Hellenic Open University

DATA SCIENCE AND MACHINE LEARNING (MSc)

DAMA51: Foundations in Computer Science

Academic Year: 2023–2024

#4 Written Assignment	
Submission Deadline	**Wednesday, 3 April 2024, 23:59:00 EET**

Guidelines

The deadline is definitive.

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

The assignment is due via the STUDY submission system. **You are expected to deliver a document (.DOC, .ODT, .PDF – if there is any specific preference from your tutor regarding the file format, you will be notified in advance through your class forum) <u>and</u> a compressed (.ZIP, .RAR) file containing all your work:**

- 1 document file (this document) with the answers to all the topics.
- 1 compressed file with 3 R scripts (.R files) that correspond to topics 2, 3, and 4.

You should not make any changes in the written assignment file other than providing your answers. You should also type all of your answers into Word and not attach any handwritten notes as pictures to 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 the last name Aggelou, the files should be named as follows: Aggelou-HW4.zip, Aggelou-HW4.doc, Aggelou-Topic2.R, Aggelou-Topic3.R, and Aggelou-Topic4.R. 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 **<u>three decimal places</u>**.

Topic	Points	Grades
1. Online QUIZ	40	
2. Prototype-based and k-means Clustering	20	
3. Hierarchical based Clustering using R	20	
4. Itemset Mining and Association Rules using R	20	
TOTAL	**100**	**/100**

Topic 1: Online QUIZ

Complete the corresponding online quiz available at:

```
https://study.eap.gr/mod/quiz/view.php?id=24568
```

You have one effort and unlimited time to complete the quiz, up to the submission deadline. **(40 points)**

Topic 2: Prototype-based and k-means Clustering

This topic will use the SOIL DATA GR dataset[1], which contains soil parameter data collected by agricultural farms in northern Greece. The dataset can be found at the following link:

https://study.eap.gr/mod/assign/view.php?id=23782

Read the Excel data using the **Import Dataset > From Excel…** functionality of RStudio, installing the **`readxl`** library if prompted to do so:

```
SOIL_DATA_GR <- read_excel("path/SOIL DATA GR.xlsx")
```

Note about reproducibility for the k-means algorithm: Since k-means will pseudo-randomly initialize its state, make sure that, **each time**, exactly before using the k-means algorithm, you call:

```
set.seed(123)
```

All the topics are expected to be answered using R unless explicitly stated otherwise. **(20 points)**

a. Prepare the data by performing the following steps: i) Remove the 1st column. ii) Count the number of records with NA values in the data. iii) Omit the records with NA values from the data by using the function `na.omit()`. iv) Scale the data. Then, fill in the requested information below. **(2 points)**

Answer:

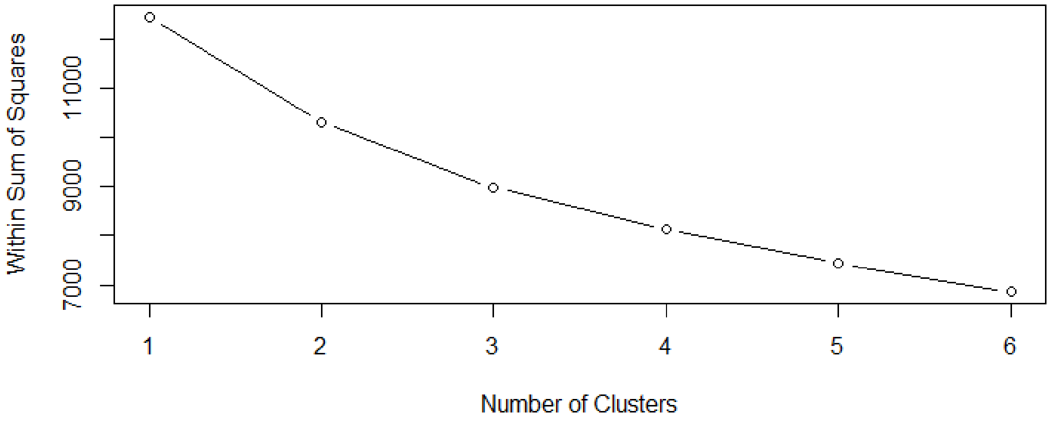
Number of omitted rows in the cleaned data	**1**
The maximum value of the **pH** attribute in the scaled data	**1.189**
The Minimum value of the **Sand %** attribute in the scaled data	**-1.96**
The median value of the **Clay %** attribute in the scaled data	**0.089**

[1] Tziachris, P., Aschonitis, V., Metaxa, E., & Bountla, A. (2022), A soil parameter dataset collected by agricultural farms in northern Greece. Data in Brief, 43, 108408. doi: 10.1016/j.dib.2022.108408)

b. Run the k-means algorithm for different values of the k parameter (from 1 to 6) and create a scree plot of the within-cluster sum of squares in relation to k. **(4 points)**

Answer:

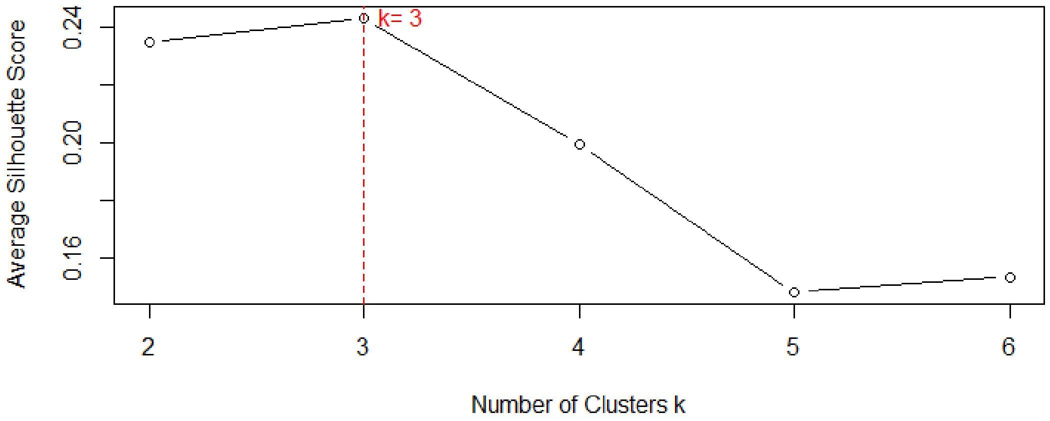
Plot:



c. Repeat the process to create a scree plot of the average silhouette score in relation to k (you will need to install the **cluster** R-package first). Which value of k is indicated as the optimal number of clusters? **(4 points)**

Answer:

Plot(3 points):

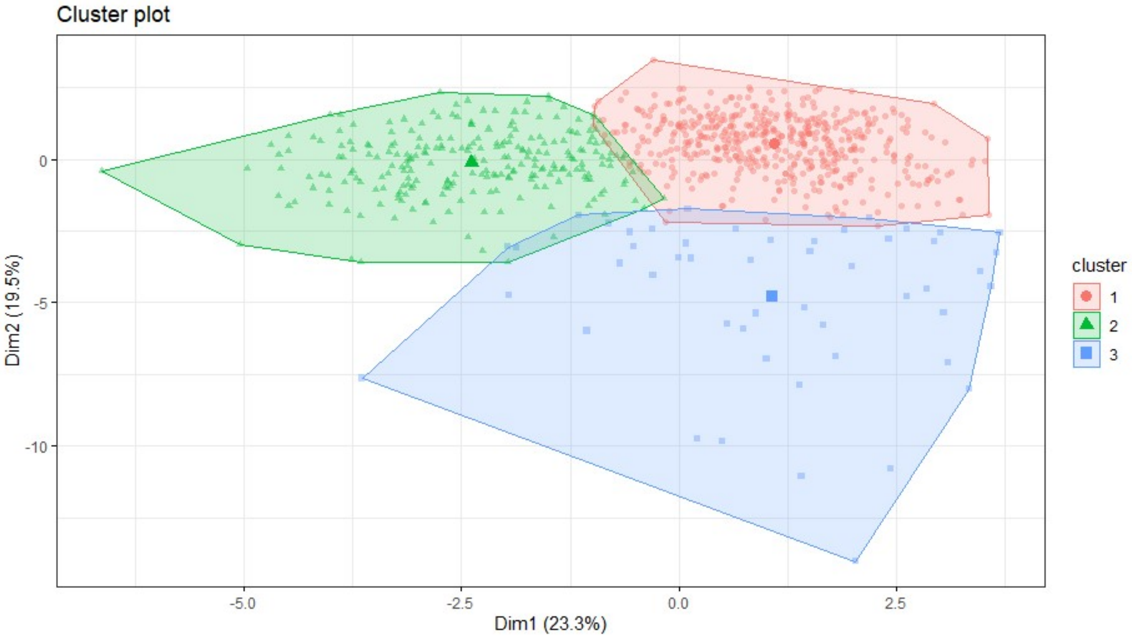


According to the plot, the optimal number of clusters is (1 point):	**3**

d. Using the function `fviz_cluster()` of the **`factoextra`** R-package (you will need to install it first, together with the **`ggplot2`** R-package), visualize the k-means clusters for k=3. Moreover, fill in the requested information in the following table. **(6 points)**

Answer:

Plot (2 points):



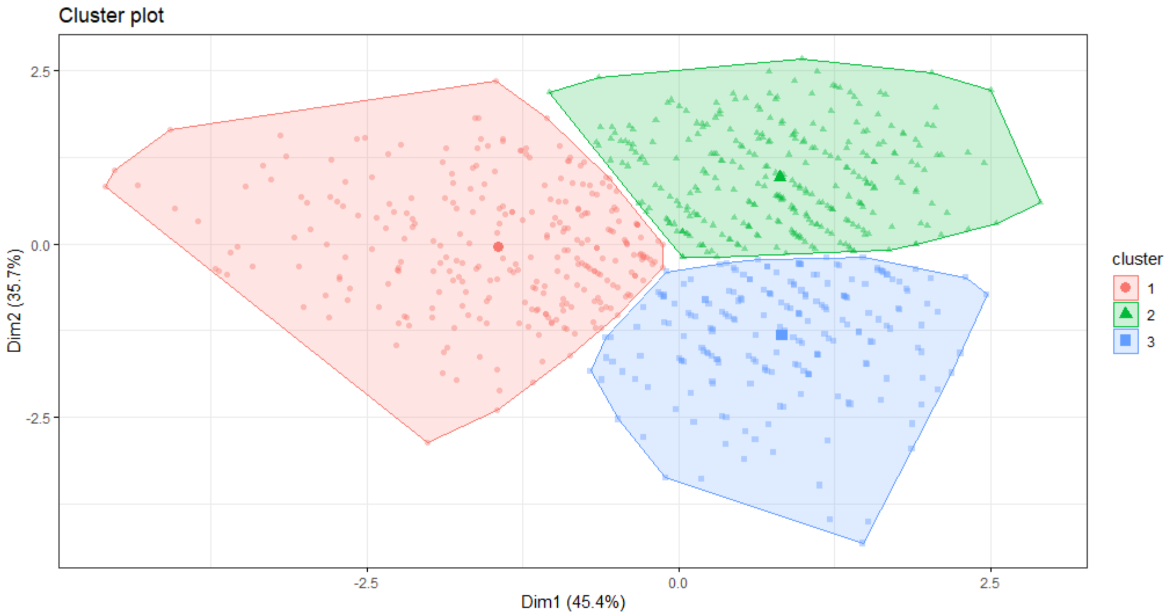
(1 point/question)

What is the value of **`pH`** for the center of cluster 1?	**0.588**
What is the value of **`Mg ppm`** for the center of cluster 2?	**0.608**
What is the number of the Cluster that the data instance of row 100 has been assigned to?	**1**
What is the number of the Cluster that the data instance of row 101 has been assigned to?	**2**

e. For this question, modify the scaled dataset so that it includes only the following 4 columns: **`Sand %`**`,` **`Clay %`**`,` **`Silt %`**, and **`pH`**. Using the function `fviz_cluster()` of the **`factoextra`** R package (make sure that have it installed first), visualize the k-means clusters for k=3. Compared with the results of question d, which clustering seems to be better? **(4 points)**

Answer:

Plot (3 points):



Better clustering result is achieved in approach (Choose between **d** or **e**) (1 point):	e

Topic 3: Hierarchical-based Clustering using R

For this topic, you will work on the ***fruits*** dataset, which can be found at the following link:

https://study.eap.gr/mod/assign/view.php?id=23782

This dataset includes records on the nutritional composition of different types of fruits. All the topics are expected to be answered using R unless explicitly stated otherwise.

a. Inspect the dataset, set the row names according to the values of the corresponding fruit name column, and then, remove this column. Proceed to scale the data and **then** fill in the requested information in the following table. **(3 points)**

Answer:

(3 points)

The Median value of the `Sugars`_g attribute	**-0.13**
The Maximum value of the Energy_kcal attribute	**2.52**
The minimum value of the Water_g attribute	**-2.554**

b. Calculate the dissimilarity distance matrices of the dataset using the Euclidean distance method. Then, fill in the following tables with the requested distances. Considering again the Euclidean distance, identify the fruit that is closest to ***Pear***. **(5 points)**

Answer:

Euclidean distance	Orange
Apple	**3.174**

Euclidean distance	Peach
Banana	**6.155**

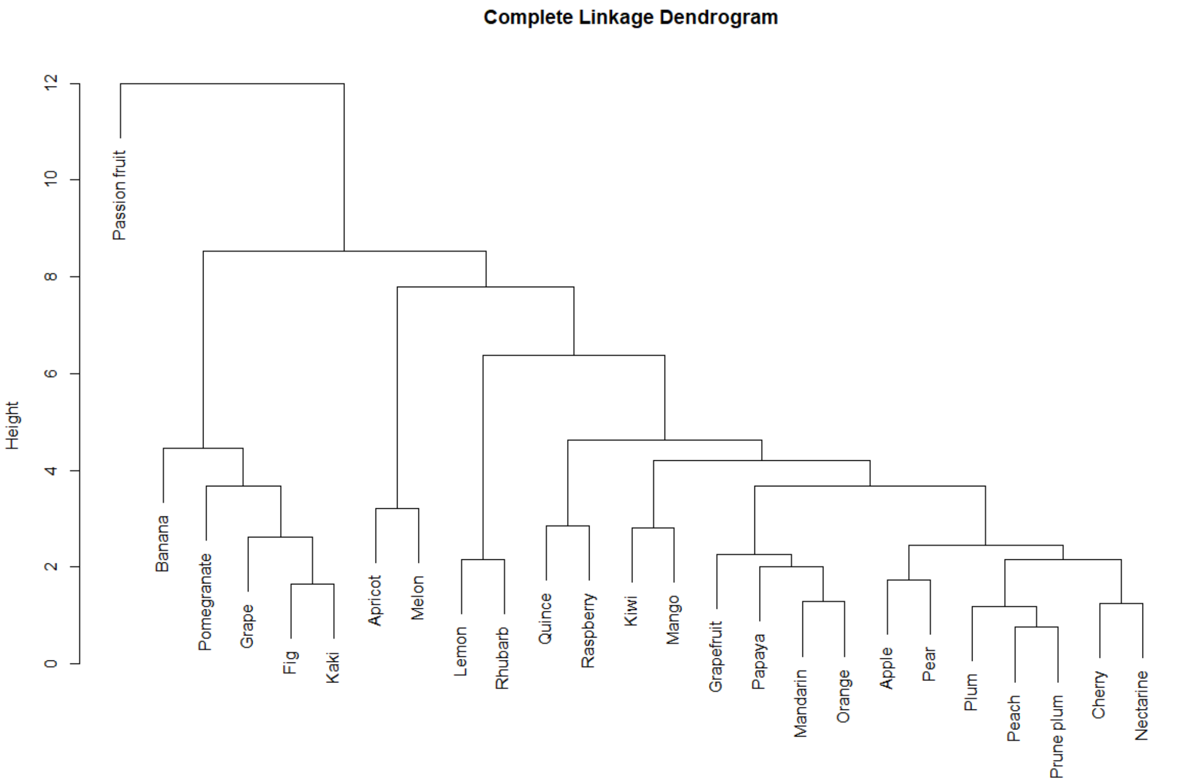
Euclidean distance	Mango
Lemon	**5.756**

	Name of closest fruit	Distance of closest fruit
The fruit closest to Pear	**Cherry**	**1.533**

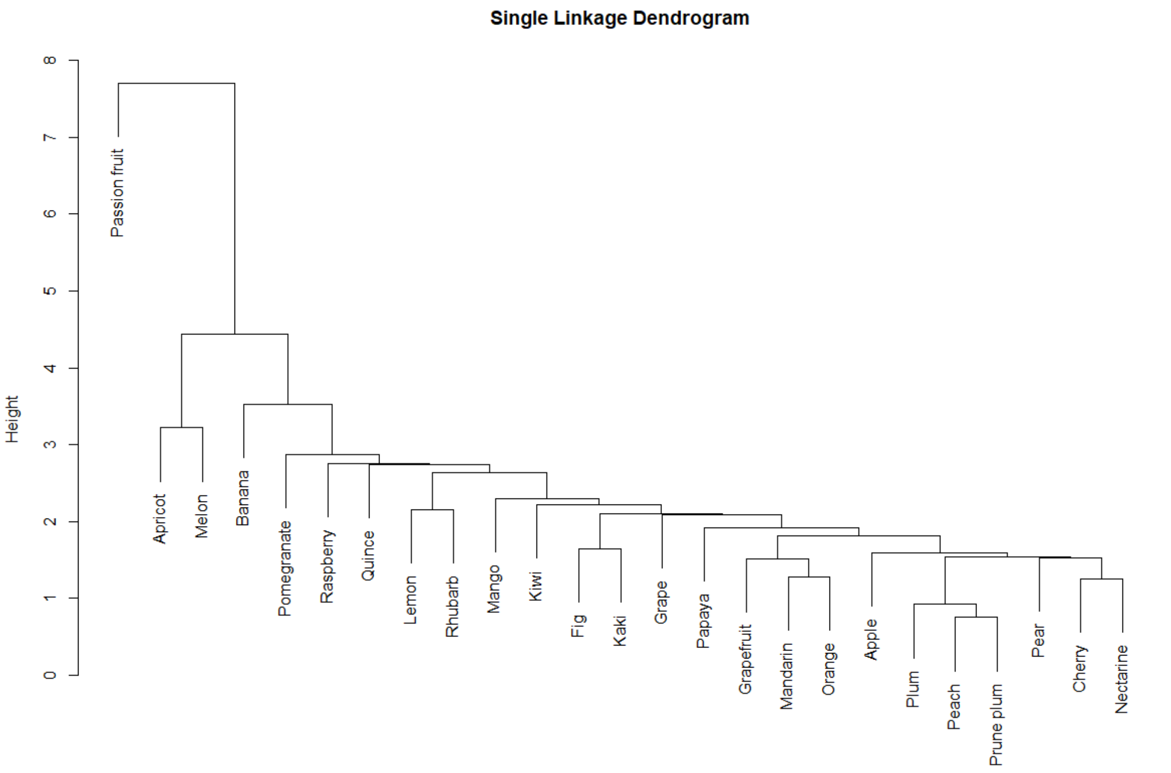
c. Perform agglomerative hierarchical clustering using the Euclidean dissimilarity distance matrix, for both complete and single linkage. Provide the dendrograms of both cases. **(4 points)**

Answer:

Plot for complete linkage (2 points):



Plot for single linkage (2 points):



d. Assume that *Orange, Grapefruit, Nectarine, Lemon*, and *Mandarin* all belong to the family of Citrus fruits. Is this also validated by your complete linkage clustering using 5 clusters, i.e. is there a cluster that includes all of these fruits? Provide the names of the fruits that have been assigned to the same cluster as *Orange*. **(3 points)**

Answer:

Mark the correct answer with an “X”:

Yes, there is a cluster that includes all of the 5 citrus fruits	
No, there is no cluster that includes all of the 5 citrus fruits	**X**

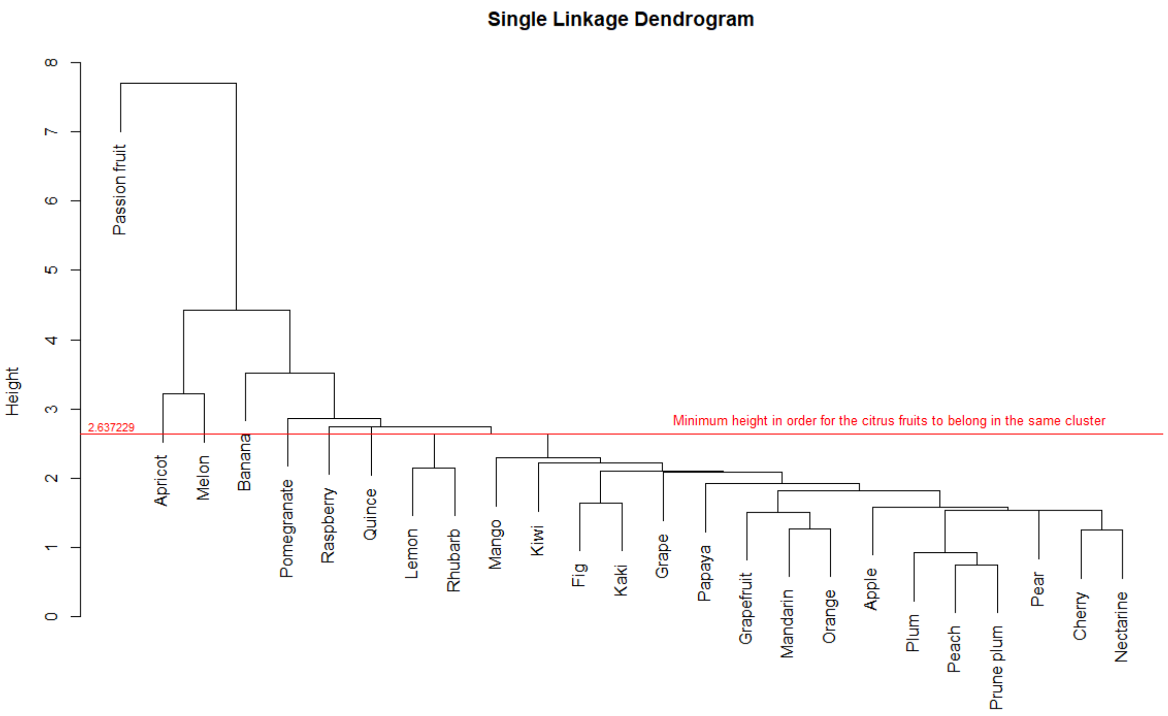
Fruits that have been assigned to the same cluster as *Orange*:
Apple, Cherry, Grapefruit, Kiwi, Mandarin, Mango, Nectarine, Orange, Papaya, Peach, Pear, Plum, Prune plum, Quince, Raspberry

e. Using **single** linkage clustering, identify the maximum number of clusters for which all of the Citrus fruits (*Orange, Grapefruit, Nectarine, Lemon*, and *Mandarin*) belong to the same cluster. Also, identify the minimum height at which the dendrogram needs to be cut, in order for all the Citrus fruits to be clustered together. Finally, plot the dendrogram depicting with a horizontal line this minimum height at which the dendrogram needs to be cut. **(5 points)**

Answer:

The maximum number of clusters is:	**8**
The corresponding height is:	**2.637**

Plot (single linkage) the dendrogram with the horizontal line (3 points):



Topic 4: Itemset Mining and Association Rules using R

For this topic, you will work on the ***countries*** dataset which can be found at the following link:

https://study.eap.gr/mod/assign/view.php?id=23782

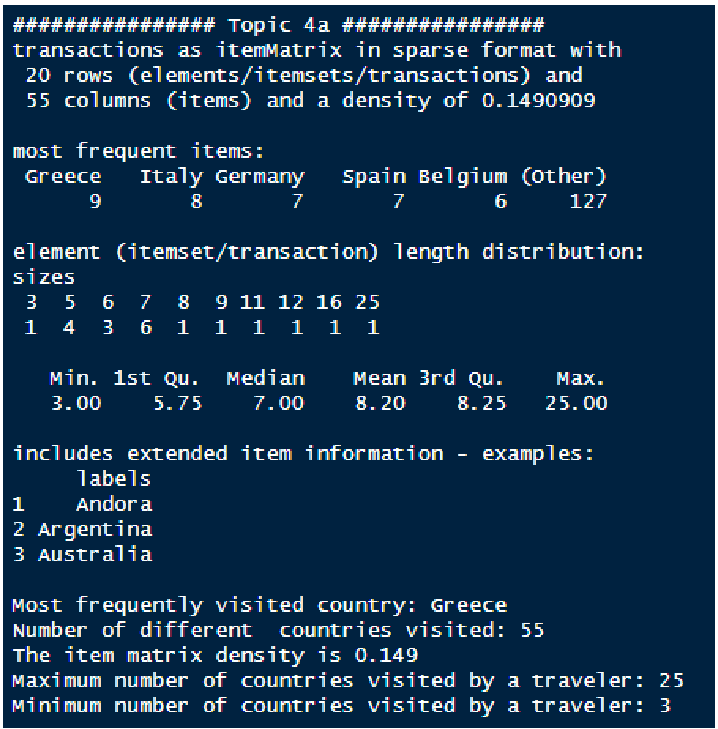
This dataset includes records of the countries that different travelers have visited. Each record (transaction) includes the set of countries visited by each traveler.

After first installing and loading the **arules** library, read the file using the **read.transactions** function as follows:

```
visits <- read.transactions("path/countries.csv", format="basket",
header=FALSE, sep=",", rm.duplicates=FALSE)
```

a. Inspect the dataset and provide the information requested in the following table. (**5 points**)

Answer:

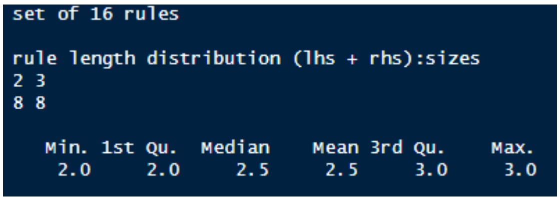


```
################ Topic 4a ################
transactions as itemMatrix in sparse format with
 20 rows (elements/itemsets/transactions) and
 55 columns (items) and a density of 0.1490909

most frequent items:
 Greece   Italy Germany   Spain Belgium (Other)
      9       8       7       7       6     127

element (itemset/transaction) length distribution:
sizes
 3  5  6  7  8  9 11 12 16 25
 1  4  3  6  1  1  1  1  1  1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.00    5.75    7.00    8.20    8.25   25.00

includes extended item information - examples:
     labels
1    Andora
2 Argentina
3 Australia

Most frequently visited country: Greece
Number of different  countries visited: 55
The item matrix density is 0.149
Maximum number of countries visited by a traveler: 25
Minimum number of countries visited by a traveler: 3
```

Most frequently visited country	**Greece**
Number of different countries visited	**55**
Item matrix density	**0.149**
Maximum number of countries visited by a traveler	**25**
Minimum number of countries visited by a traveler	**3**

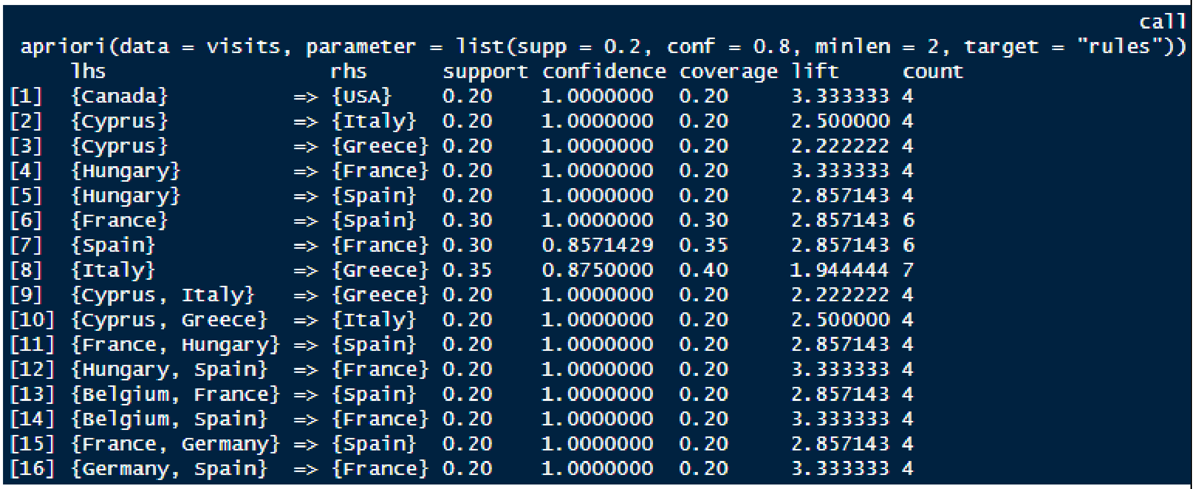
b. Run the Apriori algorithm for a minimum support threshold of 0.2, a minimum confidence threshold of 0.8, and a minimum of 2 items involved in a rule. Fill in the information in the following table. Then, inspect the rules identified and fill in the missing information denoted with a question mark in the next table. In case a rule does not exist, write “N/A” in place of the question mark. **(7 points**)

Answer:



```
set of 16 rules

rule length distribution (lhs + rhs):sizes
2 3
8 8

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     2.0     2.5     2.5     3.0     3.0
```

Number of identified rules	16
Number of rules with maximum number of items involved	8
Number of rules with minimum number of items involved	8



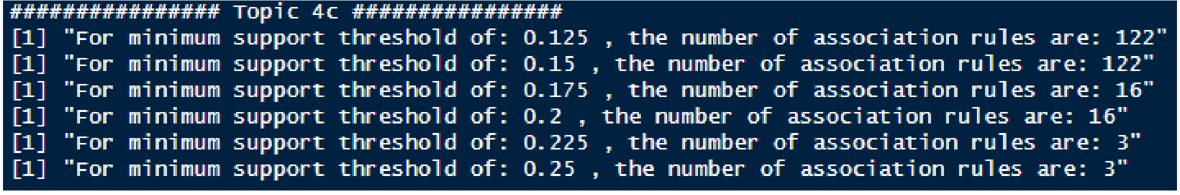
```
                                                                                  call
 apriori(data = visits, parameter = list(supp = 0.2, conf = 0.8, minlen = 2, target = "rules"))
     lhs                   rhs      support confidence coverage lift     count
[1]  {Canada}           => {USA}    0.20    1.0000000  0.20     3.333333 4
[2]  {Cyprus}           => {Italy}  0.20    1.0000000  0.20     2.500000 4
[3]  {Cyprus}           => {Greece} 0.20    1.0000000  0.20     2.222222 4
[4]  {Hungary}          => {France} 0.20    1.0000000  0.20     3.333333 4
[5]  {Hungary}          => {Spain}  0.20    1.0000000  0.20     2.857143 4
[6]  {France}           => {Spain}  0.30    1.0000000  0.30     2.857143 6
[7]  {Spain}            => {France} 0.30    0.8571429  0.35     2.857143 6
[8]  {Italy}            => {Greece} 0.35    0.8750000  0.40     1.944444 7
[9]  {Cyprus, Italy}    => {Greece} 0.20    1.0000000  0.20     2.222222 4
[10] {Cyprus, Greece}   => {Italy}  0.20    1.0000000  0.20     2.500000 4
[11] {France, Hungary}  => {Spain}  0.20    1.0000000  0.20     2.857143 4
[12] {Hungary, Spain}   => {France} 0.20    1.0000000  0.20     3.333333 4
[13] {Belgium, France}  => {Spain}  0.20    1.0000000  0.20     2.857143 4
[14] {Belgium, Spain}   => {France} 0.20    1.0000000  0.20     3.333333 4
[15] {France, Germany}  => {Spain}  0.20    1.0000000  0.20     2.857143 4
[16] {Germany, Spain}   => {France} 0.20    1.0000000  0.20     3.333333 4
```

lhs	rhs	support	confidence	coverage	lift
{Belgium, Spain}	France	0.20			
{Hungary}	{Spain, France}		N/A		
{Belgium}	{Spain}			N/A	
{Cyprus}	{Greece}				2.222

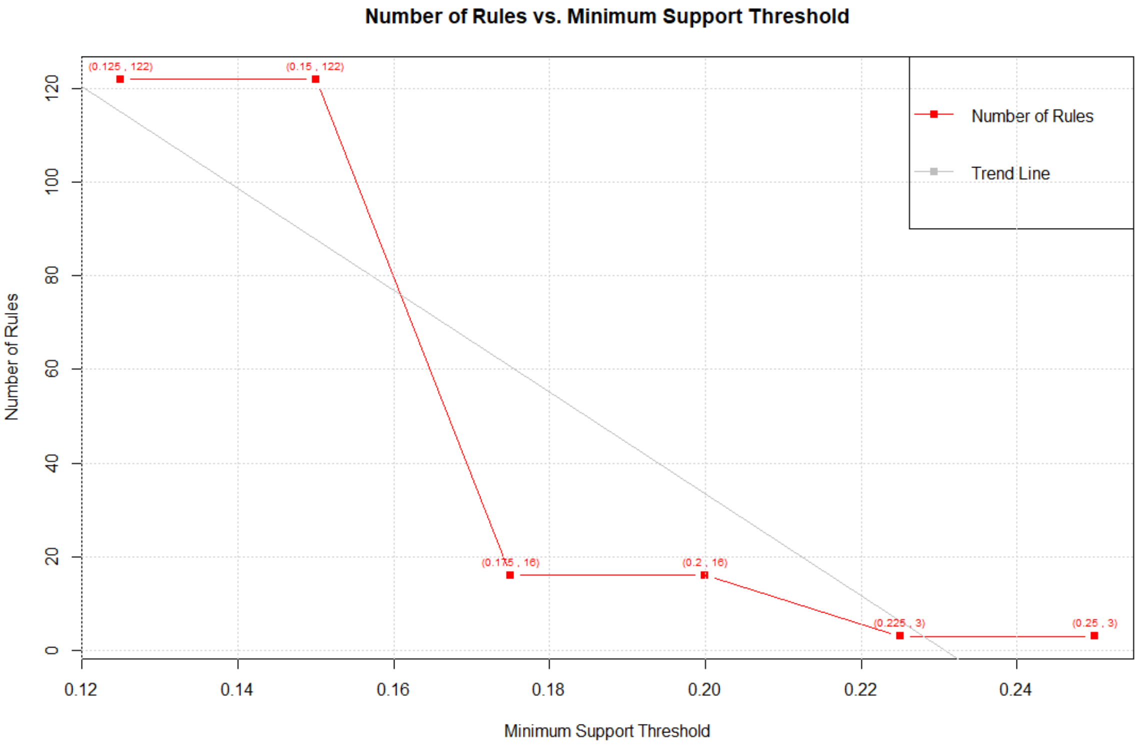
c. Run the Apriori algorithm for a minimum confidence threshold of 0.8, a minimum of 2 items involved in a rule, and for values of the minimum support threshold ranging from 0.125 up to 0.25 with a step of 0.025. How does the number of association rules change in relation to the minimum support threshold value? Explain. **(4 points)**

Answer:

The plot is as follows (3 points)**:**



```
################ Topic 4c ################
[1] "For minimum support threshold of: 0.125 , the number of association rules are: 122"
[1] "For minimum support threshold of: 0.15 , the number of association rules are: 122"
[1] "For minimum support threshold of: 0.175 , the number of association rules are: 16"
[1] "For minimum support threshold of: 0.2 , the number of association rules are: 16"
[1] "For minimum support threshold of: 0.225 , the number of association rules are: 3"
[1] "For minimum support threshold of: 0.25 , the number of association rules are: 3"
```

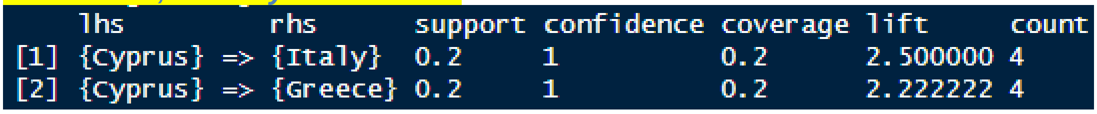


Explanation (1 point)**:**
Lowering the support threshold allows for more itemsets to be considered frequent and thus eligible for generating association rules. Therefore, we observe that less rules are being generated while we increase the minimum support threshold. That can also be confirmed by the trend line which slopes downward, indicating a negative correlation between the two variables.

d. Identify all countries that are included in the consequent in the rules where *Cyprus* is the antecedent (minimum support threshold of 0.2, a minimum confidence threshold of 0.8, and a minimum of 2 items involved in a rule). **(2 points)**

Answer:

The countries that are included in the consequent in the rules, where Cyprus is the antecedent, are Italy and Greece.



```
    lhs           rhs      support confidence coverage lift     count
[1] {Cyprus} => {Italy}  0.2     1          0.2      2.500000 4
[2] {Cyprus} => {Greece} 0.2     1          0.2      2.222222 4
```

e. How can the association rule '{Cyprus} => {Greece}' with a high confidence level be interpreted in the context of traveler patterns, and what implications does this rule have for understanding cultural or geographical affinities between countries in the dataset? **(2 points)**

Answer:

The rule suggests that there is strong indication that a significant proportion of travelers who visit Cyprus also tend to visit Greece. Geographical proximity, cultural similarities, historical ties or even shared language, traditions, lifestyle and cuisine could be leading to a natural flow of tourists between the two countries.