STA 6100 PROJECT

AUTHOR: CHUKWUNONSO AKAEME

JUNE 1, 2023

UNIVERSITY OF SOUTHERN QUEENSLAND
STUDENT ID: 0061156572

Table of Contents

PART A

TASK 1

Action 1:

Final structure of data: After cleaning the river data set that originally had 200 cases, there are now a total of 187 cases (rows/rivers) each containing 14 values (columns/variables). The first 3 variables in the data set are in characters consisting of seasons (autumn, spring, summer, and winter), river size (large, medium, and small), and fluid velocity (low, medium, and high) of the rivers. The next six variables are the chemical concentrations which are numerical in values, namely; nitrogen, nitrates, nitrites, ammonia, phosphate, and oxygen (all measured in mg/L). The last five values of each row are the amount of different kinds of algae which are numerical and are represented as: A1, A2, A3, A4 and A5.

Number of cases removed due to missing data: A total of 13 cases were removed due to missing data.

Action 2:

Frequency table showing the number of rivers in the cleaned sample:

```
[1] 187
```

Comment: there are 187 rivers in the river data set.

Frequency table showing the number of rivers measured in each season:

```
autumn spring summer winter 
    37     48     43     59 
```

Comment: 37 rivers were measured during the autumn season. 48 rivers were measured in spring, which is the second highest number among all seasons, 43 rivers were measured during the summer season and the winter season had the highest number of rivers measured, with 59 rivers included in the sample.

Frequency table showing the number of rivers in each category of river size:

```
large medium  small 
  43     84     60 
```

Comment: There are 43 rivers classified as "large" in terms of size. Majority of rivers fall into the "medium" category, with 84 rivers classified as such and lastly, there are 60 rivers in the "small" size category.

Summary Statistics of each of the chemical variable:

```
nitrogen          nitrates           nitrites          ammonia
 Min.   : 1.500    Min.   :  0.222    Min.   : 0.050    Min.   :   5.80
 1st Qu.: 7.700    1st Qu.: 11.044    1st Qu.: 1.324    1st Qu.:  46.35
 Median : 9.800    Median : 34.037    Median : 2.805    Median : 110.00
 Mean   : 9.053    Mean   : 44.232    Mean   : 3.124    Mean   : 413.90
 3rd Qu.:10.735    3rd Qu.: 58.090    3rd Qu.: 4.521    3rd Qu.: 234.35
 Max.   :13.400    Max.   :391.500    Max.   :10.416    Max.   :8777.60

   phosphate           oxygen
 Min.   :  1.25    Min.   :  2.50
 1st Qu.: 18.46    1st Qu.: 49.67
 Median : 45.00    Median :111.75
 Mean   : 77.26    Mean   :144.62
 3rd Qu.:102.35    3rd Qu.:218.95
 Max.   :564.60    Max.   :771.60
```

Comment: The table contains the summary statistics that provide an overview of the distribution of all the chemical variables in the river data set. The minimum value, first quartile(1st Qu.), ,median, mean, third quartile (3rd Qu.), and maximum values are given for each variable. For ammonia, phosphate and oxygen, the wide range between the first and third quartiles suggests significant variability in their chemical concentrations. The mean values represent the average level of the chemical concentrations in the rivers, all of which are measured in (mg/L).

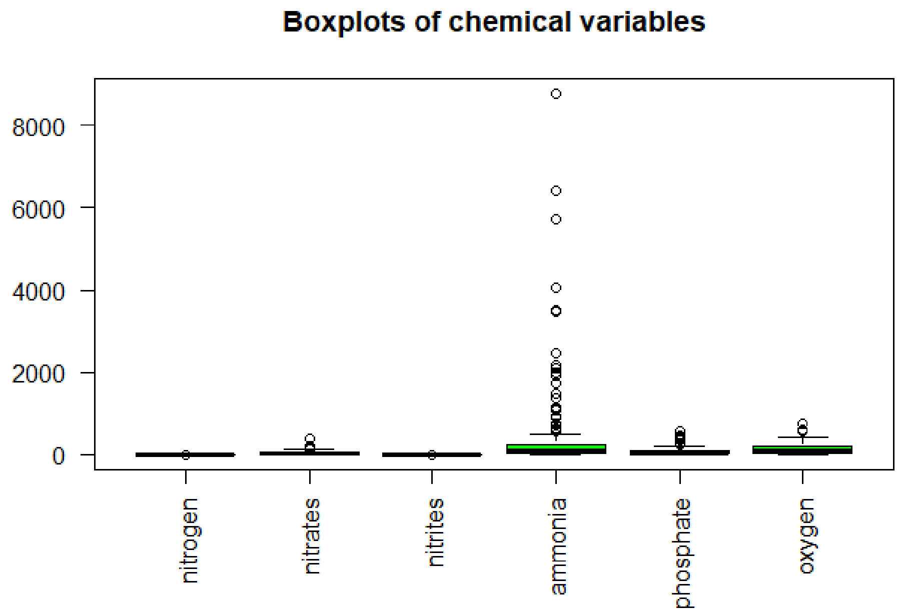
Summary Statistics of each of the algae variable:

```
 A1               A2                A3               A4
 Min.   : 0.00    Min.   : 0.000    Min.   : 0.00    Min.   : 0.000
 1st Qu.: 1.40    1st Qu.: 0.000    1st Qu.: 0.00    1st Qu.: 0.000
 Median : 5.70    Median : 3.400    Median : 1.70    Median : 0.000
 Mean   :15.46    Mean   : 7.765    Mean   : 4.56    Mean   : 1.602
 3rd Qu.:19.65    3rd Qu.:11.800    3rd Qu.: 5.45    3rd Qu.: 2.400
 Max.   :89.80    Max.   :72.600    Max.   :42.80    Max.   :12.700

 A5
 Min.   : 0.000
 1st Qu.: 0.000
 Median : 2.500
 Mean   : 5.405
 3rd Qu.: 7.850
 Max.   :44.400
```

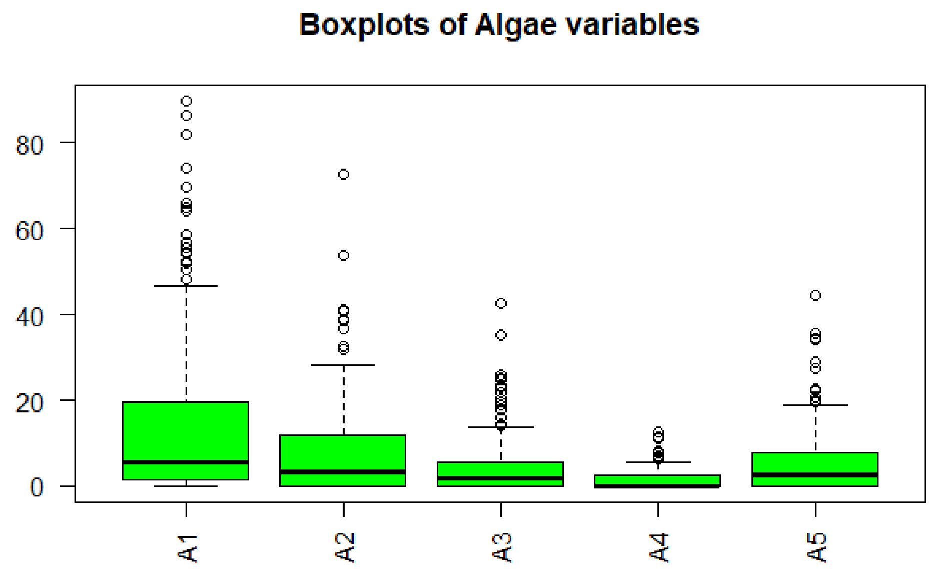
Comment: The table contains the summary statistics that provide an overview of the distribution of all the algae variables in the river data set. The minimum value, first quartile(1st Qu.), median, mean, third quartile (3rd Qu.), and maximum values are given for each variable. All of the algae variables have a minimum zero value; *A2*, *A3* and *A5* have a minimum of zero value for their minimum value and first quartile while *A4* has zero values for its minimum, first quartile and median. This indicates that at least 50% of the data for these variables is equal to zero due to low frequency.

Boxplots of chemical variables:



Interpretation: The boxplots reveal that the chemical variable, *ammonia,* has a wider range of values and multiple outliers. *oxygen* and *phosphate* also have outliers present with *oxygen* having the second highest range of values amongst the others; *nitrates* have fewer outliers and *nitrites* and *nitrogen* have the smallest range of values compared to other chemical variables indicating that there is little to no variability in their chemical concentrations.

Boxplots of algae variables:



Interpretation: The boxplots reveal that the algae variable, *A1* has the widest range of values and multiple outliers extending from its whiskers, *A2, A5,* and *A3* closely follow *A1* in terms of their frequency range and presence of outliers. *A4* has fewer outliers and the lowest range of values indicating that nearly half its value is zero due to low frequency.

TASK 2

Action 1:

Frequency table of the number of rivers in the combination of river size and fluid velocity:

```
river_size_vel

   large_high      large_low 
            7             16 
 large_medium    medium_high 
           20             34 
   medium_low medium_medium 
           15             35 
   small_high   small_medium 
           38             22 
```

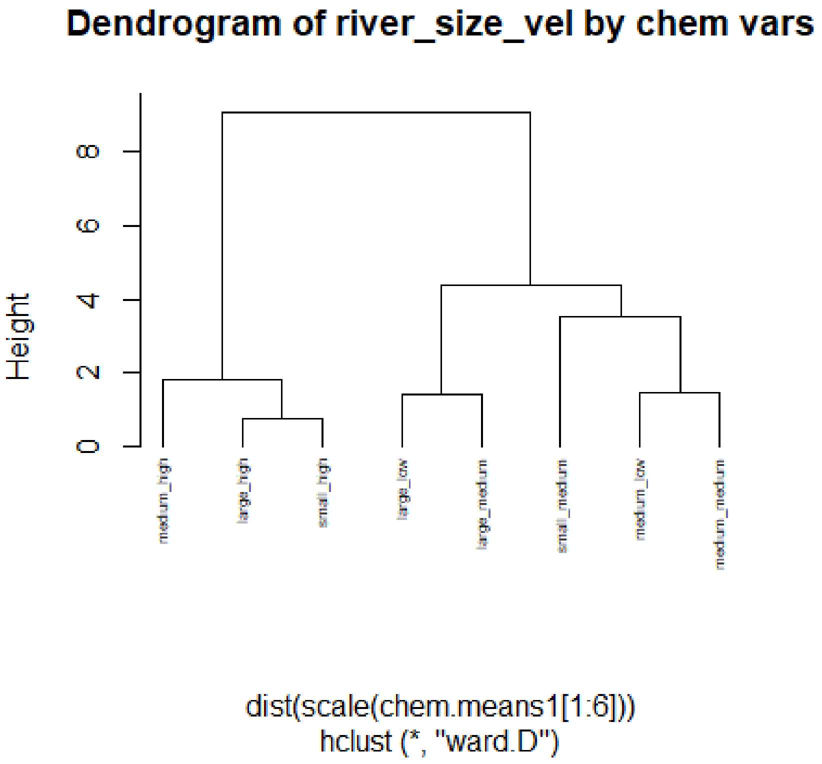
Comment: It appears that the most common combination of river size and fluid velocity in the sample is *small_high*, with 38 rivers falling into this category. The second most common combination is *medium_medium*, with 35 rivers. The least common combination is *large_high*, with only 7 rivers falling into this category.

Additionally, for large and medium-sized rivers, the most common fluid velocity is *medium*, while for small rivers, the most common fluid velocity is *high*. There are more small and medium-sized rivers than large rivers in the sample.

It's important to note that the characters of river size are large, medium, and small and the characters of fluid velocity is low, medium, and high.

Action 2:

Dendrogram representing relationships between the groups of river_size_vel:



Description of Method:

To carry out this action, we have employed cluster analysis. we first calculated the means of the chemical variables grouped by the *river_size_vel* variable and merged into a new data frame then calculated the Euclidean distance matrix between the group means using the ***dist*** function followed by the ***scale*** function to standardize it as this is an important step in cluster analysis. Next, we performed hierarchical clustering using ***ward's method*** by passing the distance matrix to the ***hclust*** function. Finally, we plotted the resulting dendrogram using the ***plot*** function.

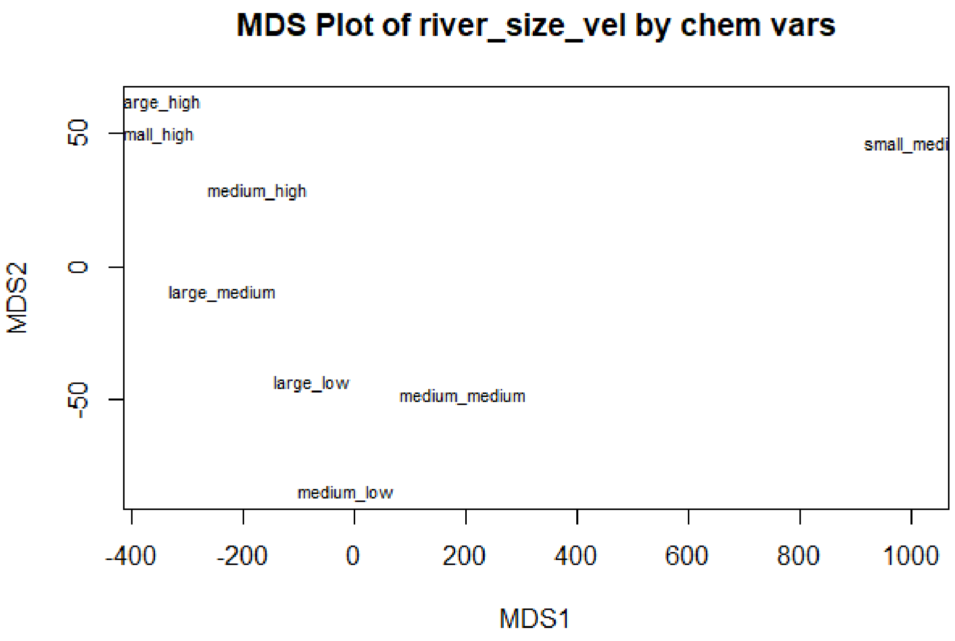
Ward's method was chosen because it minimizes the total within-cluster variable at each step to produce compact and well-separated clusters. The use of ***Euclidean*** distance was because it helps in measuring the dissimilarity between observations in multivariate analysis, it is also simple to calculate and easy to interpret.

Interpretation of Dendrogram:

The resulting dendrogram shows how similar or dissimilar each category of ***river_size_vel*** is based on their chemical concentrations. The height of each branch represents the distance between clusters and this distance is a measure of dissimilarity. Based on this dendrogram, categories with the shortest distance between them are merged. It appears the ***small_medium*** and ***medium_high*** groups are the most dissimilar from other groups of the ***river_size_vel*** variable, as they form their own cluster without any close pairs. This suggests that the chemical concentrations in this categories are significantly different from the other categories. ***large_high and small_high, large_low and _large_medium,*** and ***medium_low*** and ***medium_medium*** are the pairs of groups that are most similar indicating that each pair have similar chemical concentrations.

Action 3:

MDS plot representing relationships between the groups of river_size_vel:



Description of Method:

For this action, we have employed multidimensional scaling which is a technique that takes a matrix of dissimilarities (in this case, Euclidean distances) and returns a set of points in a lower-dimensional space (usually two dimensions for visualization purposes) such that the distances between the points are approximately equal to the dissimilarities. In other words, it tries to represent the dissimilarities between the observations as distances between points in a lower-dimensional space.

we chose to use metric scaling because it preserves the original distances as closely as possible. Non-metric scaling, on the other hand, only tries to preserve the rank order of the dissimilarities and can be used when the dissimilarities are not meaningful on an interval or ratio scale.

Interpretation of MDS plot:

The resulting MDS plot shows how similar or dissimilar each category of ***river_size_vel*** is based on their chemical concentrations.

The groups ***'large_high'*** and ***'small_high'*** are the most similar to each other compared to other categories. They appear to be closely knitted on the MDS plot, indicating significant level in their chemical concentrations. similarly, the groups ***, large_low and _large_medium, medium_low*** and ***medium_medium*** are the pairs of groups that are also closely represented on the plot also indicating that each pair have similar chemical concentrations.

By contrast, the ***medium_high*** groups are dissimilar from other groups of the ***river_size_vel*** variable in the MDS plot. They are scattered individually suggesting that they have distinct chemical concentrations compared to the other groups.

The group ***small_medium*** is the most dissimilar group in the MDS plot. It is located far away from all other groups at a dimension of 1000, indicating a high level of dissimilarity in terms of its chemical concemtrations.

Overall, the MDS plot provides a visual representation of the similarity and dissimilarity between the 8 groups of the 'river_size_vel' variable with higher dimensions the axis representing the mean range of values of the chemical concentration and indicating metric relationship between the groups

Action 4:

Reason for providing a dendrogram and MDS plot

We have provided them both in this case to help establish a deeper understanding of the connections between these groups giving both a dendrogram and an MDS plot.

This is because these approaches employ the distance matrix in different ways. A dendrogram is a type of tree diagram that displays how data points are hierarchically clustered according to their pairwise distances. It uses the distance matrix to pair up similar groups and represent their relationships in a hierarchical structure. MDS (Multidimensional Scaling), on the other hand, is a method that depicts data points as points in a low-dimensional space while trying to preserve their pairwise distances. It does this by using an iterative optimization technique to find a configuration of points in the low-dimensional space that best represents the distances between the data points.

In conclusion, A visual depiction of the pairwise distances between groups is provided by the MDS plot, whereas the dendrogram reveals the hierarchical structure of the groups. Insights into the data may be obtained using both of these in combination, hence the reason for provision.

TASK 3:

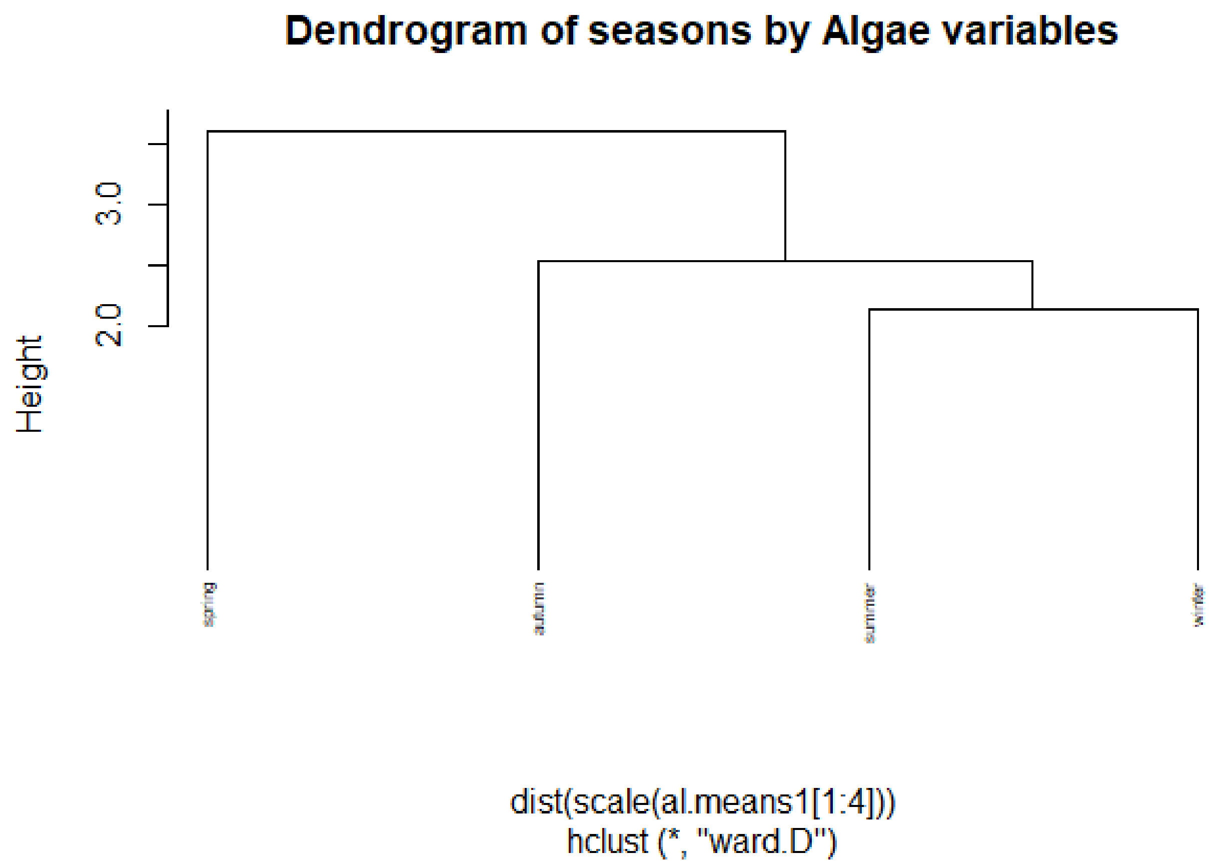
Action 1:

Structure of the cases available for algae variables:

The dataset contains 187 observations, with each observation representing a frequency of algae. Each algae variable has corresponding numeric values associated with it, indicating the frequencies of the respective algae variable.

In summary, the dataset consists of five algae variables (A1, A2, A3, A4, A5) with numeric values for each observation, providing information about the levels of these variables.

Dendrogram representing relationships between the groups of river_size_vel:



Description of Method:

To carry out this action, we have employed cluster analysis. we first calculated the means of the algae variables grouped by the *season* variable and merged into a new data frame then calculated the Euclidean distance matrix between the group means using the ***dist*** function followed by the ***scale*** function to standardize it as this is an important step in cluster analysis. Next, we performed hierarchical clustering using ***ward's method*** by passing the distance matrix to the ***hclust*** function. Finally, we plotted the resulting dendrogram using the ***plot*** function.

Ward's method was chosen because it minimizes the total within-cluster variable at each step to produce compact and well-separated clusters. The use of ***Euclidean*** distance was because it helps in measuring the dissimilarity between observations in multivariate analysis, it is also simple to calculate and easy to interpret.

Interpretation of Dendrogram:

The resulting dendrogram shows how similar or dissimilar each ***season*** is based on their algae variables. The height of each branch represents the distance between clusters and this distance is a measure of dissimilarity. Based on this dendrogram, categories with the shortest distance between them are merged. It appears the ***summer*** and ***winter*** seasons are the most similar amongst the other groups of the ***season*** variable, as they pair first forming their own cluster. The most dissimilar of the seasons is spring towering over the others and forming its own cluster without any close pairs and it is closely followed by the ***autumn*** season in terms of dissimilarity, standing alone, forming no close pair.

Action 2:

Table of mean frequencies for algae variables by each season:

```
       A1       A2       A3       A4       A5
Autumn 16.63784 9.918919 1.572973 6.643243 6.643243
Spring 12.67708 7.183333 7.616667 3.654167 3.654167
Summer 15.32093 6.732558 3.151163 7.560465 7.560465
Winter 17.07797 7.640678 4.974576 4.483051 4.483051
```

Relationship of mean frequencies with clustering seasons within dendrogram:

The table of means provides insights into the average frequencies of each algae variable (A1, A2, A3, A4, A5) for each season (Autumn, Spring, Summer, Winter). These mean values can help explain the clustering of seasons within the dendrogram by identifying patterns and differences in the algae variables across seasons.

1. Autumn: The mean frequencies for A1 and A4 are relatively high compared to the other seasons, indicating a higher abundance of these variables during autumn. This could contribute to the clustering of autumn with other seasons that have similar patterns.

2. Spring: The mean frequencies for A2, A3, and A4 are relatively high compared to the other seasons. This suggests a higher presence of these variables during spring, which can contribute to the clustering of spring with other seasons that exhibit similar patterns.

3. Summer: The mean frequencies for A1, A4, and A5 are relatively high compared to the other seasons, indicating a higher abundance of these variables during summer. This similarity in algae variable frequencies may contribute to the clustering of summer with other seasons that exhibit similar patterns.

4. Winter: The mean frequencies for A1, A2, and A3 are relatively high compared to the other seasons. This suggests a higher presence of these variables during winter, which can contribute to the clustering of winter with other seasons that exhibit similar patterns.

PART B

Question 1: Complete table of relevant features matched to the listed method of analysis

Feature	MANOVA	PCA	FA	DFA	CCA	CA	MDS
Eigen analysis	Yes	Yes	Yes	No	Yes	No	No
Distance matrix	No	Yes	No	No	No	Yes	Yes
Dimension reduction	No	Yes	Yes	No	Yes	Yes	Yes
Classification	No	No	No	Yes	No	No	No
Can be used to determine group structure	Yes	No	No	Yes	No	Yes	No
Needs independent group provided in the data	Yes	No	No	Yes	No	No	No
Ordination method	No	Yes	No	No	No	No	Yes

Question 2: Limitations of multivariate methods

- To generate reliable results using multivariate methods, it may require having a large sample size.
- Large datasets require a lot of computation when performing multivariate analyses.
- Multivariate methods assume linearity and normality in the data, which is not always the case.
- Outliers in the data can affect multivariate techniques.
- If the number of variables is greater than the sample size, multivariate techniques may suffer from overfitting.
- Statistical expertise is needed to interpret multivariate results, which can be complex and challenging.

Question 3: Mantel's test and calculation of the significance of its test statistic

The connection or correlation between two distance matrices is evaluated using Mantel's randomisation test. It is frequently used in genetic and ecological research. The test is used when comparing two distance matrices (such as dissimilarity matrices) to see if there is a relationship or similarity. it is frequently applied to determine whether two ecological groups or spatial patterns are comparable.

The significance of the test statistics is determined by randomisation or permutation. Calculated correlation between the two sets of distances serves as both the test statistic and the correlation measure supplied for the test. Randomisation yields the null distribution and significance level.

Question 4: Eigenvectors and Eigenvalues

In multivariate analysis methods like Principal Component Analysis, eigenvectors and eigenvalues are used and they also often appear in important concepts of linear algebra.

Eigenvectors are vectors that create a scaled replica of themselves when multiplied by a square matrix. They describe axes or directions that remain the same regardless of the linear transformation applied. In Principal Component Analysis, eigenvectors, which reflect the directions of the largest variance in the dataset, are frequently connected to principal components.

Eigenvalues are scalar values of the eigenvector. They serve as a representation of the linear transformation's scaling factor, which is applied to the associated eigenvector. The variance or significance of the related eigenvectors is shown by eigenvalues. Eigenvalues are used in Principal Component Analysis to calculate the percentage of variation that is accounted for by each principal component.

Question 5: Hand calculation of Euclidian distances

To calculate the Euclidean distance between individuals 1 and 3 for all X variables, we can use the following formula:

$$d_{13} = \sqrt{\{(\mathrm{x}_{11} - x_{31})^2 + (\mathrm{x}_{12} - x_{32})^2 + (\mathrm{x}_{13} - x_{33})\}^2}$$

Where $\mathrm{x}_{11}, \mathrm{x}_{12}\ and\ \mathrm{x}_{13}$ are the coordinates of individual 1 and $x_{31}, x_{32}\ and\ x_{33}$ are the coordinates of individual 3.

d_{13} = $\sqrt{\{(-0.214\ -\ \ 1.763)^2 + (-0.36\ -\ 1.37)^2\ +\ (0.03 -\ (-1.97))\}^2}$

$d_{13}\ = \sqrt{\{(-1.977)^2 + (-\ 1.73)^2\ +\ (2)\}^2}$

$d_{13}\ = \sqrt{3.909 + 2.992\ +\ 4}$

$d_{13}\ = \sqrt{10.901}$

$d_{13}\ =$ 3.301 ≅ 3.3

Therefore, the Euclidian distance between individuals 1 and 3 for all X variables is approximately **3.3.**