HELLENIC OPEN UNIVERSITY

Hellenic Open University

DATA SCIENCE AND MACHINE LEARNING (MSc)

DAMA51: Foundations in Computer Science

Academic Year: 2022–2023

#1 Written Assignment	
Submission Deadline	**Wednesday, 7 December 2022, 23:59:59 EET**

Guidelines

The deadline is definitive.

An indicative solution will be posted online along with the returning of the graded assignments.

The assignment is due via the STUDY submission system. **You are expected to deliver a document (.DOC, .ODT, .PDF) and a compressed (.ZIP, .RAR) file containing all your work:**

- 1 document file (this document) with the answers to all the topics, along with the R code and the results of the execution of the code
- 1 compressed file with 3 R scripts that correspond to topics 3, 4 and 5.

You should not make any changes in the written assignment file other than providing your own answers. You should also type all of your answers into Word and not attach any handwritten notes as pictures into your work otherwise a 5% reduction of your final grade will be applied. Make sure to name all the files (ZIP file, DOC file and R script files) with **your last name first followed by a dash and the names of each component at the end**. For example for the student with last name Aggelou the files should be named as follows: Aggelou-HW1.zip, Aggelou-HW1.doc, Aggelou-Topic2.R, Aggelou-Topic3.R and Aggelou-Topic5.R. The R script files should automatically run with the **source** command and generate the correct results. Also, please include comments before each command to explain the functionality of the command that follows. Unless otherwise stated in the question, all numerical answers should be given to **three decimal places**.

Topic	Points	Grades
1. Online QUIZ	50	
2. Article Review	10	
3. Tabular and Graphical Representations	10	
4. Correlation	15	
5. Data Frames	15	
TOTAL	**100**	**/100**

Topic 1: Online QUIZ

Complete the corresponding online quiz available at:

https://study.eap.gr/mod/quiz/view.php?id=24553

You have one effort and unlimited time to complete the quiz, up to the submission deadline. **(50 points)**

Topic 2: Article Review

The review article by de Bie et al. entitled "Automating Data Science" (available at https://dl.acm.org/doi/10.1145/3495256) resumes existing and potential automation practices in four different areas of the Data Science pipeline: (1) Model building, (2) Data Engineering, (3) Data Exploration and (4) Exploitation. Select one of the four areas and summarize in two or three brief paragraphs the main automation techniques, descripted in the article. Finally, conclude by briefly explaining what would be your personal degree of trustfulness to the descripted techniques.

(10 points)

The issue that the researchers have to overcome in the general project, is to keep in mind whether they deal with supervised, unsupervised, semi-supervised or reinforcement learning. In the first phase of the process, which is Model building, the optimal way to proceed is with supervised learning. In that way, the researchers can make the automation mechanized, which is the most optimal way to address the data in that stage.

More particularly, by that way the researchers can make this process as automated as it gets, which is never exclusively automated because the human eye must and should be overlooking the situation, to make the more theoretical decisions. Especially with the latest technology advancements, the tools that are used nowadays have the option to automate the selection of models and parameters, which saves the researchers a lot of time and gives them the opportunity to invest more effort to the most unseen parts of the project, such as logical error handling and data management.

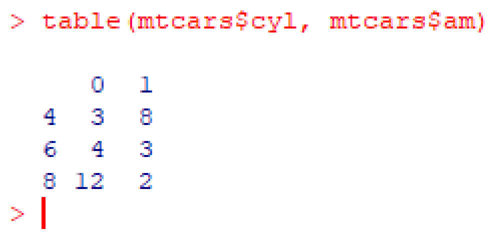
This is a very sound plan, and as mentioned above, with the help of the advanced tools that researchers have developed, the degree of trustfulness on that plan to handle this phase of the process is very large.

Topic 3: Tabular and Graphical Representations

For this topic, we'll use the built-in data set **mtcars** of R in order to answer the following points. **(10 points)**

a. Create a contingency table, using R, with absolute frequencies for the attributes `cyl` and `am`. **(4 points)**

table(mtcars$cyl, mtcars$am)

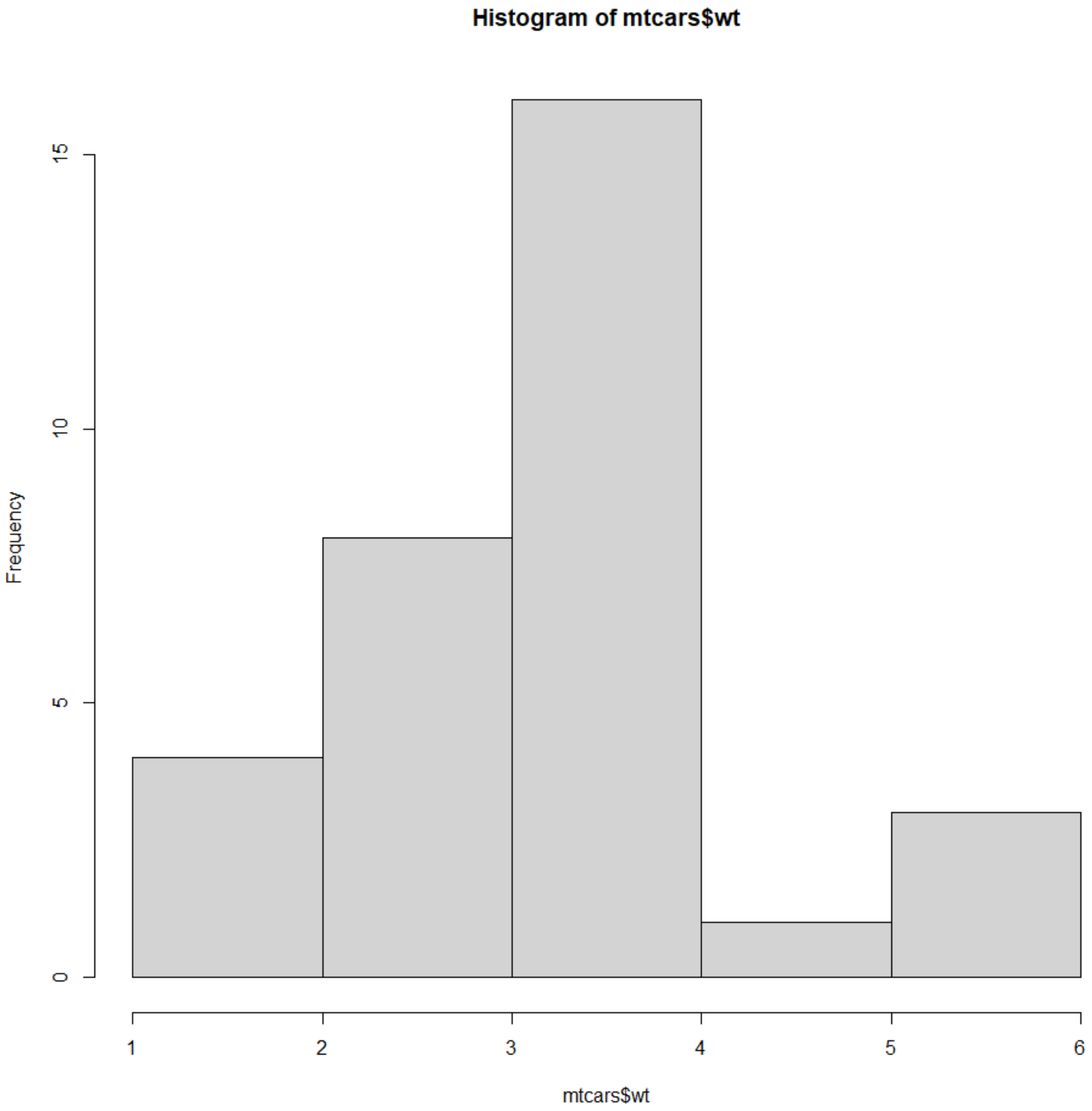


```
> table(mtcars$cyl, mtcars$am)

     0  1
  4  3  8
  6  4  3
  8 12  2
> 
```

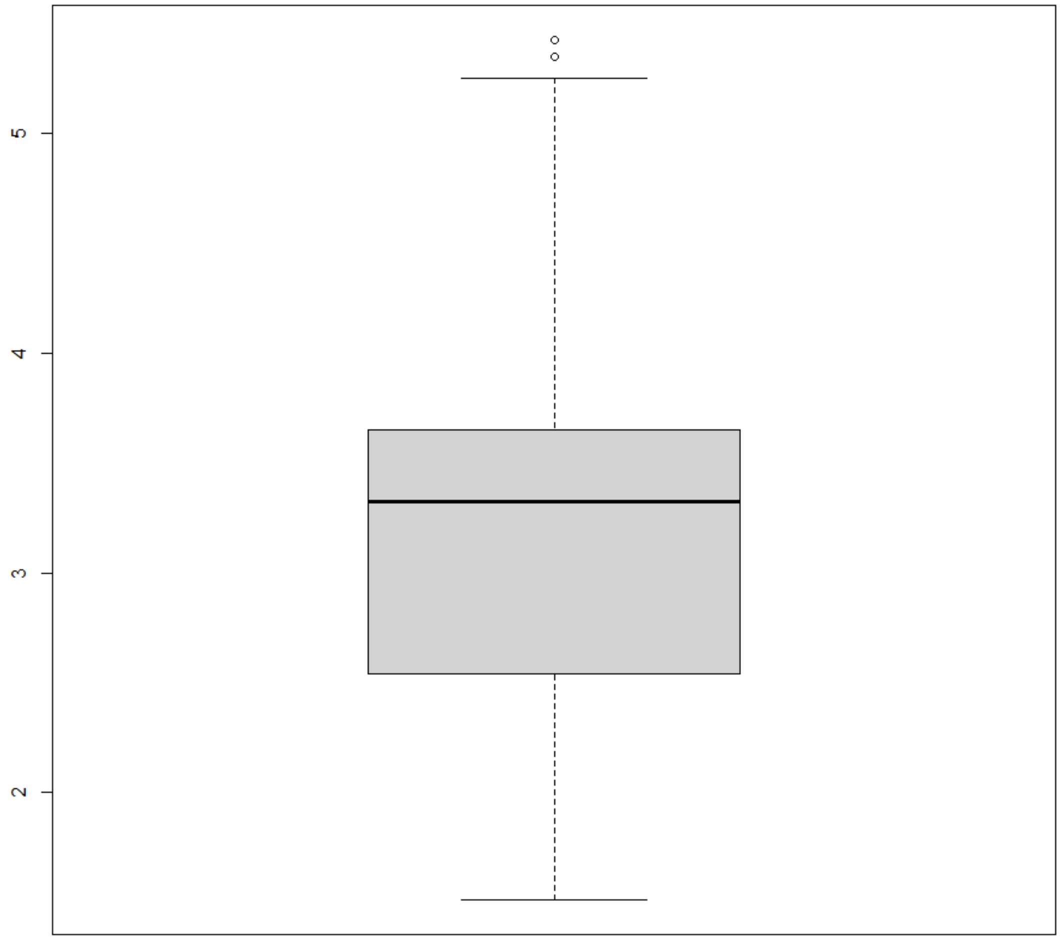
b. Create a histogram with absolute frequencies for the attribute `wt` using five bins.**(3 points)**

```
hist(mtcars$wt, breaks=5)
```



c. Create a box plot for the attribute `wt`. **(3 points)**

```
boxplot(mtcars$wt)
```



Topic 4: Correlation

You are given the following vectors `a = [11, 15, 23, 46, 52, 75]` and `w = [34, 49, 58, 62, 69, 64]`, which represent ages and weights, respectively, of a sample of people. **(15 points)**

a. Use pen and paper to calculate the correlation between `a` and `w` using the Pearson correlation coefficient. Verify your answer with R. Include your calculations, R code, and an interpretation of the result. **(5 points)**

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}$$

Solving for $n = 6,\ \bar{x} = 37,\ \bar{y} = 56,\ \ S_x = 24.923,\ S_y = 12.696$ we find $r_{xy} = 0.791$

R command: round(cor(c(11, 15, 23, 46, 52, 75), c(34, 49, 58, 62, 69, 64), method = "pearson"), digits=3)

Result: 0.791

"The larger the absolute value of the Pearson correlation coefficient, the stronger the relationship between the coefficients." Since 0.791 is closer to 1 than 0, we conclude that the relationship between the coefficients is strong.

b. Use pen and paper to calculate the correlation between `a` and `w` using the Spearman's rank correlation coefficient. Verify your answer with R. Include your calculations, R code, and an interpretation of the result. **(5 points)**

$$\rho = 1 - 6\frac{\sum_{i=1}^{n}(r(x_i) - r(y_i))^2}{n(n^2-1)}$$

Solving for $n = 6,\ r(x) = [1\ 2\ 3\ 4\ 5\ 6],\ r(y) = [1\ 2\ 3\ 4\ 6\ 5]$ we find ρ = 0.943

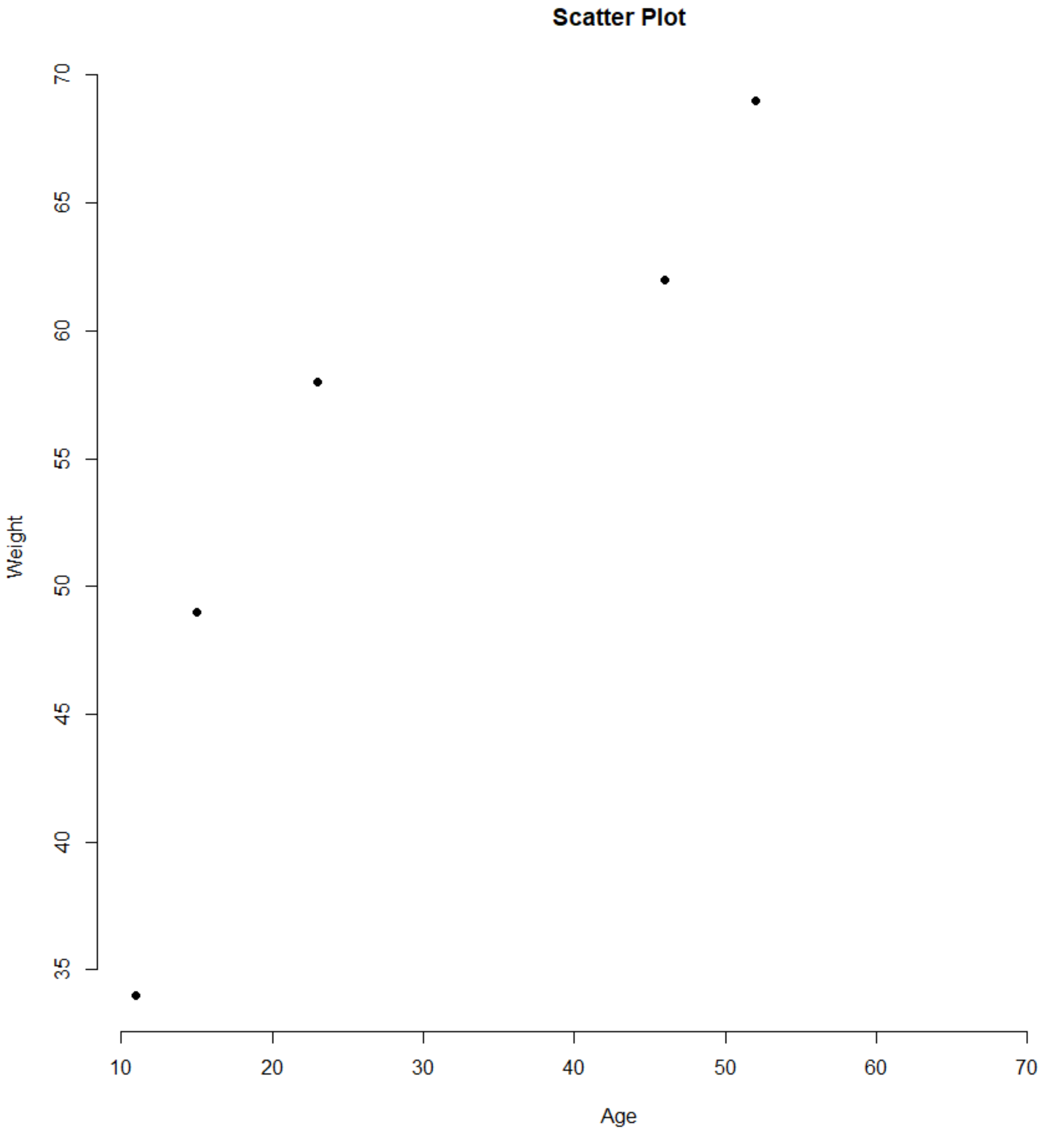
R command: round(cor(c(11, 15, 23, 46, 52, 75), c(34, 49, 58, 62, 69, 64), method = "spearman"), digits=3)

Result: 0.943

"When the rankings of the x- and y-values are exactly in the same order, Spearman's rho will yield value 1." The result being less than 1 means that the values are not in the exact same order but it being very close to 1 means that they are in a very similar order

Draw a scatter plot in R to visually check if there exists a relationship between `a` (x-axis) and `w` (y-axis). Include the labels of both axes. **(5 points)**

```
plot(c(11, 15, 23, 46, 52, 75), c(34, 49, 58, 62, 69, 64), main="Scatter Plot", xlab="Age",
ylab="Weight", pch=19, frame= FALSE)
```



Topic 5: Data Frames

For this topic, we’ll use the built-in data set **state.x77** of R in order to answer the following points. **(15 points)**

a. Check the data type of the data set and make sure it is a data frame. if not convert it into a data frame. Provide the R code in the space provided below. **(3 points)**

```
typeof(state.x77)
#returned “double” so we use the following command
myDataFrame <- as.data.frame.matrix(state.x77)
```

b. Create a new attribute called `states` and assign to it the row names of the data set. Then, remove the row names from the data set. Provide the R code in the space provided below. **(3 points)**

```
myDataFrame <- as.data.frame.matrix(state.x77)
states <- row.names(myDataFrame )
rownames(myDataFrame ) <- NULL
```

c. Find out how many states have income (per capita) of more than 4,300$ and population more than 1,000 (in thousands) people. **(3 points)**

R commands:

```
myDataFrame <- as.data.frame.matrix(state.x77)
sum(myDataFrame$Income > 4300 & myDataFrame$Population > 1000)
```

Result:

Number of states: 22

d. Print out the top-5 states, which exhibit the highest income, after having ordered the data frame in decreasing order based on attribute `Income`. **(3 points)**

R commands:

```
myDataFrame <- as.data.frame.matrix(state.x77)
sorted <- myDataFrame[order(-myDataFrame$Income),]
rownames(head(sorted, n=5))
```

Result:

1. Alaska
2. Connecticut
3. Maryland
4. New Jersey
5. Nevada

e. Create a new ordinal attribute called `frost_cat` which takes on the values `low`, `intermediate`, and `high` that correspond to the following intervals (-1, 30], (30, 90], and (90, 190]. Print out the states of the low category in the space provided below. **(3 points)**

R commands:

```
myDataFrame <- as.data.frame.matrix(state.x77)
x <- cut(myDataFrame$Frost, breaks = c(-1, 30, 90, 190), include.lowest = TRUE,
labels = c("low", "intermediate", "high"))
myDataFrame$frost_cat <- factor(x)
rownames(myDataFrame[myDataFrame$frost_cat == "low",])
```

Result:

1. Alabama
2. Arizona
3. California
4. Florida
5. Hawaii
6. Louisiana