Mini Project Report

On

BOSTAN HOUSING PRICE

Submitted in partial fulfillment of the requirements of the degree of

Bachelor of Engineering

By

Pooja Molavade [42]
Apeksha Sable [57]

Subject Incharge

Prof. Sonali Mhatre

Department of Information Technology

Bharati Vidyapeeth College of Engineering, Navi Mumbai

2020-2021

Bharati Vidyapeeth College of Engineering, Navi Mumbai

Department of Information Technology

CERTIFICATE

This is to certify that,

Pooja Molavade [42]
Apeksha Sable [57]

Class- BEIT Semester-VIII have completed the Mini Project **BOSTON HOUSING PRICE** *of the Course* ***R Programming Lab*** *Satisfactorily in the Department of Information Technology, as prescribed by the Mumbai University in the academic year 2020-2021.*

Prof. Sonali Mhatre
Subject Incharge

Prof. H.B.Sale
Head

TABLE OF CONTENTS

INTRODUCTION:

Boston is a database with informations of areas around Boston city, and the median house prices. We will use linear regression to predict the house prices.

DATA:

Data variable and there description.

crim: per capita crime rate by town.

zn: proportion of residential land zoned for lots over 25,000 sq.ft.

indus: proportion of non-retail business acres per town.

chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox: nitrogen oxides concentration (parts per 10 million).

rm: average number of rooms per dwelling.

age: proportion of owner-occupied units built prior to 1940.

dis: weighted mean of distances to five Boston employment centres.

rad: index of accessibility to radial highways.

tax: full-value property-tax rate per $10,000.

ptratio: pupil-teacher ratio by town.

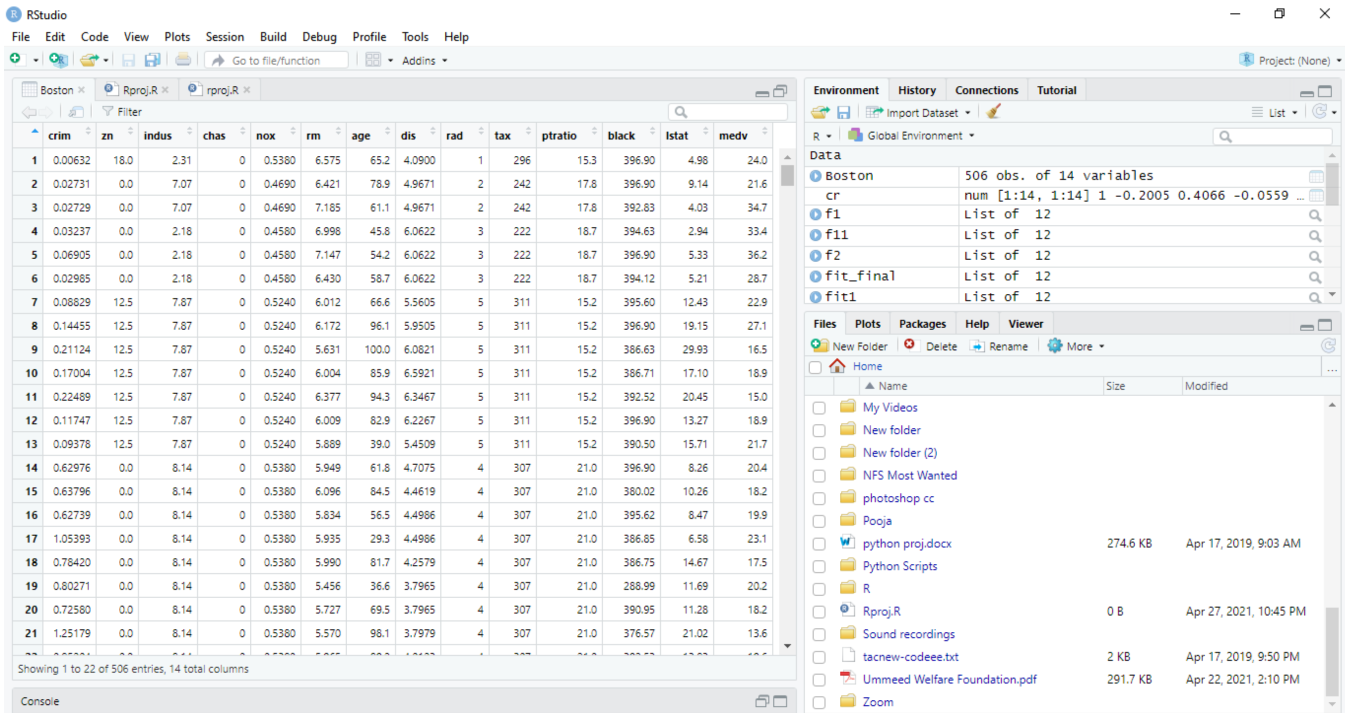
Black: 1000(Bk−0.63)21000(Bk−0.63)2 where BkBk is the proportion of blacks by town.

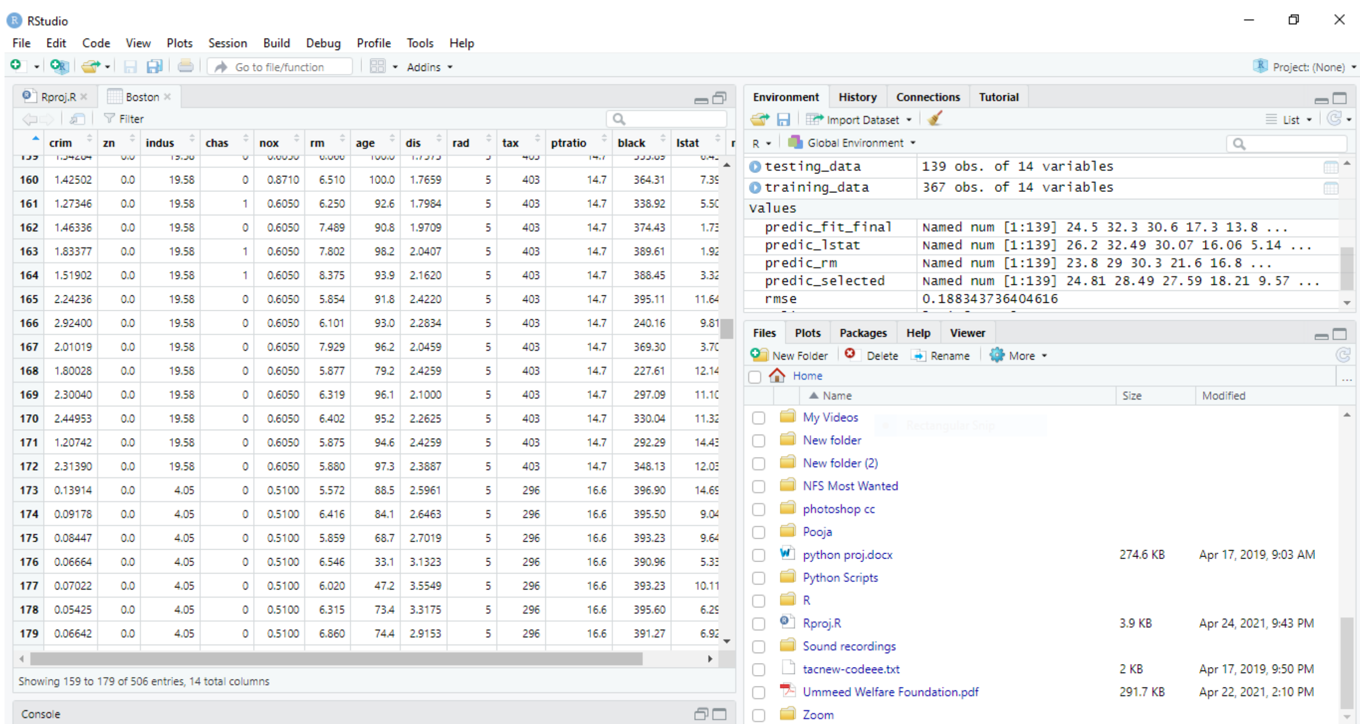
Lstat: lower status of the population (percent).

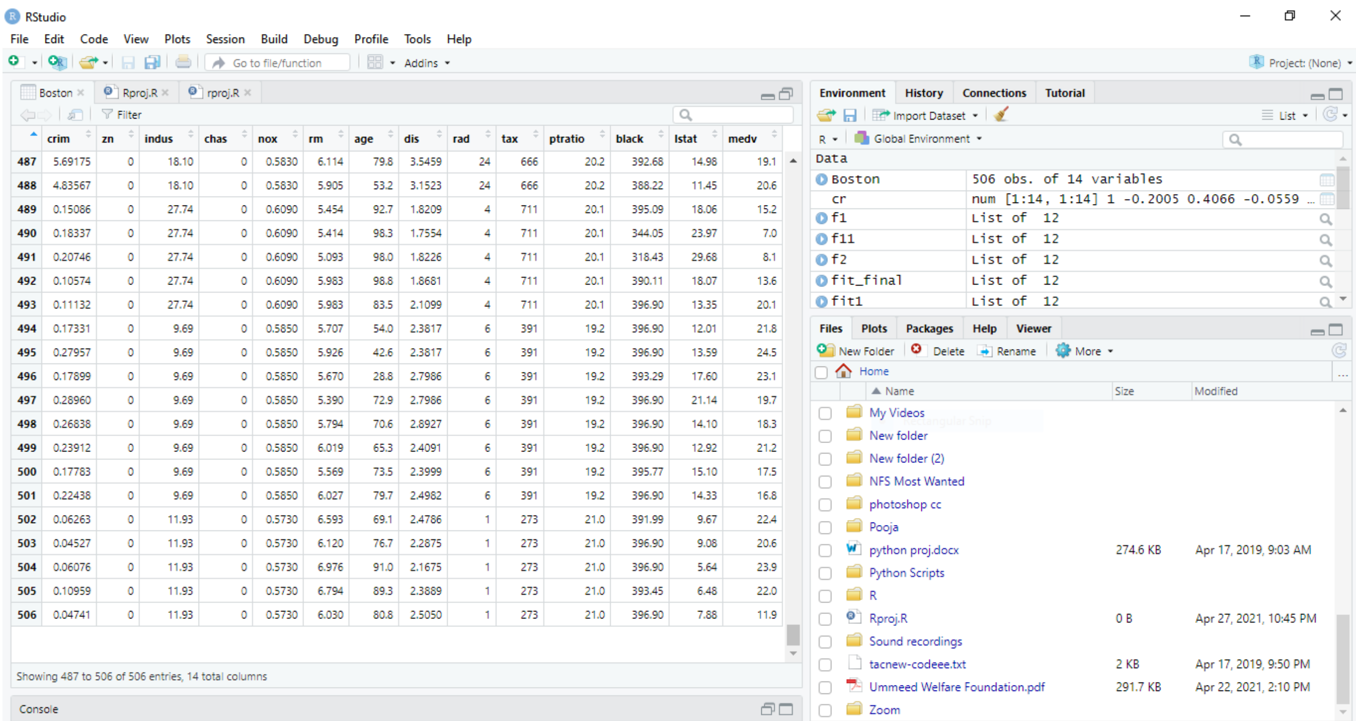
Medv: median value of owner-occupied homes in $1000s.

We have to take this data and make an easier visual representation of the eligibility process based on details. For this project we will be using the language “R”

DATASET:

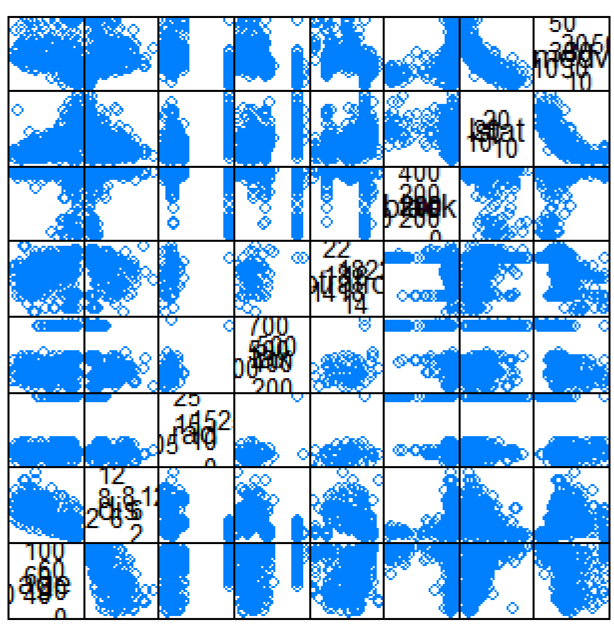

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Boston
Rproj.R
rproj.R
Filter
crim zn indus chas nox rm age dis rad tax ptratio black lstat medv
1 0.00632 18.0 2.31 0 0.5380 6.575 65.2 4.0900 1 296 15.3 396.90 4.98 24.0
2 0.02731 0.0 7.07 0 0.4690 6.421 78.9 4.9671 2 242 17.8 396.90 9.14 21.6
3 0.02729 0.0 7.07 0 0.4690 7.185 61.1 4.9671 2 242 17.8 392.83 4.03 34.7
4 0.03237 0.0 2.18 0 0.4580 6.998 45.8 6.0622 3 222 18.7 394.63 2.94 33.4
5 0.06905 0.0 2.18 0 0.4580 7.147 54.2 6.0622 3 222 18.7 396.90 5.33 36.2
6 0.02985 0.0 2.18 0 0.4580 6.430 58.7 6.0622 3 222 18.7 394.12 5.21 28.7
7 0.08829 12.5 7.87 0 0.5240 6.012 66.6 5.5605 5 311 15.2 395.60 12.43 22.9
8 0.14455 12.5 7.87 0 0.5240 6.172 96.1 5.9505 5 311 15.2 396.90 19.15 27.1
9 0.21124 12.5 7.87 0 0.5240 5.631 100.0 6.0821 5 311 15.2 386.63 29.93 16.5
10 0.17004 12.5 7.87 0 0.5240 6.004 85.9 6.5921 5 311 15.2 386.71 17.10 18.9
11 0.22489 12.5 7.87 0 0.5240 6.377 94.3 6.3467 5 311 15.2 392.52 20.45 15.0
12 0.11747 12.5 7.87 0 0.5240 6.009 82.9 6.2267 5 311 15.2 396.90 13.27 18.9
13 0.09378 12.5 7.87 0 0.5240 5.889 39.0 5.4509 5 311 15.2 390.50 15.71 21.7
14 0.62976 0.0 8.14 0 0.5380 5.949 61.8 4.7075 4 307 21.0 396.90 8.26 20.4
15 0.63796 0.0 8.14 0 0.5380 6.096 84.5 4.4619 4 307 21.0 380.02 10.26 18.2
16 0.62739 0.0 8.14 0 0.5380 5.834 56.5 4.4986 4 307 21.0 395.62 8.47 19.9
17 1.05393 0.0 8.14 0 0.5380 5.935 29.3 4.4986 4 307 21.0 386.85 6.58 23.1
18 0.78420 0.0 8.14 0 0.5380 5.990 81.7 4.2579 4 307 21.0 386.75 14.67 17.5
19 0.80271 0.0 8.14 0 0.5380 5.456 36.6 3.7965 4 307 21.0 288.99 11.69 20.2
20 0.72580 0.0 8.14 0 0.5380 5.727 69.5 3.7965 4 307 21.0 390.95 11.28 18.2
21 1.25179 0.0 8.14 0 0.5380 5.570 98.1 3.7979 4 307 21.0 376.57 21.02 13.6
Showing 1 to 22 of 506 entries, 14 total columns
Console
Environment History Connections Tutorial
Import Dataset
List
Global Environment
Data
Boston 506 obs. of 14 variables
cr num [1:14, 1:14] 1 -0.2005 0.4066 -0.0559 ...
f1 List of 12
f11 List of 12
f2 List of 12
fit_final List of 12
fit1 List of 12
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home
Name Size Modified
My Videos
New folder
New folder (2)
NFS Most Wanted
photoshop cc
Pooja
python proj.docx 274.6 KB Apr 17, 2019, 9:03 AM
Python Scripts
R
Rproj.R 0 B Apr 27, 2021, 10:45 PM
Sound recordings
tacnew-codeee.txt 2 KB Apr 17, 2019, 9:50 PM
Ummeed Welfare Foundation.pdf 291.7 KB Apr 22, 2021, 2:10 PM
Zoom


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Rproj.R
Boston
Filter
crim zn indus chas nox rm age dis rad tax ptratio black lstat
160 1.42502 0.0 19.58 0 0.8710 6.510 100.0 1.7659 5 403 14.7 364.31
161 1.27346 0.0 19.58 1 0.6050 6.250 92.6 1.7984 5 403 14.7 338.92
162 1.46336 0.0 19.58 0 0.6050 7.489 90.8 1.9709 5 403 14.7 374.43
163 1.83377 0.0 19.58 1 0.6050 7.802 98.2 2.0407 5 403 14.7 389.61
164 1.51902 0.0 19.58 1 0.6050 8.375 93.9 2.1620 5 403 14.7 388.45
165 2.24236 0.0 19.58 0 0.6050 5.854 91.8 2.4220 5 403 14.7 395.11
166 2.92400 0.0 19.58 0 0.6050 6.101 93.0 2.2834 5 403 14.7 240.16
167 2.01019 0.0 19.58 0 0.6050 7.929 96.2 2.0459 5 403 14.7 369.30
168 1.80028 0.0 19.58 0 0.6050 5.877 79.2 2.4259 5 403 14.7 227.61
169 2.30040 0.0 19.58 0 0.6050 6.319 96.1 2.1000 5 403 14.7 297.09
170 2.44953 0.0 19.58 0 0.6050 6.402 95.2 2.2625 5 403 14.7 330.04
171 1.20742 0.0 19.58 0 0.6050 5.875 94.6 2.4259 5 403 14.7 292.29
172 2.31390 0.0 19.58 0 0.6050 5.880 97.3 2.3887 5 403 14.7 348.13
173 0.13914 0.0 4.05 0 0.5100 5.572 88.5 2.5961 5 296 16.6 396.90
174 0.09178 0.0 4.05 0 0.5100 6.416 84.1 2.6463 5 296 16.6 395.50
175 0.08447 0.0 4.05 0 0.5100 5.859 68.7 2.7019 5 296 16.6 393.23
176 0.06664 0.0 4.05 0 0.5100 6.546 33.1 3.1323 5 296 16.6 390.96
177 0.07022 0.0 4.05 0 0.5100 6.020 47.2 3.5549 5 296 16.6 393.23
178 0.05425 0.0 4.05 0 0.5100 6.315 73.4 3.3175 5 296 16.6 395.60
179 0.06642 0.0 4.05 0 0.5100 6.860 74.4 2.9153 5 296 16.6 391.27
Showing 159 to 179 of 506 entries, 14 total columns
Console
Environment History Connections Tutorial
Import Dataset
List
Global Environment
testing_data 139 obs. of 14 variables
training_data 367 obs. of 14 variables
Values
predic_fit_final Named num [1:139] 24.5 32.3 30.6 17.3 13.8 ...
predic_lstat Named num [1:139] 26.2 32.49 30.07 16.06 5.14 ...
predic_rm Named num [1:139] 23.8 29 30.3 21.6 16.8 ...
predic_selected Named num [1:139] 24.81 28.49 27.59 18.21 9.57 ...
rmse 0.188343736404616
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home
Name Size Modified
My Videos
New folder
New folder (2)
NFS Most Wanted
photoshop cc
Pooja
python proj.docx 274.6 KB Apr 17, 2019, 9:03 AM
Python Scripts
R
Rproj.R 3.9 KB Apr 24, 2021, 9:43 PM
Sound recordings
tacnew-codeee.txt 2 KB Apr 17, 2019, 9:50 PM
Ummeed Welfare Foundation.pdf 291.7 KB Apr 22, 2021, 2:10 PM
Zoom

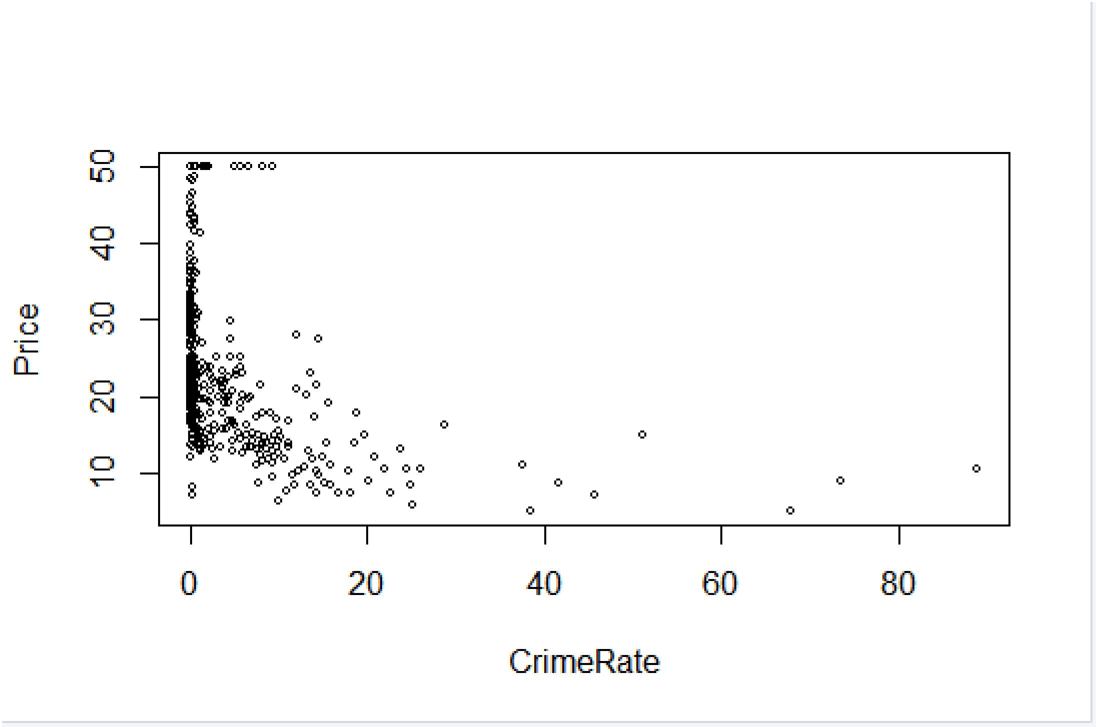


We import this dataset in R Studio in which analysis is performed and proper output is generated. With this dataset we will be visualizing, Plots, Graphs and a Linear Regression algorithm.

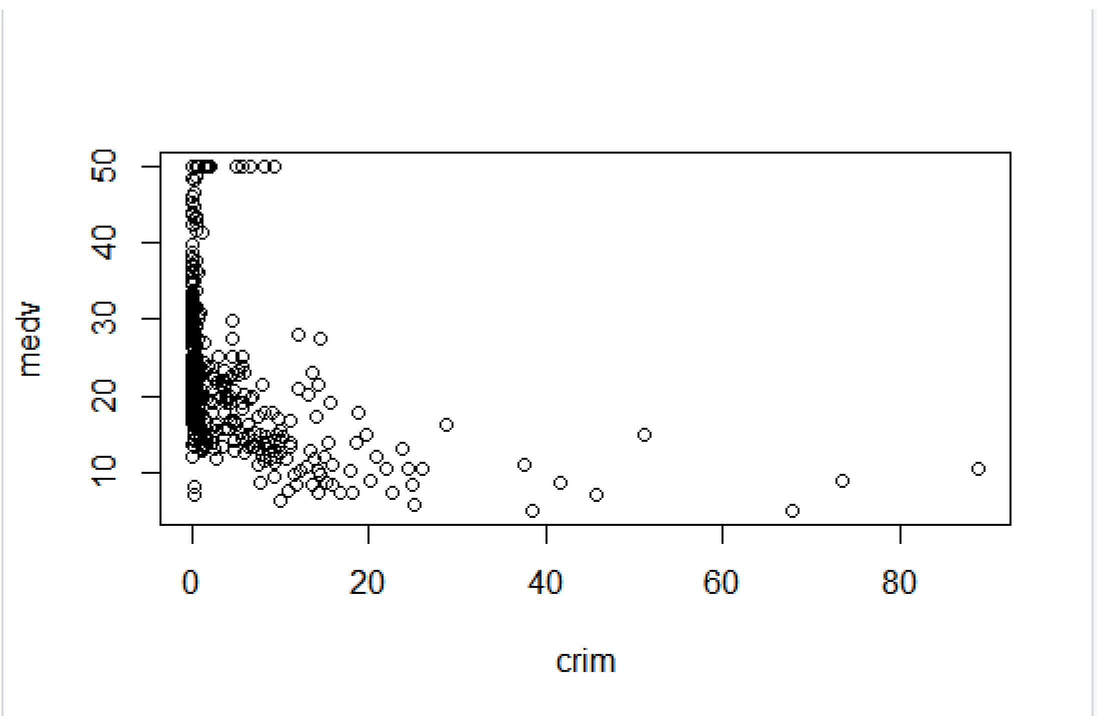
GRAPHICS:



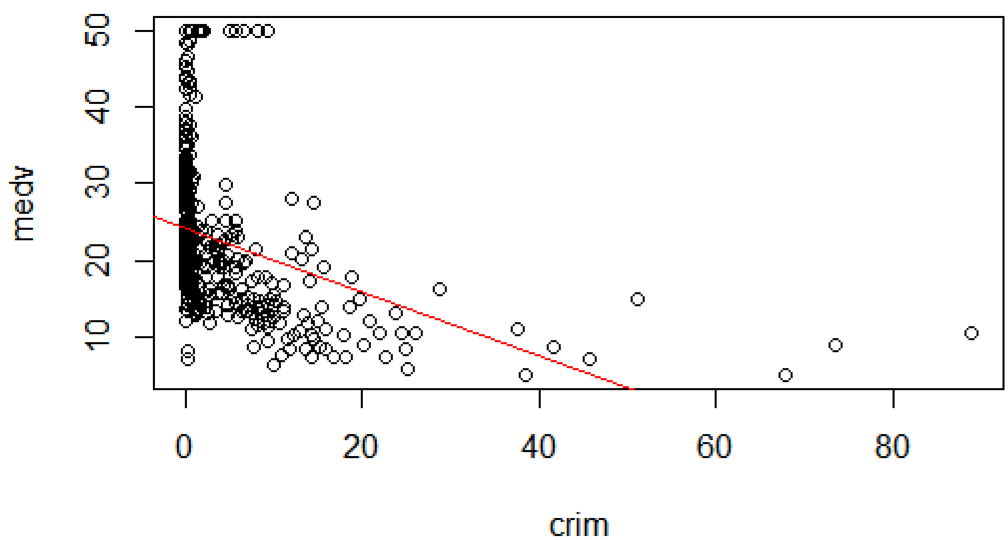
Scatter Plot Matrix



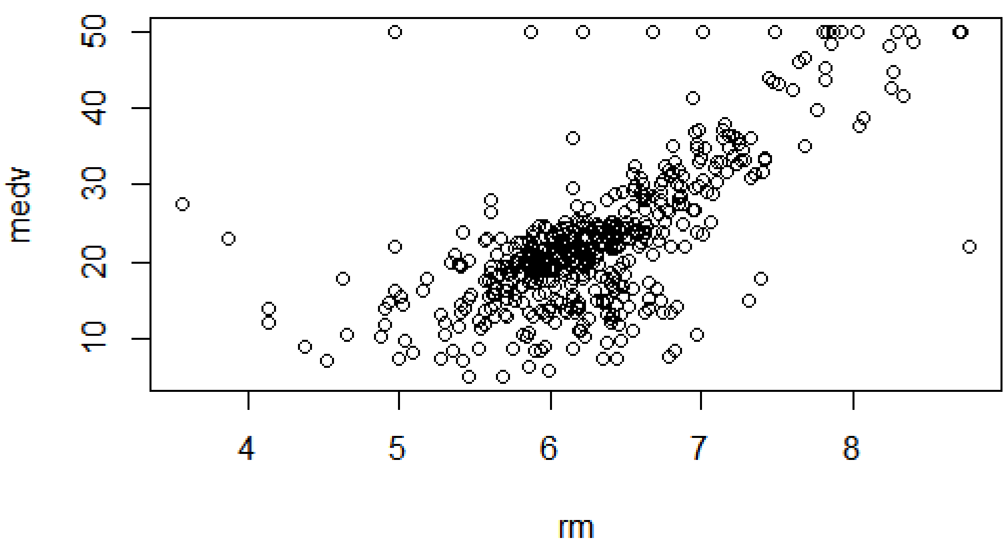
Price
CrimeRate
0
20
40
60
80
10
20
30
40
50



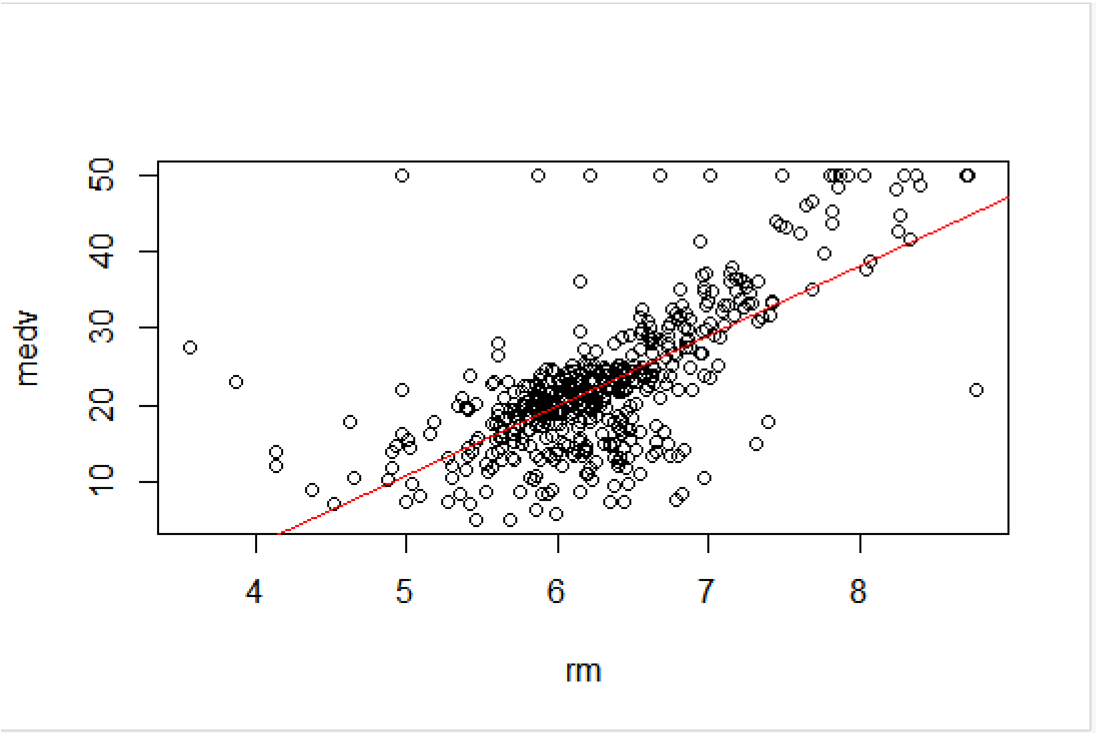
medv
crim
0
20
40
60
80
10
20
30
40
50

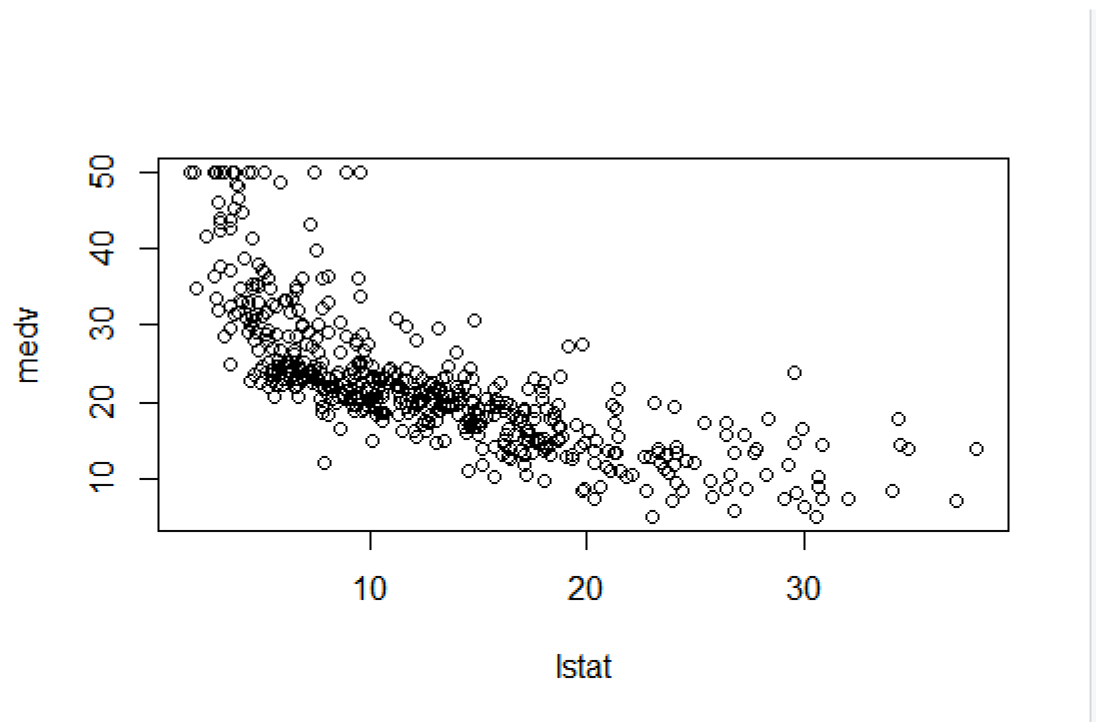


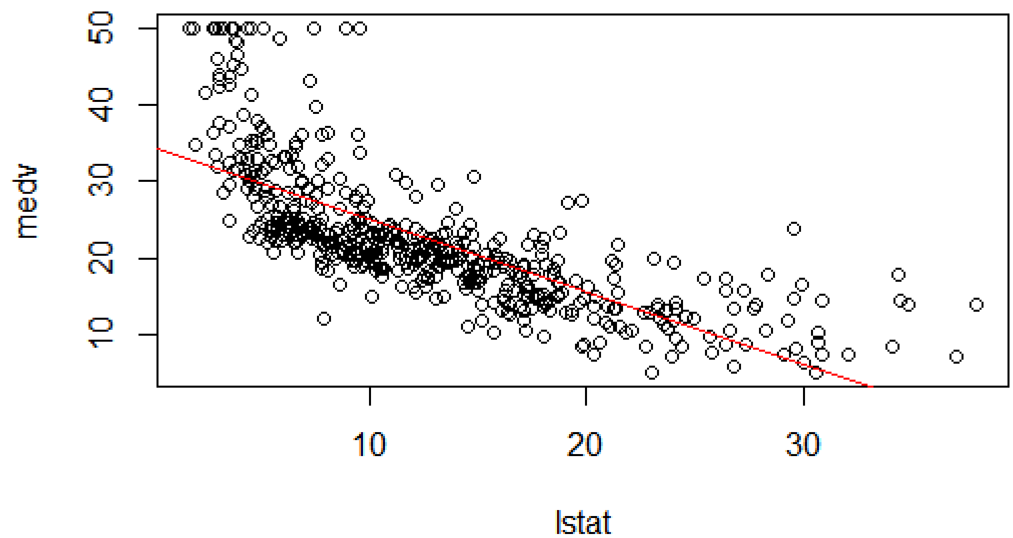
medv
crim
0
20
40
60
80
10
20
30
40
50

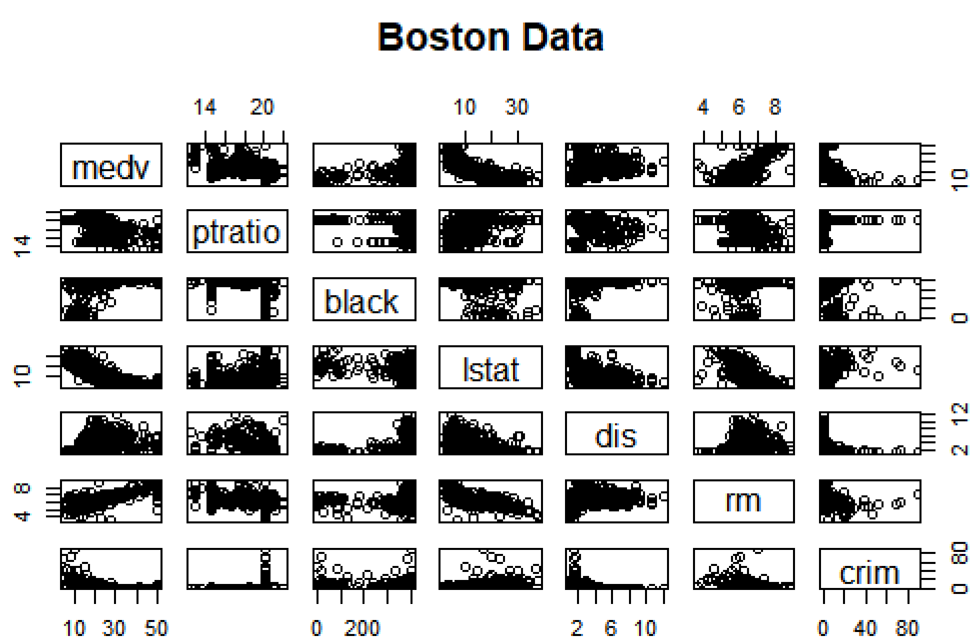


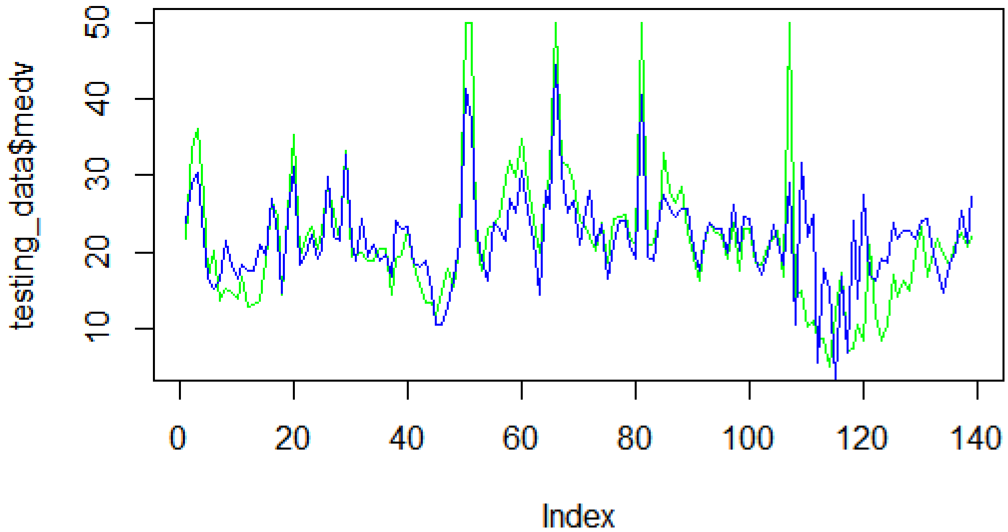
medv
rm
4
5
6
7
8
10
20
30
40
50

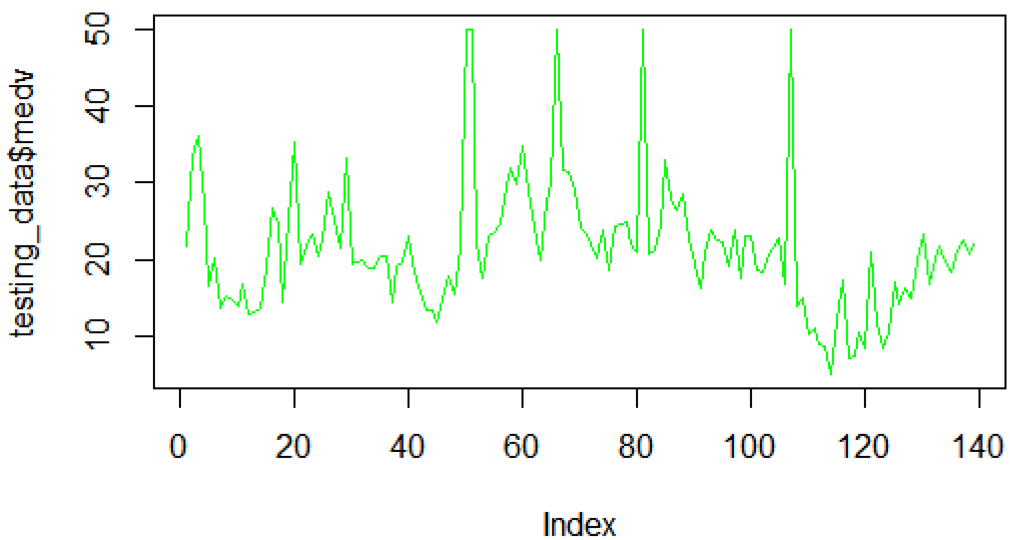

medv
10
20
30
40
50
4
5
6
7
8
rm

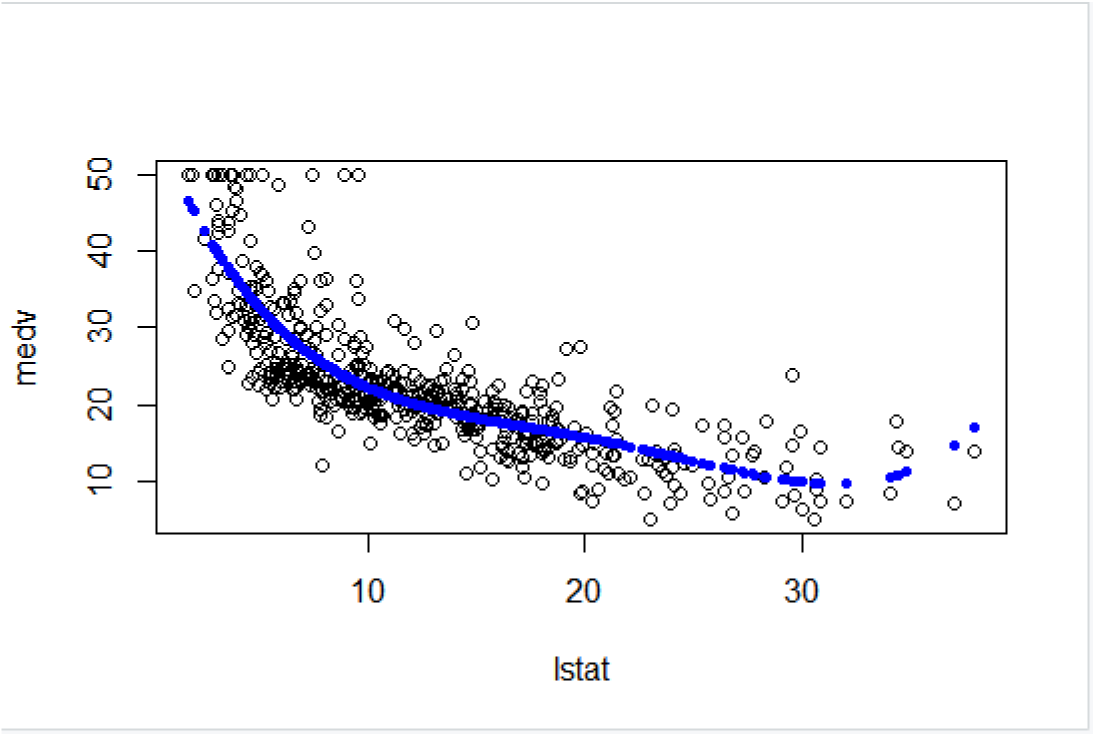

medv
10
20
30
40
50
10
20
30
lstat

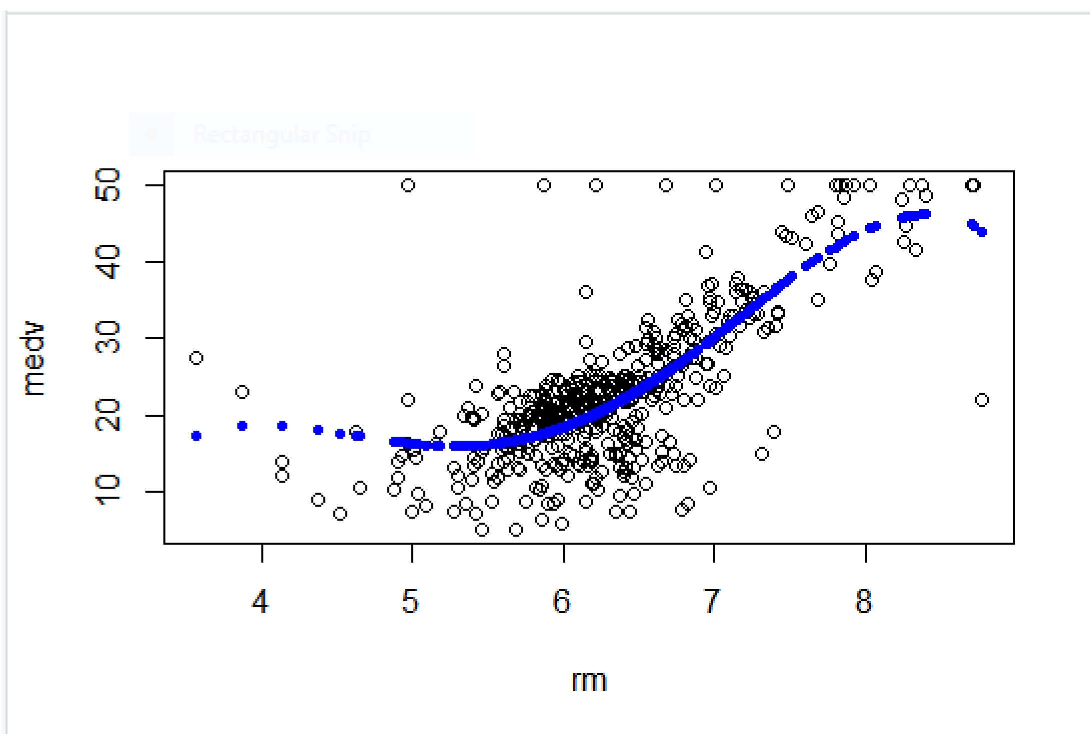

medv
50
40
30
20
10
10
20
30
lstat


Boston Data
medv
ptratio
black
lstat
dis
rm
crim


testing_data$medv
10
20
30
40
50
0
20
40
60
80
100
120
140
Index


testing_data$medv
10
20
30
40
50
0
20
40
60
80
100
120
140
Index


medv
50
40
30
20
10
10
20
30
lstat


medv
50
40
30
20
10
4
5
6
7
8
rm

DATA MINING ALGORITHM:

For this project, we are using the data mining algorithm which is Linear Regression algorithm. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables.

For example, **regression** might be used to predict the cost of a product or service, given other variables. Regression analysis is a very widely used statistical tool to establish a relationship model between two variables. One of these variable is called predictor variable whose value is gathered through experiments. The other variable is called response variable whose value is derived from the predictor variable. In Linear Regression these two variables are related through an equation, where exponent (power) of both these variables is 1. Mathematically a linear relationship represents a straight line when plotted as a graph. A non-linear relationship where the exponent of any variable is not equal to 1 creates a curve.

The general mathematical equation for a linear regression is −

$$y = ax + b$$

Following is the description of the parameters used −

- **y** is the response variable.
- **x** is the predictor variable.
- **a** and **b** are constants which are called the coefficients.

PROGRAM CODE:

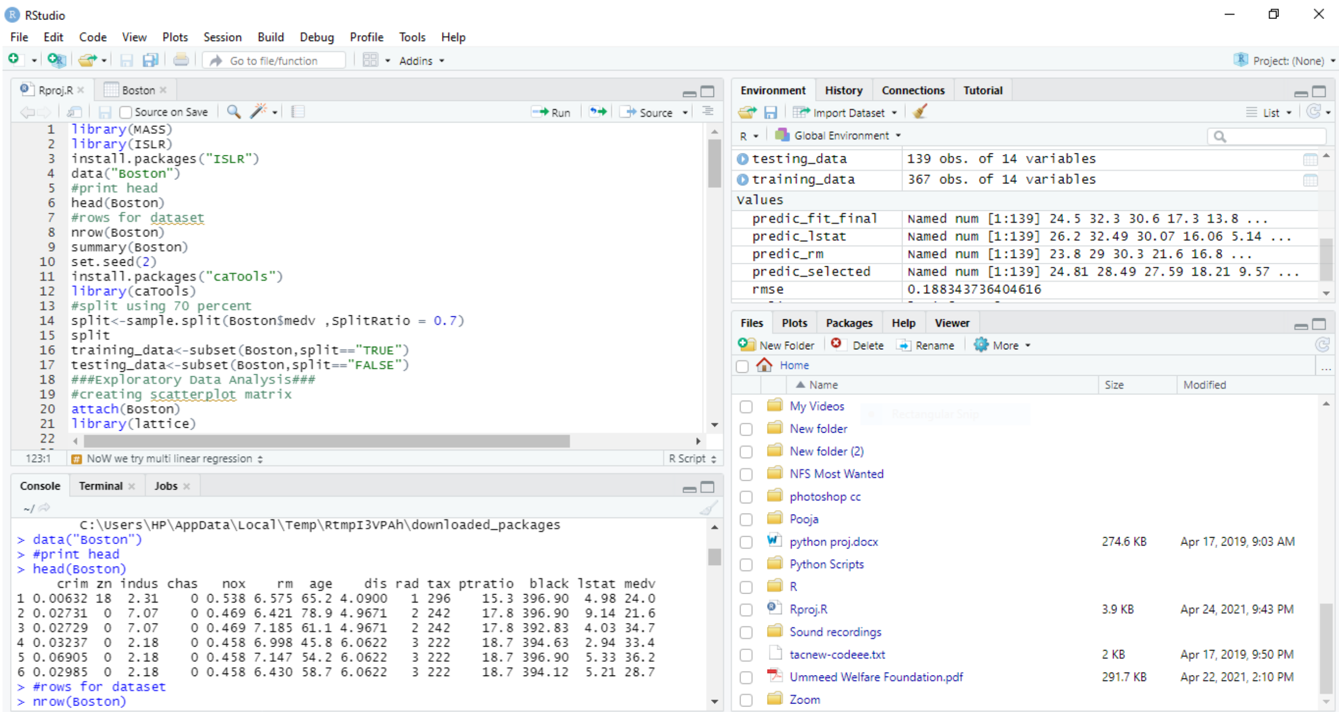
```
library(MASS)
library(ISLR)

#install.packages("ISLR")
data("Boston")

#print head
head(Boston)

#rows for dataset
nrow(Boston)

summary(Boston)

set.seed(2)
library(caTools)

#split using 70 percent
split<-sample.split(Boston$medv ,SplitRatio = 0.7)
split

training_data<-subset(Boston,split=="TRUE")
testing_data<-subset(Boston,split=="FALSE")


###Exploratory Data Analysis###

#creating scatterplot matrix
attach(Boston)
library(lattice)
splom(~Boston[c(1:6,14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
splom(~Boston[c(7:14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)

#corplot to visualize
#install.packages("corrplot")
library(corrplot)
corrplot(cr, type = "lower")
corrplot(cr, method = "number")


#to view corelation of variables
plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
cr<-cor(Boston)
pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston Data")
```

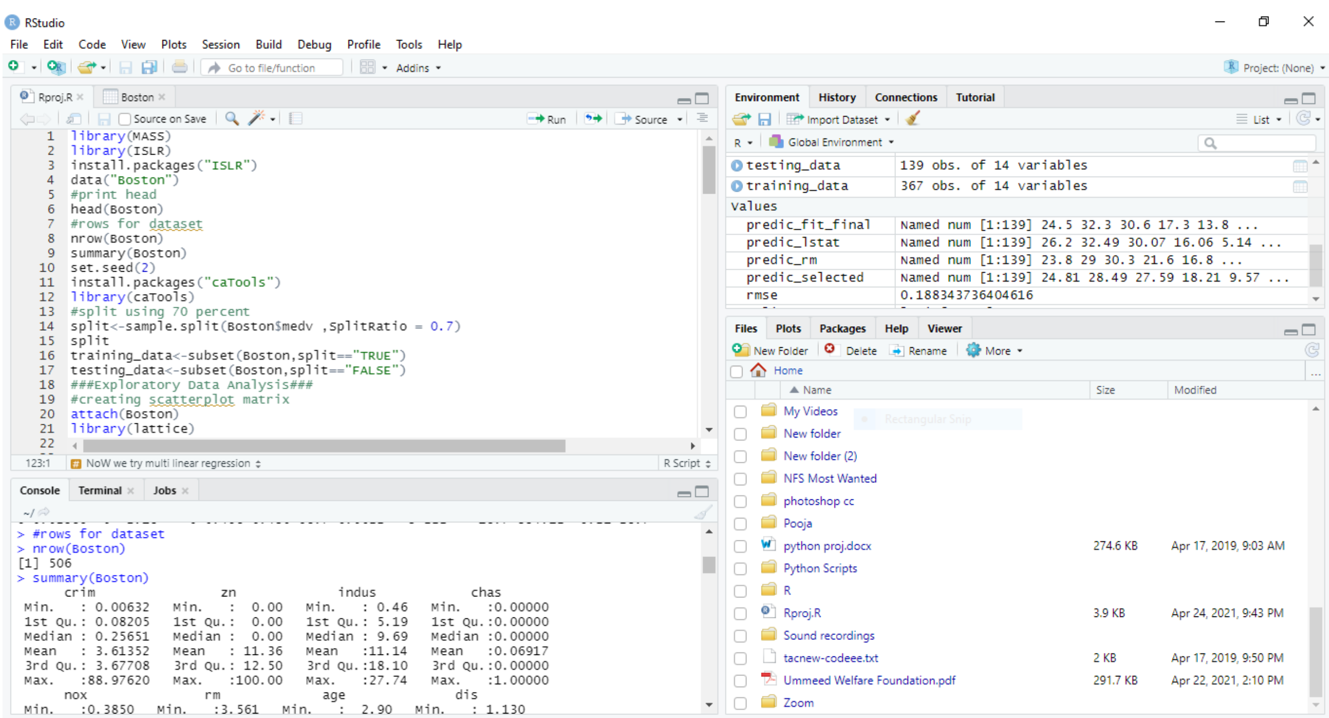
```
## crim is not acceptable to be a linear variable


#studying crim and medv
plot(crim,medv)
fit1<-lm(medv~crim, data=Boston)
abline(fit1, col="red")# regression fit line

#studying rm and medv
plot(rm,medv)
fit1<-lm(medv~rm, data=Boston)
abline(fit1, col="red")# regression fit line

#studying lstat and medv
plot(lstat,medv)
fit1<-lm(medv~lstat, data=Boston)
abline(fit1, col="red")# regression fit line




##Creating Model


####Since line is acceptable through rm and lstat variable we use rm, lstat to model to predict data
####Using rm, lstat as they are good linear variables

#Rm
model_regx_rm<-lm(medv~rm,data = training_data)
#summary
summary(model_regx_rm)
#prediction
predic_rm<-predict(model_regx_rm, testing_data)
predic_rm
#compare actual values and prediction
plot(testing_data$medv, type = "l", lty = 1.8, col = "green")
lines(predic_rm,type = "l", col = "blue")



#lstat
model_regx_lstat<-lm(medv~lstat,data = training_data)
#summary
summary(model_regx_lstat)
#prediction
predic_lstat<-predict(model_regx_lstat, testing_data)
```

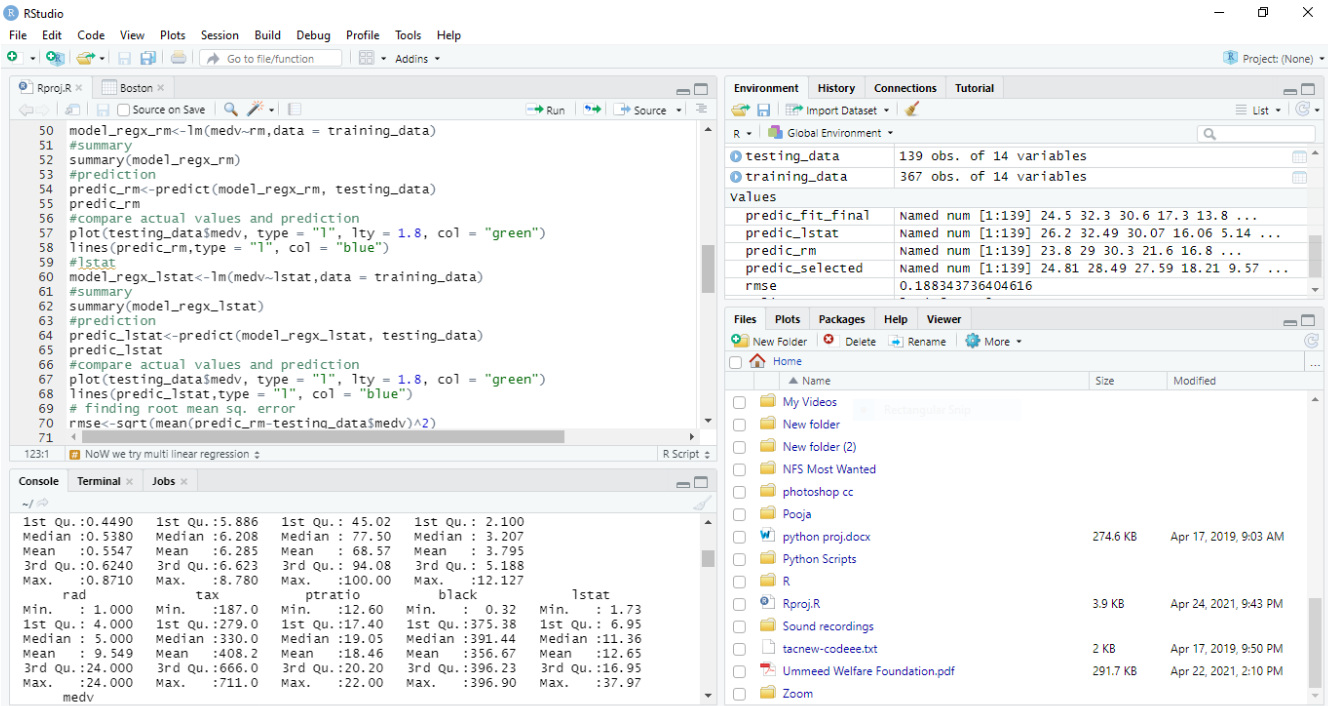
```
predic_lstat
#compare actual values and prediction
plot(testing_data$medv, type = "l", lty = 1.8, col = "green")
lines(predic_lstat,type = "l", col = "blue")


# finding root mean sq. error
rmse<-sqrt(mean(predic_rm-testing_data$medv)^2)
rmse
rmse<-sqrt(mean(predic_lstat-testing_data$medv)^2)
rmse

#### NoW we try multi linear regression ####

#selecting only variables
model_regx_ml<-lm(medv~ rm + lstat,data = Boston)
#summary
summary(model_regx_ml)


#selecting all variables
model_regx_all<-lm(medv~.,data = training_data)
#summary
summary(model_regx_all)


#removing age and indus
model_regx_selected<-lm(medv~ crim + zn + tax + chas + rm + rad + dis + nox +
                ptratio + black + lstat,data = training_data)
#summary
summary(model_regx_selected)

#prediction
predic_selected<-predict(model_regx_selected, testing_data)
predic_selected


# finding root mean sq. error
rmse<-sqrt(mean(predic_selected-testing_data$medv)^2)
rmse

#compare actual values and prediction
plot(testing_data$medv, type = "l", lty = 1.8, col = "green")
lines(predic_selected,type = "l", col = "blue")
```

```
#since rmse value is still high we need to optimize the model

f1=lm(medv~lstat +I(lstat^2),Boston)
summary(fit1)
attach(Boston)
f11=lm(medv~poly(lstat,4))
plot(medv~lstat)
points(lstat,fitted(f11),col="blue",pch=20)

f2=lm(medv~rm +I(rm^2),Boston)
summary(f2)
attach(Boston)
fit22=lm(medv~poly(rm,4))
plot(medv~rm)
points(rm,fitted(fit22),col="blue",pch=20)


#building final model
fit_final=lm(medv~lstat+crim+rm+dis+black+chas+nox+rad+tax+ptratio+I(lstat^2)+I(rm^2))
summary(fit_final)

#prediction
predic_fit_final<-predict(fit_final, testing_data)
predic_fit_final

# finding root mean sq. error
rmse<-sqrt(mean(predic_fit_final-testing_data$medv)^2)
rmse

#compare actual values and prediction
plot(testing_data$medv, type = "l", lty = 1.8, col = "green")
lines(predic_fit_final,type = "l", col = "blue")
```

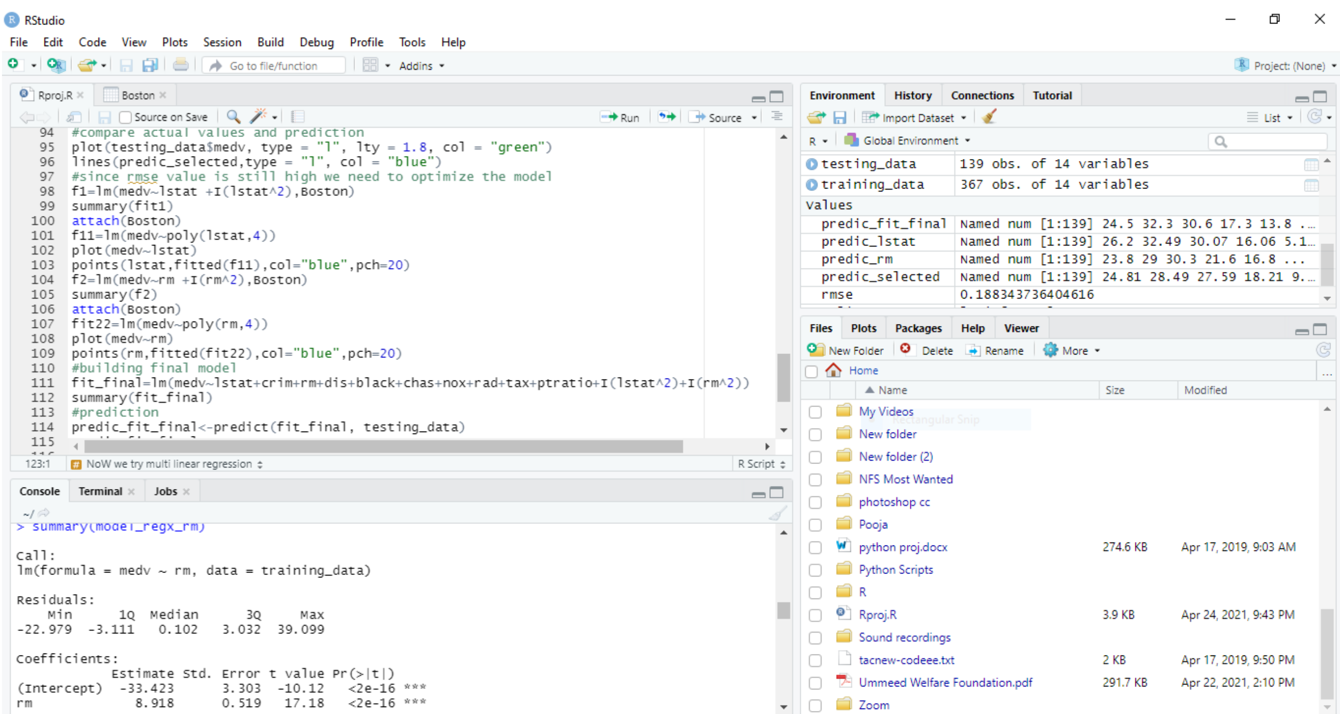
OUTPUT:


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
library(MASS)
library(ISLR)
install.packages("ISLR")
data("Boston")
#print head
head(Boston)
#rows for dataset
nrow(Boston)
summary(Boston)
set.seed(2)
install.packages("caTools")
library(caTools)
#split using 70 percent
split<-sample.split(Boston$medv ,SplitRatio = 0.7)
split
training_data<-subset(Boston,split=="TRUE")
testing_data<-subset(Boston,split=="FALSE")
###Exploratory Data Analysis###
#creating scatterplot matrix
attach(Boston)
library(lattice)
NoW we try multi linear regression
C:\Users\HP\AppData\Local\Temp\RtmpI3VPAh\downloaded_packages
> data("Boston")
> #print head
> head(Boston)
crim zn indus chas nox rm age dis rad tax ptratio black lstat medv
1 0.00632 18 2.31 0 0.538 6.575 65.2 4.0900 1 296 15.3 396.90 4.98 24.0
2 0.02731 0 7.07 0 0.469 6.421 78.9 4.9671 2 242 17.8 396.90 9.14 21.6
3 0.02729 0 7.07 0 0.469 7.185 61.1 4.9671 2 242 17.8 392.83 4.03 34.7
4 0.03237 0 2.18 0 0.458 6.998 45.8 6.0622 3 222 18.7 394.63 2.94 33.4
5 0.06905 0 2.18 0 0.458 7.147 54.2 6.0622 3 222 18.7 396.90 5.33 36.2
6 0.02985 0 2.18 0 0.458 6.430 58.7 6.0622 3 222 18.7 394.12 5.21 28.7
> #rows for dataset
> nrow(Boston)
Environment History Connections Tutorial
testing_data 139 obs. of 14 variables
training_data 367 obs. of 14 variables
Values
predic_fit_final Named num [1:139] 24.5 32.3 30.6 17.3 13.8 ...
predic_lstat Named num [1:139] 26.2 32.49 30.07 16.06 5.14 ...
predic_rm Named num [1:139] 23.8 29 30.3 21.6 16.8 ...
predic_selected Named num [1:139] 24.81 28.49 27.59 18.21 9.57 ...
rmse 0.188343736404616
Files Plots Packages Help Viewer
Name Size Modified
My Videos
New folder
New folder (2)
NFS Most Wanted
photoshop cc
Pooja
python proj.docx 274.6 KB Apr 17, 2019, 9:03 AM
Python Scripts
R
Rproj.R 3.9 KB Apr 24, 2021, 9:43 PM
Sound recordings
tacnew-codeee.txt 2 KB Apr 17, 2019, 9:50 PM
Ummeed Welfare Foundation.pdf 291.7 KB Apr 22, 2021, 2:10 PM
Zoom

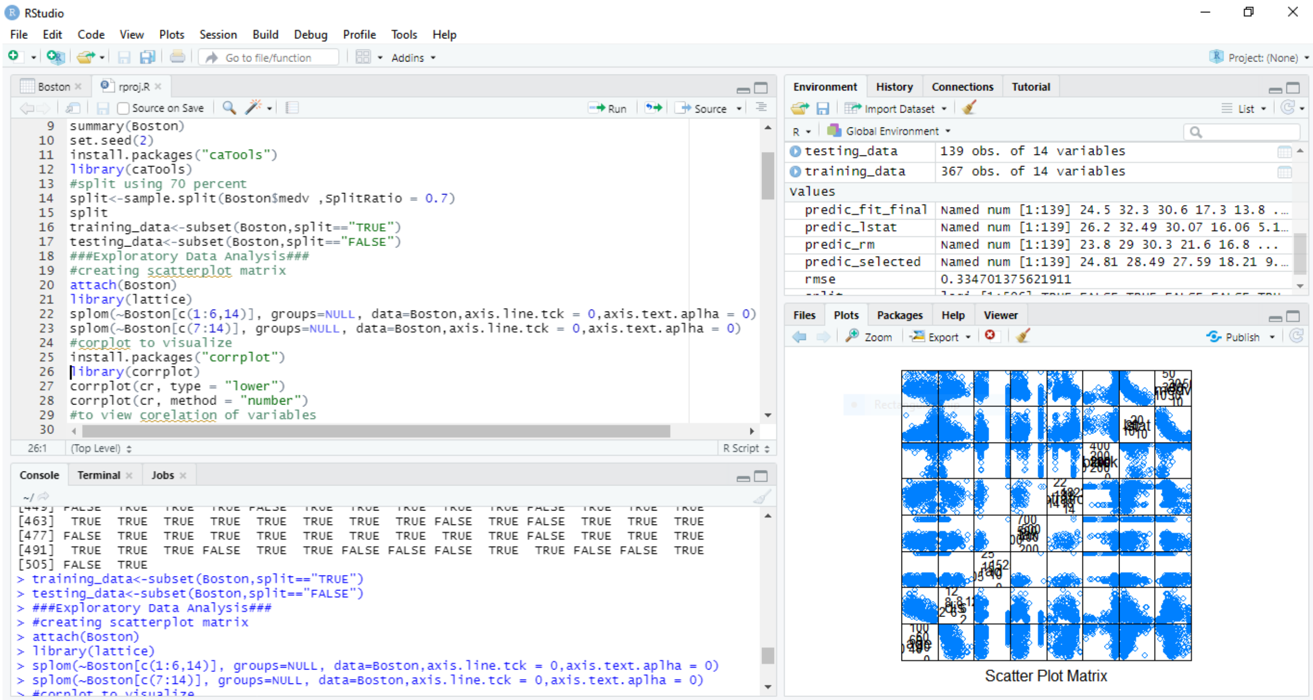

RStudio
> #rows for dataset
> nrow(Boston)
[1] 506
> summary(Boston)
crim zn indus chas
Min. : 0.00632 Min. : 0.00 Min. : 0.46 Min. :0.00000
1st Qu.: 0.08205 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.:0.00000
Median : 0.25651 Median : 0.00 Median : 9.69 Median :0.00000
Mean : 3.61352 Mean : 11.36 Mean :11.14 Mean :0.06917
3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.:18.10 3rd Qu.:0.00000
Max. :88.97620 Max. :100.00 Max. :27.74 Max. :1.00000
nox rm age dis
Min. :0.3850 Min. :3.561 Min. : 2.90 Min. : 1.130

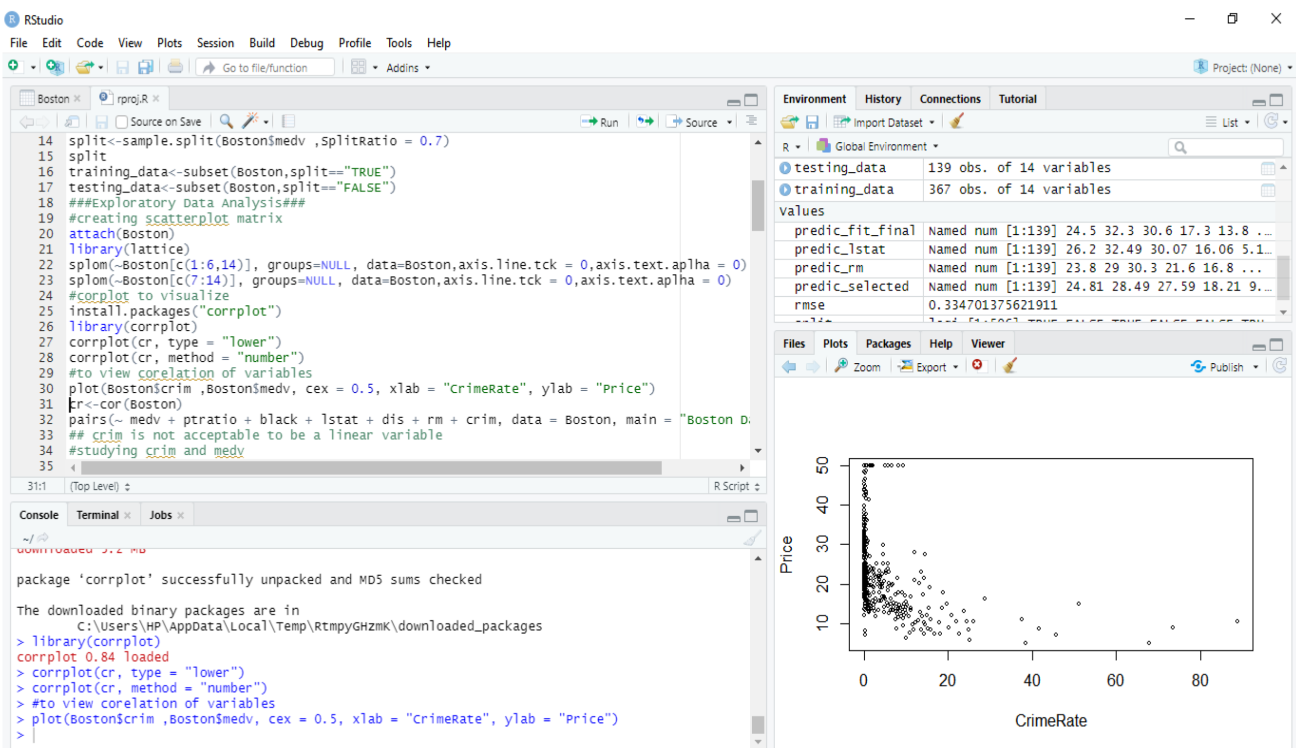


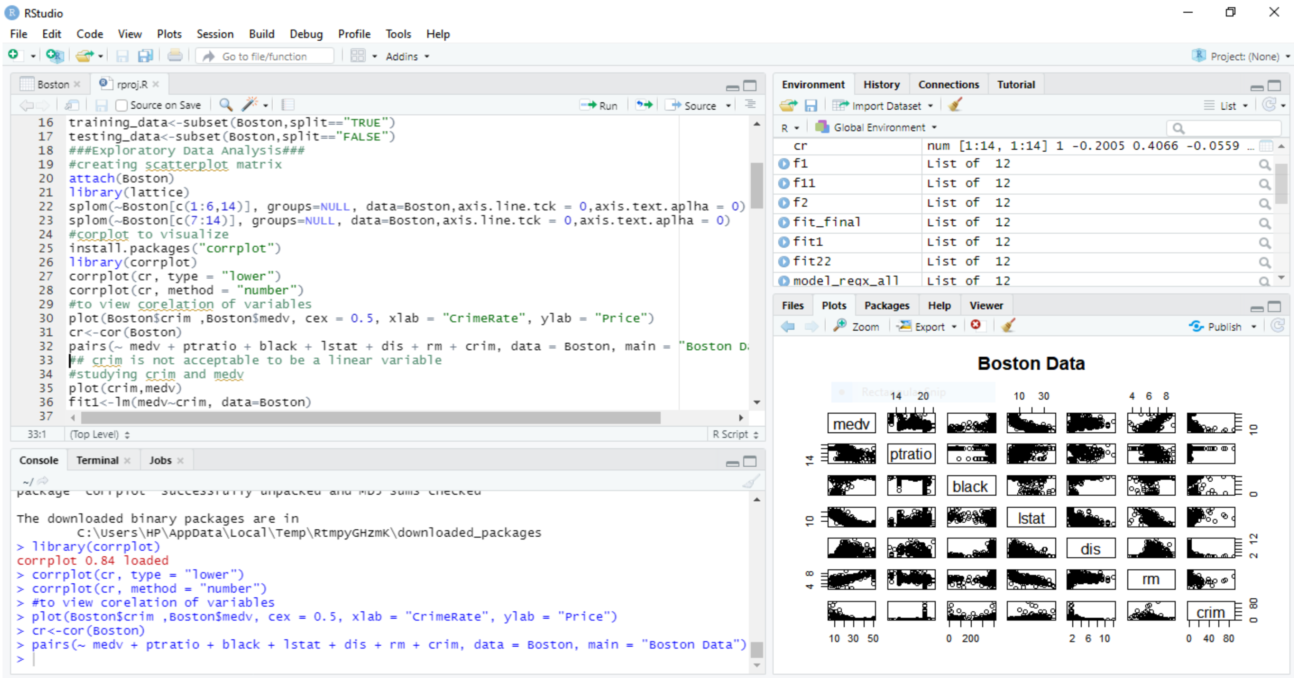
model_regx_rm<-lm(medv~rm,data = training_data)
#summary
summary(model_regx_rm)
#prediction
predic_rm<-predict(model_regx_rm, testing_data)
predic_rm
#compare actual values and prediction
plot(testing_data$medv, type = "l", lty = 1.8, col = "green")
lines(predic_rm,type = "l", col = "blue")
#lstat
model_regx_lstat<-lm(medv~lstat,data = training_data)
#summary
summary(model_regx_lstat)
#prediction
predic_lstat<-predict(model_regx_lstat, testing_data)
predic_lstat
#compare actual values and prediction
plot(testing_data$medv, type = "l", lty = 1.8, col = "green")
lines(predic_lstat,type = "l", col = "blue")
# finding root mean sq. error
rmse<-sqrt(mean(predic_rm-testing_data$medv)^2)
testing_data 139 obs. of 14 variables
training_data 367 obs. of 14 variables
predic_fit_final Named num [1:139] 24.5 32.3 30.6 17.3 13.8 ...
predic_lstat Named num [1:139] 26.2 32.49 30.07 16.06 5.14 ...
predic_rm Named num [1:139] 23.8 29 30.3 21.6 16.8 ...
predic_selected Named num [1:139] 24.81 28.49 27.59 18.21 9.57 ...
rmse 0.188343736404616
1st Qu.:0.4490 1st Qu.:5.886 1st Qu.: 45.02 1st Qu.: 2.100
Median :0.5380 Median :6.208 Median : 77.50 Median : 3.207
Mean :0.5547 Mean :6.285 Mean : 68.57 Mean : 3.795
3rd Qu.:0.6240 3rd Qu.:6.623 3rd Qu.: 94.08 3rd Qu.: 5.188
Max. :0.8710 Max. :8.780 Max. :100.00 Max. :12.127
rad tax ptratio black lstat
Min. : 1.000 Min. :187.0 Min. :12.60 Min. : 0.32 Min. : 1.73
1st Qu.: 4.000 1st Qu.:279.0 1st Qu.:17.40 1st Qu.:375.38 1st Qu.: 6.95
Median : 5.000 Median :330.0 Median :19.05 Median :391.44 Median :11.36
Mean : 9.549 Mean :408.2 Mean :18.46 Mean :356.67 Mean :12.65
3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:396.23 3rd Qu.:16.95
Max. :24.000 Max. :711.0 Max. :22.00 Max. :396.90 Max. :37.97
medv

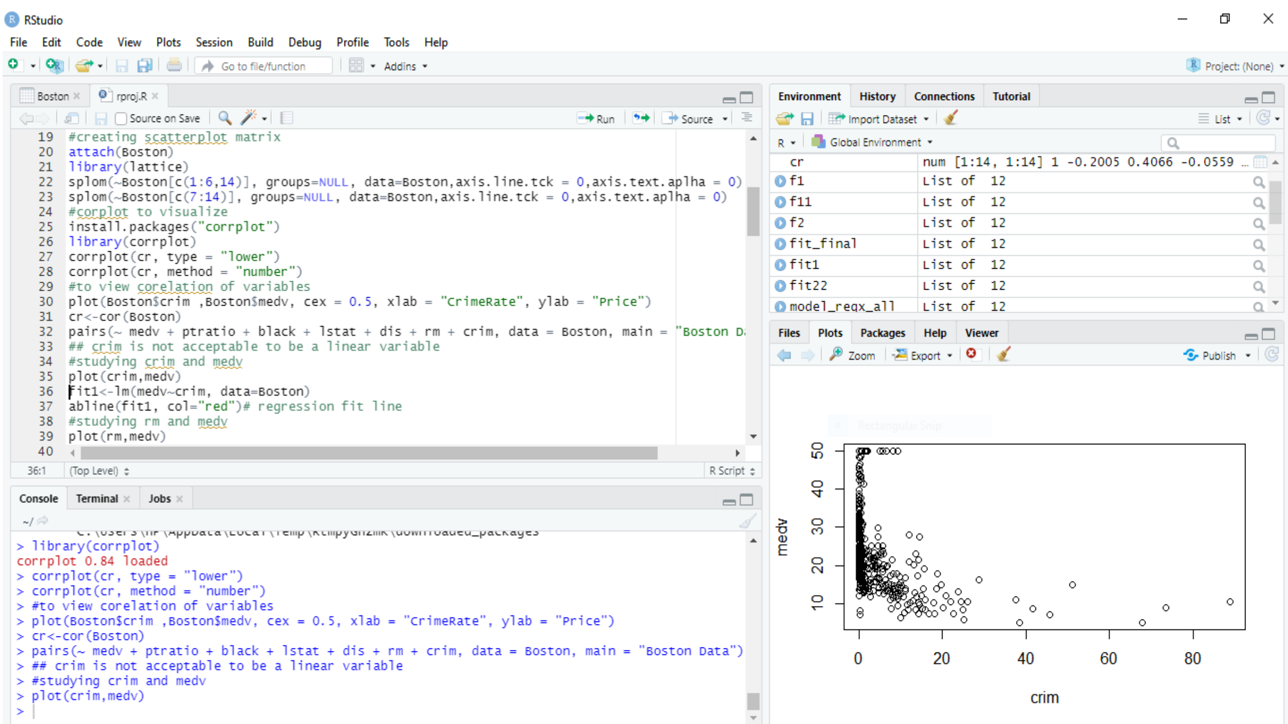


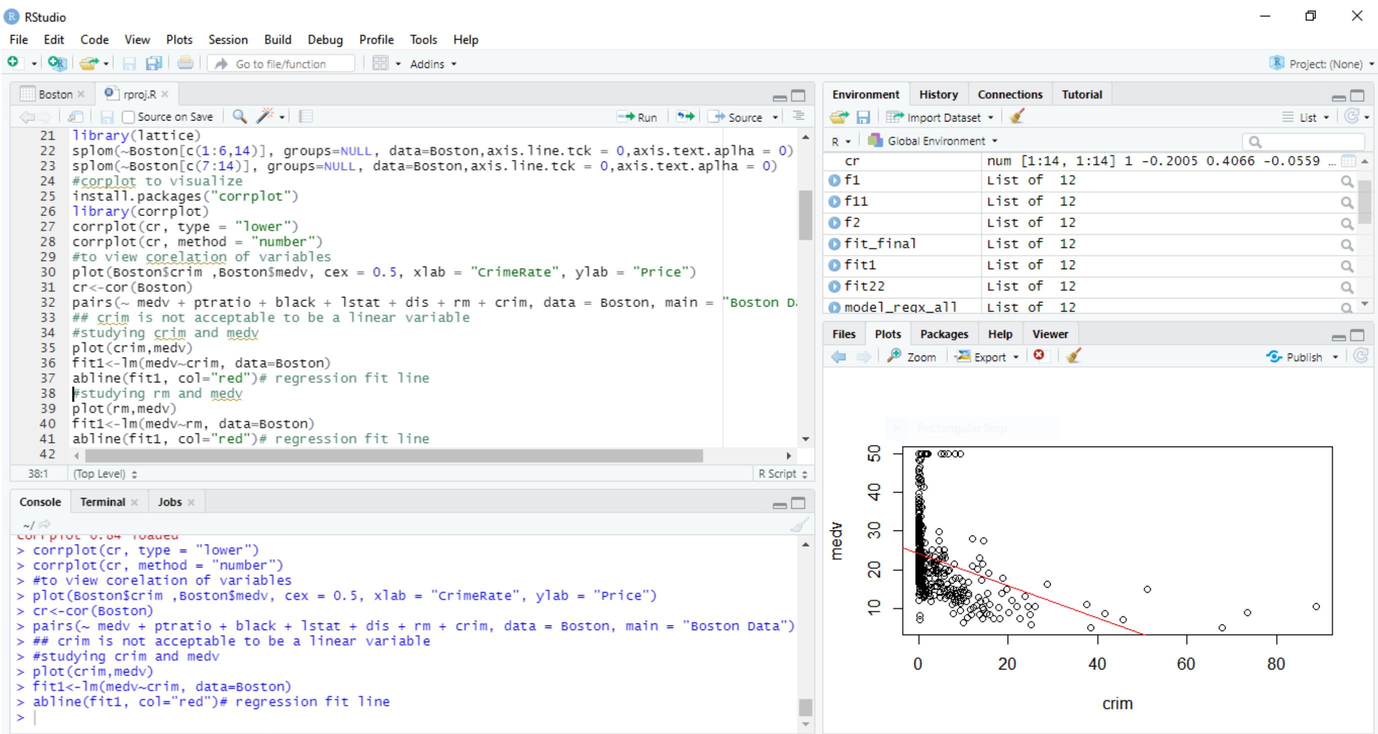
#compare actual values and prediction
plot(testing_data$medv, type = "l", lty = 1.8, col = "green")
lines(predic_selected,type = "l", col = "blue")
#since rmse value is still high we need to optimize the model
f1=lm(medv~lstat +I(lstat^2),Boston)
summary(fit1)
attach(Boston)
f11=lm(medv~poly(lstat,4))
plot(medv~lstat)
points(lstat,fitted(f11),col="blue",pch=20)
f2=lm(medv~rm +I(rm^2),Boston)
summary(f2)
attach(Boston)
fit22=lm(medv~poly(rm,4))
plot(medv~rm)
points(rm,fitted(fit22),col="blue",pch=20)
#building final model
fit_final=lm(medv~lstat+crim+rm+dis+black+chas+nox+rad+tax+ptratio+I(lstat^2)+I(rm^2))
summary(fit_final)
#prediction
predic_fit_final<-predict(fit_final, testing_data)
Call:
lm(formula = medv ~ rm, data = training_data)
Residuals:
Min 1Q Median 3Q Max
-22.979 -3.111 0.102 3.032 39.099
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -33.423 3.303 -10.12 <2e-16 ***
rm 8.918 0.519 17.18 <2e-16 ***

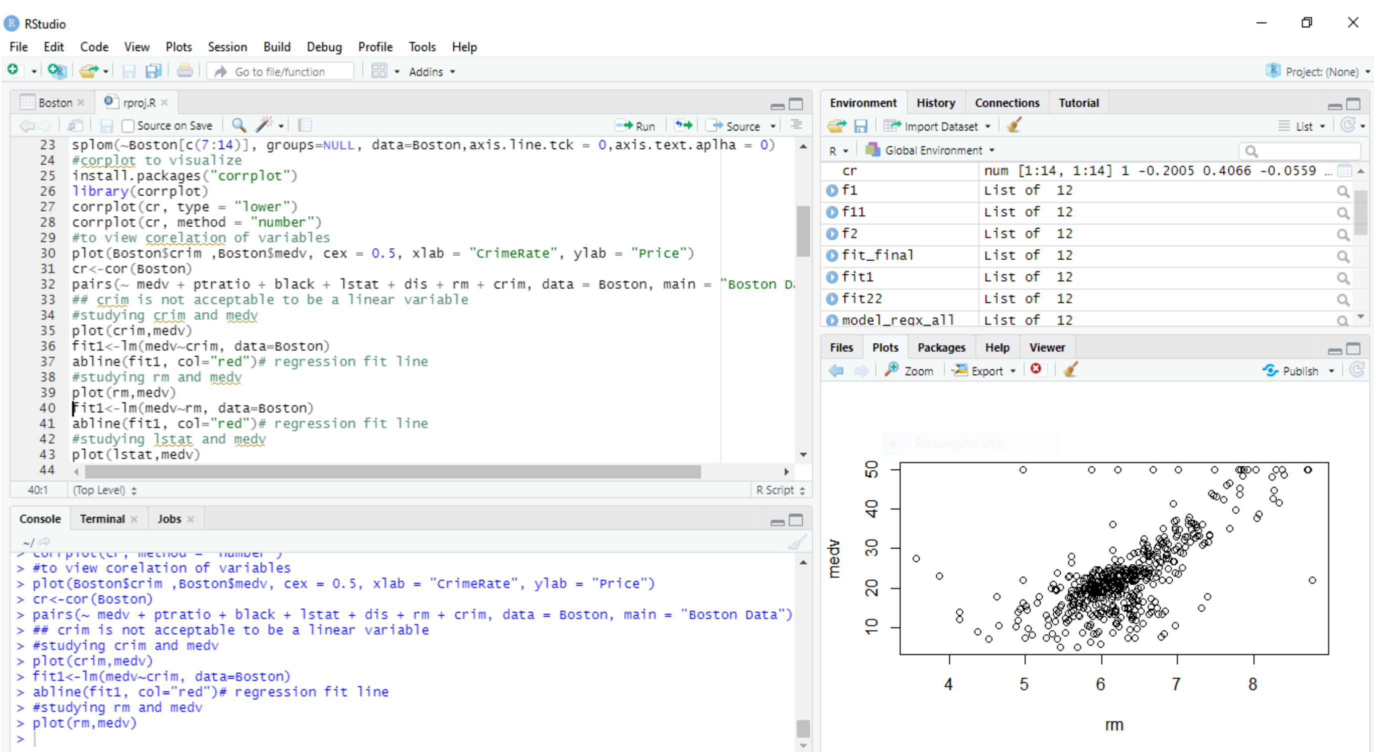

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Boston
rproj.R
Source on Save
Run
Source
summary(Boston)
set.seed(2)
install.packages("caTools")
library(caTools)
#split using 70 percent
split<-sample.split(Boston$medv ,SplitRatio = 0.7)
split
training_data<-subset(Boston,split=="TRUE")
testing_data<-subset(Boston,split=="FALSE")
###Exploratory Data Analysis###
#creating scatterplot matrix
attach(Boston)
library(lattice)
splom(~Boston[c(1:6,14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
splom(~Boston[c(7:14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
#corplot to visualize
install.packages("corrplot")
library(corrplot)
corrplot(cr, type = "lower")
corrplot(cr, method = "number")
#to view corelation of variables
26:1
(Top Level)
R Script
Console
Terminal
Jobs
[463] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE FALSE TRUE TRUE TRUE
[477] FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE
[491] TRUE TRUE TRUE FALSE TRUE TRUE FALSE FALSE FALSE TRUE TRUE FALSE FALSE TRUE
[505] FALSE TRUE
> training_data<-subset(Boston,split=="TRUE")
> testing_data<-subset(Boston,split=="FALSE")
> ###Exploratory Data Analysis###
> #creating scatterplot matrix
> attach(Boston)
> library(lattice)
> splom(~Boston[c(1:6,14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
> splom(~Boston[c(7:14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
Environment
History
Connections
Tutorial
Import Dataset
List
Global Environment
testing_data 139 obs. of 14 variables
training_data 367 obs. of 14 variables
Values
predic_fit_final Named num [1:139] 24.5 32.3 30.6 17.3 13.8 ...
predic_lstat Named num [1:139] 26.2 32.49 30.07 16.06 5.1...
predic_rm Named num [1:139] 23.8 29 30.3 21.6 16.8 ...
predic_selected Named num [1:139] 24.81 28.49 27.59 18.21 9....
rmse 0.334701375621911
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Scatter Plot Matrix

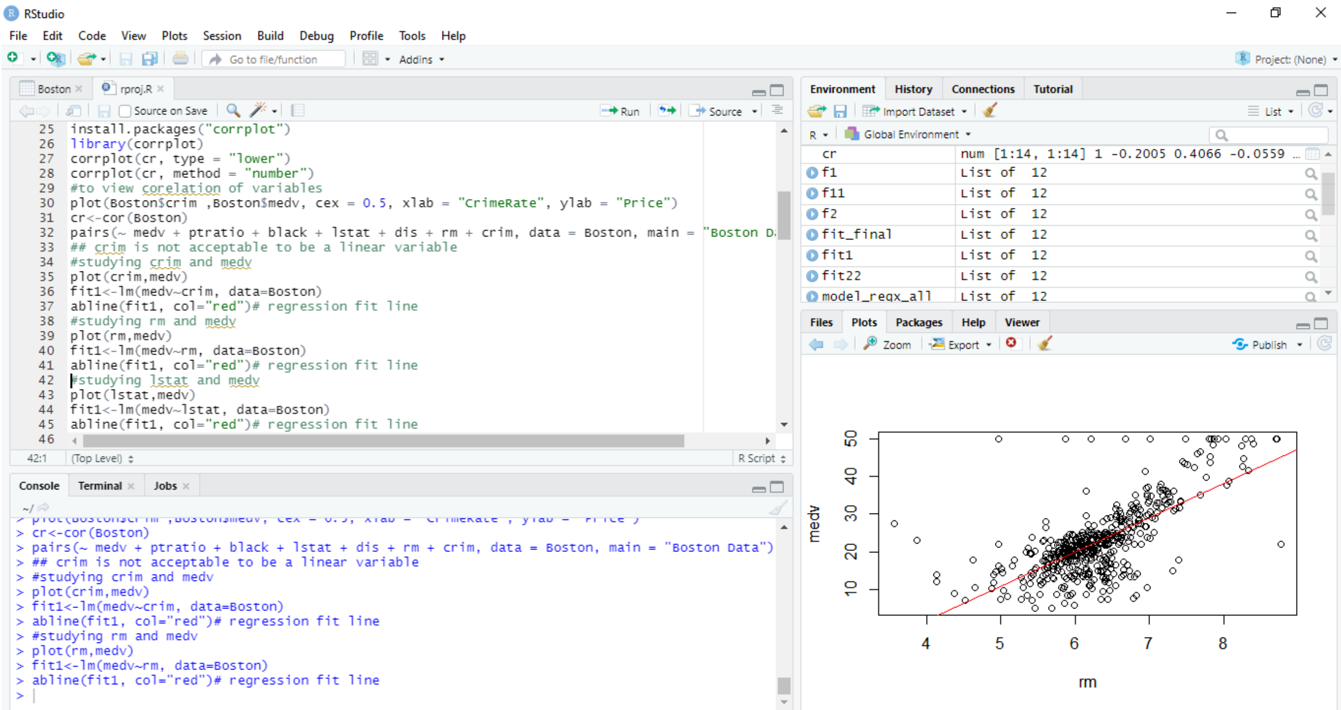

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Boston
rproj.R
Source on Save
Run
Source
split<-sample.split(Boston$medv ,SplitRatio = 0.7)
split
training_data<-subset(Boston,split=="TRUE")
testing_data<-subset(Boston,split=="FALSE")
###Exploratory Data Analysis###
#creating scatterplot matrix
attach(Boston)
library(lattice)
splom(~Boston[c(1:6,14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
splom(~Boston[c(7:14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
#corplot to visualize
install.packages("corrplot")
library(corrplot)
corrplot(cr, type = "lower")
corrplot(cr, method = "number")
#to view corelation of variables
plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
cr<-cor(Boston)
pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston D
crim is not acceptable to be a linear variable
#studying crim and medv
31:1
(Top Level)
R Script
Console
Terminal
Jobs
package ‘corrplot’ successfully unpacked and MD5 sums checked
The downloaded binary packages are in
C:\Users\HP\AppData\Local\Temp\RtmpyGHzmK\downloaded_packages
> library(corrplot)
corrplot 0.84 loaded
> corrplot(cr, type = "lower")
> corrplot(cr, method = "number")
> #to view corelation of variables
> plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
>
Environment
History
Connections
Tutorial
Import Dataset
List
Global Environment
testing_data 139 obs. of 14 variables
training_data 367 obs. of 14 variables
Values
predic_fit_final Named num [1:139] 24.5 32.3 30.6 17.3 13.8 ...
predic_lstat Named num [1:139] 26.2 32.49 30.07 16.06 5.1...
predic_rm Named num [1:139] 23.8 29 30.3 21.6 16.8 ...
predic_selected Named num [1:139] 24.81 28.49 27.59 18.21 9....
rmse 0.334701375621911
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Price
10
20
30
40
50
0
20
40
60
80
CrimeRate

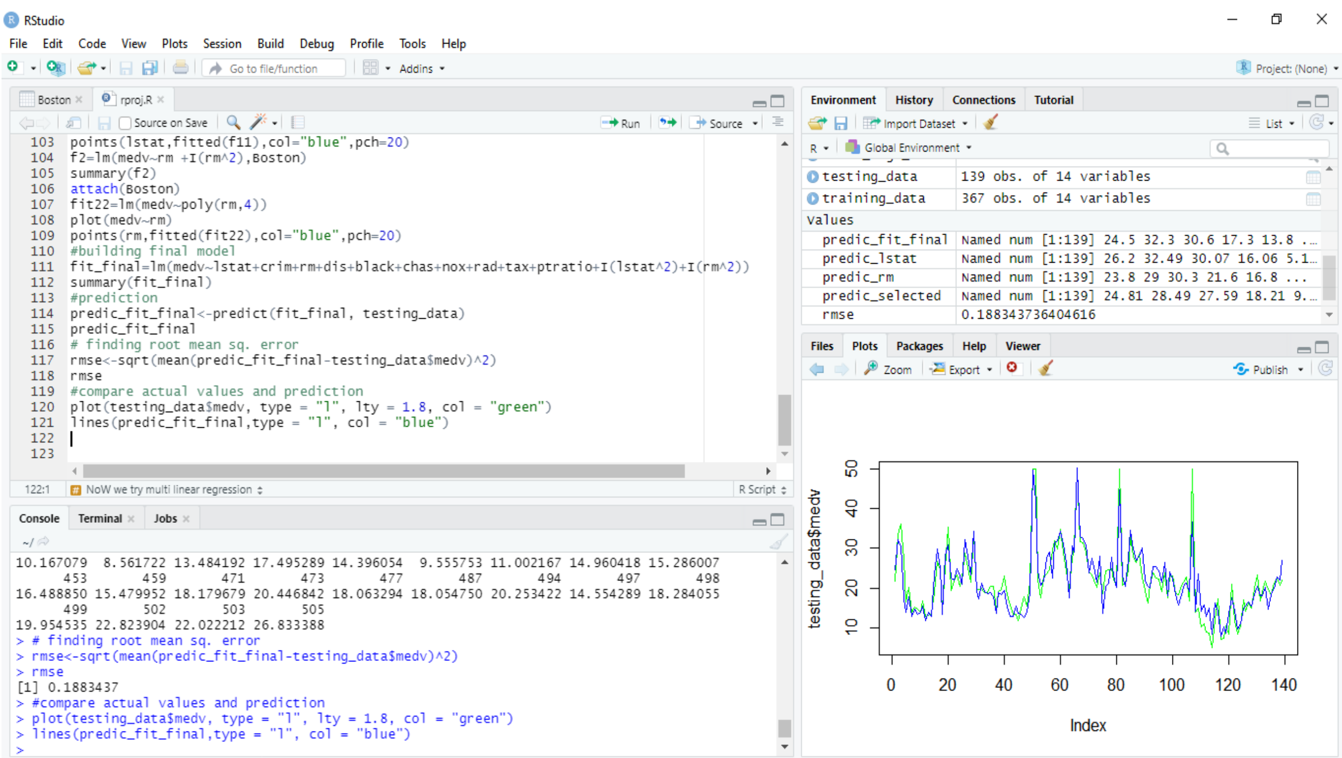

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Boston
rproj.R
Source on Save
Run
Source
training_data<-subset(Boston,split=="TRUE")
testing_data<-subset(Boston,split=="FALSE")
###Exploratory Data Analysis###
#creating scatterplot matrix
attach(Boston)
library(lattice)
splom(~Boston[c(1:6,14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
splom(~Boston[c(7:14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
#corplot to visualize
install.packages("corrplot")
library(corrplot)
corrplot(cr, type = "lower")
corrplot(cr, method = "number")
#to view corelation of variables
plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
cr<-cor(Boston)
pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston D
crim is not acceptable to be a linear variable
#studying crim and medv
plot(crim,medv)
fit1<-lm(medv~crim, data=Boston)
33:1
(Top Level)
R Script
Console
Terminal
Jobs
The downloaded binary packages are in
C:\Users\HP\AppData\Local\Temp\RtmpyGHzmK\downloaded_packages
> library(corrplot)
corrplot 0.84 loaded
> corrplot(cr, type = "lower")
> corrplot(cr, method = "number")
> #to view corelation of variables
> plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
> cr<-cor(Boston)
> pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston Data")
Environment
History
Connections
Tutorial
Import Dataset
List
Global Environment
cr num [1:14, 1:14] 1 -0.2005 0.4066 -0.0559 ...
f1 List of 12
f11 List of 12
f2 List of 12
fit_final List of 12
fit1 List of 12
fit22 List of 12
model_regx_all List of 12
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Boston Data
medv
ptratio
black
lstat
dis
rm
crim

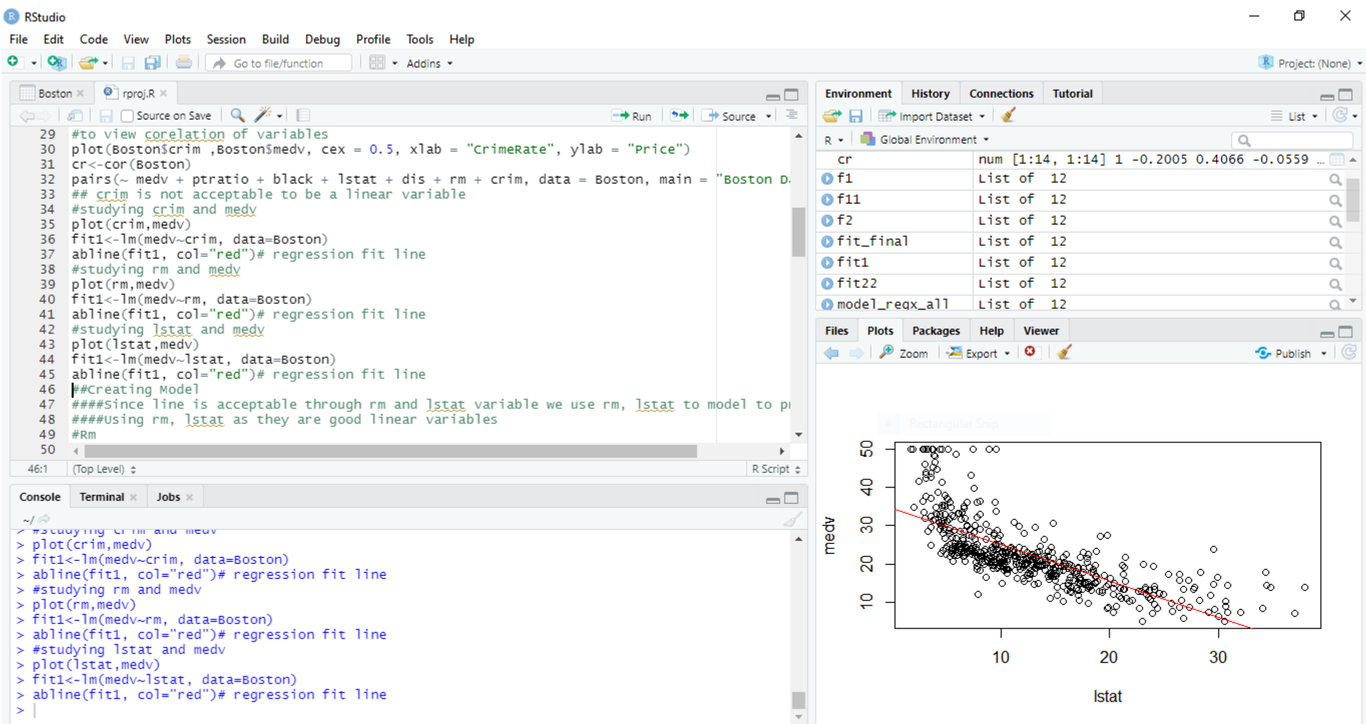

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Boston
rproj.R
Source on Save
Run
Source
#creating scatterplot matrix
attach(Boston)
library(lattice)
splom(~Boston[c(1:6,14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
splom(~Boston[c(7:14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
#corplot to visualize
install.packages("corrplot")
library(corrplot)
corrplot(cr, type = "lower")
corrplot(cr, method = "number")
#to view corelation of variables
plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
cr<-cor(Boston)
pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston D
crim is not acceptable to be a linear variable
#studying crim and medv
plot(crim,medv)
fit1<-lm(medv~crim, data=Boston)
abline(fit1, col="red")# regression fit line
#studying rm and medv
plot(rm,medv)
36:1
(Top Level)
R Script
Console
Terminal
Jobs
> library(corrplot)
corrplot 0.84 loaded
> corrplot(cr, type = "lower")
> corrplot(cr, method = "number")
> #to view corelation of variables
> plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
> cr<-cor(Boston)
> pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston Data")
> ## crim is not acceptable to be a linear variable
> #studying crim and medv
> plot(crim,medv)
Environment
History
Connections
Tutorial
Import Dataset
List
Global Environment
cr num [1:14, 1:14] 1 -0.2005 0.4066 -0.0559 ...
f1 List of 12
f11 List of 12
f2 List of 12
fit_final List of 12
fit1 List of 12
fit22 List of 12
model_regx_all List of 12
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
medv
crim
0
20
40
60
80
10
20
30
40
50

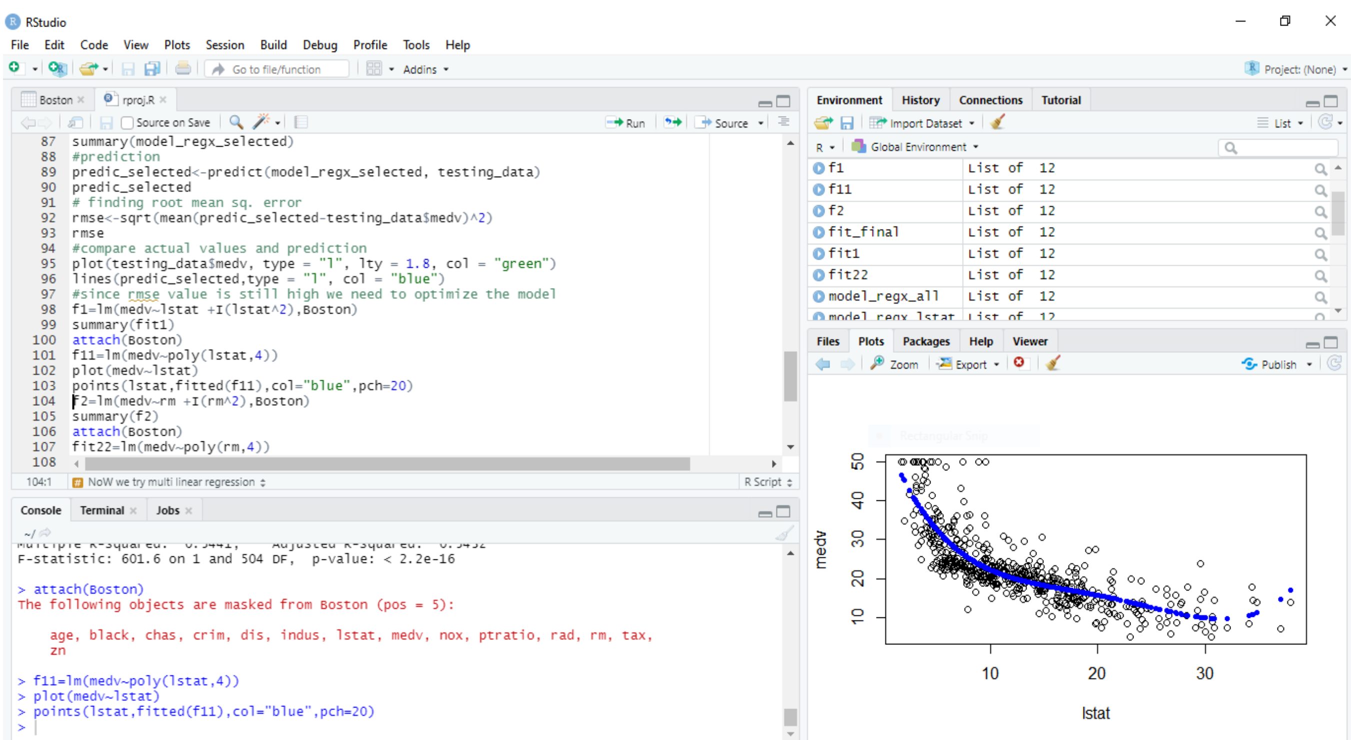

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Boston
rproj.R
Source on Save
Run
Source
21 library(lattice)
22 splom(~Boston[c(1:6,14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
23 splom(~Boston[c(7:14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
24 #corplot to visualize
25 install.packages("corrplot")
26 library(corrplot)
27 corrplot(cr, type = "lower")
28 corrplot(cr, method = "number")
29 #to view corelation of variables
30 plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
31 cr<-cor(Boston)
32 pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston D
33 ## crim is not acceptable to be a linear variable
34 #studying crim and medv
35 plot(crim,medv)
36 fit1<-lm(medv~crim, data=Boston)
37 abline(fit1, col="red")# regression fit line
38 #studying rm and medv
39 plot(rm,medv)
40 fit1<-lm(medv~rm, data=Boston)
41 abline(fit1, col="red")# regression fit line
Console Terminal Jobs
> corrplot(cr, type = "lower")
> corrplot(cr, method = "number")
> #to view corelation of variables
> plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
> cr<-cor(Boston)
> pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston Data")
> ## crim is not acceptable to be a linear variable
> #studying crim and medv
> plot(crim,medv)
> fit1<-lm(medv~crim, data=Boston)
> abline(fit1, col="red")# regression fit line
Environment History Connections Tutorial
Import Dataset
Global Environment
cr num [1:14, 1:14] 1 -0.2005 0.4066 -0.0559 ...
f1 List of 12
f11 List of 12
f2 List of 12
fit_final List of 12
fit1 List of 12
fit22 List of 12
model_regx_all List of 12
Files Plots Packages Help Viewer
Zoom Export Publish
medv
crim
0 20 40 60 80
10 20 30 40 50

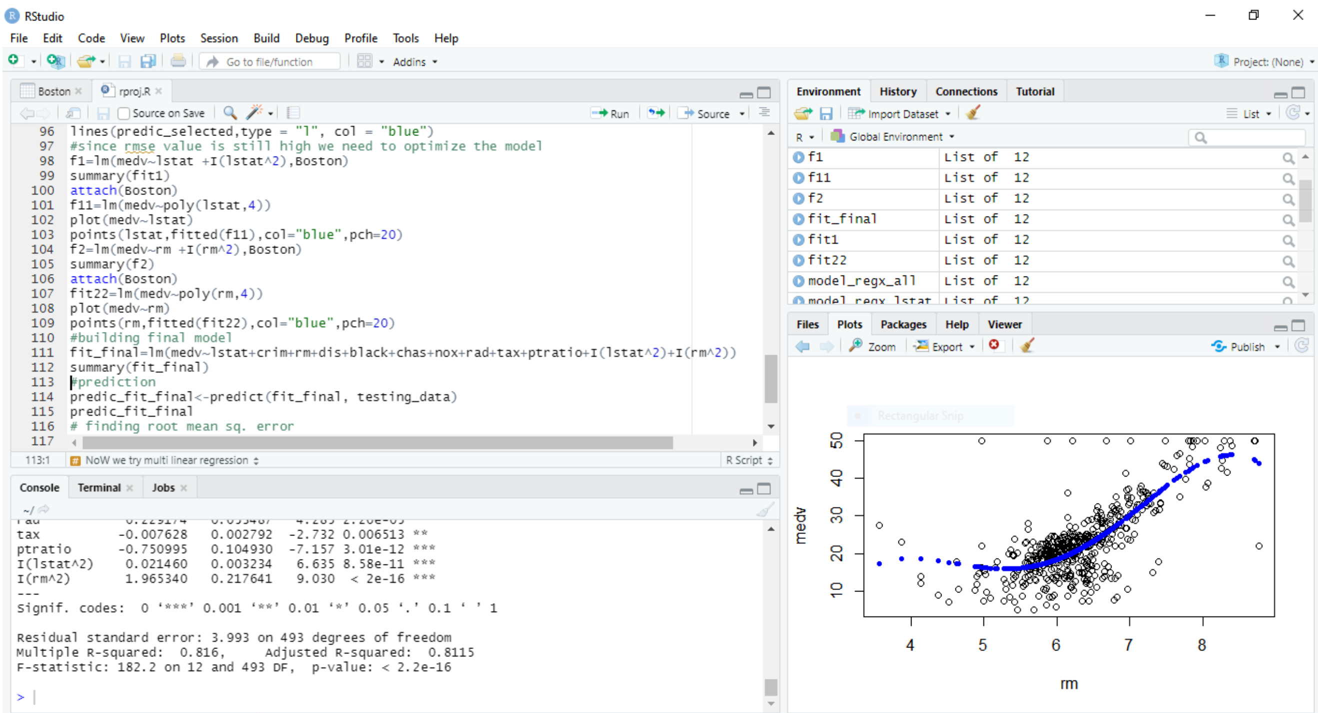

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Boston
rproj.R
Source on Save
Run
Source
23 splom(~Boston[c(7:14)], groups=NULL, data=Boston,axis.line.tck = 0,axis.text.aplha = 0)
24 #corplot to visualize
25 install.packages("corrplot")
26 library(corrplot)
27 corrplot(cr, type = "lower")
28 corrplot(cr, method = "number")
29 #to view corelation of variables
30 plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
31 cr<-cor(Boston)
32 pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston D
33 ## crim is not acceptable to be a linear variable
34 #studying crim and medv
35 plot(crim,medv)
36 fit1<-lm(medv~crim, data=Boston)
37 abline(fit1, col="red")# regression fit line
38 #studying rm and medv
39 plot(rm,medv)
40 fit1<-lm(medv~rm, data=Boston)
41 abline(fit1, col="red")# regression fit line
42 #studying lstat and medv
43 plot(lstat,medv)
Console Terminal Jobs
> #to view corelation of variables
> plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
> cr<-cor(Boston)
> pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston Data")
> ## crim is not acceptable to be a linear variable
> #studying crim and medv
> plot(crim,medv)
> fit1<-lm(medv~crim, data=Boston)
> abline(fit1, col="red")# regression fit line
> #studying rm and medv
> plot(rm,medv)
Environment History Connections Tutorial
Import Dataset
Global Environment
cr num [1:14, 1:14] 1 -0.2005 0.4066 -0.0559 ...
f1 List of 12
f11 List of 12
f2 List of 12
fit_final List of 12
fit1 List of 12
fit22 List of 12
model_regx_all List of 12
Files Plots Packages Help Viewer
Zoom Export Publish
medv
rm
4 5 6 7 8
10 20 30 40 50


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Boston
rproj.R
install.packages("corrplot")
library(corrplot)
corrplot(cr, type = "lower")
corrplot(cr, method = "number")
#to view corelation of variables
plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
cr<-cor(Boston)
pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston D
crim is not acceptable to be a linear variable
#studying crim and medv
plot(crim,medv)
fit1<-lm(medv~crim, data=Boston)
abline(fit1, col="red")# regression fit line
#studying rm and medv
plot(rm,medv)
fit1<-lm(medv~rm, data=Boston)
abline(fit1, col="red")# regression fit line
#studying lstat and medv
plot(lstat,medv)
fit1<-lm(medv~lstat, data=Boston)
abline(fit1, col="red")# regression fit line
Console Terminal Jobs
> cr<-cor(Boston)
> pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston Data")
> ## crim is not acceptable to be a linear variable
> #studying crim and medv
> plot(crim,medv)
> fit1<-lm(medv~crim, data=Boston)
> abline(fit1, col="red")# regression fit line
> #studying rm and medv
> plot(rm,medv)
> fit1<-lm(medv~rm, data=Boston)
> abline(fit1, col="red")# regression fit line
Environment History Connections Tutorial
cr num [1:14, 1:14] 1 -0.2005 0.4066 -0.0559 ...
f1 List of 12
f11 List of 12
f2 List of 12
fit_final List of 12
fit1 List of 12
fit22 List of 12
model_regx_all List of 12
Files Plots Packages Help Viewer
medv
rm


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Boston
rproj.R
points(lstat,fitted(f11),col="blue",pch=20)
f2=lm(medv~rm +I(rm^2),Boston)
summary(f2)
attach(Boston)
fit22=lm(medv~poly(rm,4))
plot(medv~rm)
points(rm,fitted(fit22),col="blue",pch=20)
#building final model
fit_final=lm(medv~lstat+crim+rm+dis+black+chas+nox+rad+tax+ptratio+I(lstat^2)+I(rm^2))
summary(fit_final)
#prediction
predic_fit_final<-predict(fit_final, testing_data)
predic_fit_final
# finding root mean sq. error
rmse<-sqrt(mean(predic_fit_final-testing_data$medv)^2)
rmse
#compare actual values and prediction
plot(testing_data$medv, type = "l", lty = 1.8, col = "green")
lines(predic_fit_final,type = "l", col = "blue")
NoW we try multi linear regression
Console Terminal Jobs
10.167079 8.561722 13.484192 17.495289 14.396054 9.555753 11.002167 14.960418 15.286007
453 459 471 473 477 487 494 497 498
16.488850 15.479952 18.179679 20.446842 18.063294 18.054750 20.253422 14.554289 18.284055
499 502 503 505
19.954535 22.823904 22.022212 26.833388
> # finding root mean sq. error
> rmse<-sqrt(mean(predic_fit_final-testing_data$medv)^2)
> rmse
[1] 0.1883437
> #compare actual values and prediction
> plot(testing_data$medv, type = "l", lty = 1.8, col = "green")
> lines(predic_fit_final,type = "l", col = "blue")
Environment History Connections Tutorial
testing_data 139 obs. of 14 variables
training_data 367 obs. of 14 variables
Values
predic_fit_final Named num [1:139] 24.5 32.3 30.6 17.3 13.8 ...
predic_lstat Named num [1:139] 26.2 32.49 30.07 16.06 5.1...
predic_rm Named num [1:139] 23.8 29 30.3 21.6 16.8 ...
predic_selected Named num [1:139] 24.81 28.49 27.59 18.21 9....
rmse 0.188343736404616
Files Plots Packages Help Viewer
testing_data$medv
Index


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
#to view corelation of variables
plot(Boston$crim ,Boston$medv, cex = 0.5, xlab = "CrimeRate", ylab = "Price")
cr<-cor(Boston)
pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston D
crim is not acceptable to be a linear variable
#studying crim and medv
plot(crim,medv)
fit1<-lm(medv~crim, data=Boston)
abline(fit1, col="red")# regression fit line
#studying rm and medv
plot(rm,medv)
fit1<-lm(medv~rm, data=Boston)
abline(fit1, col="red")# regression fit line
#studying lstat and medv
plot(lstat,medv)
fit1<-lm(medv~lstat, data=Boston)
abline(fit1, col="red")# regression fit line
##Creating Model
####Since line is acceptable through rm and lstat variable we use rm, lstat to model to p
####Using rm, lstat as they are good linear variables
#Rm
cr num [1:14, 1:14] 1 -0.2005 0.4066 -0.0559 ...
f1 List of 12
f11 List of 12
f2 List of 12
fit_final List of 12
fit1 List of 12
fit22 List of 12
model_regx_all List of 12
medv
lstat


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
summary(model_regx_selected)
#prediction
predic_selected<-predict(model_regx_selected, testing_data)
predic_selected
# finding root mean sq. error
rmse<-sqrt(mean(predic_selected-testing_data$medv)^2)
rmse
#compare actual values and prediction
plot(testing_data$medv, type = "l", lty = 1.8, col = "green")
lines(predic_selected,type = "l", col = "blue")
#since rmse value is still high we need to optimize the model
f1=lm(medv~lstat +I(lstat^2),Boston)
summary(f1)
attach(Boston)
f11=lm(medv~poly(lstat,4))
plot(medv~lstat)
points(lstat,fitted(f11),col="blue",pch=20)
f2=lm(medv~rm +I(rm^2),Boston)
summary(f2)
attach(Boston)
fit22=lm(medv~poly(rm,4))
F-statistic: 601.6 on 1 and 504 DF, p-value: < 2.2e-16
> attach(Boston)
The following objects are masked from Boston (pos = 5):
age, black, chas, crim, dis, indus, lstat, medv, nox, ptratio, rad, rm, tax, zn
> f11=lm(medv~poly(lstat,4))
> plot(medv~lstat)
> points(lstat,fitted(f11),col="blue",pch=20)
medv
lstat



CONCLUSION:

By using the Boston Housing Price data we got the outcome, which was the visualization and because of the data mining algorithm we got an easier classification of the data.

BIBLIOGRAPHY

https://www.r-graph-gallery.com/

http://www.sthda.com/english/wiki/r-basics-quick-and-easy

http://www.tutorialpoint.com/r