DS-6030 Homework Module 1

Tom Lever

05/25/2023

DS 6030 | Spring 2022 | University of Virginia

1. Flexible vs Inflexible Methods

   For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

   (a) The sample size n is extremely large, and the number of predictors p is small.

   We would expect generally the performance of a flexible statistical learning method to be worse than the performance of an inflexible method as there are few predictors and parameters and ample data to characterize a trend.

   (b) The number of predictors p is extremely large, and the number of observations n is small.

   We would expect generally the performance of a flexible statistical learning method to be better than the performance of an inflexible method as there.

   (c) The relationship between the predictors and response is highly non-linear.

   (d) The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   This scenario is an inference regression problem as we are interested in understanding which predictors affect continuous response CEO salary. The number of samples and top firms $n = 500$. The number of predictors, including profit, number of employees, and industry, $p = 3$.

   (b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

   This scenario is a prediction classification problem as we wish to know whether a new product will be a success or a failure. The number of samples and similar products that were previously launched $n = 20$. The number of predictors, including price charged for the product, marketing budget, competition price, and ten other variables, $p = 13$.

   (c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

   This scenario is a prediction regression problem as we are interested in predicting the percent change in the US Dollar / Euro exchange rate in relation to the weekly changes in the world stock markets in 2012. The number of samples and weekly records $n = 52$. The number of predictors, including the percent change in the US market, the percent change in the British market, and the percent change in the German market, $p = 3$.

3. Describe the differences between a parametric and a non-parametric statistical learning approach.

   What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

   A parametric approach to regression and classification has the advantages of involving a explicit, understandable formula with generally a relatively small number of parameters with values that are trained. A parametric approach generally is simple, may be used for prediction and inference, is inflexible, and may have low variance among predicted values given different training data sets. A parametric approach may be too simple, inflexible, and have a high bias. Variance

$$v = E\left(\left\{\hat{f}(x) - E\left[\hat{f}(x)\right]\right\}^2\right)$$

   where $\hat{f}(x)$ is the predicted value of a statistical learning model at x and $E\left[\hat{f}(x)\right]$ is the expected predicted value of a statistical learning model at x given different training data sets. Bias

$$b = \left\{E\left[\hat{f}(x)\right] - y(x)\right\}^2$$

   where $y(x)$ is the ground-truth value at x corresponding to the predicted value.

   A non-parametric approach to regression and classification has the advantage of modeling generally nonlinear relationships using a relatively large number of parameters with values that are trained. A non-parametric approach is flexible and has a low bias. A non-parametric approach may be used for prediction. A non-parametric approach generally is a complex black box that may not be used for inference. A non-parametric approach may have high variance among predicted values given different training data sets.

4. This exercise relates to the College data set, which can be found in the file `College.csv` on the book website.

   It contains a number of variables for 777 different universities and colleges in the US. The variables are

   - `Private` : Public/private indicator
   - `Apps` : Number of applications received
   - `Accept` : Number of applicants accepted
   - `Enroll` : Number of new students enrolled
   - `Top10perc` : New students from top 10 % of high school class
   - `Top25perc` : New students from top 25 % of high school class
   - `F.Undergrad` : Number of full-time undergraduates
   - `P.Undergrad` : Number of part-time undergraduates
   - `Outstate` : Out-of-state tuition
   - `Room.Board` : Room and board costs
   - `Books` : Estimated book costs
   - `Personal` : Estimated personal spending
   - `PhD` : Percent of faculty with Ph.D.'s
   - `Terminal` : Percent of faculty with terminal degree • S.F.Ratio : Student/faculty ratio
   - `perc.alumni` : Percent of alumni who donate
   - `Expend` : Instructional expenditure per student
   - `Grad.Rate` : Graduation rate

   Before reading the data into R, it can be viewed in Excel or a text editor.

(a) Use the `read.csv()` function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

```
college = read.csv("Colleges.csv")
```

(b) Look at the data using the `View()` function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

```
rownames(college) <- college[, 1]
```

```
View(college)
```

Now you should see that the first data column is `Indicator_Of_Whether_College_Is_Private`. Note that another column labeled `row names` now appears before the `Indicator_Of_Whether_College_Is_Private` column. However, this is not a data column but rather the name that R is giving to each row.

```
college <- college[, -1]
```

```
View(college)
```

Now you should see that the first data column with name `Name_Of_College` has been removed from the data frame. Note that another column labeled `row names` now appears. However, this is not a data column but rather a column of the names that R is giving to each row.

(c) See below.

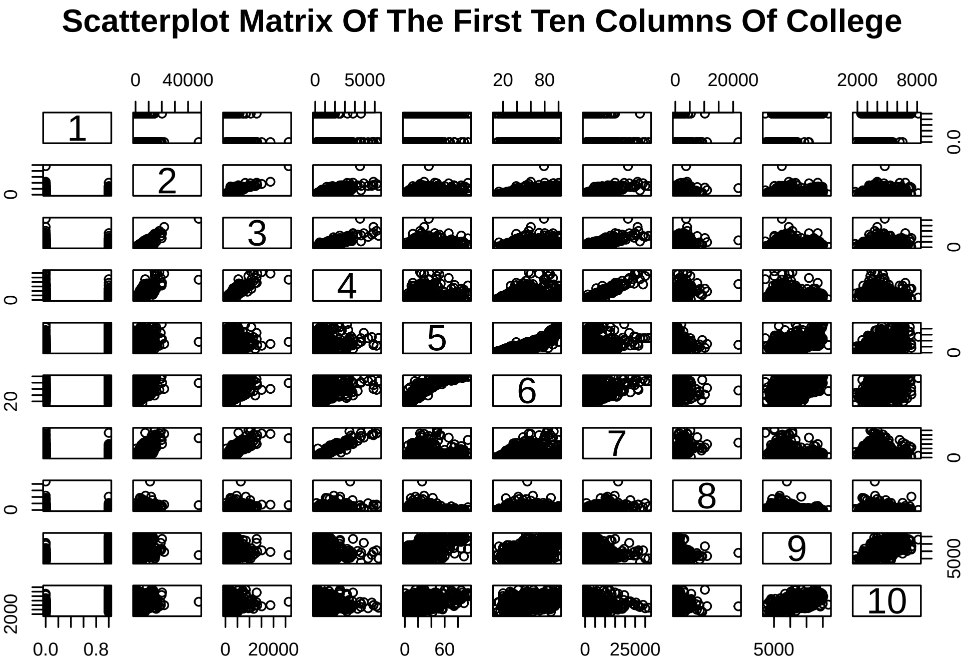
i. Use the `summary()` function to produce a numerical summary of the variables in the data set.

```
summary(college)
```

```
#  Indicator_Of_Whether_College_Is_Private Number_Of_Applications
#  Length:777                              Min.   :   81
#  Class :character                        1st Qu.:  776
#  Mode  :character                        Median : 1558
#                                          Mean   : 3002
#                                          3rd Qu.: 3624
#                                          Max.   :48094
#  Number_Of_Acceptances Number_Of_Enrollments
#  Min.   :   72         Min.   :  35
#  1st Qu.:  604         1st Qu.: 242
#  Median : 1110         Median : 434
#  Mean   : 2019         Mean   : 780
#  3rd Qu.: 2424         3rd Qu.: 902
#  Max.   :26330         Max.   :6392
#  Percent_Of_New_Students_Who_Were_In_Top_10_Percent_Of_High_School_Class
#  Min.   : 1.00
#  1st Qu.:15.00
#  Median :23.00
#  Mean   :27.56
#  3rd Qu.:35.00
#  Max.   :96.00
#  Percent_Of_New_Students_Who_Were_In_Top_25_Percent_Of_High_School_Class
#  Min.   :  9.0
#  1st Qu.: 41.0
#  Median : 54.0
#  Mean   : 55.8
#  3rd Qu.: 69.0
#  Max.   :100.0
#  Number_Of_Full_Time_Undergraduates Number_Of_Part_Time_Undergraduates
#  Min.   :  139                      Min.   :    1.0
#  1st Qu.:  992                      1st Qu.:   95.0
#  Median : 1707                      Median :  353.0
#  Mean   : 3700                      Mean   :  855.3
#  3rd Qu.: 4005                      3rd Qu.:  967.0
#  Max.   :31643                      Max.   :21836.0
#  Out_Of_State_Tuition Cost_Of_Room_And_Board Cost_Of_Books    Personal_Spending
#  Min.   : 2340        Min.   :1780           Min.   :  96.0   Min.   : 250
#  1st Qu.: 7320        1st Qu.:3597           1st Qu.: 470.0   1st Qu.: 850
#  Median : 9990        Median :4200           Median : 500.0   Median :1200
#  Mean   :10441        Mean   :4358           Mean   : 549.4   Mean   :1341
#  3rd Qu.:12925        3rd Qu.:5050           3rd Qu.: 600.0   3rd Qu.:1700
#  Max.   :21700        Max.   :8124           Max.   :2340.0   Max.   :6800
#  Percent_Of_Faculty_With_PhDs Percent_Of_Faculty_With_Terminal_Degrees
#  Min.   :  8.00               Min.   : 24.0
#  1st Qu.: 62.00               1st Qu.: 71.0
#  Median : 75.00               Median : 82.0
#  Mean   : 72.66               Mean   : 79.7
#  3rd Qu.: 85.00               3rd Qu.: 92.0
#  Max.   :103.00               Max.   :100.0
#  Student_Faculty_Ratio Percent_Of_Alumni_Who_Donate
#  Min.   : 2.50         Min.   : 0.00
#  1st Qu.:11.50         1st Qu.:13.00
#  Median :13.60         Median :21.00
#  Mean   :14.09         Mean   :22.74
#  3rd Qu.:16.50         3rd Qu.:31.00
#  Max.   :39.80         Max.   :64.00
#  Instructional_Expenditure_Per_Student Graduation_Rate
#  Min.   : 3186                         Min.   : 10.00
#  1st Qu.: 6751                         1st Qu.: 53.00
#  Median : 8377                         Median : 65.00
#  Mean   : 9660                         Mean   : 65.46
#  3rd Qu.:10830                         3rd Qu.: 78.00
#  Max.   :56233                         Max.   :118.00
```

ii. Use the `pairs()` function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix `A` using `A[,1:10]`.

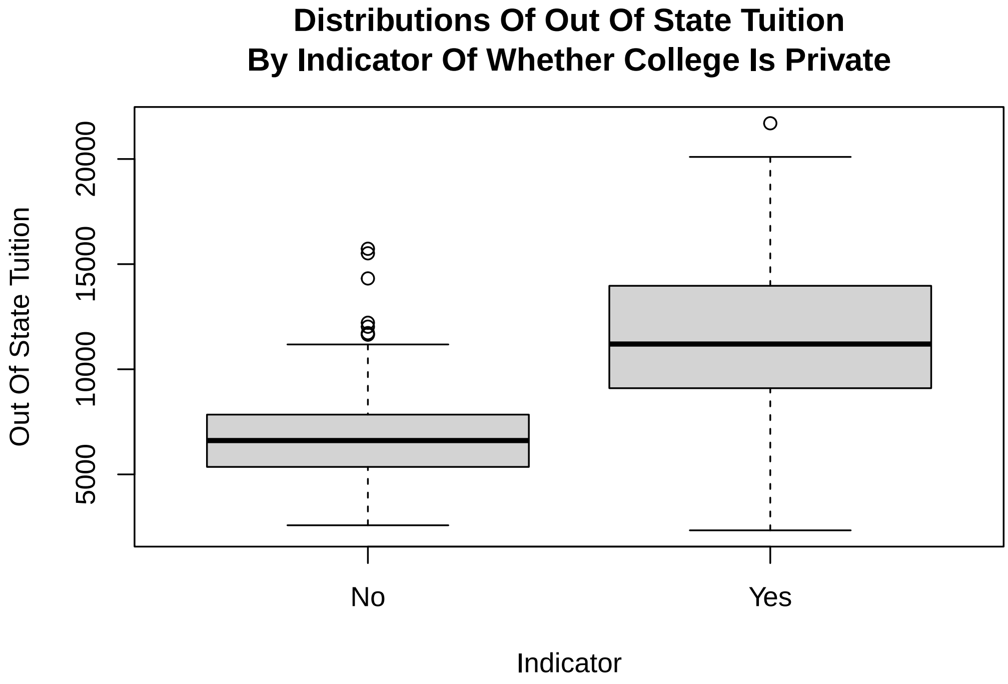
```
number_of_rows_in_college <- nrow(college)
column_of_numerical_representations <- rep(0, number_of_rows_in_college)
condition <- college$Indicator_Of_Whether_College_Is_Private == "Yes"
column_of_numerical_representations[condition] <- 1
library(TomLeversRPackage)
index_Of_Indicator_Of_Whether_College_Is_Private <- college |>
    get_column_index("Indicator_Of_Whether_College_Is_Private")
data_frame_of_second_through_tenth_columns <- college[, 2:10]
data_frame_representing_first_ten_columns_of_college <- data.frame(
    column_of_numerical_representations,
    data_frame_of_second_through_tenth_columns
)
pairs(
    data_frame_representing_first_ten_columns_of_college,
    main = "Scatterplot Matrix Of The First Ten Columns Of College",
    labels = seq(1, 10)
)
```



```
#library(GGally)
#ggpairs(data_frame_representing_first_ten_columns_of_college)
```

iii. Use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Private`.

```
factor_Indicator_Of_Whether_College_Is_Private <- as.factor(
    college$Indicator_Of_Whether_College_Is_Private
)
plot(
    x = factor_Indicator_Of_Whether_College_Is_Private,
    y = college$Out_Of_State_Tuition,
    main = "Distributions Of Out Of State Tuition\nBy Indicator Of Whether College Is Priva
    xlab = "Indicator",
    ylab = "Out Of State Tuition"
)
```



iv. Create a new qualitative variable, called `Elite`, by binning the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

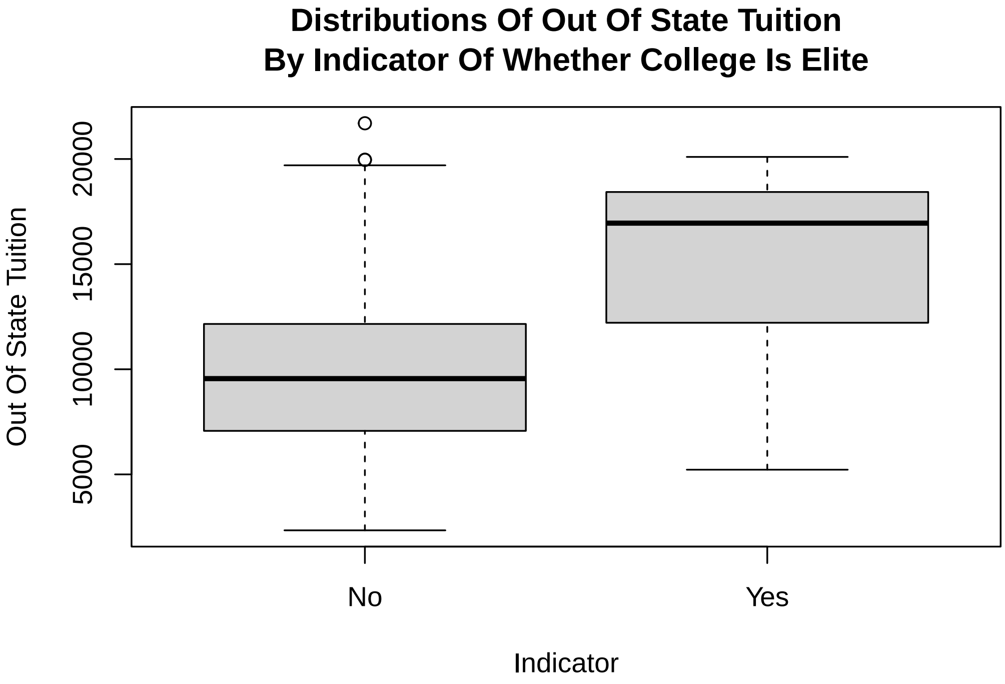
```
number_of_rows_in_college <- nrow(college)
Elite <- rep("No", number_of_rows_in_college)
column_of_percents <- college$
    Percent_Of_New_Students_Who_Were_In_Top_10_Percent_Of_High_School_Class
condition <- column_of_percents > 50
Elite[condition] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college, Elite)
```

Use the `summary()` function to see how many elite universities there are. Now use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Elite`.

```
library(TomLeversRPackage)
index_of_column_Elite <- get_column_index(college, "Elite")
column_Elite <- college[, index_of_column_Elite]
summary(column_Elite)
```

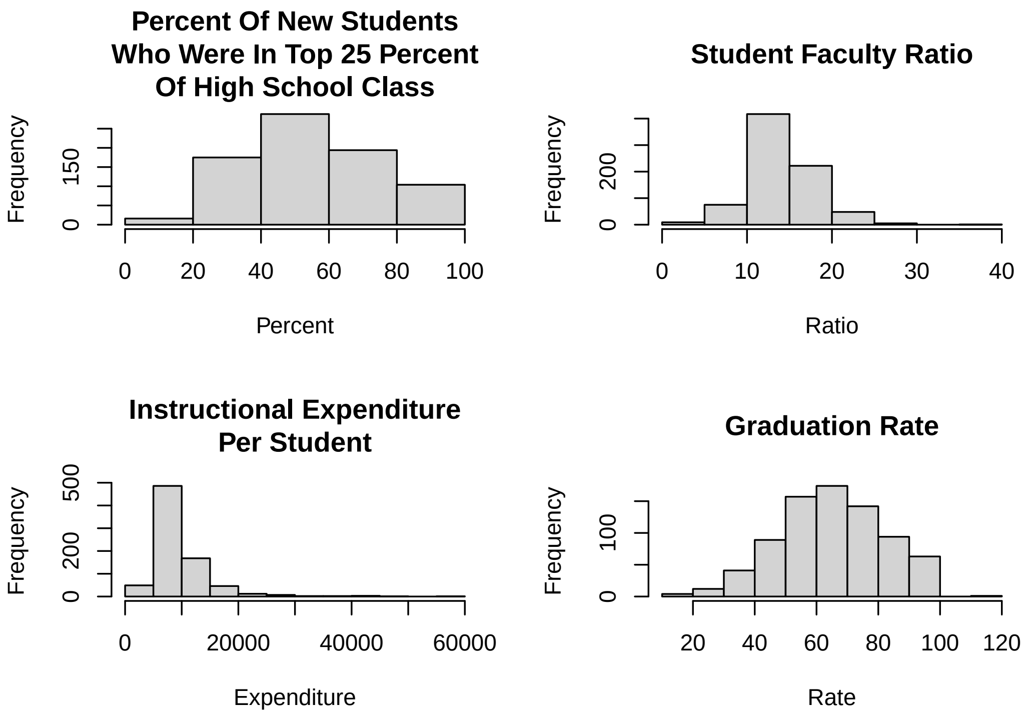
```
#  No Yes
# 699  78
```

```
plot(
    x = Elite,
    y = college$Out_Of_State_Tuition,
    main = "Distributions Of Out Of State Tuition\nBy Indicator Of Whether College Is Elite
    xlab = "Indicator",
    ylab = "Out Of State Tuition"
)
```



v. Use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command `par(mfrow = c(2, 2))` useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

```
par(mfrow = c(2, 2))
column_of_percents <- college$
    Percent_Of_New_Students_Who_Were_In_Top_25_Percent_Of_High_School_Class
hist(
    x = column_of_percents,
    breaks = 6,
    main = "Percent Of New Students\nWho Were In Top 25 Percent\nOf High School Class",
    xlab = "Percent"
)
hist(
    x = college$Student_Faculty_Ratio,
    breaks = 7,
    main = "Student Faculty Ratio",
    xlab = "Ratio"
)
hist(
    x = college$Instructional_Expenditure_Per_Student,
    breaks = 8,
    main = "Instructional Expenditure\nPer Student",
    xlab = "Expenditure"
)
hist(
    x = college$Graduation_Rate,
    breaks = 9,
    main = "Graduation Rate",
    xlab = "Rate"
)
```



vi. Continue exploring the data, and provide a brief summary of what you discover.

5. This exercise involves the Boston housing data set.

(a) To begin, load in the `Boston` data set. The Boston data set is part of the ISLR2 library.

```
library(ISLR2)
```

Now the data set is contained in the object Boston.

```
head(Boston)
```

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

Read about the data set:

```
?Boston
```

How many rows are in this data set? How many columns? What do the rows and columns represent?

```
nrow(Boston)
```

```
# [1] 506
```

```
ncol(Boston)
```

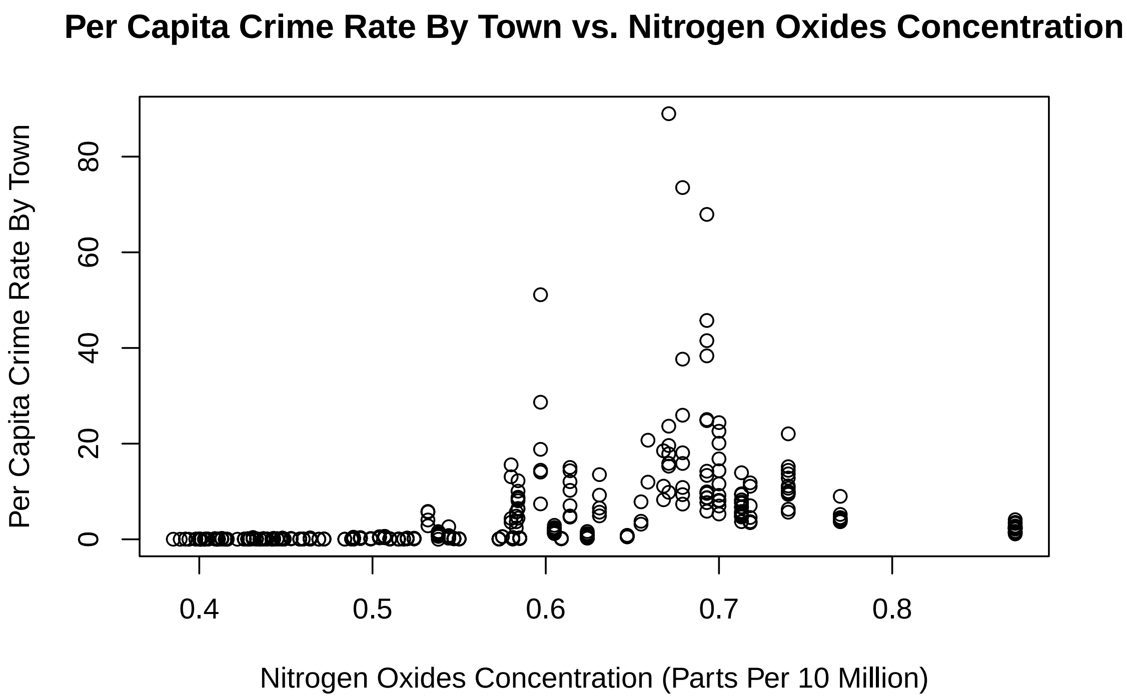
```
# [1] 13
```

Rows are records containing “housing values in 506 suburbs of Boston” (chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://cran.r-project.org/web/packages/ISLR2/ISLR2.pdf). Columns represent features of suburbs of Boston and include:

- `crim`: per capita crime rate by town
- `zn`: proportion of residential land zoned for lots over $25,000$ square feet
- `indus`: proportion of non-retail business acres per town
- `chas`: Charles River dummy variable that is 1 if tract bounds river and 0 otherwise
- `nox`: nitrogen oxides concentration in 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 centers
- `rad`: index of accessibility to radial highways
- `tax`: full-value property-tax rate per 10,000 dollars
- `ptratio`: pupil-teacher ratio by town
- `lstat`: lower status of the population in percent
- `medv`: median value of owner-occupied homes in thousands of dollars

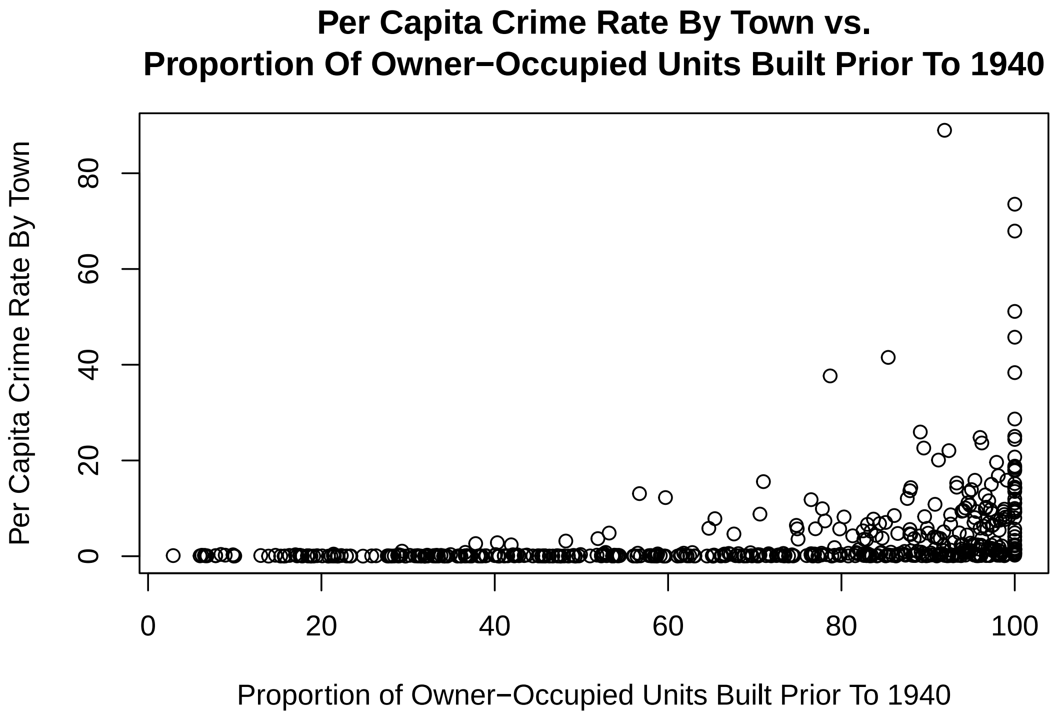
(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

```
plot(
    x = Boston$nox,
    y = Boston$crim,
    main = "Per Capita Crime Rate By Town vs. Nitrogen Oxides Concentration",
    xlab = "Nitrogen Oxides Concentration (Parts Per 10 Million)",
    ylab = "Per Capita Crime Rate By Town"
)
```



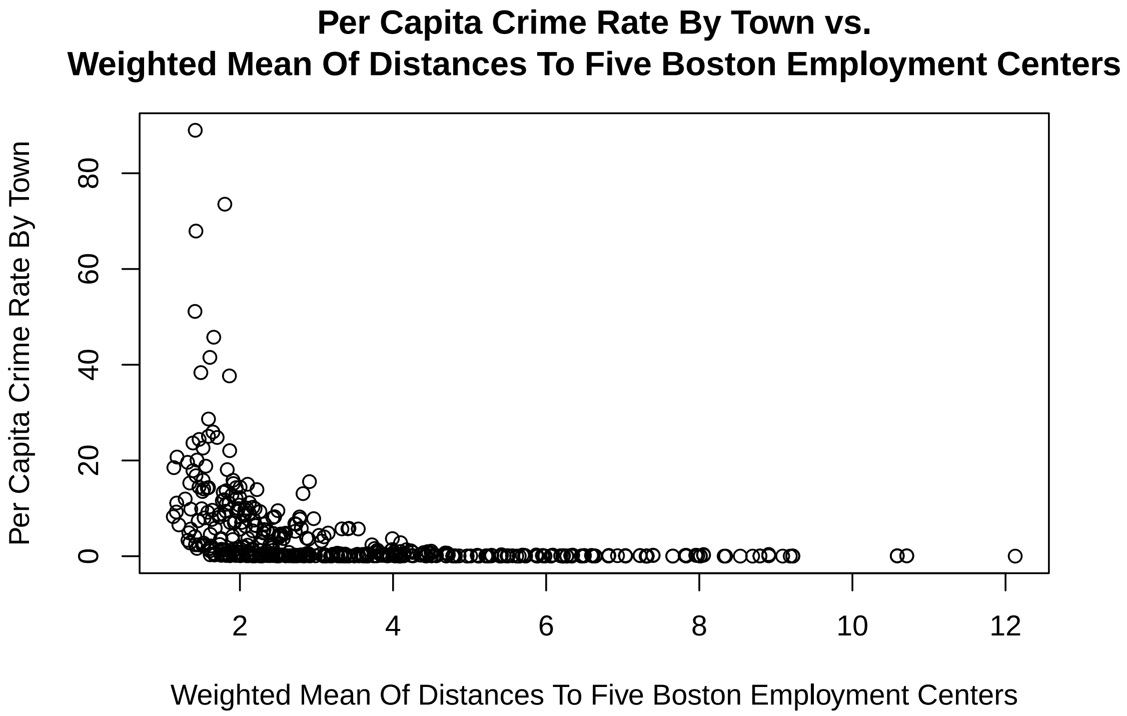
Generally, as nitrogen oxides concentration increases, per capita crime rate by town and its variance increase exponentially. Nitrogen oxides concentration may be a proxy for reduction in quiet green space allowing an individual to be themselves, industrialization, people being on top of each other, and reduction in health.

```
plot(
    x = Boston$age,
    y = Boston$crim,
    main = "Per Capita Crime Rate By Town vs.\nProportion Of Owner-Occupied Units Built Prior T
    xlab = "Proportion of Owner-Occupied Units Built Prior To 1940",
    ylab = "Per Capita Crime Rate By Town"
)
```



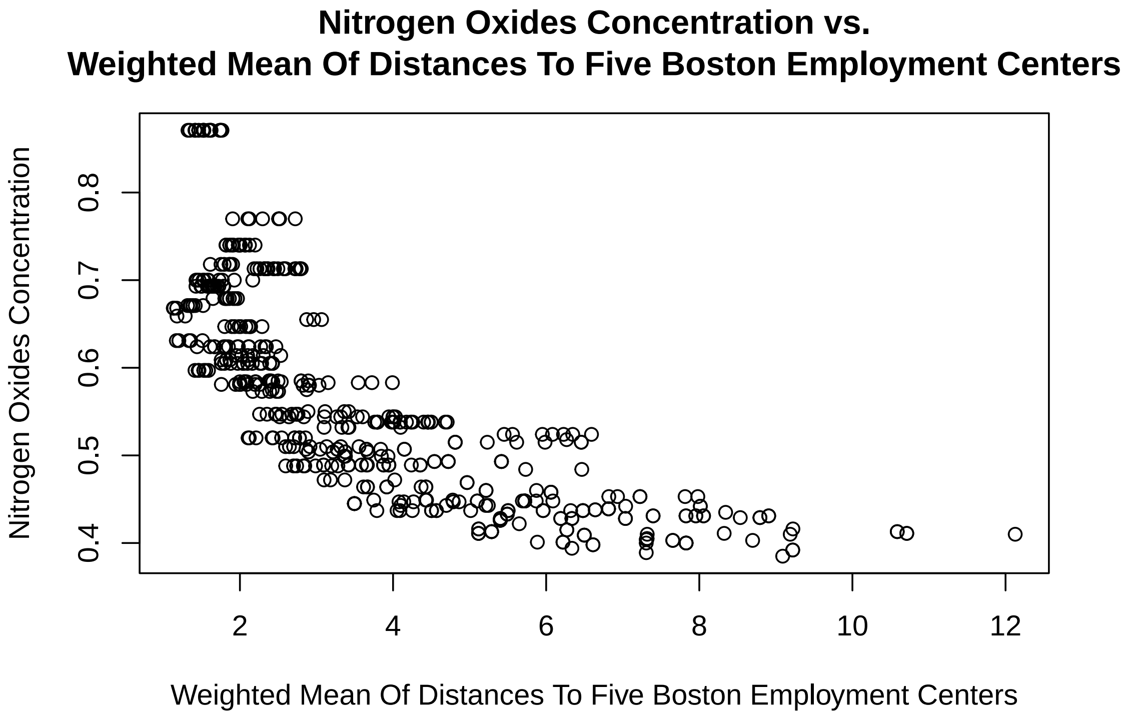
Generally, as the proportion of owner-occupied units built prior to 1940 increases, per capita crime rate by town and its variance increase exponentially. Proportion of owner-occupied units built prior to 1940 may be a proxy for proportion of substandard housing and urban decay.

```
plot(
    x = Boston$dis,
    y = Boston$crim,
    main = "Per Capita Crime Rate By Town vs.\nWeighted Mean Of Distances To Five Boston Employ
    xlab = "Weighted Mean Of Distances To Five Boston Employment Centers",
    ylab = "Per Capita Crime Rate By Town"
)
```



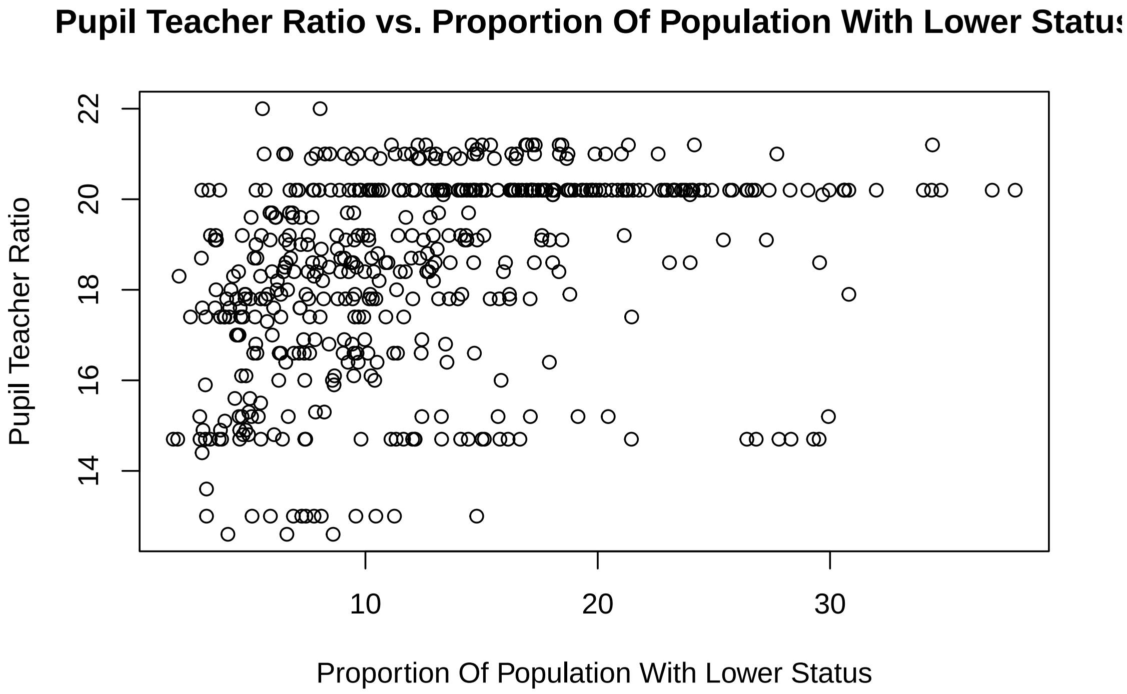
Generally, as the weighted mean of distances to five Boston employment centers increases, per capita crime rate by town and its variance decrease exponentially. Weighted mean of distances to five Boston employment centers may be a proxy for white flight or reduction in people being on top of each other.

```
plot(
    x = Boston$dis,
    y = Boston$nox,
    main = "Nitrogen Oxides Concentration vs.\nWeighted Mean Of Distances To Five Boston Emplo
    xlab = "Weighted Mean Of Distances To Five Boston Employment Centers",
    ylab = "Nitrogen Oxides Concentration"
)
```



Generally, as the weighted mean of distances to five Boston employment centers increases, nitrogen oxides concentration decreases. Boston employment centers may be relatively industrialized and distant suburbs may be relatively green.

```
plot(
    x = Boston$lstat,
    y = Boston$ptratio,
    main = "Pupil Teacher Ratio vs. Proportion Of Population With Lower Status",
    xlab = "Proportion Of Population With Lower Status",
    ylab = "Pupil Teacher Ratio"
)
```



Generally, as proportion of population with lower status increases, pupil teacher ratio increases. Pupil teacher ratio may be a proxy for lack of individualized education.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

```r
correlation_matrix <- cor(Boston)
analyze_correlation_matrix(correlation_matrix)
```

```
# crim
#     V+:  crim
#     V-:
#     H+:
#     H-:
#     M+:  rad, tax
#     M-:
#     L+:  indus, nox, age, lstat
#     L-:  dis, medv
#     N:  zn, chas, rm, ptratio
# zn
#     V+:  zn
#     V-:
#     H+:
#     H-:
#     M+:  dis
#     M-:  indus, nox, age
#     L+:  rm, medv
#     L-:  rad, tax, ptratio, lstat
#     N:  crim, chas
# indus
#     V+:  indus
#     V-:
#     H+:  nox, tax
#     H-:  dis
#     M+:  age, rad, lstat
#     M-:  zn
#     L+:  crim, ptratio
#     L-:  rm, medv
#     N:  chas
# chas
#     V+:  chas
#     V-:
#     H+:
#     H-:
#     M+:
#     M-:
#     L+:
#     L-:
#     N:  crim, zn, indus, nox, rm, age, dis, rad, tax, ptratio, lstat, medv
# nox
#     V+:  nox
#     V-:
#     H+:  indus, age
#     H-:  dis
#     M+:  rad, tax, lstat
#     M-:  zn
#     L+:  crim
#     L-:  rm, medv
#     N:  chas, ptratio
# rm
#     V+:  rm
#     V-:
#     H+:
#     H-:
#     M+:  medv
#     M-:  lstat
#     L+:  zn
#     L-:  indus, nox, ptratio
#     N:  crim, chas, age, dis, rad, tax
# age
#     V+:  age
#     V-:
#     H+:  nox
#     H-:  dis
#     M+:  indus, tax, lstat
#     M-:  zn
#     L+:  crim, rad
#     L-:  medv
#     N:  chas, rm, ptratio
# dis
#     V+:  dis
#     V-:
#     H+:
#     H-:  indus, nox, age
#     M+:  zn
#     M-:  tax
#     L+:
#     L-:  crim, rad, lstat
#     N:  chas, rm, ptratio, medv
# rad
#     V+:  rad, tax
#     V-:
#     H+:
#     H-:
#     M+:  crim, indus, nox
#     M-:
#     L+:  age, ptratio, lstat
#     L-:  zn, dis, medv
#     N:  chas, rm
# tax
#     V+:  rad, tax
#     V-:
#     H+:  indus
#     H-:
#     M+:  crim, nox, age, lstat
#     M-:  dis
#     L+:  ptratio
#     L-:  zn, medv
#     N:  chas, rm
# ptratio
#     V+:  ptratio
#     V-:
#     H+:
#     H-:
#     M+:
#     M-:  medv
#     L+:  indus, rad, tax, lstat
#     L-:  zn, rm
#     N:  crim, chas, nox, age, dis
# lstat
#     V+:  lstat
#     V-:
#     H+:
#     H-:  medv
#     M+:  indus, nox, age, tax
#     M-:  rm
#     L+:  crim, rad, ptratio
#     L-:  zn, dis
#     N:  chas
# medv
#     V+:  medv
#     V-:
#     H+:
#     H-:  lstat
#     M+:  rm
#     M-:  ptratio
#     L+:  zn
#     L-:  crim, indus, nox, age, rad, tax
#     N:  chas, dis
```

Per capita crime rate is very highly positively associated with itself; highly correlated with no predictors; moderately positively correlated with index of accessibility of radial highways and full-value property-tax rate per 10,000 dollars; lowly positively correlated with proportion of non-retail business acres per town, nitrogen oxides concentration in parts per 10 million, proportion of owner-occupied units built prior to 1940, and lower status of the population in percent; and lowly negatively correlated with weighted mean of distances to five Boston employment centers and median value of owner-occupied homes in thousands of dollars.

(d) Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

(e) How many of the census tracts in this data set bound the Charles river?

(f) What is the median pupil-teacher ratio among the towns in this data set?

(g) Which census tract of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

(h) In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.