HOUSING PREDICTION USING MULTIPLE LINEAR REGRESSION

Table of Contents

[INTRODUCTION 3](#_Toc40043966)

[RESEARCH QUESTION 3](#_Toc40043967)

[RATIONALE 3](#_Toc40043968)

[SCOPE OF THE ANALYSIS 3](#_Toc40043969)

[PRESENTATION OF DATASET 5](#_Toc40043970)

[PRACTICAL ANALYSIS 7](#_Toc40043971)

[DATA PREPARATION 8](#_Toc40043972)

[HANDLING CATEGORICAL VALUES 8](#_Toc40043973)

[FINDING MISSING VALUES 9](#_Toc40043974)

[HANDLING MISSING VALUES 9](#_Toc40043975)

[DATA ANALYSIS 10](#_Toc40043976)

[CORRELATION ANALYSIS 10](#_Toc40043977)

[SPLITTING TRAIN AND TEST DATA 12](#_Toc40043978)

[APPLYING LINEAR REGRESSION 12](#_Toc40043979)

[PREDICTING THE HOUSE VALUE 12](#_Toc40043980)

[COMPUTING ACCURACY 12](#_Toc40043981)

[RESULTS EVALUATION 12](#_Toc40043982)

[CONCLUSION 20](#_Toc40043983)

[LIMITATIONS OF THE STUDY 21](#_Toc40043984)

[FUTURE STUDY 21](#_Toc40043985)

[REFERENCES 21](#_Toc40043986)

# INTRODUCTION

Statistics is a mathematical science that deals with collection. organizing and analysing data. Applying statistical models using machine learning algorithms can help capture patterns and predict values.  
Statistics can be applied to social, political or economic problems to study the status quo and can bring about interesting patterns and solutions which can be beneficial for the system.  
Census data be it in any form tells us a lot of important information. It is the basis for not only renowned statistical institutions but also several other businesses and public organizations to come up with interesting theories and conclusions which aid in decision-making.  
One such census data is the housing data which is being taken as a dataset for this research. This contains the housing prices in a small geographical location in the United States from a particular period. There are other information such as latitude, longitude, total number of rooms and other independent variables. The dependent variable here is the median house value.

## RESEARCH QUESTION

1. To predict the housing prices using Multiple Linear Regression based on a number of independent variables.
2. To verify if multiple linear regression model is a good fit for the data.

## RATIONALE

Predicting housing prices can help real estate developers, agents and sellers to decide the selling price for houses.  
This can also benefit the customers who are interested in purchasing a property by helping them in decision making.  
It also acts as a basis for statistical institutes and public organisations.  
It helps in risk analysis for policy makers in the housing domian.

## SCOPE OF THE ANALYSIS

In this research, we analyse the relationship between dependent and independent variables and then apply multiple linear regression model on the dataset and evaluate them using the error metrics - MAE, MSE, RMSE, MAPE and diagnostic plots - Residuals vs Fitted, Normal Q-Q, Scale Location, Residuals vs Leverage.

##### **SETUP**

In order to carry out the analysis the following libraries need to be installed.  
1. library(tidyverse) : It is used for manipulation and analysis of data.  
2. library(ggplot2) : It is used to graphically visualise complex data by plotting charts and graphs.  
3. library(dplyr) : It is data manipulation package in R used to manipulate any data.  
4. library(plotly) : It is similar to ggplot, but with the help of this the plots can be made interactive.  
5. library(scales) : Scale functions for visualization.  
6. librabry(RColorBrewer) : It provides different ColorBrewer palettes.  
7. library(GGally) : It is an extension to ggplot.  
8. library(Amelia) : A program for missing data.  
9. library(DMwR) : Functions and data for data mining with R.

library(tidyverse)

## -- Attaching packages ------------------------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 0.8.5  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ---------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(dplyr)  
library(plotly)

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

library(scales)

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

library(RColorBrewer)  
library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

library(Amelia)

## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.7.6, built: 2019-11-24)  
## ## Copyright (C) 2005-2020 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

#library(tinytex)  
library(DMwR)

## Loading required package: lattice

## Loading required package: grid

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

# PRESENTATION OF DATASET

housingdata <- read.csv(file="C:/Users/guest23455/Desktop/Statistics/housing.csv", header = TRUE, stringsAsFactors = FALSE) # Reading the csv file into avariable

str(housingdata) # To see the structure of the dataframe

## '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 : chr "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...

The dataset contains **20640** observations of **10** variables. Below is the list of variables and its description.

**longitude** : A measure of how far west a house is, a higher value is farther west

**latitude** : A measure of how far north a house is, a higher value is farther north

**housing\_median\_age** : Median age of a house within a block, a lower number is a newer building

**total\_rooms** : Total number of rooms within a block

**total\_bedrooms** : Total number of bedrooms within a block

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

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

**median\_income** : Median income for households within a block of houses (measured in tens of thousands of US Dollars)

**median\_house\_value** : Median house value for households within a block (measured in US Dollars)

**ocean\_proximity** : Location of the house with respect to ocean/sea

head(housingdata) # To get 6 rows of the data

## longitude latitude housing\_median\_age total\_rooms total\_bedrooms population  
## 1 -122.23 37.88 41 880 129 322  
## 2 -122.22 37.86 21 7099 1106 2401  
## 3 -122.24 37.85 52 1467 190 496  
## 4 -122.25 37.85 52 1274 235 558  
## 5 -122.25 37.85 52 1627 280 565  
## 6 -122.25 37.85 52 919 213 413  
## households median\_income median\_house\_value ocean\_proximity  
## 1 126 8.3252 452600 NEAR BAY  
## 2 1138 8.3014 358500 NEAR BAY  
## 3 177 7.2574 352100 NEAR BAY  
## 4 219 5.6431 341300 NEAR BAY  
## 5 259 3.8462 342200 NEAR BAY  
## 6 193 4.0368 269700 NEAR BAY

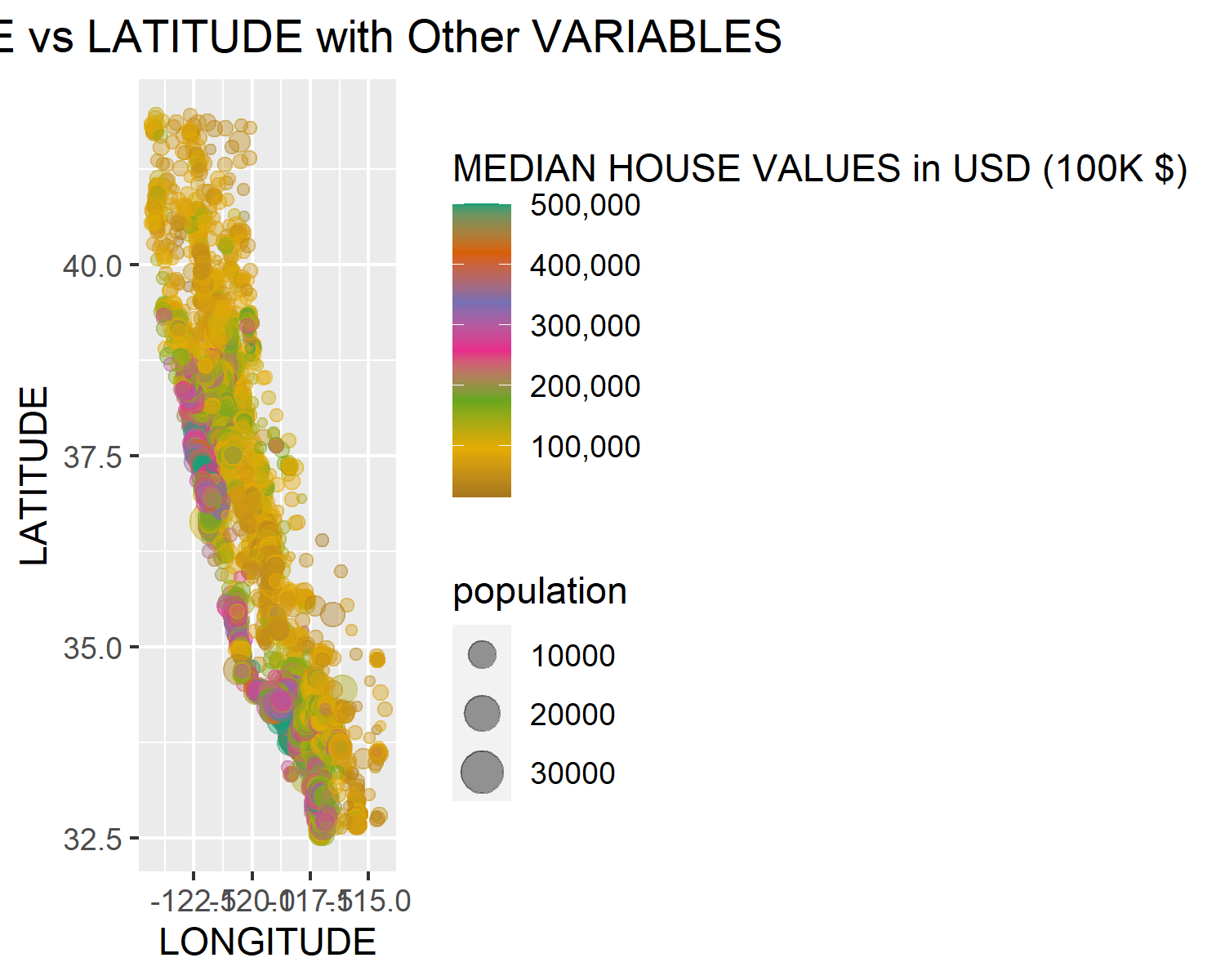
summary(housingdata) # To get the summary of the dataset

## 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 Length:20640   
## 1st Qu.:119600 Class :character   
## Median :179700 Mode :character   
## Mean :206856   
## 3rd Qu.:264725   
## Max. :500001   
##

# PRACTICAL ANALYSIS

The below visualization represents all the data points in the dataset, with Longitude on the x-axis and Latitude on the y-axis and the Median House Value as the color codes.

plot <- ggplot(housingdata, aes(x = longitude, y = latitude, color = median\_house\_value, age = housing\_median\_age, rooms = total\_rooms, broom = total\_bedrooms, holds = households, income = median\_income)) +   
 geom\_point(aes(size = population), alpha = 0.4) +  
 xlab("LONGITUDE") + ylab("LATITUDE") +   
 ggtitle("LONGITUDE vs LATITUDE with Other VARIABLES") +  
 theme(plot.title = element\_text(hjust = 0.5)) +   
 scale\_color\_distiller(palette = "Dark2", labels = comma) +   
 labs(color = "MEDIAN HOUSE VALUES in USD (100K $)", size = "population")  
plot



ggplotly(plot)

## DATA PREPARATION

### HANDLING CATEGORICAL VALUES

We must handle categorical values by converting them from text-based to a factor, so that it is easy to understand while predicting.

housingdata$ocean\_proximity = as.factor(housingdata$ocean\_proximity)

str(housingdata)

## '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 ...

### FINDING MISSING VALUES

With the help of missmap() from the Amelia package we can see the missing values.  
There are 207 NA values in total\_bedrooms.

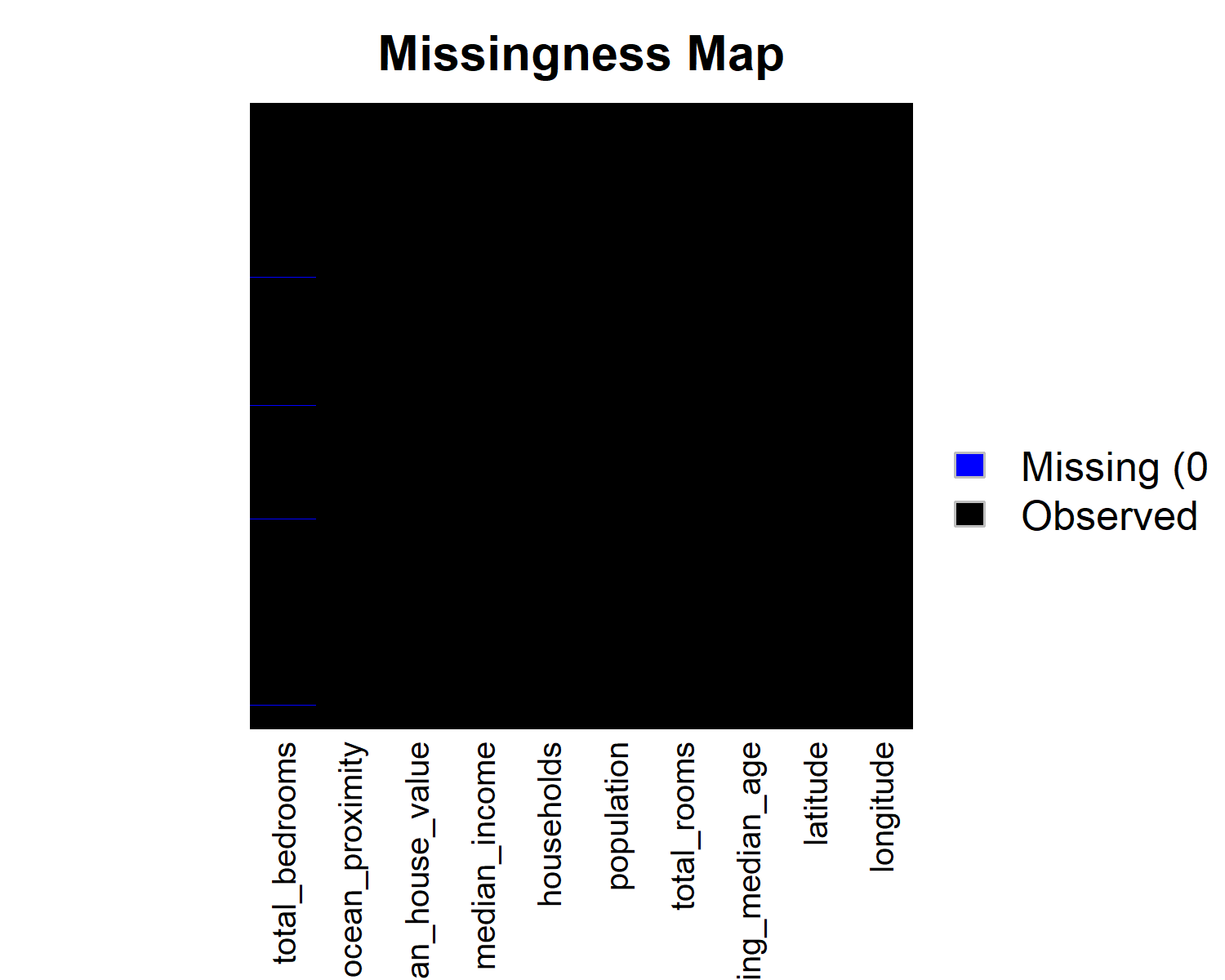
sum(is.null(housingdata))

## [1] 0

sum(is.na(housingdata))

## [1] 207

missmap(housingdata,col=c('blue','black'),y.at=1,y.labels='',legend=TRUE)



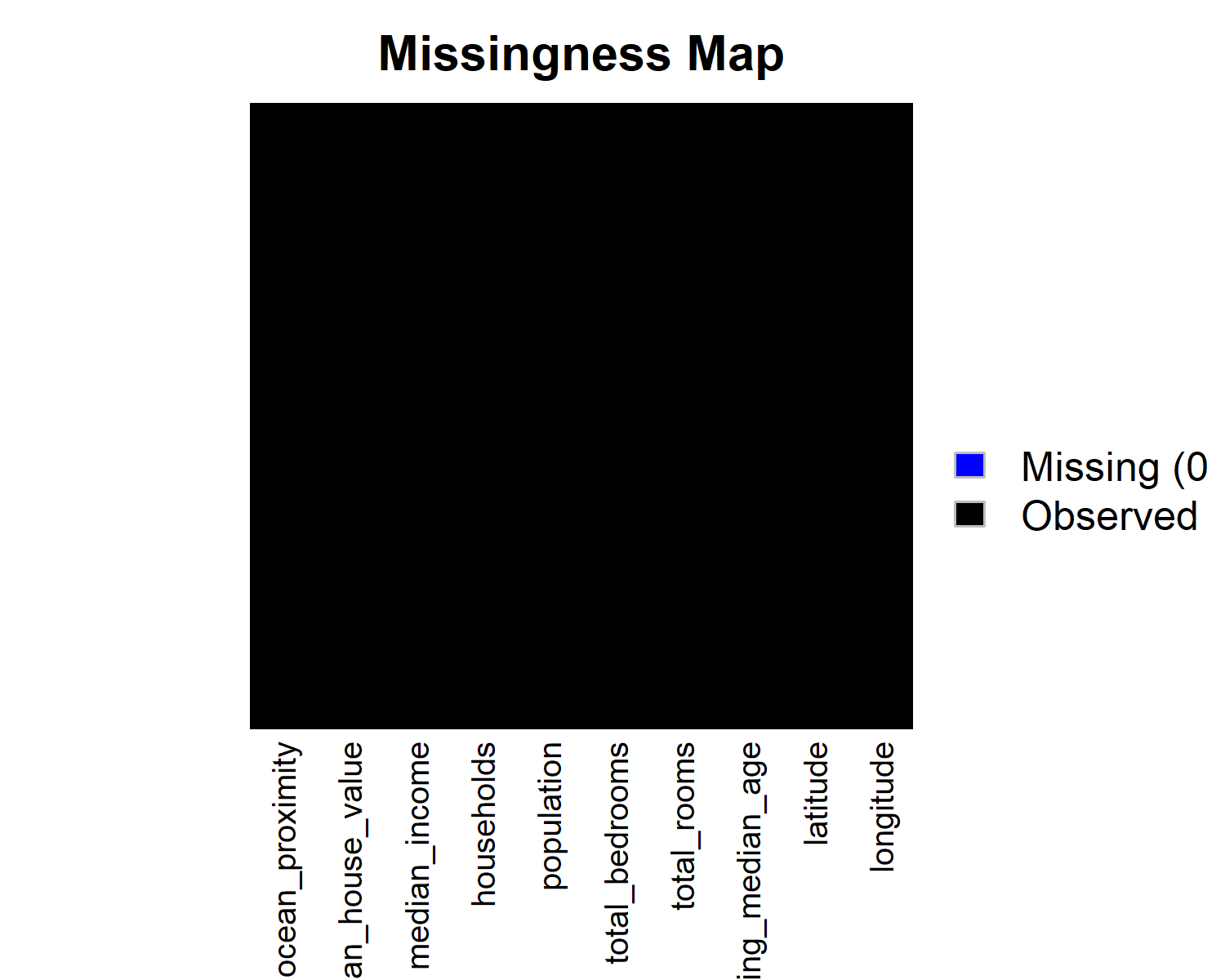
### HANDLING MISSING VALUES

Now the dataset is free of missing values, which was achieved by omitting NA values.

housingdata <- na.omit(housingdata)  
sum(is.na(housingdata))

## [1] 0

missmap(housingdata,col=c('blue','black'),y.at=1,y.labels='',legend=TRUE)

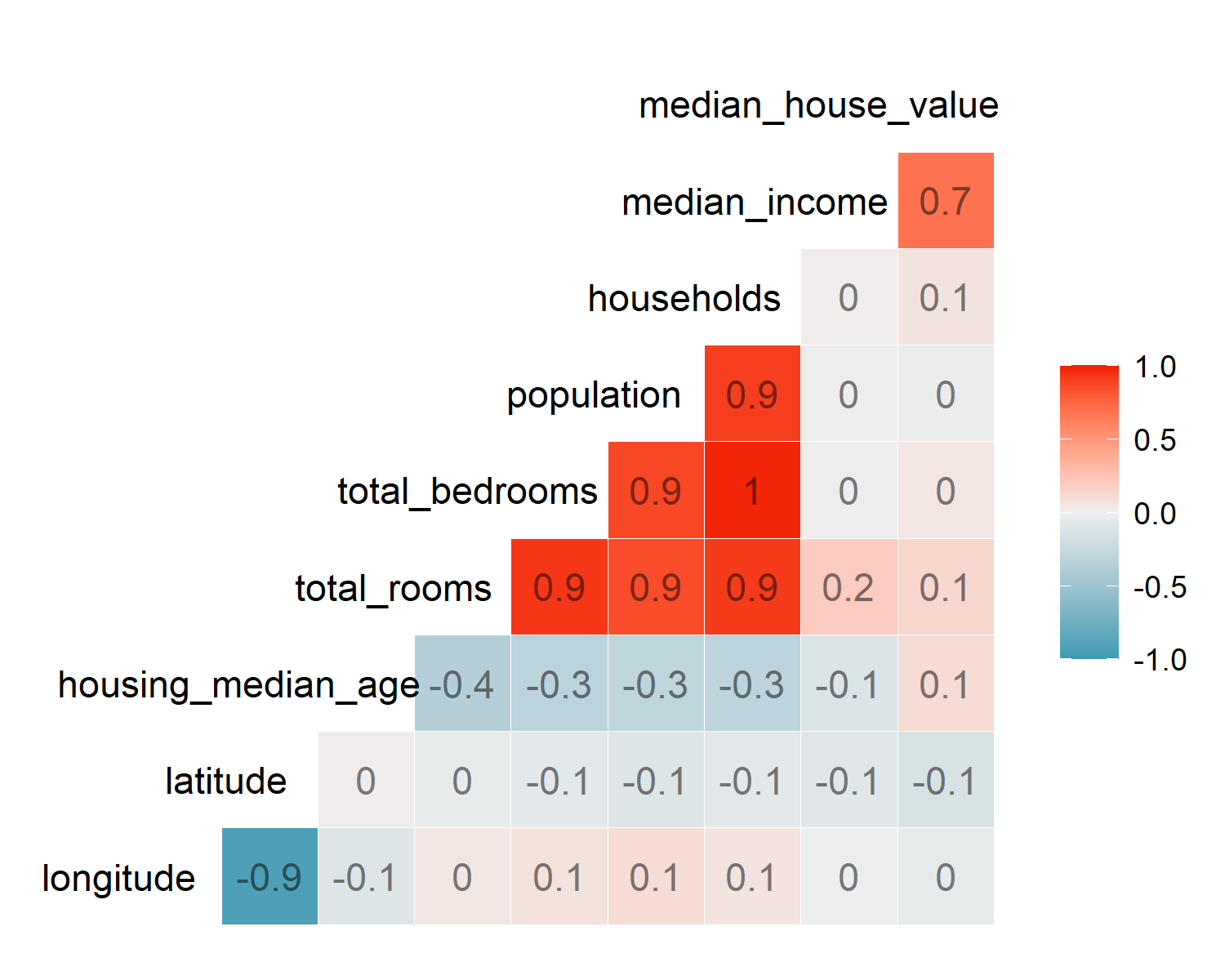


## DATA ANALYSIS

### CORRELATION ANALYSIS

The strength of a relationship between the target variable and other independent variables can be understood with the help of correlation analysis. The correlation coefficient is computed between any two variables.  
Correlation is a statistical measure, which shows the degree of linear dependence between any two variables. It can take values between -1 to +1.  
If values are close to +1,that is if the value of one variable increases with increase in value of an other, means that the two variables have a strong positive correlation and are highly correlated, whereas, values closer to -1 indicates a negative correlation, that is when one value decreases when the other increases. Values near to 0 imply weak correlations between variables.

ggcorr(housingdata,  
 label = TRUE,  
 label\_alpha = 0.5,  
 layout.exp = 1,  
 hjust = 0.85)



##### **FEATURE ENGINEERING**

To make the values of Median House Value more readable and easy to understand, it is represented in 100K.

housingdata$median\_house\_value <- housingdata$median\_house\_value / 100000

head(housingdata)

## longitude latitude housing\_median\_age total\_rooms total\_bedrooms population  
## 1 -122.23 37.88 41 880 129 322  
## 2 -122.22 37.86 21 7099 1106 2401  
## 3 -122.24 37.85 52 1467 190 496  
## 4 -122.25 37.85 52 1274 235 558  
## 5 -122.25 37.85 52 1627 280 565  
## 6 -122.25 37.85 52 919 213 413  
## households median\_income median\_house\_value ocean\_proximity  
## 1 126 8.3252 4.526 NEAR BAY  
## 2 1138 8.3014 3.585 NEAR BAY  
## 3 177 7.2574 3.521 NEAR BAY  
## 4 219 5.6431 3.413 NEAR BAY  
## 5 259 3.8462 3.422 NEAR BAY  
## 6 193 4.0368 2.697 NEAR BAY

### SPLITTING TRAIN AND TEST DATA

Let us divide this dataset into training and testing datasets.

set.seed(400) # setting seed to reproduce results of random sampling

trainingRowIndex <- sample(1:nrow(housingdata), 0.8\*nrow(housingdata)) # row indices for training data  
trainingData <- housingdata[trainingRowIndex, ] # model training data  
testData <- housingdata[-trainingRowIndex, ] # test data

### APPLYING LINEAR REGRESSION

Let us fit a linear model against training data.

fit <- lm(median\_house\_value ~., data=trainingData)

### PREDICTING THE HOUSE VALUE

PricesPred <- predict(fit, testData) # predict prices

Actual\_Pred <- data.frame(cbind(Actual=testData$median\_house\_value,   
 Predicted=PricesPred))

### COMPUTING ACCURACY

correlation\_accuracy <- cor(Actual\_Pred)

correlation\_accuracy

## Actual Predicted  
## Actual 1.0000000 0.8010302  
## Predicted 0.8010302 1.0000000

## RESULTS EVALUATION

Review Summary information on the model.  
The model summary helps us understand if the model is statistically significant or not, to further use it to make predictions.

summary(fit)

##   
## Call:  
## lm(formula = median\_house\_value ~ ., data = trainingData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.5467 -0.4236 -0.1053 0.2841 7.7944   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.203e+01 9.842e-01 -22.380 < 2e-16 \*\*\*  
## longitude -2.602e-01 1.140e-02 -22.828 < 2e-16 \*\*\*  
## latitude -2.466e-01 1.122e-02 -21.984 < 2e-16 \*\*\*  
## housing\_median\_age 1.042e-02 4.892e-04 21.295 < 2e-16 \*\*\*  
## total\_rooms -5.850e-05 8.995e-06 -6.503 8.11e-11 \*\*\*  
## total\_bedrooms 9.831e-04 7.780e-05 12.637 < 2e-16 \*\*\*  
## population -3.795e-04 1.195e-05 -31.756 < 2e-16 \*\*\*  
## households 4.965e-04 8.371e-05 5.931 3.08e-09 \*\*\*  
## median\_income 3.910e-01 3.779e-03 103.458 < 2e-16 \*\*\*  
## ocean\_proximityINLAND -4.098e-01 1.948e-02 -21.031 < 2e-16 \*\*\*  
## ocean\_proximityISLAND 1.548e+00 4.841e-01 3.197 0.001390 \*\*   
## ocean\_proximityNEAR BAY -5.505e-02 2.126e-02 -2.589 0.009641 \*\*   
## ocean\_proximityNEAR OCEAN 5.940e-02 1.753e-02 3.388 0.000705 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6843 on 16333 degrees of freedom  
## Multiple R-squared: 0.6476, Adjusted R-squared: 0.6474   
## F-statistic: 2501 on 12 and 16333 DF, p-value: < 2.2e-16

##### **Using p-values to check if the model is statistically significant**

The p-values are very important to tell if a model is statistically significant or not.  
A model is statistically significant only when the **p-values** are less than the pre-determined statistical significance level of **0.05**.  
The significance stars at the end of each row against each X variable helps us understand the significance visually. The more the stars beside a variable’s p-value, the more significant the variable.

From the model summary, **summary(fit)**, the p-values of the variables are less than the significance level, which implies that the model is satitically significant.

##### **Multiple R-Squared and Adjusted R-Squared values**

The proportion of the variation in the dependent/ target variable is given by R-Squared (R2).  
The R-squared statistic tells how well the model fits the actual data.  
R2 gives the linear relationship between the independent variables and the target variable.  
The value of R2 always lies between 0 and 1, that is if the value is closer to **0**, it represents a regression that shows no variance, whereas if the value is closer to **1**, it shows variance that was observed in the target variable.  
When applying multiple regression, the R2 value always increases as and when more variables are included, which is when, **Adjusted R2** values are preferred, as it adjusts the values for the number of variables considered.

From the model summary, the Multiple R2 and Adjusted R2 values are nearer to 1, which shows variance in the data. Since this is a multiple regression setting, Adjusted R2 is preferred and the value is **0.6474**, nearly **65%** variance in the observed variable.

##### **F-STATISTIC**

F-statistic indicates if there is a relationship between the independent and target variables.  
The further the value of F-statistic from 1, the better the model is.

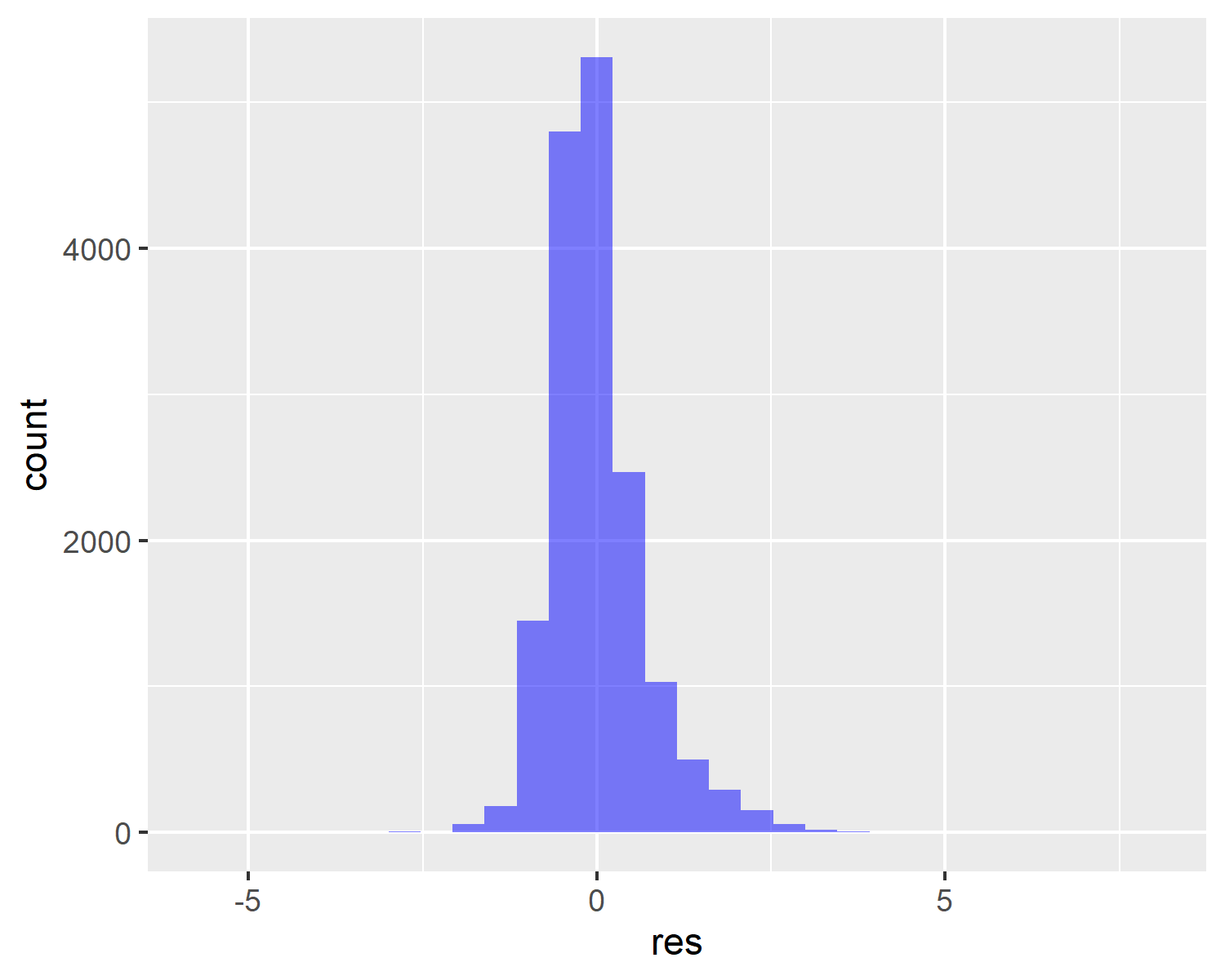
From the model summary, the observed value of F-statistic is **2501** which is relatively greater then 1, implies a good fit of the model.

##### **RESIDUALS**

The difference between the actual observed target variable values and the target variable values predicted by the model is the residuals.  
In the summary of the model, the residuals output is divided into five points.  
To assess how well a model fits the actual data, there should be a symmetrical distribution across the observed data points on the mean value 0.

res <- residuals(fit)  
res <- as.data.frame(res)

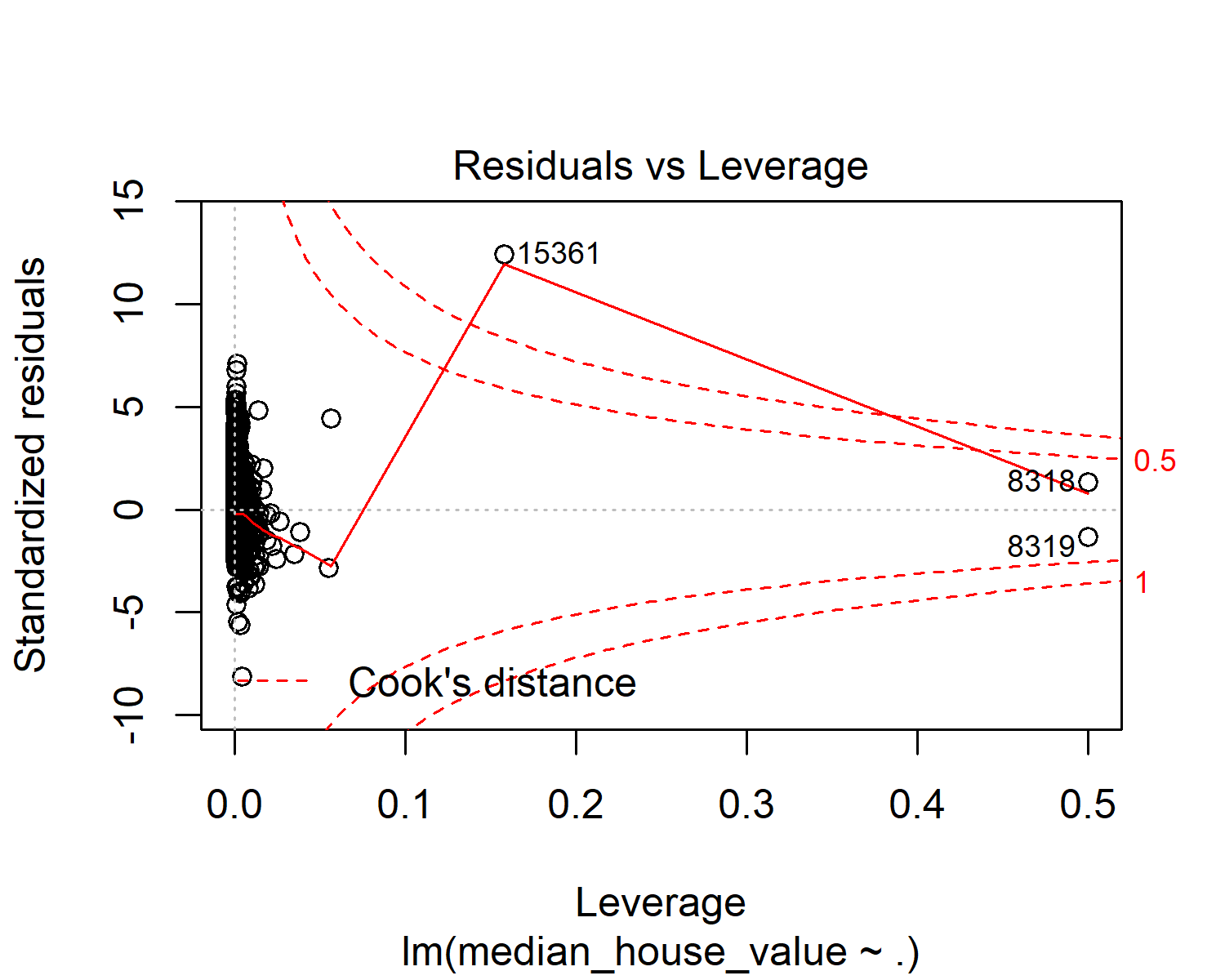
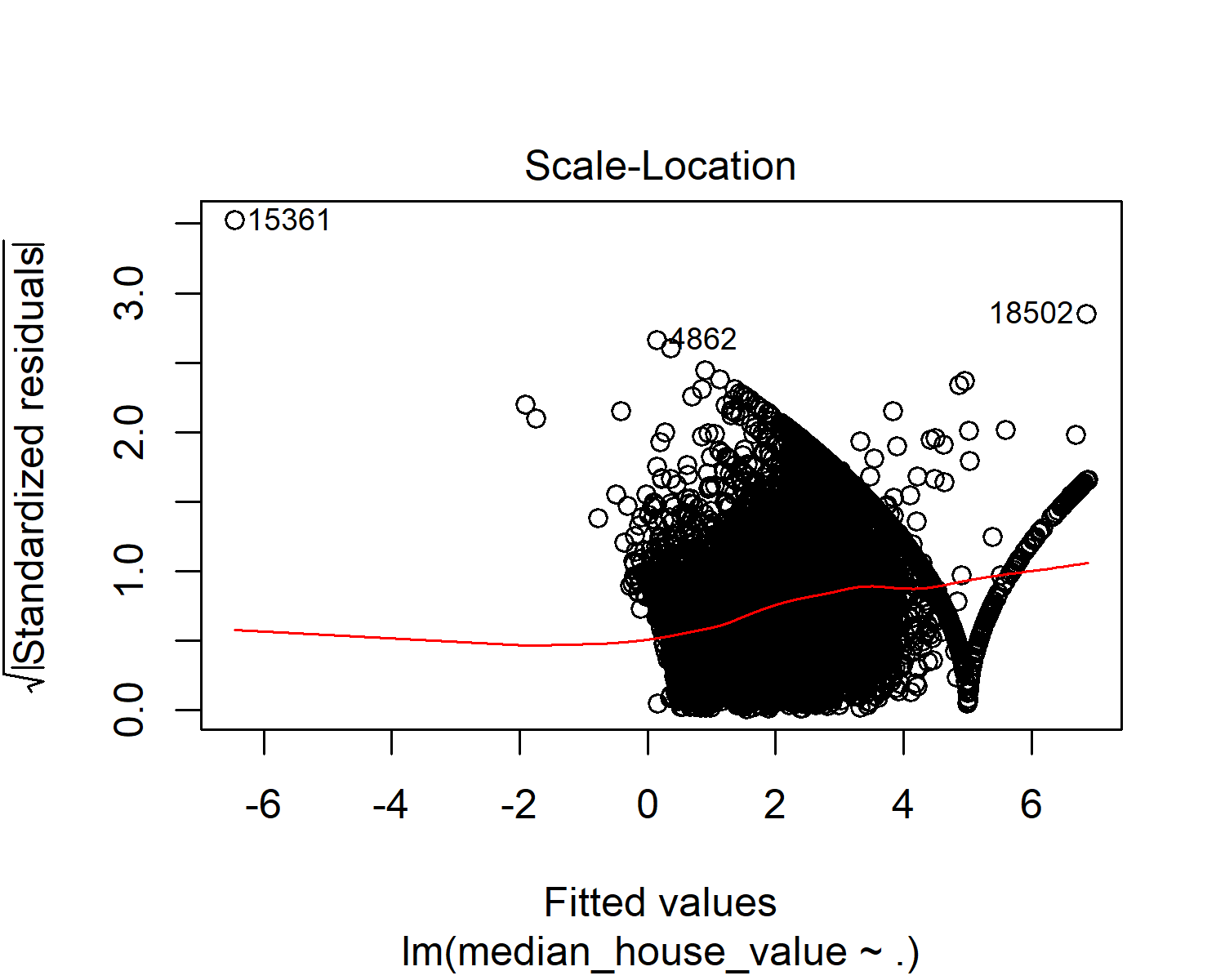
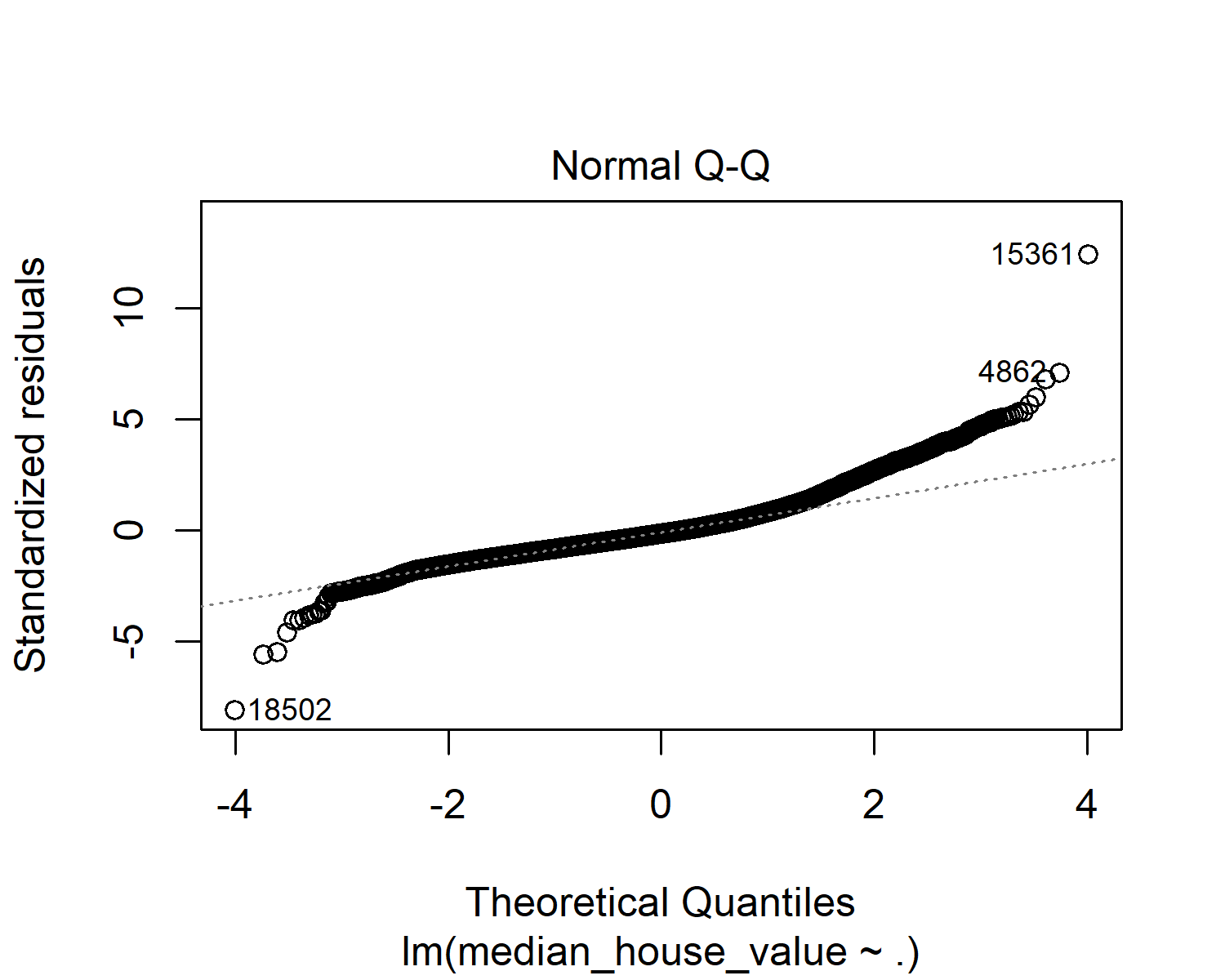
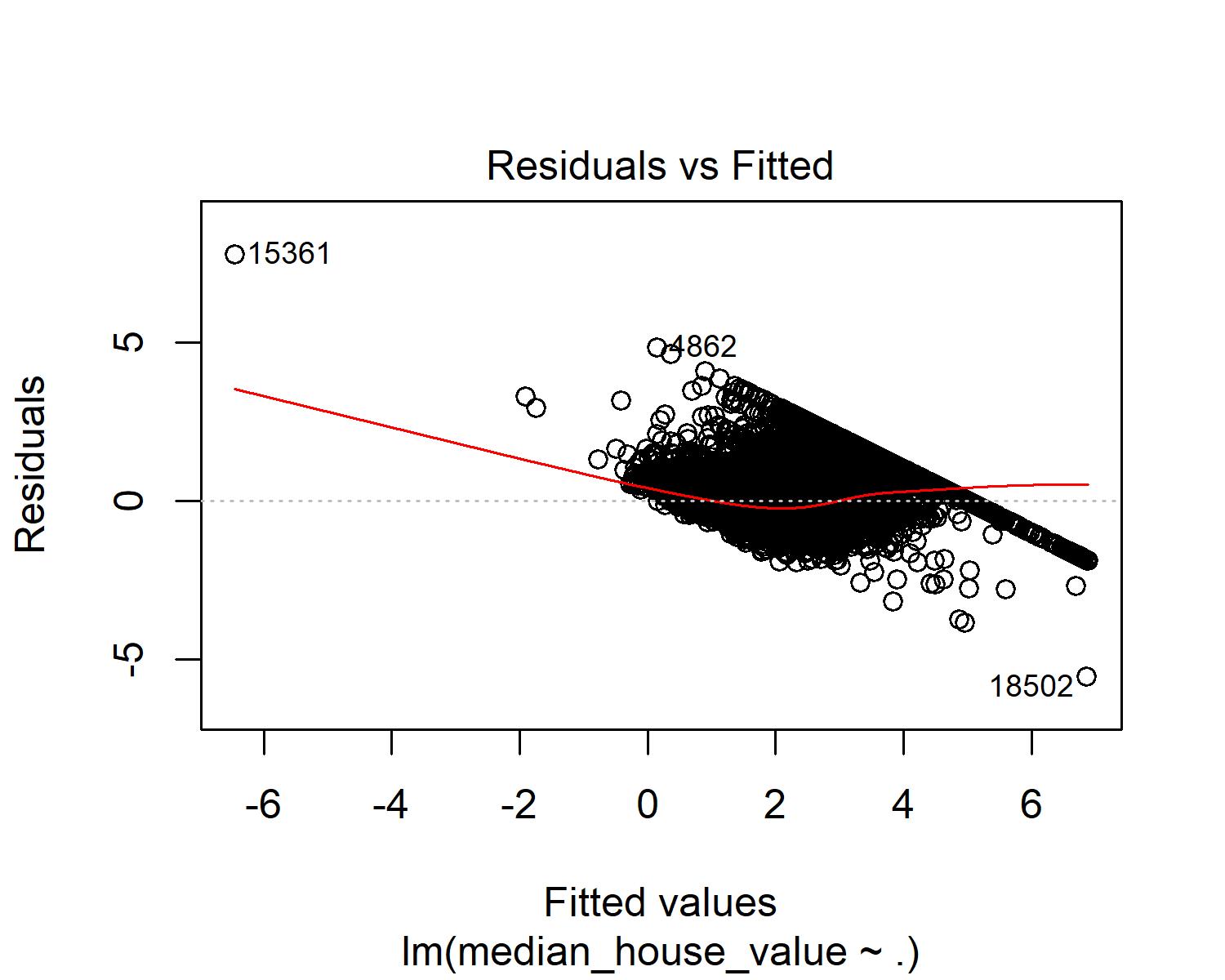
res1 <- ggplot(res,aes(res)) + geom\_histogram(fill='blue',alpha=0.5, bins = 30)  
res1



ggplotly(res1)

The above graph represents the distribution of the residuals.  
From the graph, the distribution of residuals is not highly symmetrical, which implies that few points that fall away from the actual observed points were predicted by the model.

plot(fit)



Although error metrics such as MAE, RMSE, MSE, MAPE help understand if the model fits the data. However it does not give a full picture. Diagnostic plots can help us understand the error better. Residuals are the errors between the actual and predicted values and the diagnostic plots using residuals could show how the model represents the data.

1. **RESIDUALS vs FITTED**

This graph shows whether there is a non-linear relationship between the dependent and independent variables. In the above graph, we can see that the trend is almost a horizontal line which shows that the residuals follow a linear pattern. This means that applied model fits the data very well as there is no non-linear pattern.

1. **NORMAL Q-Q**

This plot shows whether there is normal distribution followed by the residuals. If the residuals follows a straight line, then it follows normal distribution. In this case, the trend follows almost a straight line and deviating a little towards the end. This can be concluded that the residuals follow a normal distribution for most of the part although there are few outliers in this case.

1. **SCALE-LOCATION**

This graph shows that whether the residuals are spread equally among the range of independent variables. This basically shows the equal variance as this is the plot of square root of standardized residuals vs the fitted values. In the above graph we can see that it is almost a horizontal line which is linear.

1. **RESIDUALS vs LEVERAGE**

This plot help us understand if there are any influential cases in the dataset. This means that when these cases are removed from the dataset, the results will have a significant difference. In our case, #15361 is lying outside the cooks distance which means when this data is removed from the data set, it can alter the results. However, since it is not in the top right corner, the effect will not be much significant as the leverage is only between 0.1 and 0.2. We can also see that #8318 and #8319 are outliers in the dataset, however as they remain within the cook’s distance they are not influential.

##### **COMPARING ACTUAL AND PREDICTED VALUES**

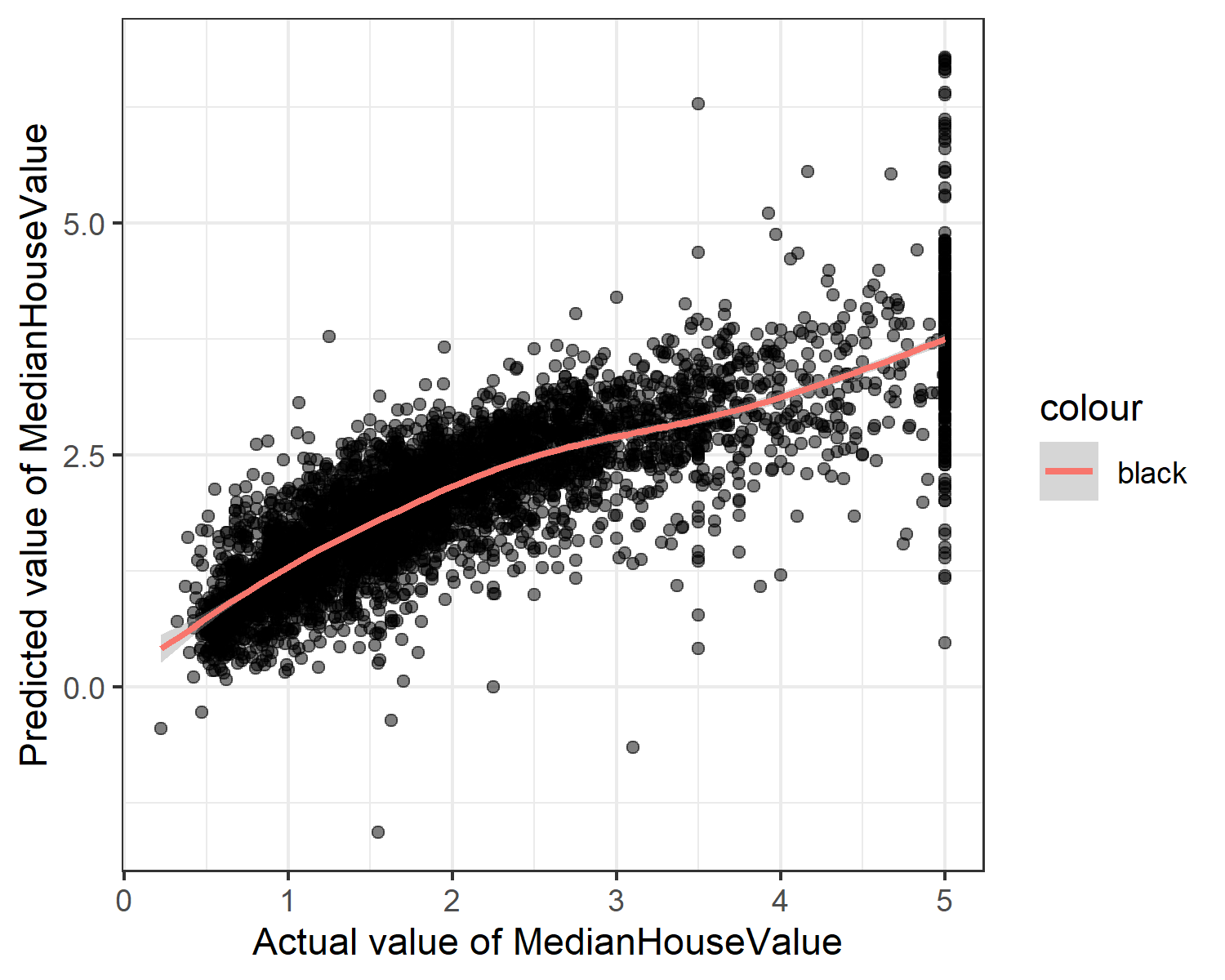
A visual look at how predicted values compare with actuals.

head(Actual\_Pred)

## Actual Predicted  
## 1 4.526 4.081776  
## 4 3.413 3.198101  
## 10 2.611 2.637264  
## 21 1.475 1.413494  
## 23 1.139 1.861877  
## 29 1.089 1.737376

plot <- Actual\_Pred %>%  
 ggplot(aes(Actual, Predicted)) +  
 geom\_point(alpha=0.5) +   
 stat\_smooth(aes(colour='black')) +  
 xlab('Actual value of MedianHouseValue ') +  
 ylab('Predicted value of MedianHouseValue') +  
 theme\_bw()  
  
plot

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



ggplotly(plot)

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

The above graph represents the distribution of actual and predicted median house values.  
The x-axis represents the actual values of the median house value and the y-axis represents the predicted values for the median house values.

##### **CALCULATING THE ERROR METRICS : MAE, MSE, RMSE, MAPE**

DMwR::regr.eval(Actual\_Pred$Actual, Actual\_Pred$Predicted)

## mae mse rmse mape   
## 0.5028420 0.4841630 0.6958182 0.2851152

The values of error metrics help us understand if the model fits the actual data are not. In other words, it tells how well the model performs.

1. **MEAN ABSOLUTE ERROR (MAE)**:

MAE is the average prediction error, which is the average of the difference between the predicted and actual values in the test data.  
There is no ideal value for MAE, but lesser the value the better the model fits the data. If the value is 0 it means the model fits the data perfectly.  
The MAE value observed for this model is **0.5028**, which is near to 0, implies the model is a very good fit for the data.

1. **MEAN SQUARED ERROR (MSE)**:

MSE is the average squared difference between the predicted and actual values. In other words, it is the sum of all the data points, of the square of the difference between the predicted and actual values divided by the number of data points.  
To tell if a model is a good fit or not, the value should as low as possible, if the value is 0, it means it is a perfect fit.  
The MSE value observed for this model is **0.4841**, which is near to 0, implies the model is a very good fit for the data.

1. **ROOT MEAN SQUARE ERROR (RMSE)**:

RMSE is the square root of MSE and is measured in the same units as the target variable.  
Similar to MSE, the value for RMSE also should be less, near to 0, if it is 0, the model is a perfect fit.  
The RMSE value observed for this model is **0.6958**, which is near to 0, implies the model is a very good fit for the data.

1. **MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)** :

MAPE is a statistic used to access how a accurate the forecast model is. The accuracy is measured as a percentage.  
There is no ideal value for MAPE, but lesser the value the better the model fits the data.  
The MAPE value observed for this model is **0.2851**, which is near to 0, implies the model is a very good fit for the data.

# CONCLUSION

The following dataframe shows the metrics used in this analysis.  
This shows that the applied model to make predictions is a good fit for the data.

error\_metrics <- read.csv(file="C:/Users/guest23455/Desktop/Statistics/errormetrics.csv", header = TRUE, stringsAsFactors = FALSE)  
error\_metrics

## STATISTIC CRITERIA  
## 1 R-Squared (R2) Higher the better  
## 2 Adjusted R-Squared (R2) Higher the better  
## 3 F-Statistic Higher the better  
## 4 Std.error Closer to zero, the better  
## 5 Mean Absolute Error (MAE) Lower the better  
## 6 Mean Squared Error (MSE) Lower the better  
## 7 Root Mean Squared Error (RMSE) Lower the better  
## 8 Mean Absolute Percentage Error (MAPE) Lower the better  
## OBSERVED\_VALUES RESULT  
## 1 0.6476 Meets the criteria, a good fit  
## 2 0.6474 Meets the criteria, a good fit  
## 3 2501.0000 Meets the criteria, a good fit  
## 4 0.6843 Meets the criteria, a good fit  
## 5 0.5028 Meets the criteria, a good fit  
## 6 0.4841 Meets the criteria, a good fit  
## 7 0.6958 Meets the criteria, a good fit  
## 8 0.2851 Meets the criteria, a good fit

In this research, we have predicted the housing prices based on parameters such as population, number of rooms, households, income, number of bedrooms, etc, using multiple linear regression.  
Using the metrics and the diagnostic plots, we can see that the model is a fair fit for the given dataset.

## LIMITATIONS OF THE STUDY

The analysis is limited to the independent variables present in the dataset such as population, number of rooms, median household income, etc. Regarding the model itself, only multiple linear regression analysis is applied in this research.

## FUTURE STUDY

Further variables such as size of the rooms, rent price,etc can be applied on the model for more accurate prediction. Also other statistical models not only limited to linear regression such as ANOVA and Ridge regression can be applied and best model can be used to predict the housing price.  
Normal Q-Q diagnostic plot for this model shows it follows a straight line for most of the part. There may be other models which can fit the data better and which follows the straight line better than this. This is a potential scope for further study in this area.

# REFERENCES

1. Kaggle.com. 2020. Kaggle: Your Machine Learning And Data Science Community. [online] Available at: <https://www.kaggle.com/> [Accessed 29 April 2020].
2. Feliperego.github.io. 2020. Quick Guide: Interpreting Simple Linear Model Output In R. [online] Available at: <https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R> [Accessed 8 May 2020].
3. Machine Learning Plus. 2020. Complete Introduction To Linear Regression In R. [online] Available at: <https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/> [Accessed 8 May 2020].
4. Upton, G.; Cook, I. (2014): A Dictionary of Statistics 3e: OUP Oxford (Oxford Paperback Reference). Available online at <https://books.google.de/books?id=4WygAwAAQBAJ>