**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 |  |
| 1. **Article Review** | 10 |  |
| 1. **Tabular and Graphical Representations** | 10 |  |
| 1. **Correlation** | 15 |  |
| 1. **Data Frames** | 15 |  |
| **TOTAL** | **100** | **/100** |

# Topic 1: Online QUIZ

Complete the corresponding online quiz available at:

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)**

|  |
| --- |
|  |

# 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)**

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

|  |
| --- |
| data(mtcars)  > table(mtcars$cyl,mtcars$am)    0 1  4 3 8  6 4 3  8 12 2 |

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

|  |
| --- |
| **hist(mtcars$wt, col= "skyblue3", breaks=5)** |

1. 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)**

1. 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)**

|  |
| --- |
| a<- c(11, 15, 23, 46, 52, 75)  > w<-c(34, 49, 58, 62, 69, 64)  > cov(a,w)  [1] 250.2  > sd(a)  [1] 24.92388  > sd(w)  [1] 12.69646  > r= cov(a,w)/(sd(a)\*sd(w))  > r  [1] 0.7906587  The r Pearson Correlation is 0.79 which means that the a and w have positive correlation. When one variable changes, the other variable changes in the same direction. |

1. 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 = cor(a,w, method= "spearman")  > rho  [1] 0.9428571  The rho Spearman correlation is equal to 0.94 which means that the association is monotonically increasing of a and w. |

1. 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)**

|  |
| --- |
|  |

# 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)**

1. 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)**

|  |
| --- |
| class(state.x77)  [1] "matrix"matrix=state.x77  dataframe=as.data.frame(matrix)  > class(dataframe)  [1] "data.frame" |

1. 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)**

|  |
| --- |
| X$states<-row.names(x)  row.names(x)<- NULL |

1. 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)**

|  |
| --- |
| Number of states: 22  Rcode: nrow(x[x$Income > 4300 & x$Population > 1000,]) |

1. 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)**

|  |
| --- |
| 1. “Alaska” 2. “Connecticut” 3. “Maryland” 4. “New Jersey” 5. “Nevada” |

1. 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)**

|  |  |
| --- | --- |
| Rcode: x$frost\_cat <- cut(x$Frost, breaks = c(-1, 30, 90, 190), labels = c("low", "intermediate", "high"))  x[x$frost\_cat == "low",]   |  | | --- | | [1] "Alabama" "Arizona" "California" "Florida" "Hawaii" "Louisiana" | |