House Price Prediction

CS 6301: Practical Aspects of Data Science

Summer 2020: Assignment 1

Team Member: -
Rajarshi Chattopadhyay (RXC170010)

Introduction

We predict real-valued or continuous output variable using Regression. Here, we have used California Housing Dataset to perform regression analysis to predict house prices.

Data

The data is about the houses found in a given California district with stats about them based on the 1990 census data. https://www.kaggle.com/camnugent/california-housing-prices

Number of records: 20640

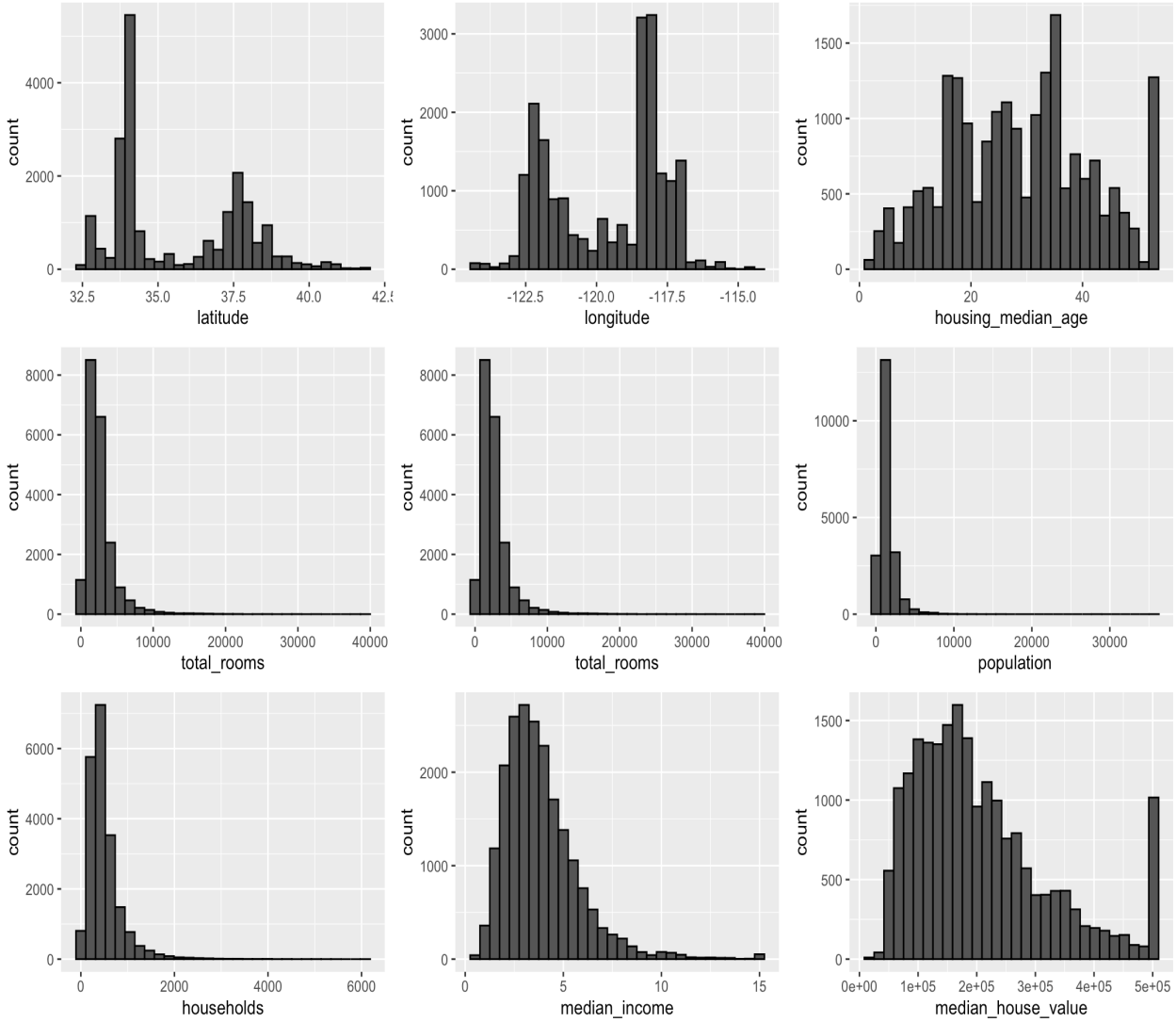
Number of Columns : 10
"longitude" "latitude" "housing_median_age"
"total_rooms" "total_bedrooms" "population"
"households" "median_income" "median_house_value"
"ocean_proximity"

Summary of 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.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 ...
```

Explore attribute values



Handling attributes

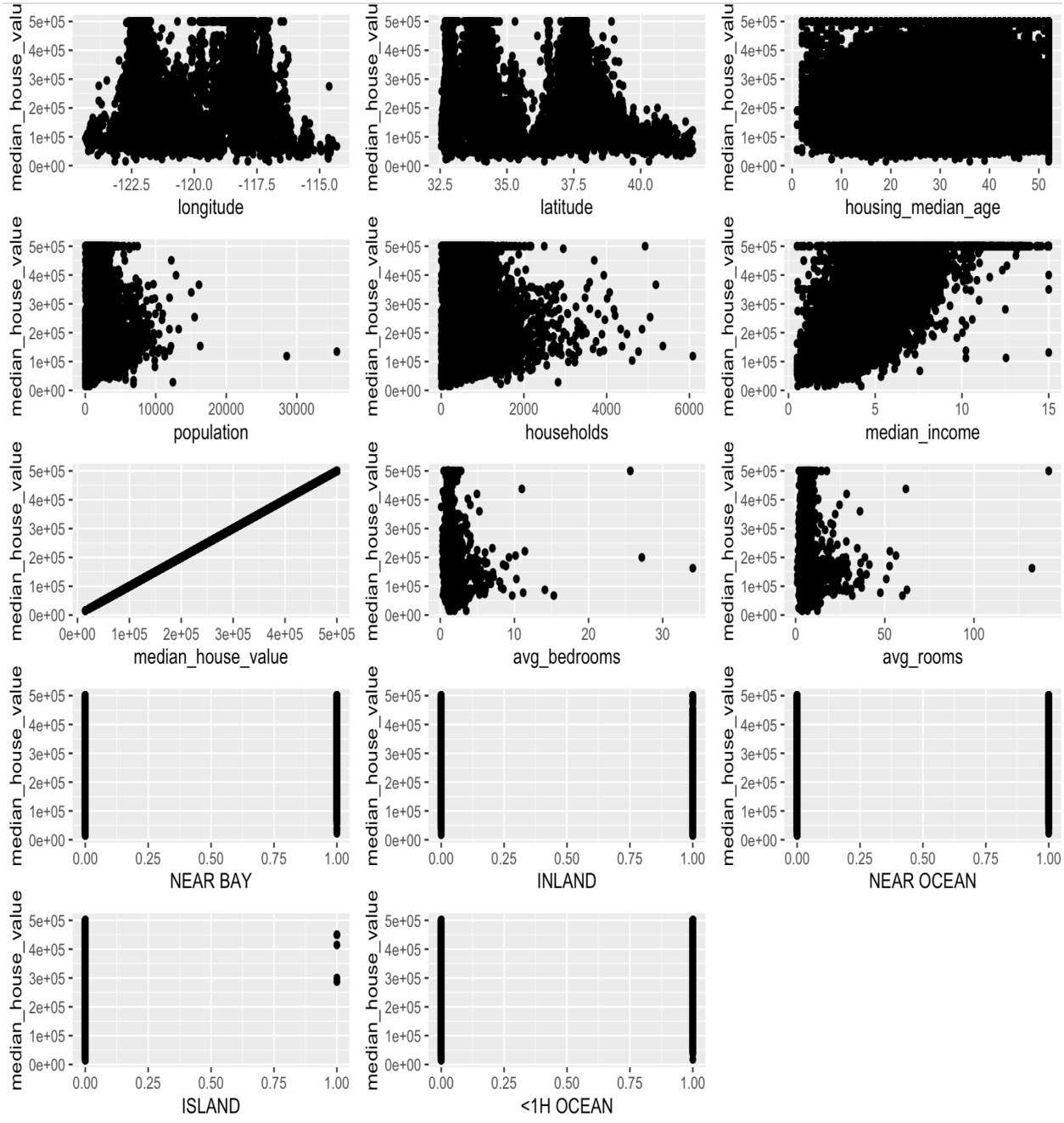
- Replace total_bedrooms and total_rooms with avg_bedrooms and avg_rooms
- Split the ocean_proximity into separate boolean category columns

Updated columns

"longitude" "latitude" "housing_median_age" "population" "households" "median_income" "median_house_value" "avg_bedrooms" "avg_rooms" "NEAR BAY" "<1H OCEAN" "INLAND" "NEAR OCEAN" "ISLAND"

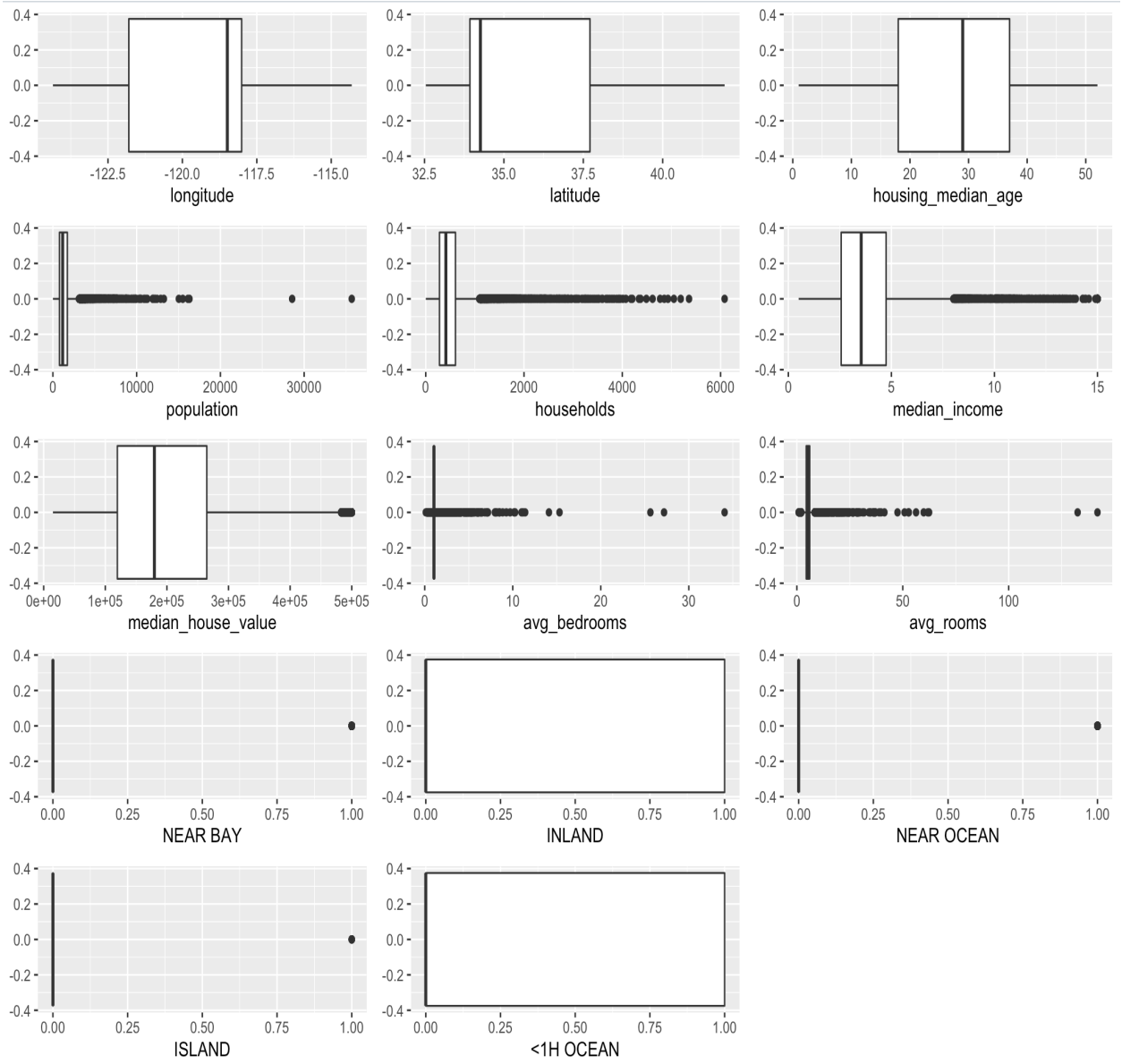
'data.frame': 20640 obs. of 14 variables:
$ longitude : num -122 -122 -122 -122 -122 ...
$ latitude : num 37.9 37.9 37.9 37.9 37.9 ...
$ housing_median_age: num 41 21 52 52 52 52 52 52 42 52 ...
$ population : num 322 2401 496 558 565 ...
$ households : num 126 1138 177 219 259 ...
$ median_income : num 8.33 8.3 7.26 5.64 3.85 ...
$ median_house_value: num 452600 358500 352100 341300 342200 ...
$ avg_bedrooms : num 1.024 0.972 1.073 1.073 1.081 ...
$ avg_rooms : num 6.98 6.24 8.29 5.82 6.28 ...
$ NEAR BAY : num 1 1 1 1 1 1 1 1 1 1 ...
$ <1H OCEAN : num 0 0 0 0 0 0 0 0 0 0 ...
$ INLAND : num 0 0 0 0 0 0 0 0 0 0 ...
$ NEAR OCEAN : num 0 0 0 0 0 0 0 0 0 0 ...
$ ISLAND : num 0 0 0 0 0 0 0 0 0 0 ...

Check relationship of the attributes with median_house_value



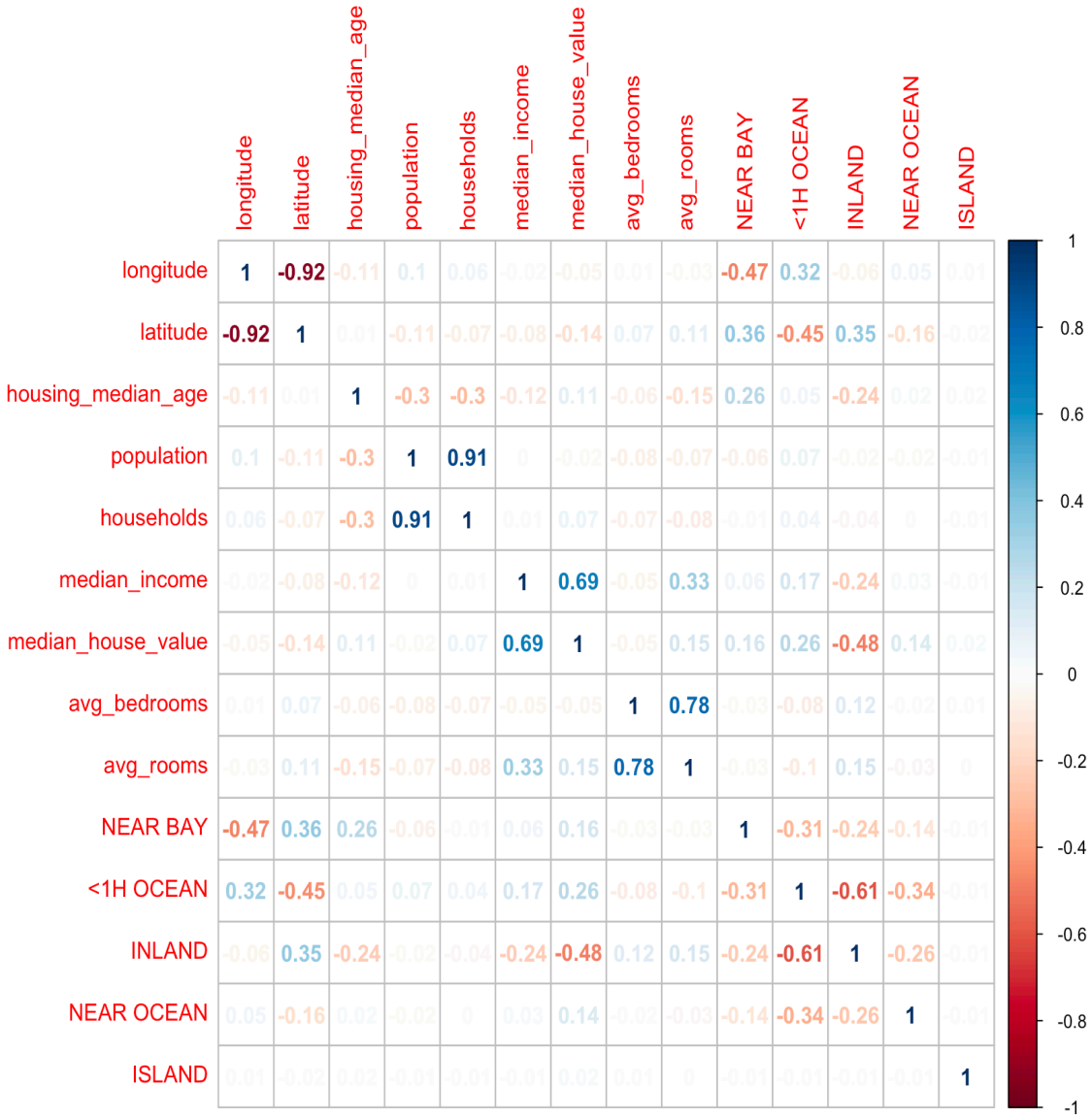
The feature median_income is the only one with a linear relationship with median_house_value.

Check for outliers in the attribute values



There are a lot of outliers.

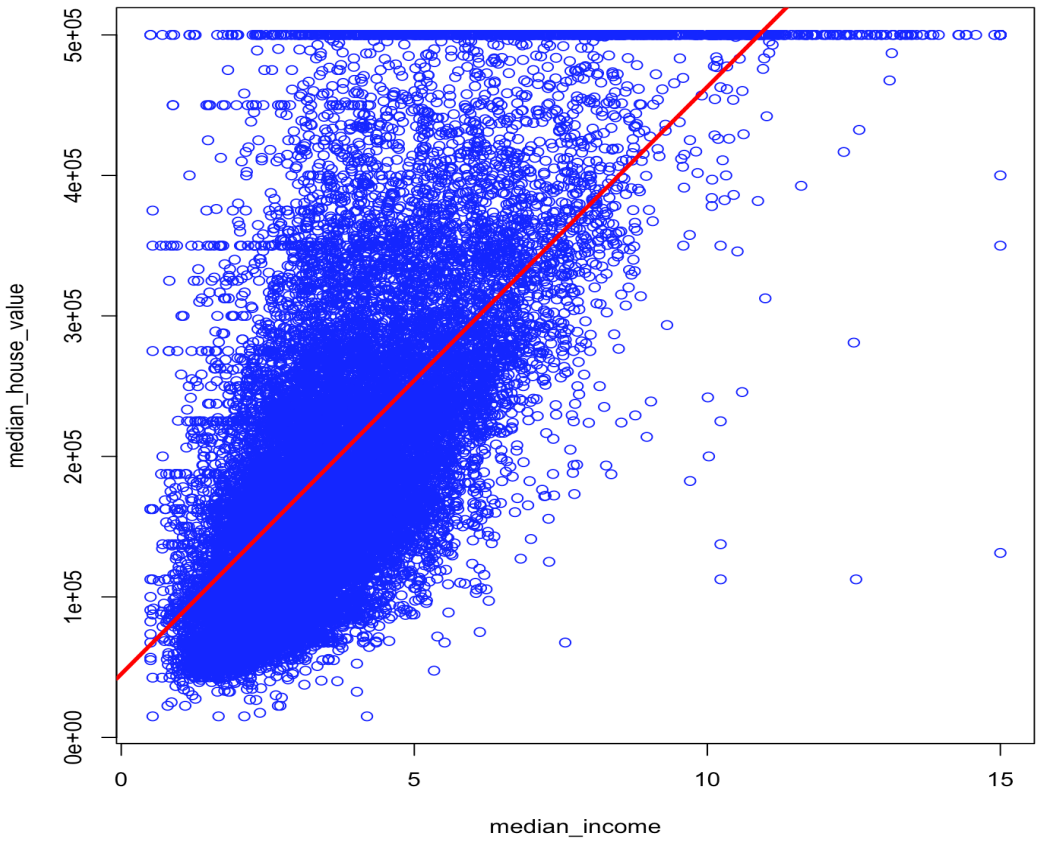
Check correlation of the attributes



We use the attribute as the predictor which has >60 % correlation with the median_house_value. Here, it is median_income. Further, it also has linear relationship with median_house_value.

Create model to predict median_house_value using median_income

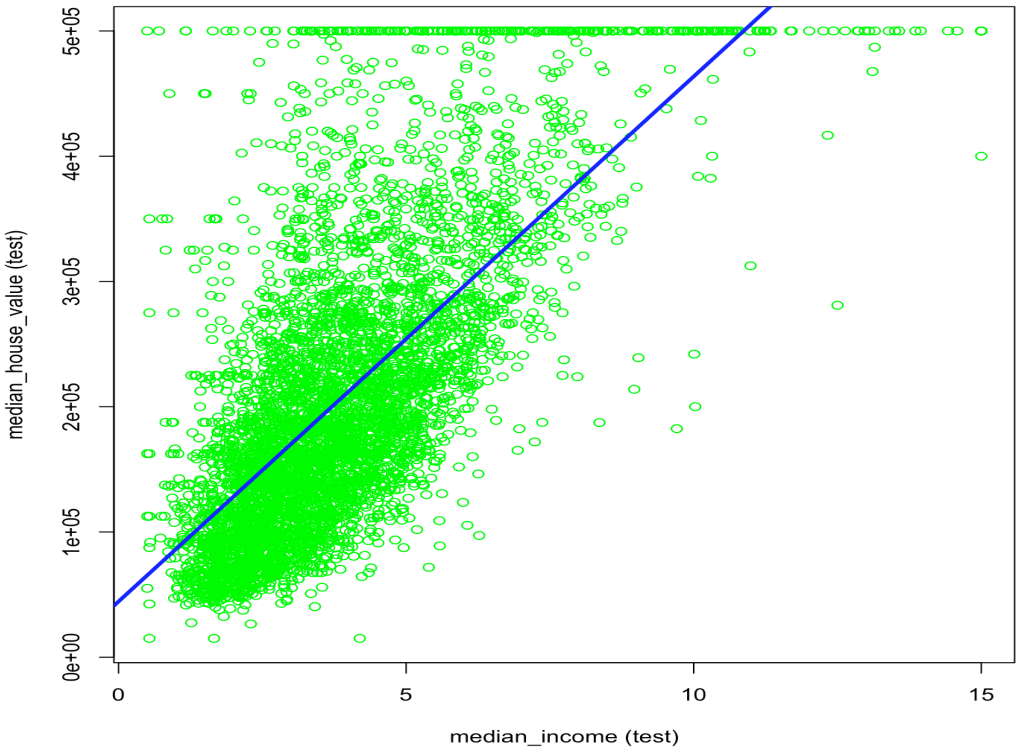
```
Residuals:
    Min      1Q  Median      3Q     Max
-540697  -55950  -16979   36978  434023

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

                  2.5 %   97.5 %
(Intercept)   42492.64 47678.51
median_income 41192.49 42395.21
```



Upgrading model

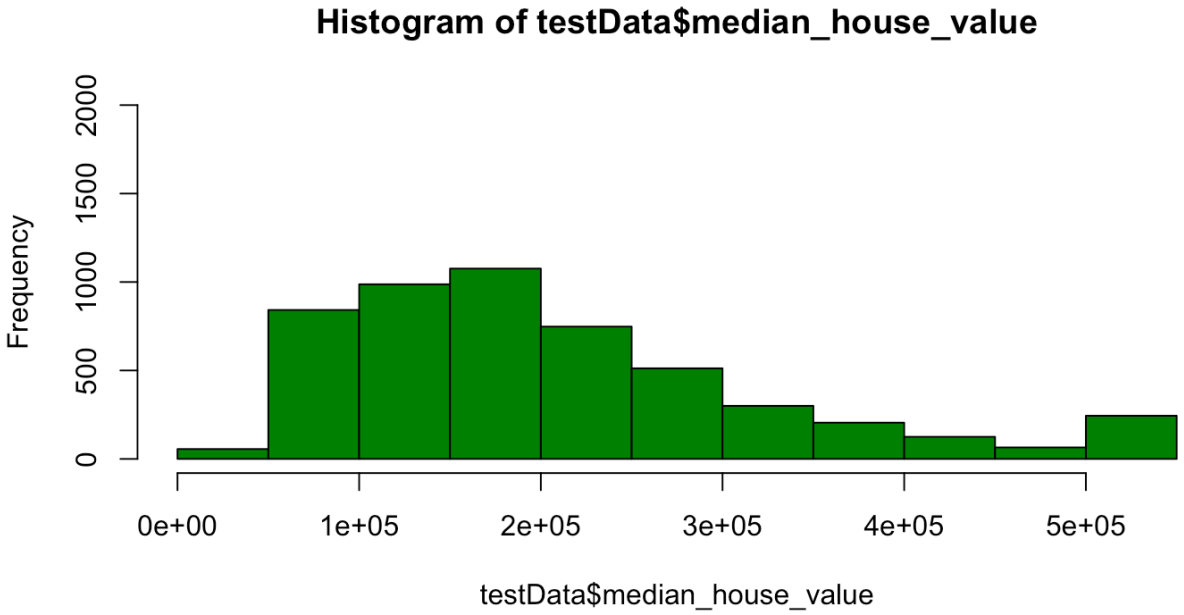
- Randomize and Split data to training (75%) and test (25%) sets
- Train model with training data
- Predict using test data

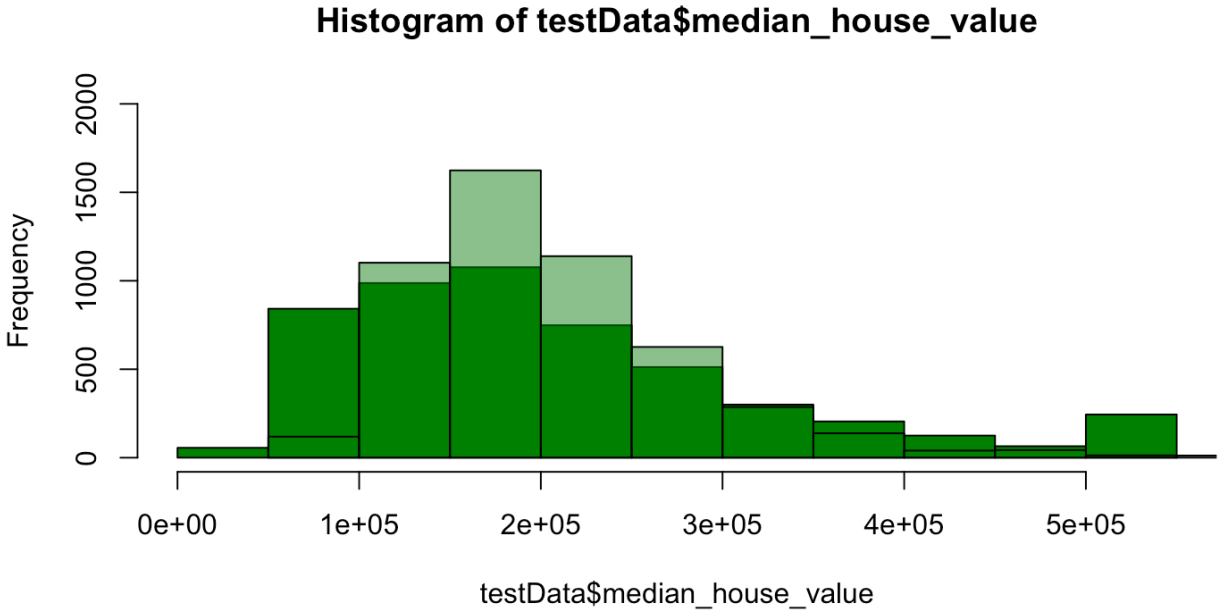
```
Residuals:
   Min     1Q  Median     3Q    Max
-541892  -56147  -16972   36727  434575

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    44473.2     1532.0   29.03   <2e-16 ***
median_income  41914.3      355.3  117.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 83930 on 15478 degrees of freedom
Multiple R-squared:  0.4735,  Adjusted R-squared:  0.4734
F-statistic: 1.392e+04 on 1 and 15478 DF,  p-value: < 2.2e-16

                  2.5 %   97.5 %
(Intercept)   41470.32 47476.15
median_income 41217.92 42610.69
```



Comparing the test data with predictions

RMSE = 142073.1





Conclusion

In both the models p-value is less than 0.05, which implies there is some relationship between the dependent variable median_house_value and the independent variable median_income that have been used in our model.

References

- Exploring Data: https://cran.r-project.org/web/packages/driftR/vignettes/ExploringData.html

- Outlier finding: http://r-statistics.co/Outlier-Treatment-With-R.html