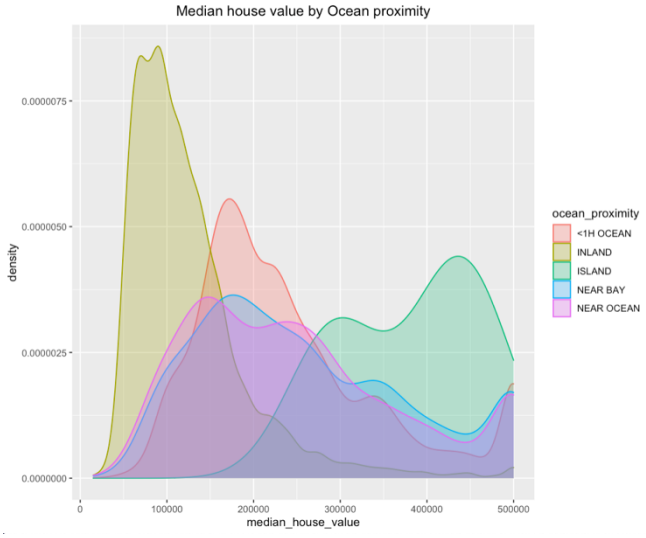
From this data set, housing there is the highest volume of housing that close to ocean (less than 1 hour from ocean) followed by inland, near ocean, near bay and island respectively



Here are histogram of all numeric variables to see overview of data distribution



Median age: Most of houses was built around 18- 37 years ago
Total rooms: Total rooms in data are mostly less than 5,000 rooms within a block
Total bedrooms: Total bedrooms are mostly less than 2000 rooms within a block
Population: Number of Population resided in a block are less than 5000
Median income: Distribution of data are mostly located between 2-5 (10K USD)
Households: Total group of household within a block are less than 1,000
Median House value: Data is concentrated between 100,000 -200,000 but have its peak data at 500,000



Here is density plot of median house values by ocean proximity. We can see that Island has the highest price while Inland data is crowded in lower side. And less than hour from ocean house and bay house disperse the most.

2.2 Data transformation

Missing value manipulation

Method to manage NA data for total bedrooms is that I used median of all others total bedrooms data, calculate median and then replace NA data with Median data.
Here is a result before and after

Before

```
> summary(cahousing$total_bedrooms)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    1.0   296.0   435.0   537.9   647.0  6445.0     207
```

After

```
>    summary(cahousing$total_bedrooms)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0   297.0   435.0   536.8   643.2  6445.0
```

Ocean proximity

In the earlier section, we saw from the California map that there is potentially that ocean proximity could influencer median house value. So we splitted all ocean proximity , create new column with binary category. Here is result after manipulated.

We now have 14 variables with the same 20,640 observations.

```
> summary(cahousing_new)
   longitude         latitude     housing_median_age  total_rooms    total_bedrooms     population      households
 Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2   Min.   :   1.0   Min.   :    3   Min.   :   1.0
 1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448   1st Qu.: 297.0   1st Qu.:  787   1st Qu.: 280.0
 Median :-118.5   Median :34.26   Median :29.00      Median : 2127   Median : 435.0   Median : 1166   Median : 409.0
 Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636   Mean   : 536.8   Mean   : 1425   Mean   : 499.5
 3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148   3rd Qu.: 643.2   3rd Qu.: 1725   3rd Qu.: 605.0
 Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320   Max.   :6445.0   Max.   :35682   Max.   :6082.0
 median_income     median_house_value   ONEH_OCEAN         INLAND           ISLAND            NEAR_BAY        NEAR_OCEAN
 Min.   : 0.4999   Min.   : 14999     Min.   :0.0000   Min.   :0.0000   Min.   :0.0000000   Min.   :0.0000   Min.   :0.0000
 1st Qu.: 2.5634   1st Qu.:119600     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000000   1st Qu.:0.0000   1st Qu.:0.0000
 Median : 3.5348   Median :179700     Median :0.0000   Median :0.0000   Median :0.0000000   Median :0.0000   Median :0.0000
 Mean   : 3.8707   Mean   :206856     Mean   :0.4426   Mean   :0.3174   Mean   :0.0002422   Mean   :0.1109   Mean   :0.1288
 3rd Qu.: 4.7432   3rd Qu.:264725     3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :15.0001   Max.   :500001     Max.   :1.0000   Max.   :1.0000   Max.   :1.0000000   Max.   :1.0000   Max.   :1.0000
```

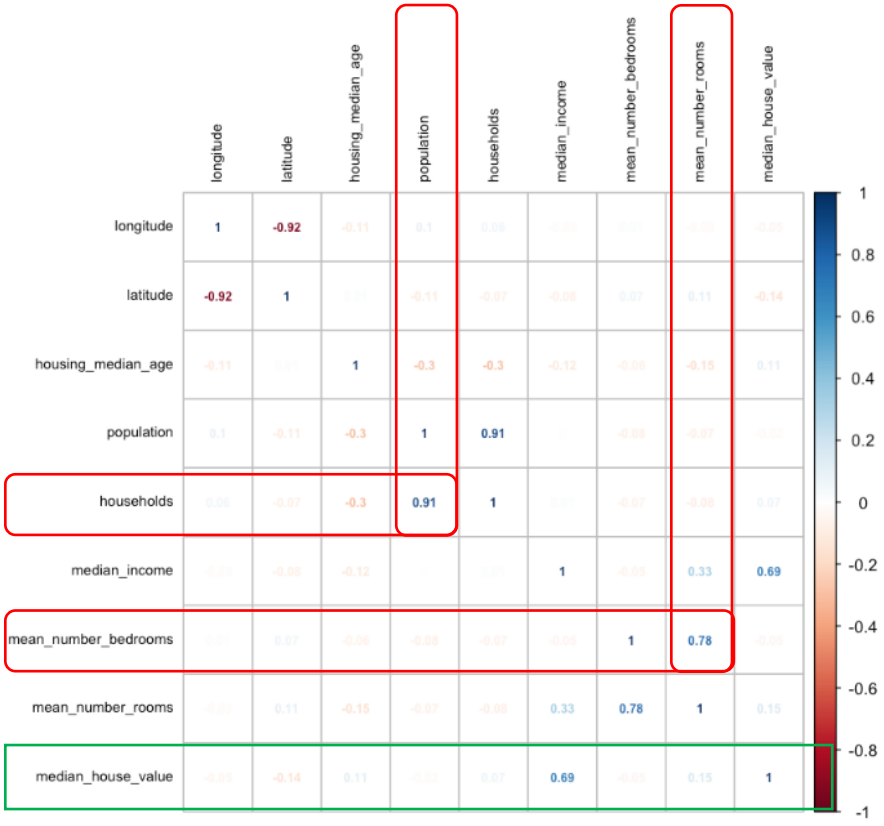
Next we will go through prediction section.

3. Price prediction model

First, we decided to go through correlation plot for 2 objectives.

1. To see if any variable is correlated to our dependent variables; median house value
2. To see if any potential of multicollinearity events of highly correlated of independent variables

Here is the correlation matrix:

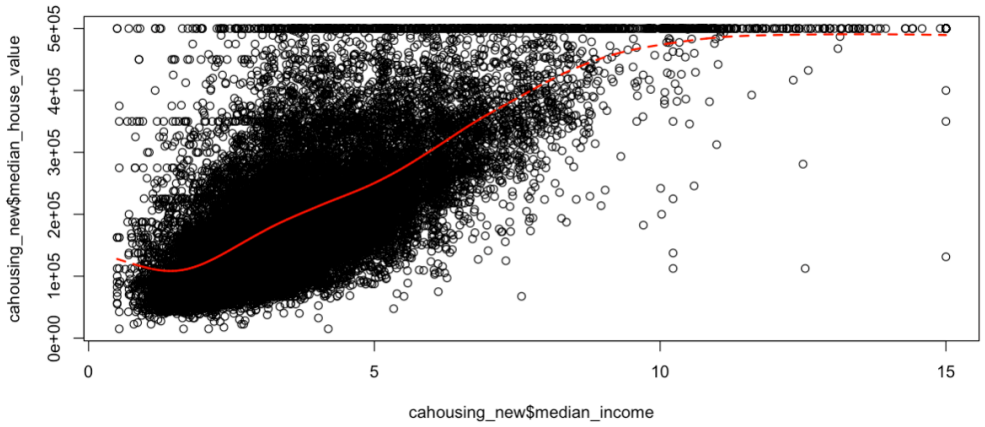


The green frame is showing correlation of median house value with others variables. Finding that median income is quite highly correlated to median house value. So we can definitely use this variable for our prediction model.

The red frames are showing relationship between 2 independent variables , so we should aware of this and should not use these pair together in the same prediction model; household & population, mean number of bedrooms & mean number of rooms.

Model 1

Here is plot of median house value & median income adding smooth line to see trends line



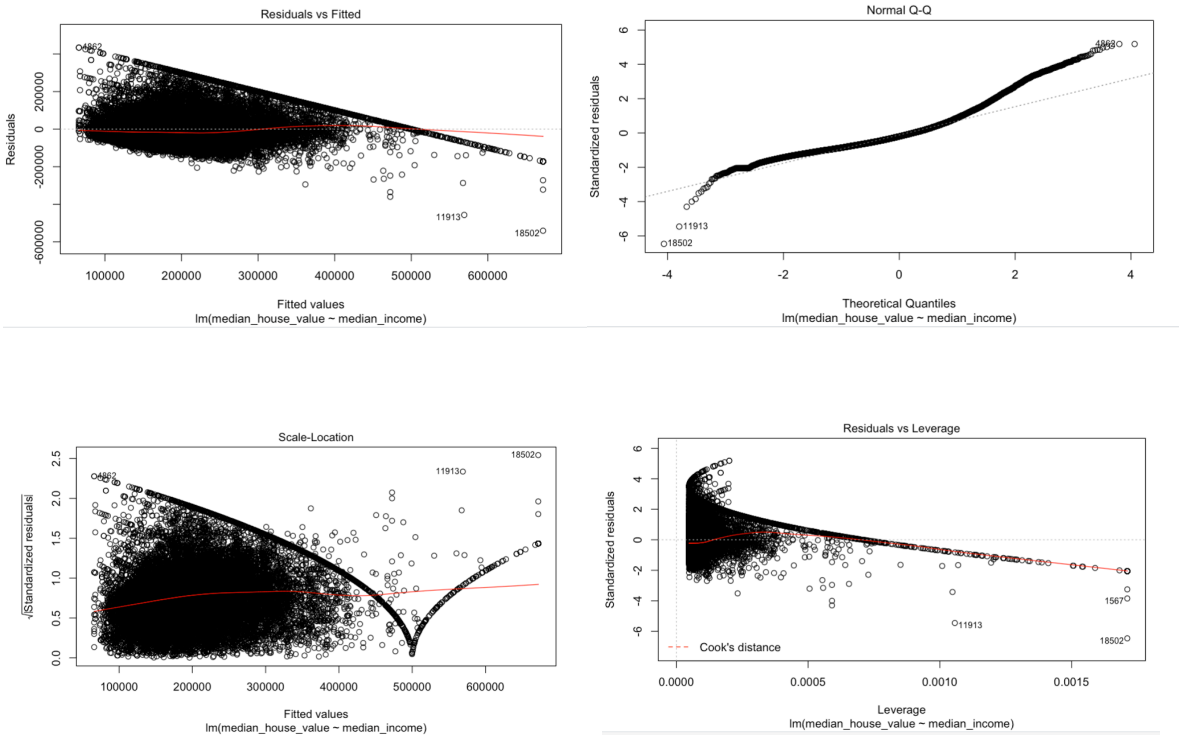
For model 1 we will try to use only median income variable to predict median house value as it shown as highest correlated to median house value compared to others variable. Let's see how is the result of this model.

```
Call:
lm(formula = median_house_value ~ median_income, data = cahousing_new)

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
```

The results shown p-value less than 0.05 that mean median income is variables that significant to use to predict median house value. And also meaning that 47% of median house value can be explained by median income. However, there are others 53% leftover that we need to find out in order to make model more efficient predicted.

Now let see diagnostic plot for this model.

Here we can see that red line of all 3 plots are almost straight line and do not have influence value that pull the line to cooking distance area which is good. However in QQ plot, even most of data are lined on the QQ line but there are some curve at the higher side.



Let's try to come up with another model and compare with this one.

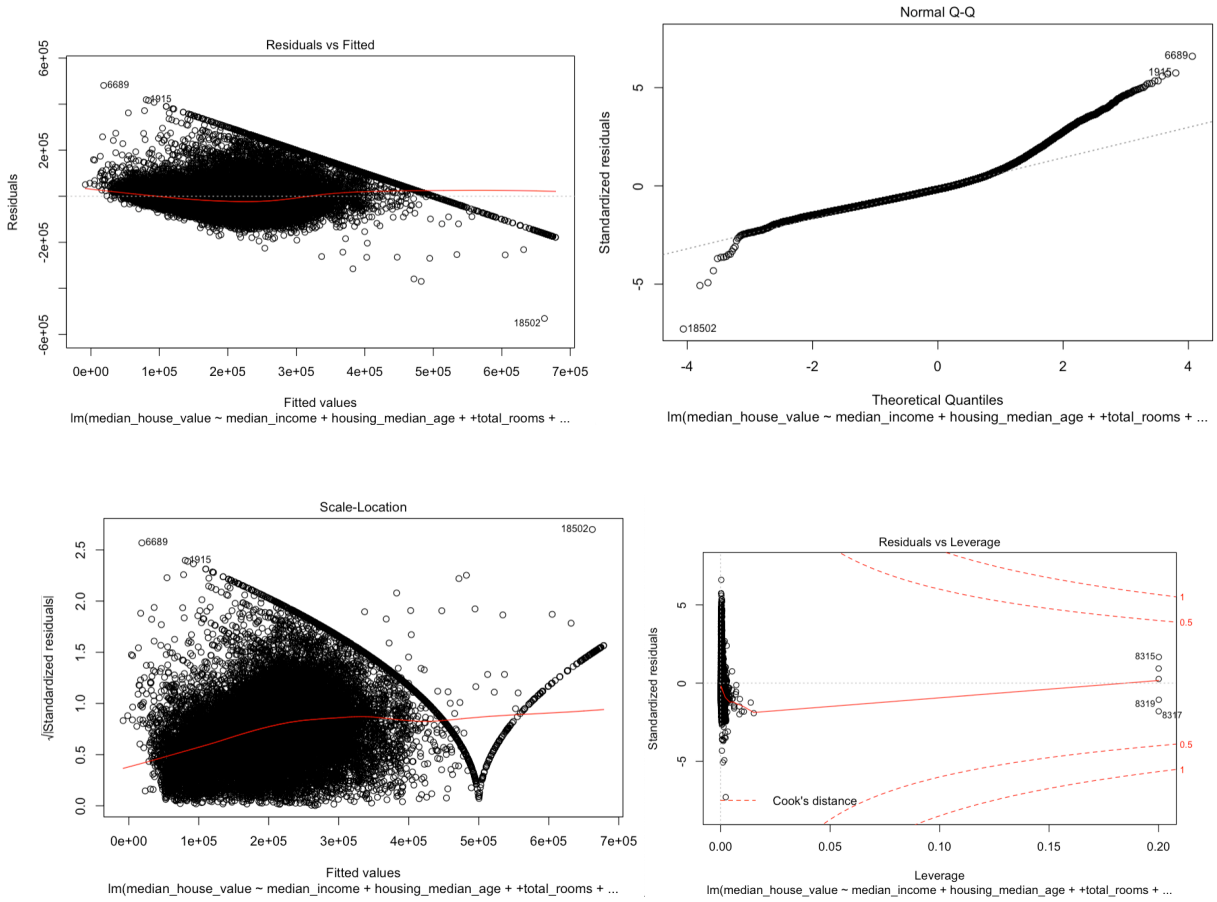
Model 2

```
Call:
lm(formula = median_house_value ~ median_income + housing_median_age +
    +total_rooms + ONEH_OCEAN + INLAND + ISLAND + NEAR_BAY, data = cahousing_new)

Residuals:
    Min      1Q  Median      3Q     Max
-530949  -46155  -12434   29871  481291

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.644e+04  2.513e+03  22.464  < 2e-16 ***
median_income       3.756e+04  2.838e+02 132.351  < 2e-16 ***
housing_median_age  1.144e+03  4.608e+01  24.818  < 2e-16 ***
total_rooms         3.467e+00  2.537e-01  13.666  < 2e-16 ***
ONEH_OCEAN         -1.787e+04  1.609e+03 -11.110  < 2e-16 ***
INLAND             -8.936e+04  1.712e+03 -52.193  < 2e-16 ***
ISLAND              1.670e+05  3.266e+04   5.112 3.21e-07 ***
NEAR_BAY           -5.773e+03  2.116e+03  -2.728  0.00637 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72950 on 20632 degrees of freedom
Multiple R-squared:  0.6005,    Adjusted R-squared:  0.6004
F-statistic:  4431 on 7 and 20632 DF,  p-value: < 2.2e-16
```

This model, I used all variables including binary variables of ocean proximity. Also removing data that not significant and data that have potentially of multicollinearity situations

The result shown that this model has higher R-squared equal to 60% that mean these variable can help better to predict median house value. And we can use these variables to predict 60% of median house value.

Diagnostic plot of this model is not look much different from the previous one. There are still have hugh curve of QQ plot on the high value side. Residual & Fitted look almost straight horizontal line. Scale location & Leverage is not the best one, not in straight line.



4. Conclusion and next step

We can use median income, housing median age, total rooms and ocean proximity to predict median housing value based on Model 2 which we can use to predict around 60% of median housing value.

There still have another 40% portion that cannot predicted by these variables, so next step need to more research to find out an additional variable that could potentially help to increase prediction model more reliable.