1. This question is worth 50 marks.

The dataset Chile is available from the carData package.

a. Install the carData package, load the library, and access the Chile dataset contained in it. Load the help file and read about the dataset. In what year were the data collected? [6 marks]

install.packages('carData')

library(carData)

data("Chile")

?Chile

str(Chile)

'data.frame': 2700 obs. of 8 variables:

$ region : Factor w/ 5 levels "C","M","N","S",..: 3 3 3 3 3 3 3 3 3 3 ...

$ population: int 175000 175000 175000 175000 175000 175000 175000 175000 175000 175000 ...

$ sex : Factor w/ 2 levels "F","M": 2 2 1 1 1 1 2 1 1 2 ...

$ age : int 65 29 38 49 23 28 26 24 41 41 ...

$ education : Factor w/ 3 levels "P","PS","S": 1 2 1 1 3 1 2 3 1 1 ...

$ income : int 35000 7500 15000 35000 35000 7500 35000 15000 15000 15000 ...

$ statusquo : num 1.01 -1.3 1.23 -1.03 -1.1 ...

$ vote : Factor w/ 4 levels "A","N","U","Y": 4 2 4 2 2 2 2 2 3 2 ...

The data were collect in 1988.

b. Look at the structure of the dataset and describe it briefly. What unit is the income variable measured in? [6 marks]

The dataset contains records on 2700 voting intentions, with 8 variables. They show the region of the country, sex, age, level of education, monthly income (in Pesos), scale of support for the status-quo and intention of vote.

The unit for income variable is Pesos.

c. Find the mean income of all respondents for which data is available. [6 marks]

mean(Chile$income, na.rm = TRUE)

[1] 33875.86

d. Create a two-way table to examine the relationship between region and sex. Describe the results. How many female respondents from the city of Santiago are there? [5 marks]

table(Chile$region, Chile$sex)

F M

C 300 300

M 51 49

N 168 154

S 362 356

SA 498 462

There are 498 female respondents from the city of Santiago.

e. Remove all rows with missing values to create a new dataset Chile2. What is the size of the dataset now? Work with this reduced dataset for the remainder of Question 1. [5 marks]

Chile2 <- na.omit(Chile)

nrow(Chile2)

[1] 2431

The Chile2 dataset has 2431 observations.

f. Find the mean and standard deviation of the age of the respondents, grouped by the sex variable. Describe your findings. [6 marks]

aggregate(Chile2[ , 4], list(Chile2$sex), mean)

Group.1 x

1 F 37.98080

2 M 38.61727

aggregate(Chile2[ , 4], list(Chile2$sex), sd)

Group.1 x

1 F 14.02227

2 M 15.31940

We can see that in average the male respondents are older than the females ones.

g. The education variable is a factor, but it is not ordered. Convert it into a sensibly ordered factor (primary education is ‘less than’ secondary education which is ‘less than’ post-secondary education). [8 marks]

Chile2$education <- factor(Chile2$education, ordered = TRUE, levels = c("Primary", "Secondary", "Post-secondary"))

str(Chile2$education)

Ord.factor w/ 3 levels "Primary"<"Secondary"<..: NA NA NA NA NA NA NA NA NA NA ...

h. Use the aggregate function to aggregate the income and age variables by the factors of sex and vote, returning a single object.

Which sex/vote combination has the highest mean income? What is the mean age of this group? You must use code to find these answers. [8 marks]

df <- aggregate(Chile2[ , c(4,6)], list(Chile2$sex, Chile2$vote), mean)

Group.1 Group.2 age income

1 F A 33.89691 31984.54

2 M A 33.58750 34093.75

3 F N 36.22535 37697.18

4 M N 35.84180 36679.69

5 F U 39.80531 25066.37

6 M U 41.47642 26603.77

7 F Y 38.85403 33001.09

8 M Y 41.84615 40915.12

df[ which.max(df$income) , ]

Group.1 Group.2 age income

8 M Y 41.84615 40915.12

df[ which.max(df$income) , ]$Group.1 # sex

[1] M

df[ which.max(df$income) , ]$Group.2 # vote

[1] Y

df[ which.max(df$income) , ]$age

[1] 41.84615

The group with the highest mean income is female who vote yes and the mean age is 41.84.

2. This question is worth 25 marks.

The dataset Cars93 is available from the MASS library.

a. Load the MASS library and access the Cars93 dataset contained in it. Load the help file and read about the dataset. How many variables does the dataset contain? [4 marks]

install.packages('MASS')

library(MASS)

data("Cars93")

?Cars93

str(Cars93)

'data.frame': 93 obs. of 27 variables:

$ Manufacturer : Factor w/ 32 levels "Acura","Audi",..: 1 1 2 2 3 4 4 4 4 5 ...

$ Model : Factor w/ 93 levels "100","190E","240",..: 49 56 9 1 6 24 54 74 73 35 ...

$ Type : Factor w/ 6 levels "Compact","Large",..: 4 3 1 3 3 3 2 2 3 2 ...

$ Min.Price : num 12.9 29.2 25.9 30.8 23.7 14.2 19.9 22.6 26.3 33 ...

$ Price : num 15.9 33.9 29.1 37.7 30 15.7 20.8 23.7 26.3 34.7 ...

$ Max.Price : num 18.8 38.7 32.3 44.6 36.2 17.3 21.7 24.9 26.3 36.3 ...

$ MPG.city : int 25 18 20 19 22 22 19 16 19 16 ...

$ MPG.highway : int 31 25 26 26 30 31 28 25 27 25 ...

$ AirBags : Factor w/ 3 levels "Driver & Passenger",..: 3 1 2 1 2 2 2 2 2 2 ...

$ DriveTrain : Factor w/ 3 levels "4WD","Front",..: 2 2 2 2 3 2 2 3 2 2 ...

$ Cylinders : Factor w/ 6 levels "3","4","5","6",..: 2 4 4 4 2 2 4 4 4 5 ...

$ EngineSize : num 1.8 3.2 2.8 2.8 3.5 2.2 3.8 5.7 3.8 4.9 ...

$ Horsepower : int 140 200 172 172 208 110 170 180 170 200 ...

$ RPM : int 6300 5500 5500 5500 5700 5200 4800 4000 4800 4100 ...

$ Rev.per.mile : int 2890 2335 2280 2535 2545 2565 1570 1320 1690 1510 ...

$ Man.trans.avail : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 1 1 1 1 ...

$ Fuel.tank.capacity: num 13.2 18 16.9 21.1 21.1 16.4 18 23 18.8 18 ...

$ Passengers : int 5 5 5 6 4 6 6 6 5 6 ...

$ Length : int 177 195 180 193 186 189 200 216 198 206 ...

$ Wheelbase : int 102 115 102 106 109 105 111 116 108 114 ...

$ Width : int 68 71 67 70 69 69 74 78 73 73 ...

$ Turn.circle : int 37 38 37 37 39 41 42 45 41 43 ...

$ Rear.seat.room : num 26.5 30 28 31 27 28 30.5 30.5 26.5 35 ...

$ Luggage.room : int 11 15 14 17 13 16 17 21 14 18 ...

$ Weight : int 2705 3560 3375 3405 3640 2880 3470 4105 3495 3620 ...

$ Origin : Factor w/ 2 levels "USA","non-USA": 2 2 2 2 2 1 1 1 1 1 ...

$ Make : Factor w/ 93 levels "Acura Integra",..: 1 2 4 3 5 6 7 9 8 10 ...

The dataset contains 27 variables.

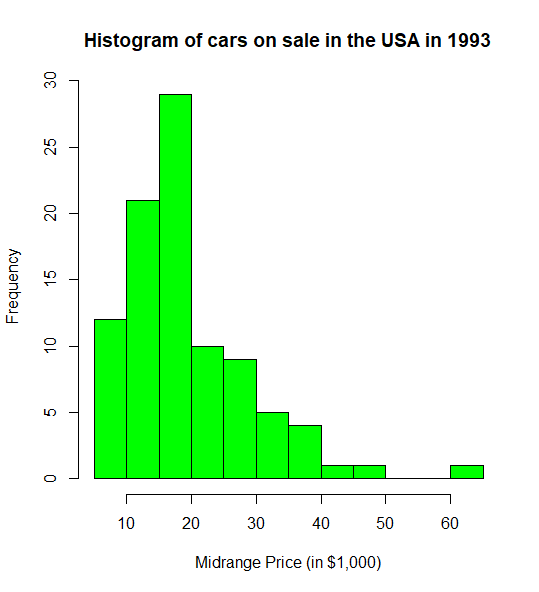
b. Produce a histogram of the Price variable, and make the graph look neat and presentable (paying particular attention to labels, colours, titles etc.). Comment on the resulting graph. [9 marks]

hist(Cars93$Price, breaks = 10,

xlab = "Midrange Price (in $1,000)",

main = "Histogram of cars on sale in the USA in 1993",

col = "green")



We can notice that the most frequent price of the cars in sale are between $15,000 and $20,000

c. Use the type='n' argument (or otherwise) to help you to create a scatterplot of the Length variable against the Price variable where there are three distinct groups in the plot, depending on the DriveTrain type of the cars.

You should:

● Select a different plotting character than the default pch

● Colour the three groups differently

● Include a legend to explain these groups

● Include sensible x- and y-axis labels and a main title

● Rotate the numbers on the y-axis so they appear horizontal

Comment on the resulting graph. [12 marks]

install.packages("car")

library(car)

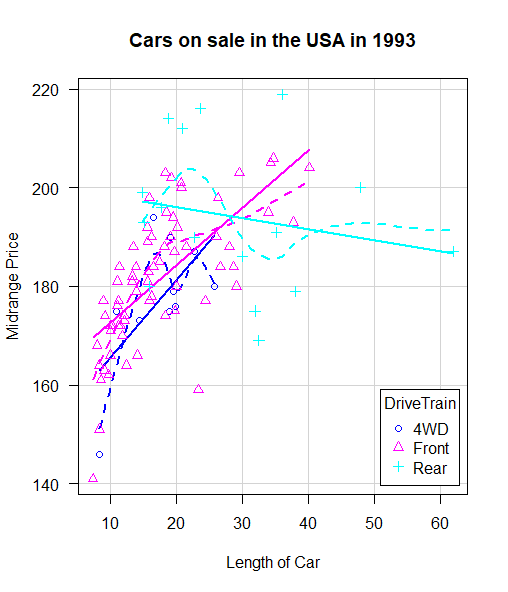
scatterplot(Length ~ Price | DriveTrain, data=Cars93,

xlab="Length of Car", ylab="Midrange Price",

main="Cars on sale in the USA in 1993",

las = 1, # position of the labes (2-vertical/1-horizontal)

legend = c(title="DriveTrain", coords="bottomright"))



3. This question is worth 25 marks.

The code below is used to take a matrix of numeric data, find the rows in this matrix which have at least one positive number, and then return the index of these rows.

(To save you time, you can find the code in the file Q3.R contained in the Exam folder.)

fun1 <- function(mat) {

out <- NULL

for(i in 1:nrow(mat)) {

if(any(mat[i,] > 0)) {

out <- c(out, i) } }

return(out) }

set.seed(13)

my.mat <- matrix(rnorm(20), 10, 2)

my.mat

res1 <- fun1(my.mat)

res1

a. Check how the function fun1() works, and add comments to explain what is happening in each line. [10 marks]

b. Comment on the relative advantages and disadvantages of using the benchmark() function (from the rbenchmark library) vs. the Rprof() function. [5 marks]

c. The function fun1() is quite slow. Write a faster version fun2() and confirm that it produces the same output. Benchmark fun1() and fun2() in order to compare their performance, and comment on the results. Copy and paste the benchmark output into your script, in order to support your answer. [10 marks]

END OF EXAMINATION