Modern Applied Regression Methods Assignment 1 - Answer Key

June 2025

Question & Answer

1. [2+2+2 = 6] For each of the following cases, explain with proper reasoning whether a flexible or an inflexible statistical learning procedure is to be preferred.

- **i. The true relationship between the response and predictors is highly non-linear.**

  Reasoning: Flexible statistical learning methods are more adapted to non-linear relationships than inflexible methods. The flexible method has better options to approximate the real distributio.

- **ii. The sample size n is extremely large while the number of predictors p is small.**

  Reasoning: In this situation the performance of a flexible statistical learning method would be better than that of an inflexible one, due to the fact that the high number of samples would avoid overfitting and therefore it would be close to the real distribution.

- **iii. The sample size n is small while the number of predictors p is extremely large.**

  Reasoning: In this case, the inflexible model performs better, because the flexible methods would try to follow the observations (which are few) too closely, which could result in finding relationships that do not exist or that, in this small sample, happened to be only by unaccountable factors (a.k.a. irreducible errors).

2. [3+3+3 = 9] Explain with proper reasoning whether each of the following scenarios represent a classification or regression problem and also whether those relate to inference or prediction.

- **i. We are interested in predicting the gold prices in Indian markets for the week starting 1st July based on the supply and demand of gold, USD/Rupee exchange rate, oil prices and inflation. Towards that end, we create a dataset containing weekly averages of the above variables for January to June 2025.**

  Reasoning: This is a **regression** problem because the response variable (gold price) is continuous. The goal is **prediction**, since we are using past data to predict future values.This is Regression and prediction problem.

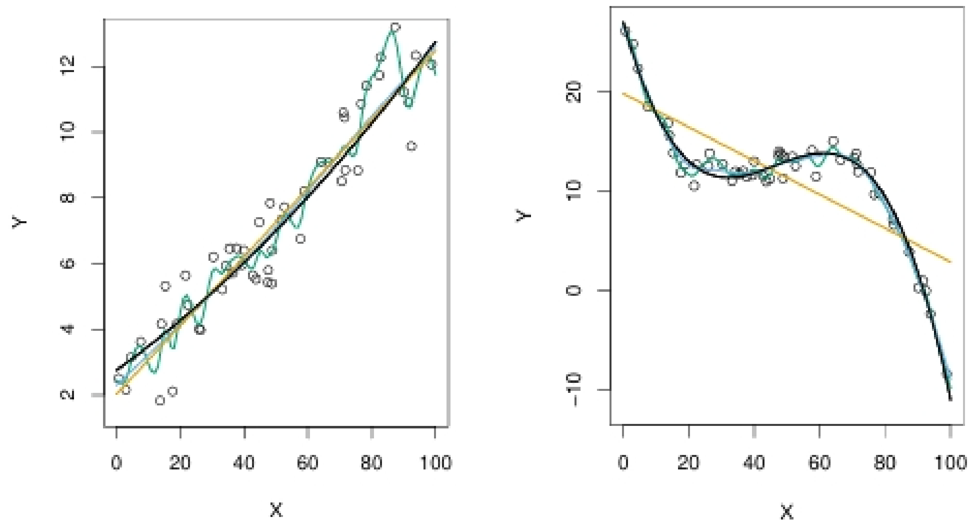
- **ii. For each of the top 300 Indian firms, data is collected on their domain of operation, yearly revenue, number of employees and CEO salary. Intent is on understanding the factors that affect CEO salary.**

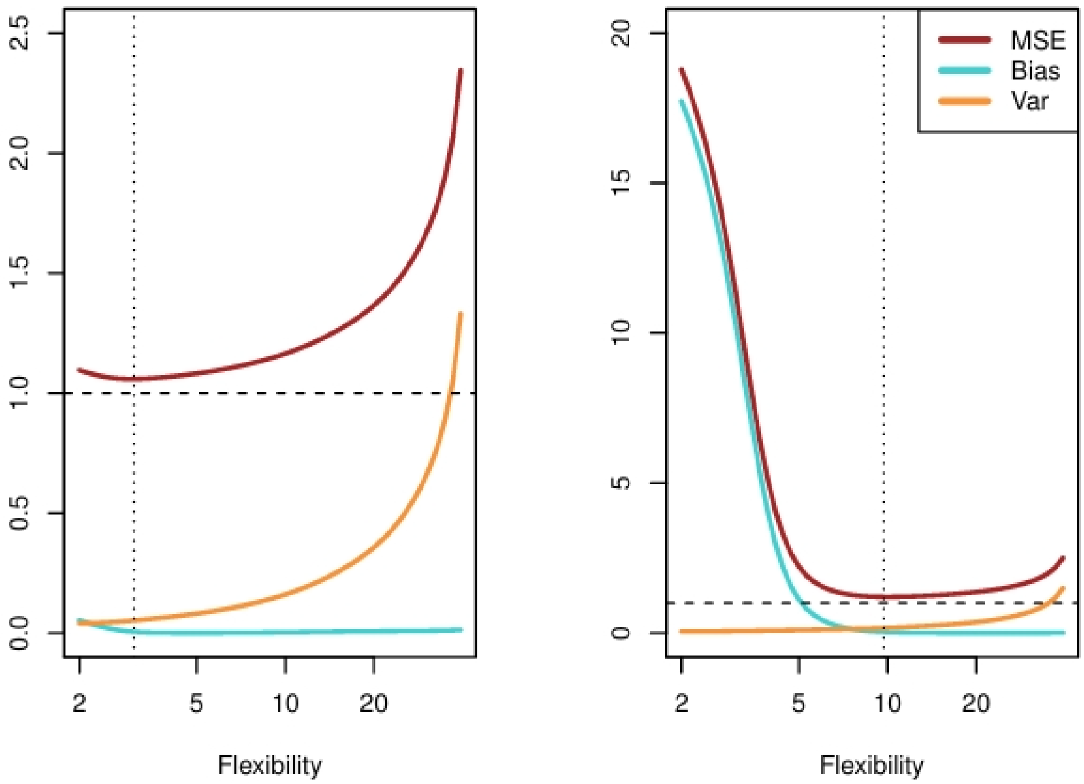
  Reasoning: CEO salary (the response variable) is a continuous numerical variable (regression). We are interested in understanding the relationship between the factors that affects CEO salary (domain of operation, yearly revenue and number of employees) and the CEO salary (inference). This is a regression/inference problem.

- **iii. LIC is considering launching a new scheme and wish to know whether it will be a success or a failure. Accordingly, data is collected on 25 comparable schemes which are in operation. For each scheme, data pertains to whether it was a success or failure, premium amount, maturity period and amount, marketing budget and other related variables.**

  Reasoning: New scheme's success or failure (response variable) is categorical variable (classification). We are interested in knowing whether the new scheme will be a success a failure (prediction) by comparing it with 25 other schemes for different variables (premium amount, maturity period and amount, marketing budget and other related variables) This is a classification/prediction problem.

3. [5+5=10] **For the following figures discussed in class, provide a hand-drawn sketch of the squared bias, variance, irreducible error and test MSE as a function of the flexibility of the statistical learning method used for estimating the true (black) curve. Make sure to label each curve.**


Y
X
Y
X


Flexibility
MSE
Bias
Var
Flexibility

4.[43] The following questions relate to the Boston housing dataset. First load the Boston data set from the MASS library in R as follows and answer the questions below:

```
library(MASS)
head(Boston)

##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7

attach(Boston)
```

(i) [1+1+1+3 = 6] Learn about the variables in the dataset from the data description file:
`> ?Boston`

- **Number of Variables:** 14
- **Sample Size:** 506
- **What Do the Rows Represent?** Each row represents a town or suburb in the Boston area.
- **Variable Description:**
  1. `crim` : per capita crime rate by town
  2. `zn` : proportion of residential land zoned for lots over 25,000 sq.ft
  3. `indus` : proportion of non-retail business acres per town
  4. `chas` : Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
  5. `nox` : nitrogen oxides concentration (parts per 10 million)
  6. `rm` : average number of rooms per dwelling
  7. `age` : proportion of owner-occupied units built prior to 1940

8. `dis` : weighted mean of distances to five Boston employment centres
9. `rad` : index of accessibility to radial highways
10. `tax` : full-value property-tax rate per $10,000
11. `ptratio` : pupil-teacher ratio by town
12. `black` : $1000(B_k - 0.63)^2$, where B_k is the proportion of Black residents by town
13. `lstat` : percentage of lower status population
14. `medv` : median value of owner-occupied homes in $1000s

- **Variable Types:**
  - **Categorical:** `chas` (binary), `rad` (treated as discrete index)
  - **Continuous:** All remaining variables

(ii) [10] Evaluate and report the mean, median, standard deviation and range of each of the quantitative vari ables.

```
# Load dataset
library(MASS)
data(Boston)
summary_stats <- function(x) {
  c(
    Mean = mean(x),
    Median = median(x),
    SD = sd(x),
    Range = paste0(range(x)[1], " to ", range(x)[2])
  )
}
quantitative_summary <- sapply(Boston, summary_stats)
quantitative_summary <- t(quantitative_summary)
quantitative_summary

##         Mean                 Median    SD                   Range
## crim    "3.61352355731225"   "0.25651" "8.60154510533249"   "0.00632 to 88.9762"
## zn      "11.3636363636364"   "0"       "23.3224529945151"   "0 to 100"
## indus   "11.1367786561265"   "9.69"    "6.86035294089759"   "0.46 to 27.74"
## chas    "0.0691699604743083" "0"       "0.25399404134041"   "0 to 1"
## nox     "0.554695059288538"  "0.538"   "0.115877675667556"  "0.385 to 0.871"
## rm      "6.28463438735178"   "6.2085"  "0.702617143415323"  "3.561 to 8.78"
## age     "68.5749011857708"   "77.5"    "28.1488614069036"   "2.9 to 100"
## dis     "3.79504268774704"   "3.20745" "2.10571012662761"   "1.1296 to 12.1265"
## rad     "9.54940711462451"   "5"       "8.70725938423937"   "1 to 24"
```

```
## tax     "408.237154150198"   "330"     "168.537116054959"  "187 to 711"
## ptratio "18.4555335968379"   "19.05"   "2.16494552371444"  "12.6 to 22"
## black   "356.674031620553"   "391.44"  "91.2948643841578"  "0.32 to 396.9"
## lstat   "12.6530632411067"   "11.36"   "7.14106151134857"  "1.73 to 37.97"
## medv    "22.5328063241107"   "21.2"    "9.19710408737982"  "5 to 50"
```

(iii) [3+3+3=9] Identify three suburbs of Boston which have the highest crime rates, tax rates and pupil-teacher ratios.

```r
# Load required data
library(MASS)
data(Boston)
Boston$Suburb_ID <- 1:nrow(Boston)
top_crime <- Boston[order(-Boston$crim), ][1:3, c("Suburb_ID", "crim")]
top_tax <- Boston[order(-Boston$tax), ][1:3, c("Suburb_ID", "tax")]
top_ptratio <- Boston[order(-Boston$ptratio), ][1:3, c("Suburb_ID", "ptratio")]

# Display results
cat("Top 3 suburbs with highest CRIME rate:\n")

## Top 3 suburbs with highest CRIME rate:

print(top_crime)

##     Suburb_ID    crim
## 381       381 88.9762
## 419       419 73.5341
## 406       406 67.9208

cat("\nTop 3 suburbs with highest TAX rate:\n")

##
## Top 3 suburbs with highest TAX rate:

print(top_tax)

##     Suburb_ID tax
## 489       489 711
## 490       490 711
## 491       491 711

cat("\nTop 3 suburbs with highest PUPIL-TEACHER ratio:\n")

##
## Top 3 suburbs with highest PUPIL-TEACHER ratio:
```

```r
print(top_ptratio)

##     Suburb_ID ptratio
## 355       355    22.0
## 356       356    22.0
## 128       128    21.2
```

(iv) [2] How many of Boston suburbs lie around the Charles river ?

```r
 length(chas[chas==1])

## [1] 35
```

we have 35 suburbs which lie around the Charles River

(v) [2] What is the median pupil-teacher ratio among the Boston towns/-suburbs ?

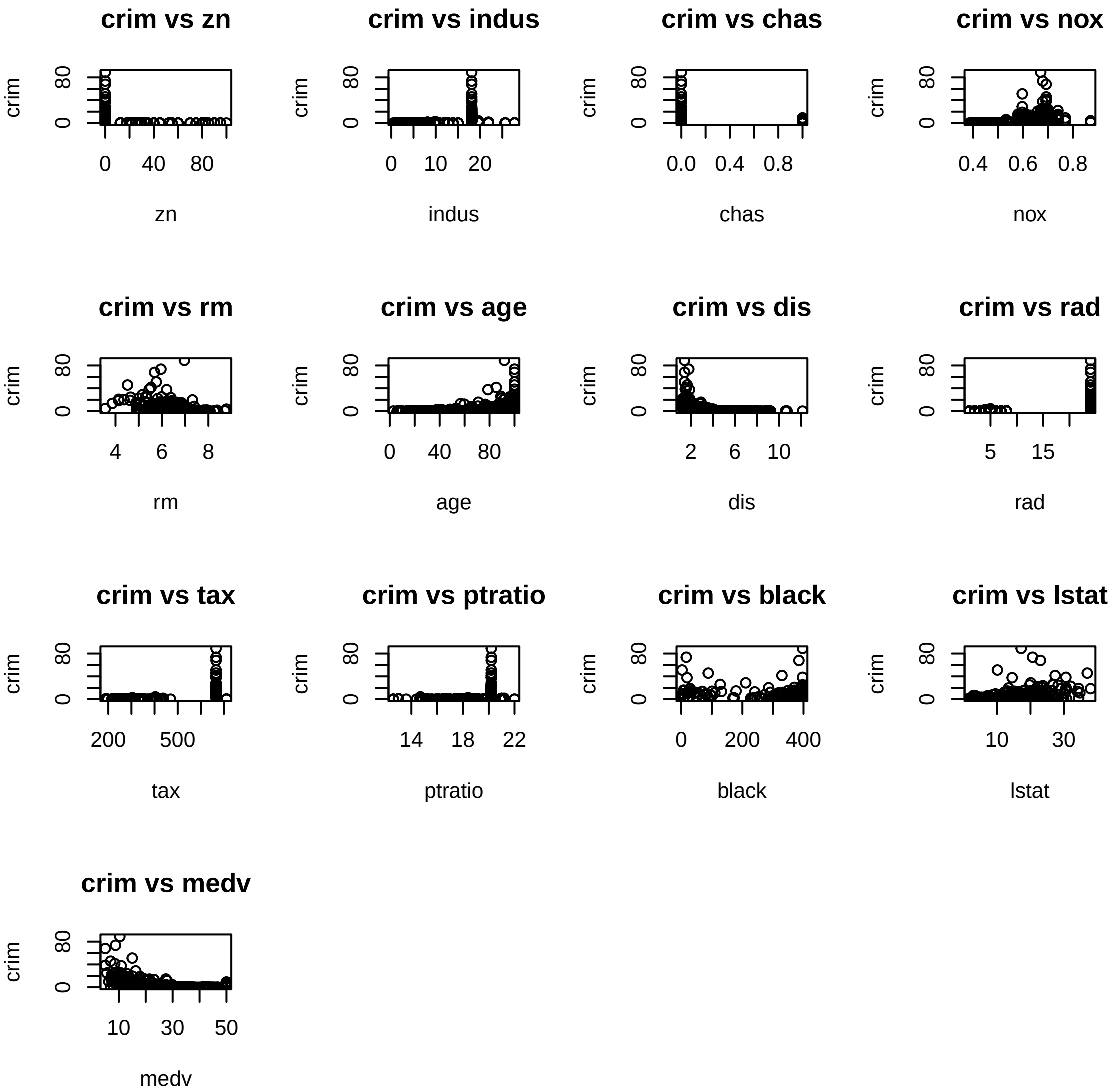
```r
median(ptratio)

## [1] 19.05
```

The median pupil-teacher ratio among all suburbs is 19.

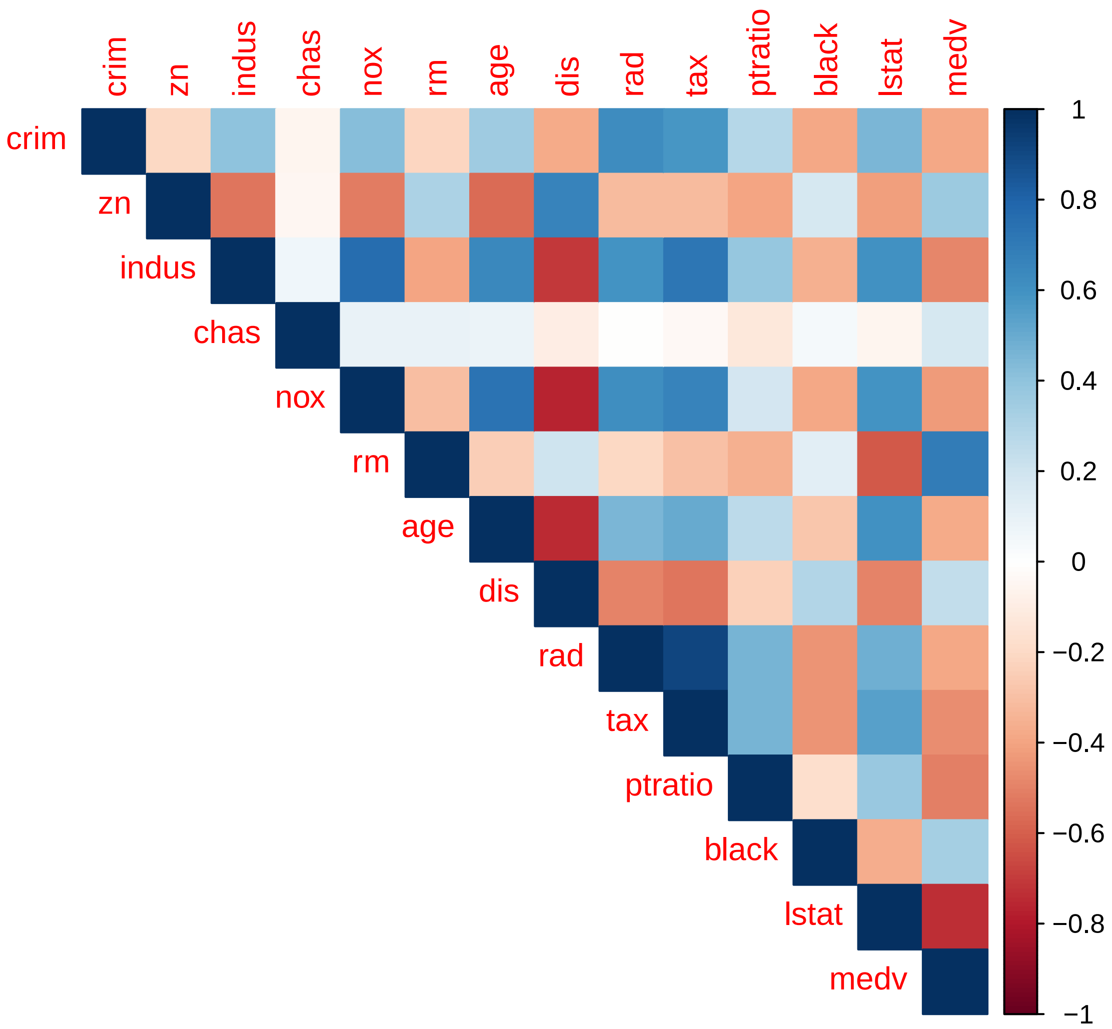
(vi) [5+5+3+2=15] Suppose our interest is on predicting per-capita crime rate based on the other variables. Ac cordingly, explore the association between the predictors themselves and that with the response graphically using scatterplots or any other tools of your choice. Comment on your findings. Which of the variables can be considered as predictors of crime rate ? Justify with reasons.

```r
Boston_clean <- Boston[ , !(names(Boston) %in% "Suburb_ID")]

# Set the layout for 3x4 scatterplots
par(mfrow = c(4, 4))

# Plot 'crim' against each predictor (excluding 'crim' itself)
for (var in names(Boston_clean)[-1]) {
  plot(Boston_clean[[var]], Boston_clean$crim,
       main = paste("crim vs", var),
       xlab = var, ylab = "crim")
}
cor(Boston[ , !(names(Boston) %in% "Suburb_ID")])

##                crim          zn       indus         chas         nox
## crim     1.00000000 -0.20046922  0.40658341 -0.055891582  0.42097171
```

```
## zn      -0.20046922  1.00000000 -0.53382819 -0.042696719 -0.51660371
## indus    0.40658341 -0.53382819  1.00000000  0.062938027  0.76365145
## chas    -0.05589158 -0.04269672  0.06293803  1.000000000  0.09120281
## nox      0.42097171 -0.51660371  0.76365145  0.091202807  1.00000000
## rm      -0.21924670  0.31199059 -0.39167585  0.091251225 -0.30218819
## age      0.35273425 -0.56953734  0.64477851  0.086517774  0.73147010
## dis     -0.37967009  0.66440822 -0.70802699 -0.099175780 -0.76923011
## rad      0.62550515 -0.31194783  0.59512927 -0.007368241  0.61144056
## tax      0.58276431 -0.31456332  0.72076018 -0.035586518  0.66802320
## ptratio  0.28994558 -0.39167855  0.38324756 -0.121515174  0.18893268
## black   -0.38506394  0.17552032 -0.35697654  0.048788485 -0.38005064
## lstat    0.45562148 -0.41299457  0.60379972 -0.053929298  0.59087892
## medv    -0.38830461  0.36044534 -0.48372516  0.175260177 -0.42732077
##                  rm         age         dis          rad         tax    ptratio
## crim    -0.21924670  0.35273425 -0.37967009  0.625505145  0.58276431  0.2899456
## zn       0.31199059 -0.56953734  0.66440822 -0.311947826 -0.31456332 -0.3916785
## indus   -0.39167585  0.64477851 -0.70802699  0.595129275  0.72076018  0.3832476
## chas     0.09125123  0.08651777 -0.09917578 -0.007368241 -0.03558652 -0.1215152
## nox     -0.30218819  0.73147010 -0.76923011  0.611440563  0.66802320  0.1889327
## rm       1.00000000 -0.24026493  0.20524621 -0.209846668 -0.29204783 -0.3555015
## age     -0.24026493  1.00000000 -0.74788054  0.456022452  0.50645559  0.2615150
## dis      0.20524621 -0.74788054  1.00000000 -0.494587930 -0.53443158 -0.2324705
## rad     -0.20984667  0.45602245 -0.49458793  1.000000000  0.91022819  0.4647412
## tax     -0.29204783  0.50645559 -0.53443158  0.910228189  1.00000000  0.4608530
## ptratio -0.35550149  0.26151501 -0.23247054  0.464741179  0.46085304  1.0000000
## black    0.12806864 -0.27353398  0.29151167 -0.444412816 -0.44180801 -0.1773833
## lstat   -0.61380827  0.60233853 -0.49699583  0.488676335  0.54399341  0.3740443
## medv     0.69535995 -0.37695457  0.24992873 -0.381626231 -0.46853593 -0.5077867
##               black      lstat       medv
## crim    -0.38506394  0.4556215 -0.3883046
## zn       0.17552032 -0.4129946  0.3604453
## indus   -0.35697654  0.6037997 -0.4837252
## chas     0.04878848 -0.0539293  0.1752602
## nox     -0.38005064  0.5908789 -0.4273208
## rm       0.12806864 -0.6138083  0.6953599
## age     -0.27353398  0.6023385 -0.3769546
## dis      0.29151167 -0.4969958  0.2499287
## rad     -0.44441282  0.4886763 -0.3816262
## tax     -0.44180801  0.5439934 -0.4685359
## ptratio -0.17738330  0.3740443 -0.5077867
## black    1.00000000 -0.3660869  0.3334608
## lstat   -0.36608690  1.0000000 -0.7376627
## medv     0.33346082 -0.7376627  1.0000000
```




```r
library(corrplot)
corrplot(cor(Boston_clean),method="color",type="upper")
```




```r
##Identify strongest correlation with crim
crim_corr=sort(cor(Boston_clean)[,"crim"],decreasing=TRUE)
print(round(crim_corr,3))
```

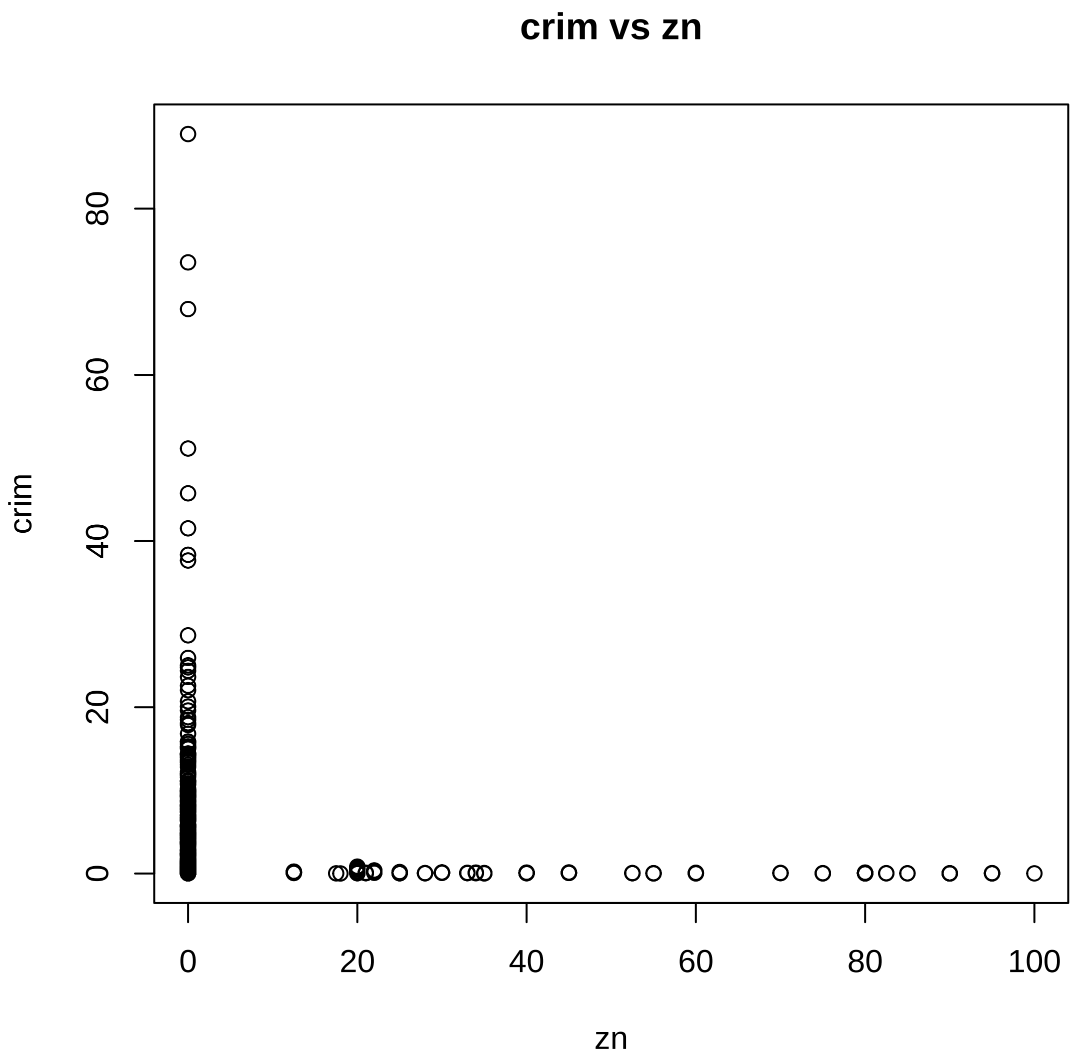
```
##     crim      rad      tax    lstat      nox    indus      age  ptratio     chas       zn
##    1.000    0.626    0.583    0.456    0.421    0.407    0.353    0.290   -0.056   -0.200
##       rm      dis    black     medv
##   -0.219   -0.380   -0.385   -0.388
```

Based on the scatterplots, correlation matrix, and the heatmap, we observe the following:

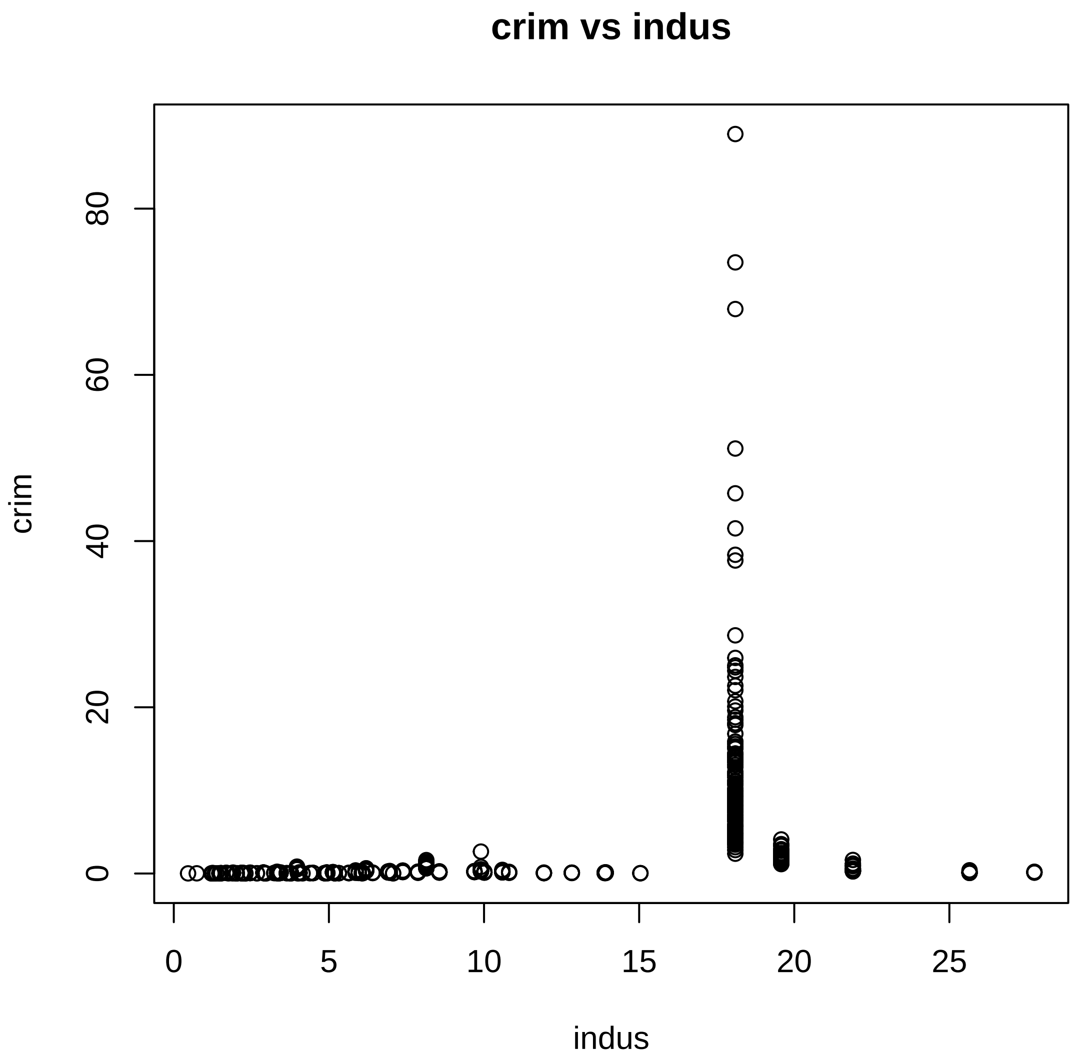
- **crim** has a positive linear association with **indus**, **nox**, **age**, **rad**, **tax**, **ptratio**, and **lstat**. This implies that as these variables increase, the per capita crime rate tends to increase.

- **crim** has a negative linear association with **zn**, **rm**, **dis**, **black**, and **medv**. This indicates that as these variables increase, the per capita crime rate tends to decrease.
- The correlation heatmap shows strong positive and negative linear associations among several predictor variables. This is indicative of **multicollinearity** among the predictors.
- The **Charles River dummy variable (chas)** shows a positive linear association with **medv** and a negative linear association with **ptratio**. However, it exhibits only weak linear associations with **crim**, **zn**, **indus**, **nox**, **rm**, **age**, **dis**, **rad**, **tax,black and lstat**.

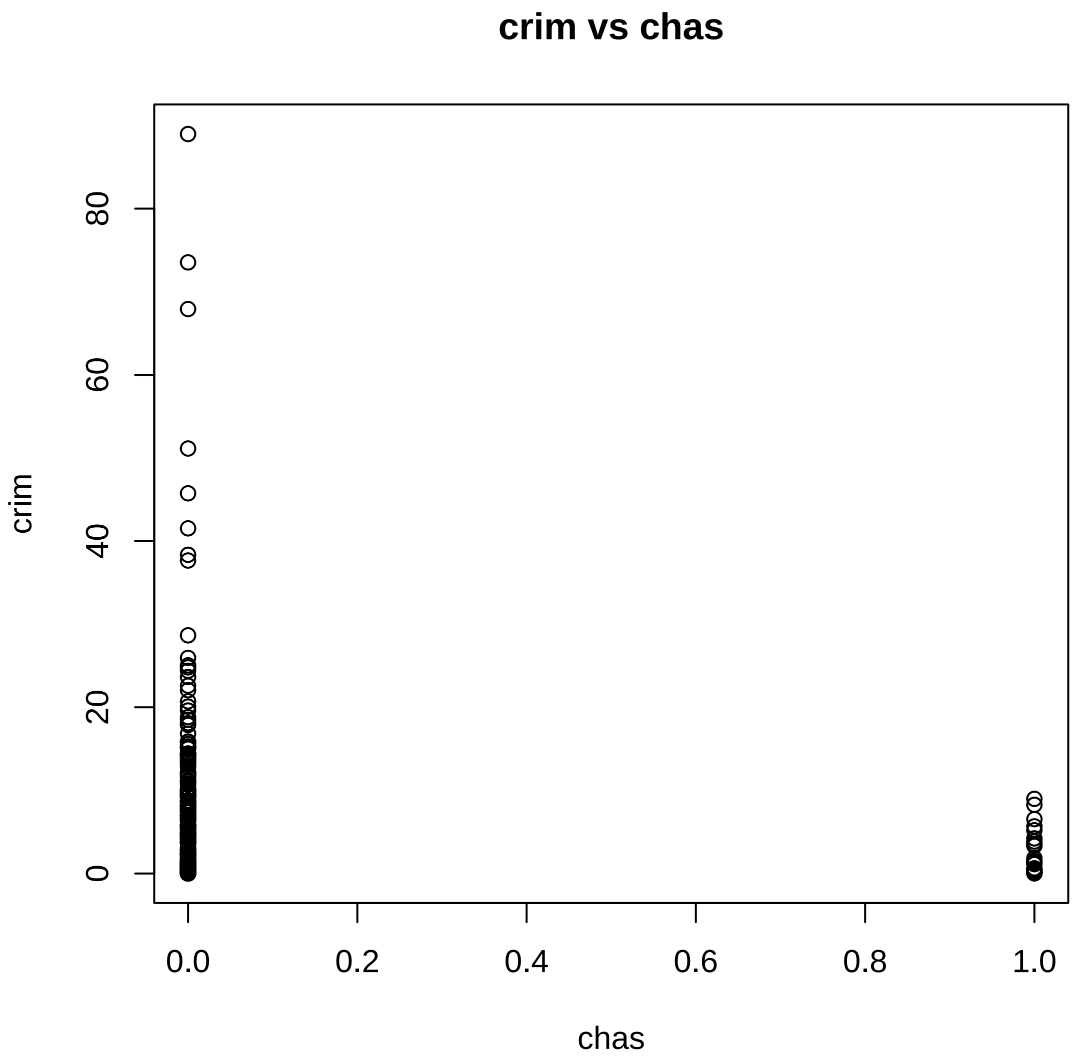
```
# Ensure Suburb_ID is removed
Boston_clean <- Boston[ , !(names(Boston) %in% "Suburb_ID")]

# Plot layout: one plot at a time
par(mfrow = c(1, 1))

# 1. crim vs zn
plot(Boston_clean$zn, Boston_clean$crim,
     main = "crim vs zn", xlab = "zn", ylab = "crim")
```



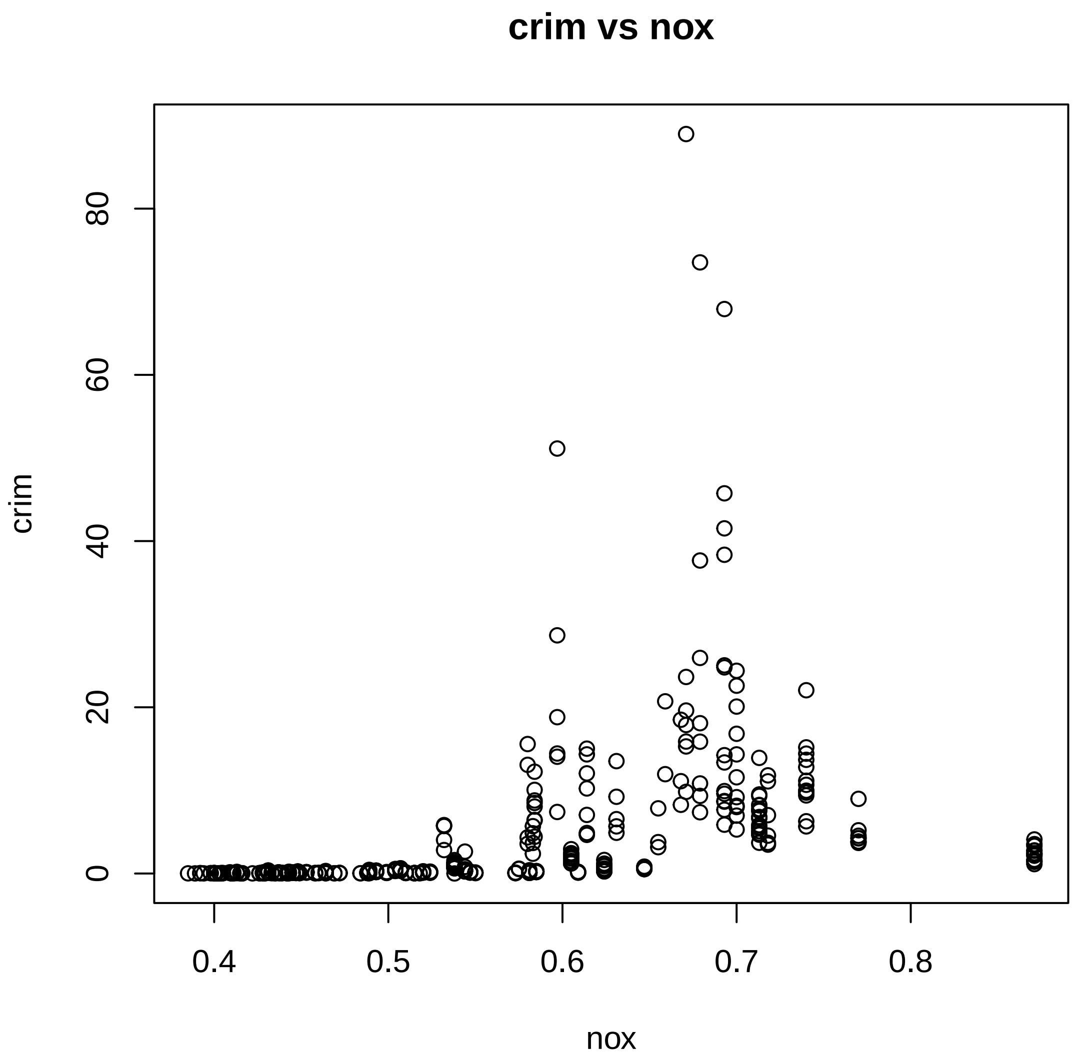

```r
# 2. crim vs indus
plot(Boston_clean$indus, Boston_clean$crim,
     main = "crim vs indus", xlab = "indus", ylab = "crim")
```



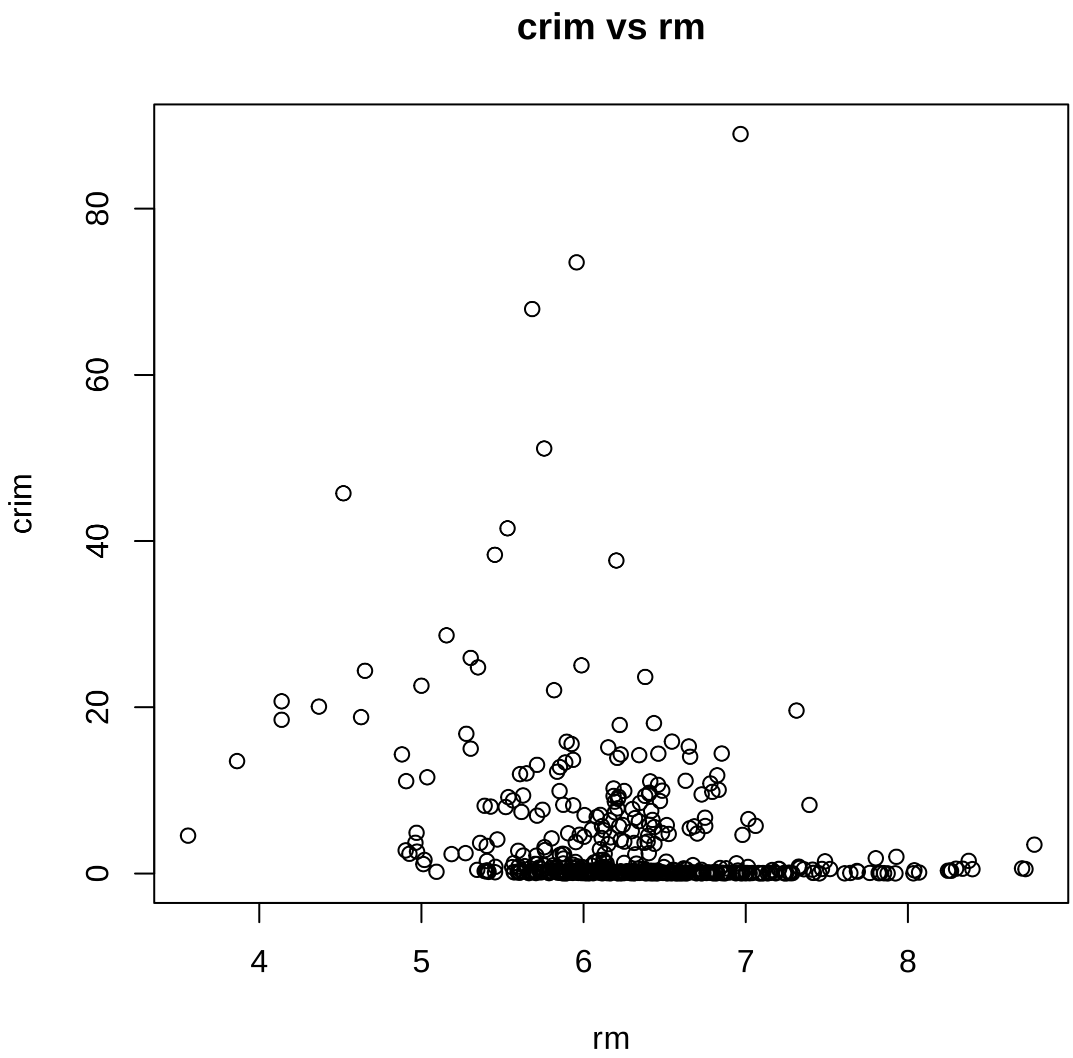

```r
# 3. crim vs chas
plot(Boston_clean$chas, Boston_clean$crim,
     main = "crim vs chas", xlab = "chas", ylab = "crim")
```




```r
# 4. crim vs nox
plot(Boston_clean$nox, Boston_clean$crim,
     main = "crim vs nox", xlab = "nox", ylab = "crim")
```

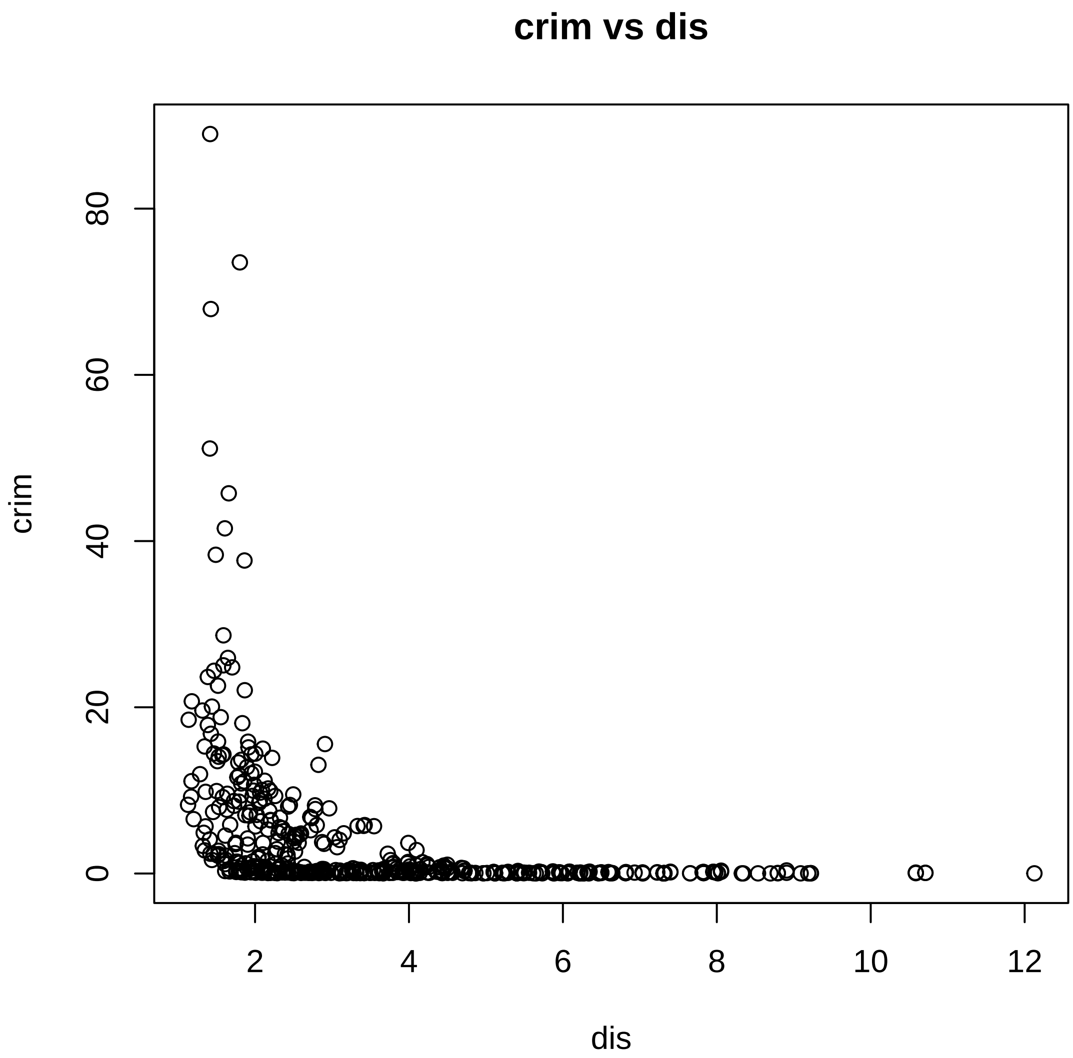



```r
# 5. crim vs rm
plot(Boston_clean$rm, Boston_clean$crim,
     main = "crim vs rm", xlab = "rm", ylab = "crim")
```

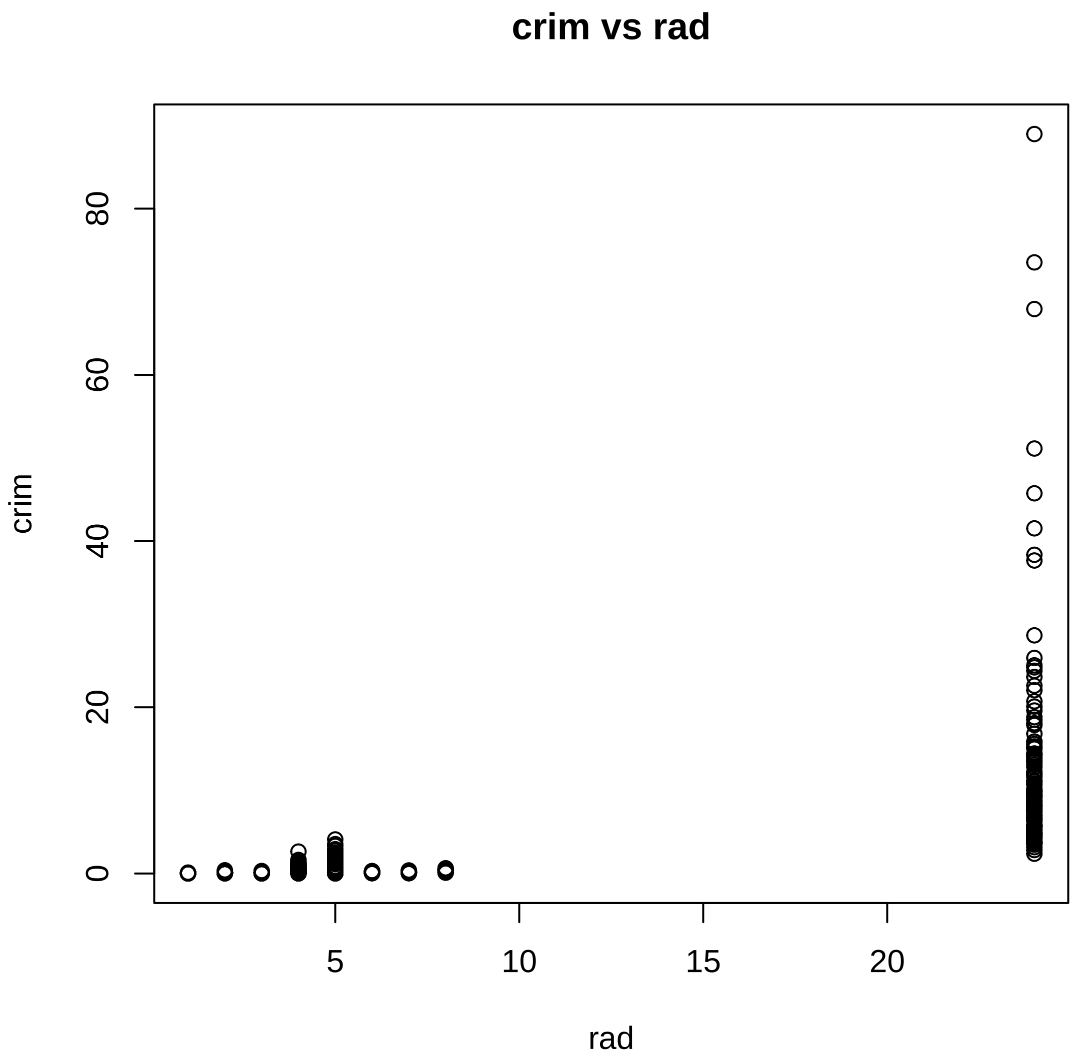



```r
# 6. crim vs age
plot(Boston_clean$age, Boston_clean$crim,
     main = "crim vs age", xlab = "age", ylab = "crim")
```

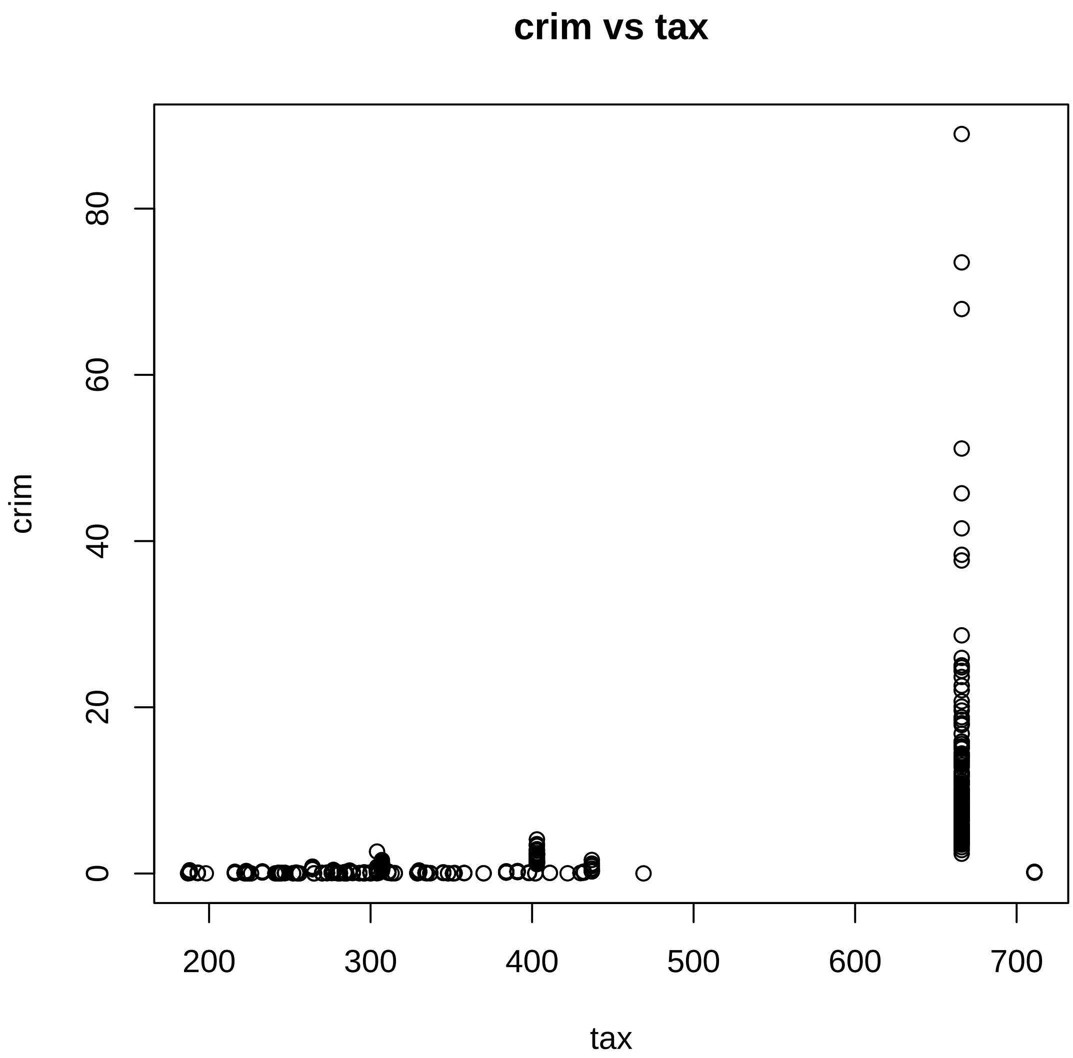

```r
# 7. crim vs dis
plot(Boston_clean$dis, Boston_clean$crim,
     main = "crim vs dis", xlab = "dis", ylab = "crim")
```



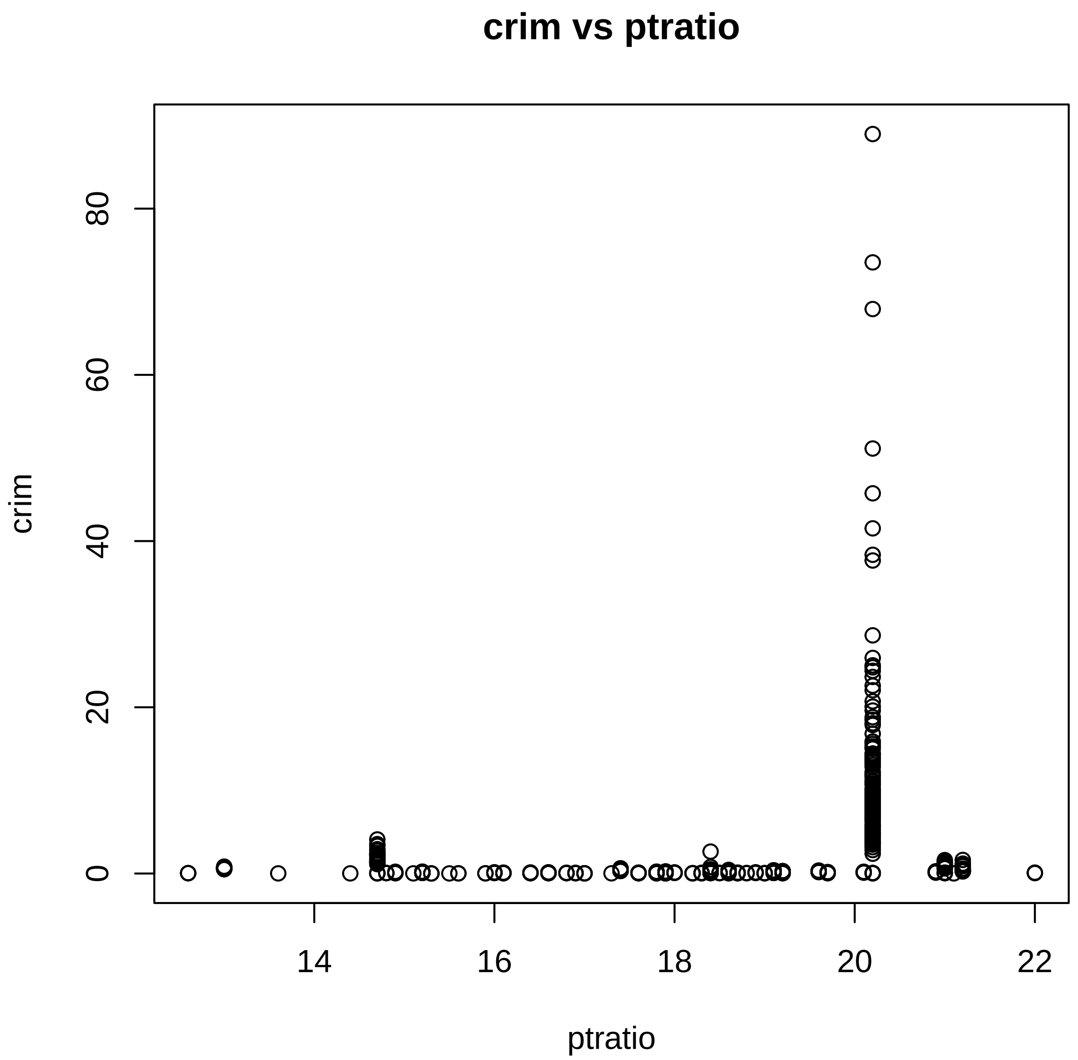

```r
# 8. crim vs rad
plot(Boston_clean$rad, Boston_clean$crim,
     main = "crim vs rad", xlab = "rad", ylab = "crim")
```



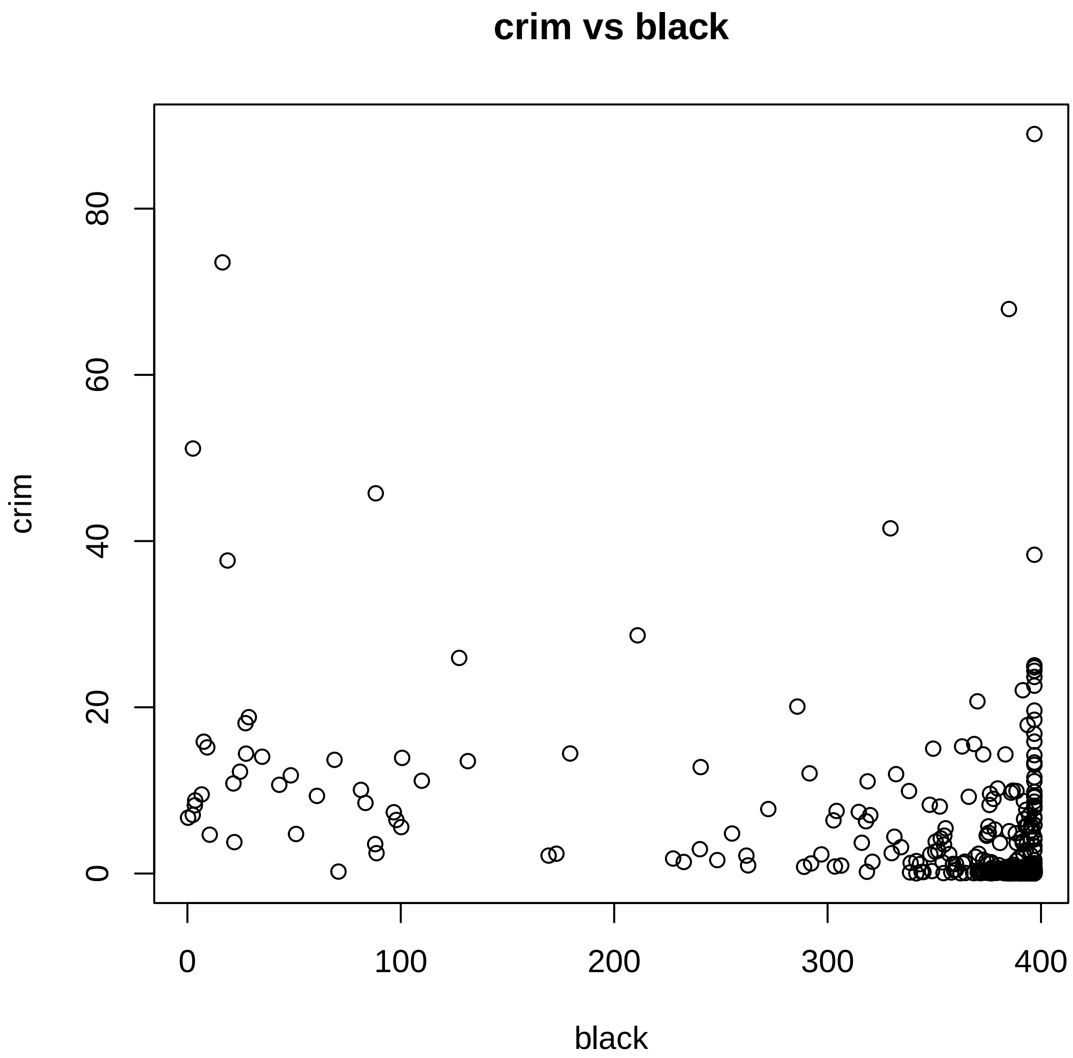

```r
# 9. crim vs tax
plot(Boston_clean$tax, Boston_clean$crim,
     main = "crim vs tax", xlab = "tax", ylab = "crim")
```



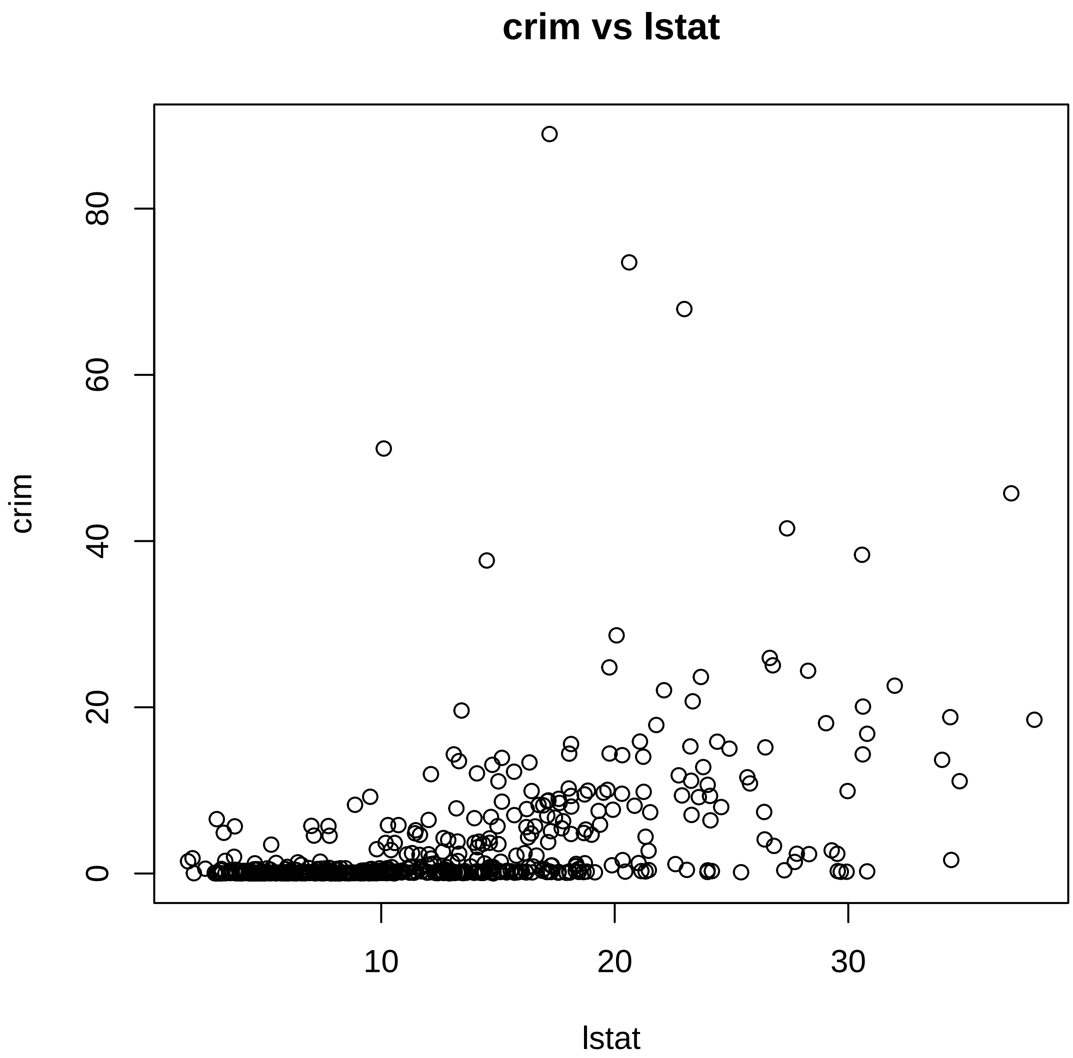

```
# 10. crim vs ptratio
plot(Boston_clean$ptratio, Boston_clean$crim,
     main = "crim vs ptratio", xlab = "ptratio", ylab = "crim")
```



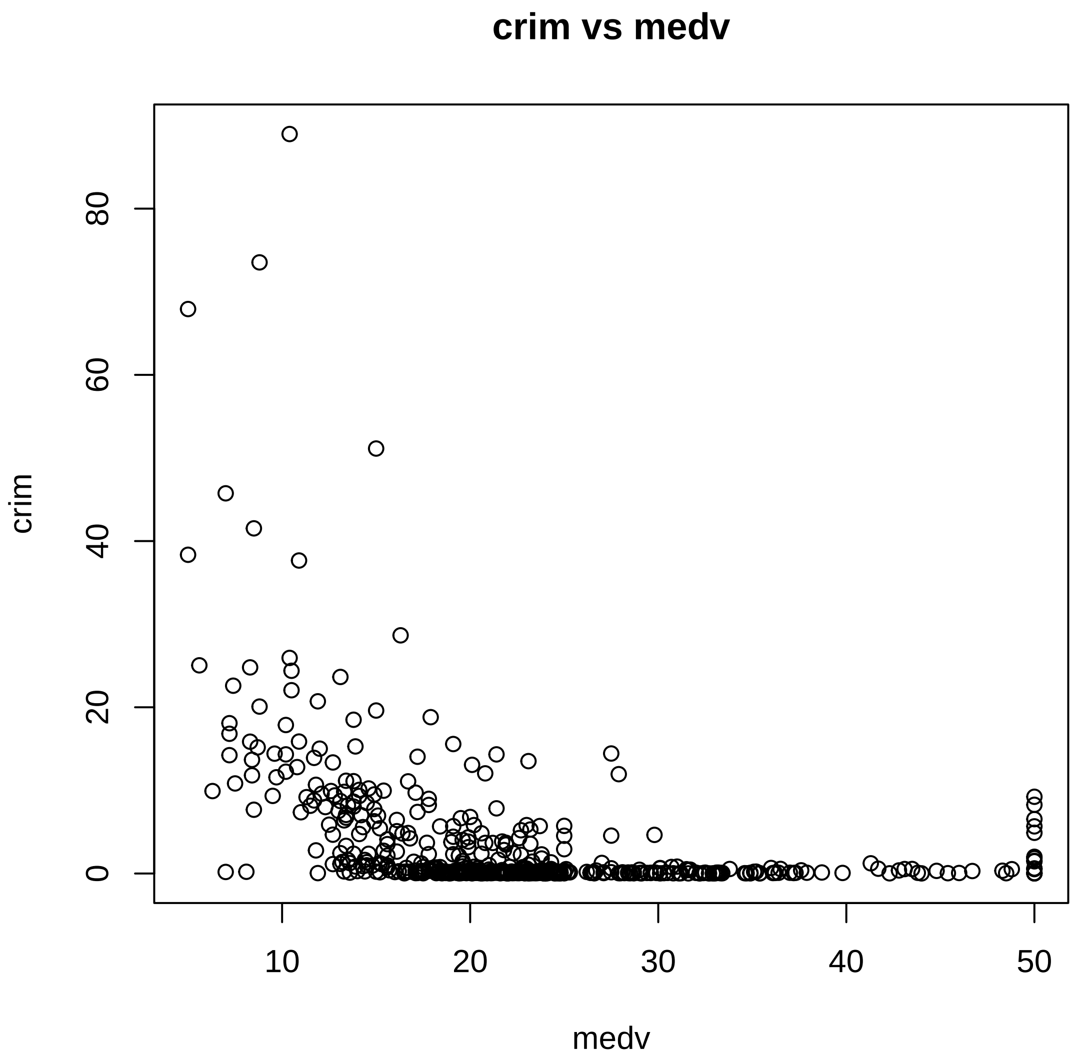

```r
# 11. crim vs black
plot(Boston_clean$black, Boston_clean$crim,
     main = "crim vs black", xlab = "black", ylab = "crim")
```




```
# 12. crim vs lstat
plot(Boston_clean$lstat, Boston_clean$crim,
     main = "crim vs lstat", xlab = "lstat", ylab = "crim")
```




```r
# 13. crim vs medv
plot(Boston_clean$medv, Boston_clean$crim,
     main = "crim vs medv", xlab = "medv", ylab = "crim")
```




```
##Linear regression model
lmodel=lm(crim~.,data=Boston_clean)
summary(lmodel)

##
## Call:
## lm(formula = crim ~ ., data = Boston_clean)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -9.924 -2.120 -0.353  1.019 75.051
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454, Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

Regression Analysis of Per Capita Crime Rate

From the correlation values, we can say that the dependent variable, per capita crime rate (**crim**), has a higher association with predictor variables like accessibility to highways (**rad**), full-value property tax rate (**tax**), percentage of lower status population (**lstat**), and median value of owner-occupied homes (**medv**) etc.

Based on this criterion, the following variables were found to be statistically significant and can be considered as predictors of per capita crime rate:

- **zn**: Proportion of residential land zoned for lots over 25,000 sq.ft.
- **dis**: Weighted mean of distances to five Boston employment centres.
- **rad**: Index of accessibility to radial highways.
- **black**: $1000(Bk - 0.63)^2$, where Bk is the proportion of Black residents.
- **medv**: Median value of owner-occupied homes in \$1000s.

All of these variables have p-values less than 5%, indicating that they are statistically significant predictors of **crim**. This means that we can confidently (at the 95% confidence level) reject the null hypothesis that these predictors have no association with the dependent variable. Therefore, they serve as reliable indicators in predicting per capita crime rate in the Boston dataset.

(vii) [2] Which suburb of Boston has the lowest (highest) median value of owner occupied homes ?

```
min(Boston$medv)

## [1] 5
```

```
max(Boston$medv)

## [1] 50
```

Following suburb of Boston has the lowest = 5 (highest = 50) median value of owner-occupied homes,

(Viii) [2+2=4] How many suburbs average more than eight rooms per dwelling ? Are there any particular charac- teristics of these suburbs that you would like to hightlight ? Number of Suburbs with rm greater than 8 are,

```
nrow(subset(Boston, rm > 8))

## [1] 13
```

```
 subset(Boston_clean, rm > 8)

##         crim zn indus chas    nox    rm  age    dis rad tax ptratio  black lstat
## 98   0.12083  0  2.89    0 0.4450 8.069 76.0 3.4952   2 276    18.0 396.90  4.21
## 164  1.51902  0 19.58    1 0.6050 8.375 93.9 2.1620   5 403    14.7 388.45  3.32
## 205  0.02009 95  2.68    0 0.4161 8.034 31.9 5.1180   4 224    14.7 390.55  2.88
## 225  0.31533  0  6.20    0 0.5040 8.266 78.3 2.8944   8 307    17.4 385.05  4.14
## 226  0.52693  0  6.20    0 0.5040 8.725 83.0 2.8944   8 307    17.4 382.00  4.63
## 227  0.38214  0  6.20    0 0.5040 8.040 86.5 3.2157   8 307    17.4 387.38  3.13
## 233  0.57529  0  6.20    0 0.5070 8.337 73.3 3.8384   8 307    17.4 385.91  2.47
## 234  0.33147  0  6.20    0 0.5070 8.247 70.4 3.6519   8 307    17.4 378.95  3.95
## 254  0.36894 22  5.86    0 0.4310 8.259  8.4 8.9067   7 330    19.1 396.90  3.54
## 258  0.61154 20  3.97    0 0.6470 8.704 86.9 1.8010   5 264    13.0 389.70  5.12
## 263  0.52014 20  3.97    0 0.6470 8.398 91.5 2.2885   5 264    13.0 386.86  5.91
## 268  0.57834 20  3.97    0 0.5750 8.297 67.0 2.4216   5 264    13.0 384.54  7.44
## 365  3.47428  0 18.10    1 0.7180 8.780 82.9 1.9047  24 666    20.2 354.55  5.29
```

```
##     medv
## 98  38.7
## 164 50.0
## 205 50.0
## 225 44.8
## 226 50.0
## 227 37.6
## 233 41.7
## 234 48.3
## 254 42.8
## 258 50.0
## 263 48.8
## 268 50.0
## 365 21.9
```

Analysis of Selected Boston Suburbs

- **High Median Home Values (medv):** Most of these suburbs have median home values close to the maximum of 50.
- **Low Crime Rates (crim):** Generally, crime rates are low, though Suburbs 164 and 365 are exceptions with rates exceeding 1.
- **Lower Pupil-Teacher Ratios (ptratio):** These suburbs typically have lower ptratios, suggesting better educational resources.
- **Older Housing Stock (age):** Many of these areas feature older houses with high age values, indicating a higher proportion built before 1940, except for Suburb 254.
- **Higher Proportion of Residential Land (zn):** These suburbs have a higher average proportion of residential land zoned for lots over 25,000 sq. ft.
- **Lower Industrial Proportion (indus):** The average proportion of non-retail business acres per town is lower.
- **Lower Nitric Oxides Concentration (nox):** These suburbs have a lower average concentration of nitric oxides.
- **Better Accessibility (dis):** They enjoy better accessibility to radial highways.
- **Lower Property Tax Rates (tax):** Property tax rates are lower on average.
- **Higher Black Population (black):** These suburbs have a higher average proportion of Black residents.