SAINTGITS COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER APPLICATIONS

RLMCAB (2017 – 2020) Batch

RLMCA208 – Introduction to Machine Learning

Assignment #1 [Dos: 20.03.2019]

Q1. We will use a small, sample dataset. The dataset is a fabrication of some survey data from software workshops, where the variables are:

- **workshop** – *software introduced at the workshop.*
- **gender** – *gender of participant.*
- **q1** – *The instructor was well prepared.*
- **q2** – *The instructor communicated well.*
- **q3** – *The course materials were helpful.*
- **q4** – *Overall, I found this workshop useful.*

Answer the following questions.

a) We create the data in R by entering:

```
workshop <- c("R", " SPSS ", NA , " SPSS ", " STATA ", " SPSS ")
gender <- factor (c(" Female ", " Male ", NA , " Female ", " Female ," Female "))
q1 <- c(4, 3, 3, 5, 4, 5)
q2 <- c(3, 4, 2, 4, 4, 4)
q3 <- c(4, 3, NA , 5, 3, 3)
q4 <- c(5, 4, 3, 3, 4, 5)
df <- data.frame(workshop , gender , q1 , q2 , q3 , q4)
```

b) Write the R statement for creating a data frame consisting of only the first two columns.

c) Write the R statement for creating a data frame consisting of only the first and last row.

d) What happens when you enter **as.list(df)**?

e) Create a data frame called **df2** where every entry in the q3 and q4 columns is 0.

Q2. Do the following tasks.

a) Create the following data using Excel (Only the four columns need to be created. The serial numbers will automatically be added in R)

	height	shoesize	gender	population
1	181	44	male	kuopio
2	160	38	female	kuopio
3	174	42	female	kuopio
4	170	43	male	kuopio
5	172	43	male	kuopio
6	165	39	female	kuopio
7	161	38	female	kuopio
8	167	38	female	tampere
9	164	39	female	tampere
10	166	38	female	tampere
11	162	37	female	tampere
12	158	36	female	tampere
13	175	42	male	tampere
14	181	44	male	tampere
15	180	43	male	tampere
16	177	43	male	tampere
17	173	41	male	tampere

[**Remarks:** *Do not write anything on paper corresponding to this question, as this is a laboratory activity*].

b) Save the file as **students.csv** (Select the type of the file from the drop-down list at the time of saving the file) in the "My Documents" folder (this is the default working folder of R in most installations). [**Remarks:** *Do not write anything on paper corresponding to this question, as this is a laboratory activity*].

c) Open the file **students.csv** in R, and assign the file to a variable **student**. [**Remarks**: *Do not write anything on paper corresponding to this question, as this is a laboratory activity*].

d) Check if the variable student has the data, created above. [**Remarks:** *Give the command, as well as the output. Not every line needs to be written – only the first two rows and the last one row.*]

e) Print the column headings of the data set **student**. [**Remarks:** *Give the command, as well as the output.*]

f) What is the mean *height* of the students? **Ans: 169.7647** [**Remarks:** *Give the command, as well as the output.*]

g) What is the mean *shoesize*? **Ans: 40.47059** [**Remarks:** *Give the command, as well as the output.*]

h) What is the standard deviation of *height* values? **Ans: 7.578996** [**Remarks:** *Give the command, as well as the output.*]

i) What is the standard deviation of *shoesize* values? **Ans: 2.695312** [**Remarks:** *Give the command, as well as the output.*]

j) Apply the table() to the *gender* column of the data frame, and see the results. [**Remarks:** *Give the command, as well as the output.*]

k) Apply the table() to the population column. [**Remarks:** *Give the command, as well as the output.*]

l) Apply the table() to the *gender* as well as *population* columns in order to establish cross-tabulations. [**Remarks:** *Give the command, as well as the output.*]

m) Plot the histogram of the *height* column. [**Remarks:** *Give the command, as well as the plot.*]

n) Prepare a scatterplot by plotting the *height* along the X – axis and *shoesize* along the Y – axis. (**Remarks:** *Give the R command, as well as the plot.*)

o) Use the **ifelse** command to make a vector of "M", "F" values from the column *gender* of student data frame. Assign this vector to a variable **sym**.
[**Remarks:** *Give the command, and note the response from R.*]

p) Use the **ifelse** command to make a vector of "Blue","Red" values from the column population of student data frame. Assign this vector to a variable **col**.

[**Remarks:** *Give the command, and note the response from R.*]

q) Make a new data frame **student.new** by including the ***student$height***, ***student$shoesize***, ***sym*** and ***col*** values.

[**Remarks:** *Give the command, and note the response.*]

r) Print the new data frame **student.new**. [**Remarks:** *Give the command, as well as the output. Not every line needs to be written – only the first two rows and the last one row.*]

s) Print the data frame **student**. [**Remarks:** *Do not write anything on paper corresponding to this question, as this is a laboratory activity*].

t) Check to see the output of the following command.

which(student$*gender*=='male') [**Remarks:** *Give the command, as well as the output.*]

u) Check to see the output of the following command.

which(student$*gender*=='female') [**Remarks:** *Give the command, as well as the output.*]

v) Make a new data frame using the variable **student.male** to contain only the male students of the data frame student.

[**Remarks:** *Give the command, and note the response.*]

w) Print the data frame **student.male**.

[**Remarks:** *Give the command, as well as the output.*]

x) Make a new data frame using the variable **student.female** to contain only the female students of the data frame student.

[**Remarks:** *Give the command, and note the response.*]

y) Print the data frame **student.female**.

[**Remarks:** *Give the command, as well as the output.*]

Q3. Try the following commands:

> A <- **matrix**(*c(1,4,2,7,5,3), nrow=2, ncol=3)*
> A
> dim(A)

Notice how the matrix as default is filled column-wise. Try instead

> B <- **matrix**(*c(1,4,2,7,5,3), nrow=2, ncol=3, byrow=T)*
> B
> dim(B)

Make the matrices C and D in R, and check that you get the right thing:

$$C = \begin{matrix} 1 & -1 \\ -1 & 3 \end{matrix} \qquad D = \begin{matrix} 6 & 4 \\ 3 & 2 \end{matrix}$$

[**Remarks:** *Give the commands, write the outputs of A, dim(A), B, dim(B), C and D.*]

Q4. Suppose

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 3 \\ 5 & 2 & 6 \\ -2 & -1 & -3 \end{bmatrix}$$

a) Check that $\mathbf{A}^3 = \mathbf{0}$ where **0** is a 3 *x* 3 matrix with every entry equal to 0.

b) Replace the third column of A by the sum of the second and third columns.

c) Find the transpose of the matrix A using the function ***t()***.

Q5. Assume that we have the following three observations of temperature: 23°C, 27°C, 19°C. Make a vector with these values. Recall the relation between the Celsius and Fahrenheit temperature scale:

$$\text{degress in Fahrenheit} = \text{degrees in Celcisus} \cdot \frac{9}{5} + 32$$

Make a new vector with the temperatures in Fahrenheit.

Q6. Assume that you are interested in cone-shaped structures, and have measured the height and radius of 6 cones. Make vectors with these values as follows:

R <- c(2.27, 1.98, 1.69, 1.88, 1.64, 2.14)

H <- c(8.28, 8.04, 9.06, 8.70, 7.58, 8.34)

Recall that the volume of a cone with radius ***R*** and height ***H*** is given by $\frac{1}{3}\pi R^2 H$

Make a vector with the volumes of the 6 cones.

Q7. Compute the mean, median and standard deviation of the cone volumes.

Q8. Explain: the three measures of central tendency and the measures of spread with the help of suitable numeric examples. Specify the R functions that are used for these measures.

Q9. What is meant by distribution? Explain how uniform distribution differs from normal distribution.

Q10. Compare, Supervised Learning, Unsupervised Learning, Semi-supervised Learning.

Q11. How do machines learn? Explain the steps in detail with the help of relevant diagram.

Q12. Brief some applications of Principal Component Analysis.