Multivariate Statistical Analysis-2

Mini Project

Amrita V:21060641004
Samruddhi Hindlekar:21060641041
Shruti Deshmukh:21060641047
Vamsikrishna A:21060641055

Steps:

Data Manipulation

Unneccesary columns were removed and converted into proper data type

Clustering

The clustering of the data is done

PCA

The Principle Component Analysis of the data is done

Interpretation

The obtained results and graphs are interpreted.

Introduction

Clustering

- K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters.
- K defines the number of pre-defined clusters that need to be created in the process

PCA

- Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning.
- It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.

Data Introduction

- The given dataset is of codon usage frequencies in the genomic coding DNA of a large sample of diverse organisms from different taxa as tabulated in the CUTG database.

- Total 69 columns &

- total of 13028 entries in the dataset.

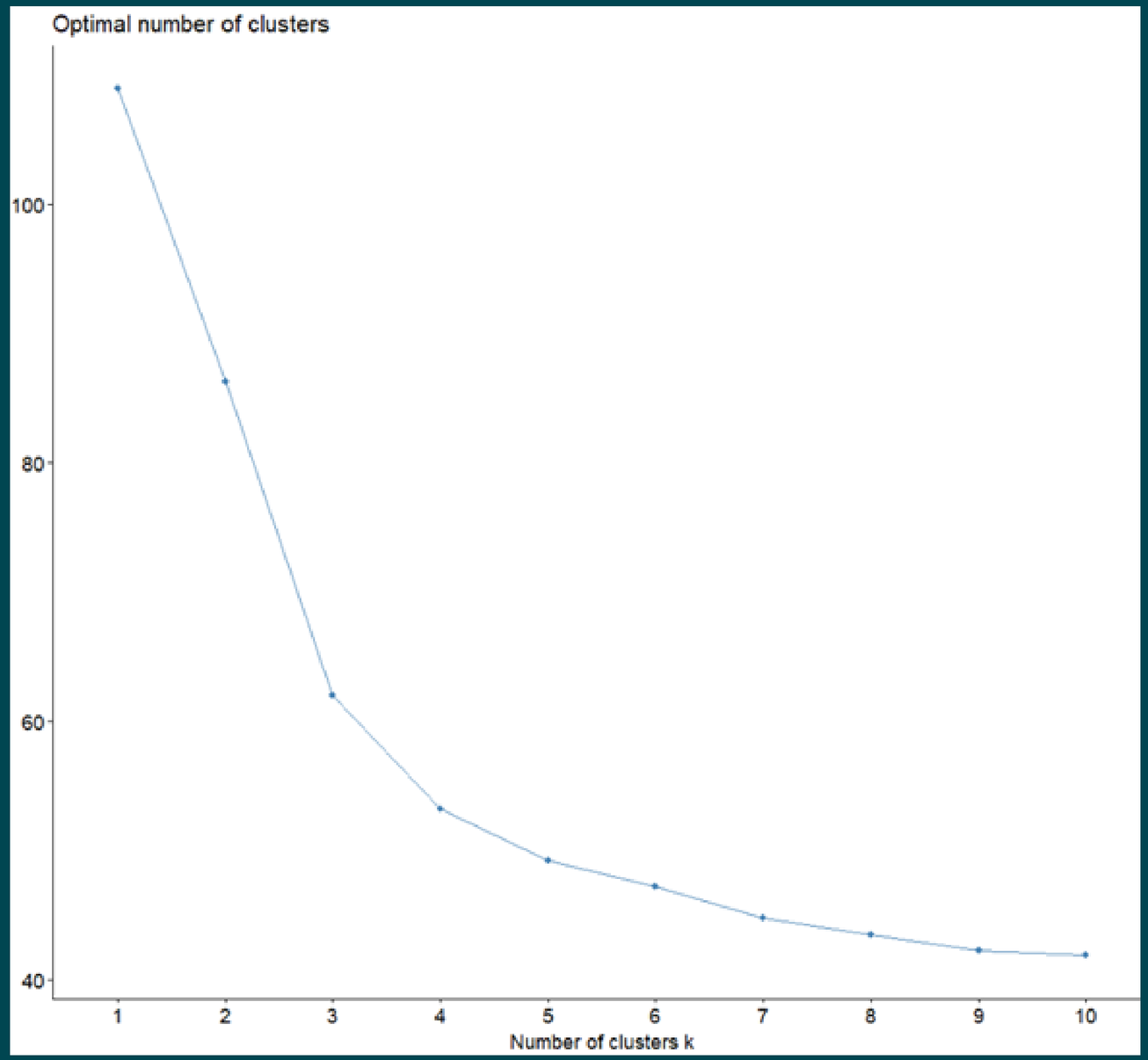
Data Manipulation

- The column “Kingdom”, “DNAtype”, “SpeciesID”, “Ncodons” and “SpeciesName” were removed as they are needed to perform the PCA or clustering.

- The columns “UUU” and “UUC” are not in numeric format, these columns are converted to numeric format.

- All entries with missing values are dropped.

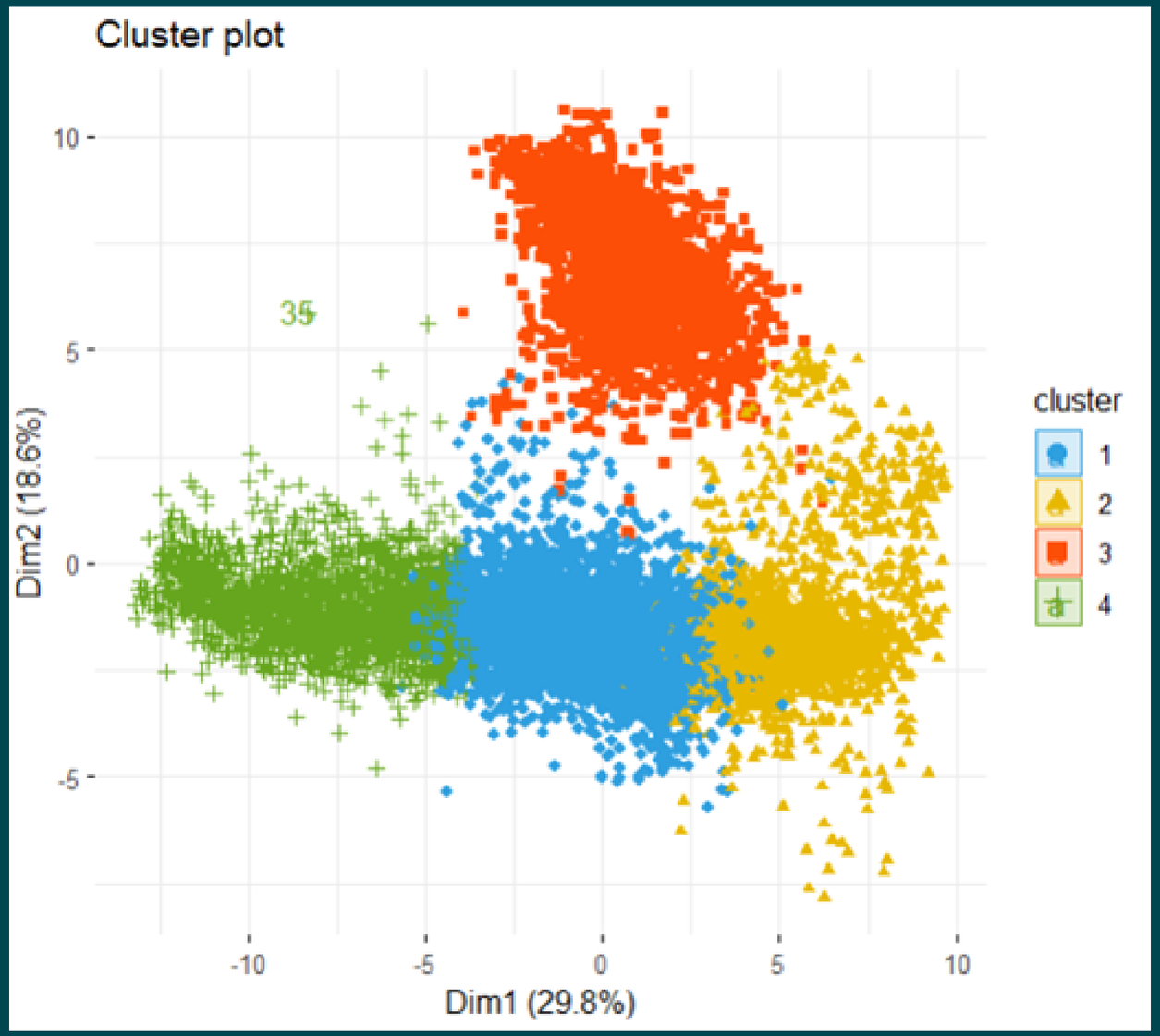
Question 1:



Elbow curve.

- The optimal clusters obtained is four as seen from the elbow curve .

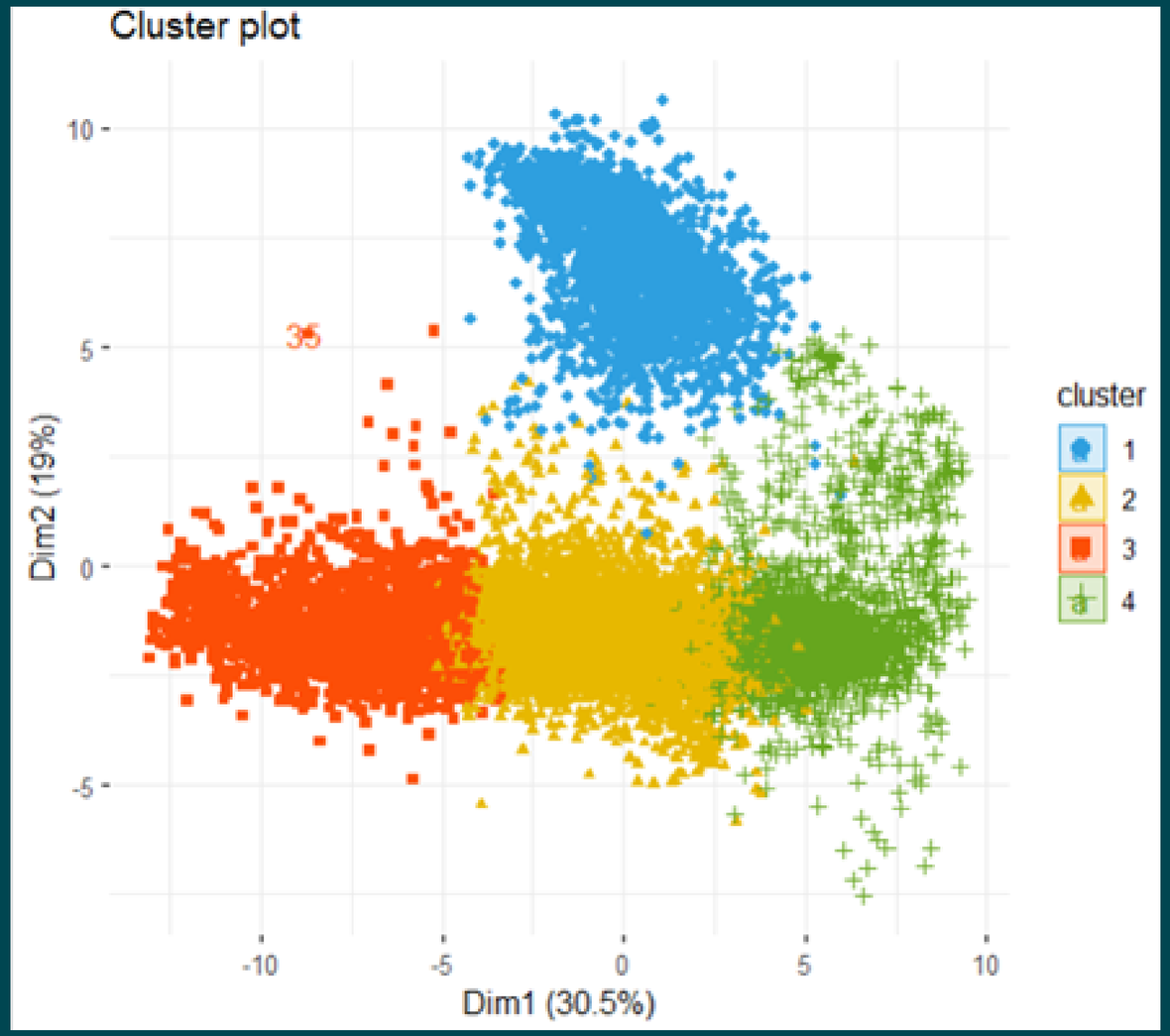
Question 2:



Visualization of the four clusters

- K-means clustering is performed with optimal clusters as 4

Question 3:



Visualization of the four clusters

- From the ANOVA, it is found that out of 64 features, only 61 features are significant.
- Considering only these 61 features, the clustering is performed.
- Clustering is performed after removing the features that are not significant and it can be visualize

Question 4:

- The Misclassification table is used as a major criteria to compare the two clusters formed in the previous 3 questions.
- Misclassification Table

Cluster 1: Using Entire dataset

	1	2	3	4
Animalia	1218	566	2314	291
Monera	764	991	0	1308
Plant	1188	1026	3	306
Virus	2250	615	2	184

Cluster 2: After removing significan variables

	1	2	3	4
Animalia	2313	1248	267	561
Monera	0	765	1304	994
Plant	3	1199	301	1020
Virus	2	2248	180	621

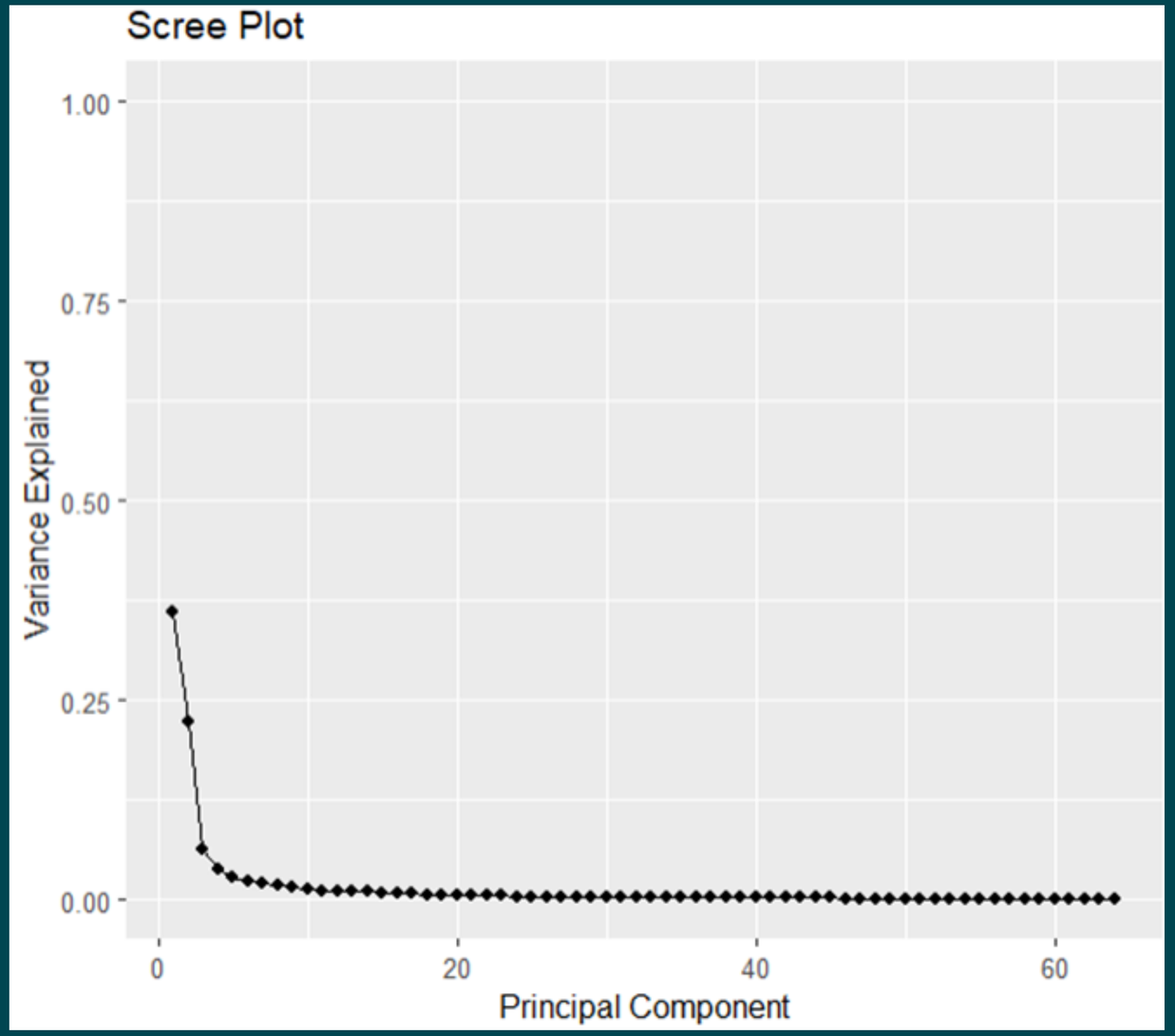
Question 4:

- Misclassification Table

Clustering Method comparison Table:

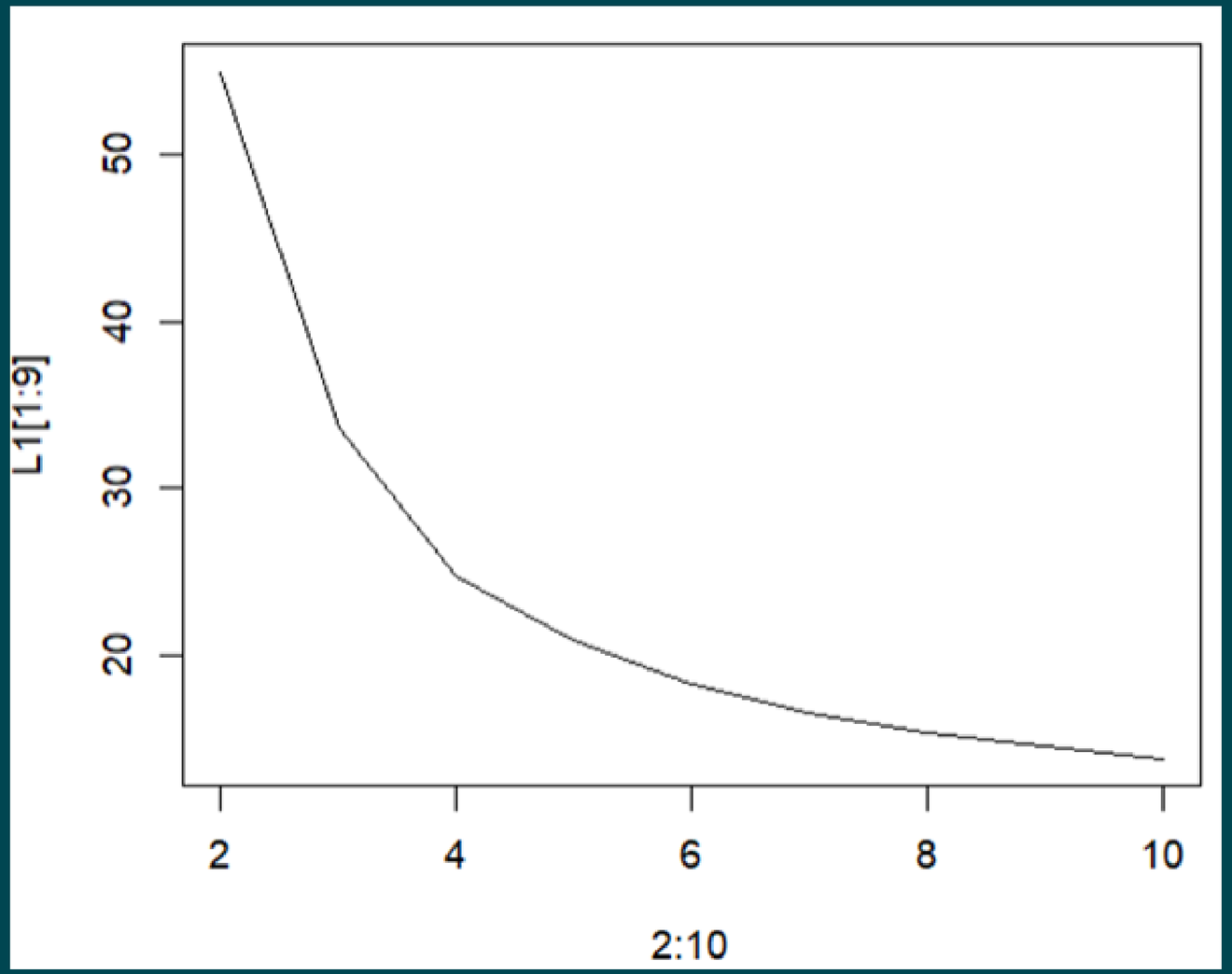
	1	2	3	4
Animalia	0	5406	0	14
Monera	0	16	0	3182
Plant	2318	1	0	0
Virus	0	37	2052	0

Question 5:



- The principal component analysis is carried out for the complete data.
- The first 6 principal components contribute to almost 74% of the variance in the data.
- The scree plot is observed to start flattening after the 6th principal component and hence first 6 principal components may be considered for further analysis

Question 6:



Elbow curve (x-axis is number of clusters and y-axis is within sum of squares)

- For various values of k, within sum of squares is calculated.
- From the elbow curve it is very evident that the optimal number of clusters is 4.
- There are naturally four groups leading to 4 optimal clusters in the data.

Question 7:

- kmeans function in R was used to perform the same. Number of clusters were chosen as 4.
- For various nstart values, the results observed were still same.

Question 8:

- For ANOVA, each block will represent each cluster. The null and alternate hypothesis is as follows.

H0: Means of all clusters are equal
H1: Means of all clusters are not equal

PC	PC1	PC2	PC3	PC4	PC5	PC6
P-values	0	2.84e-158	0	3.7e-172	1.96e-32	6.02e-41

From this we observe that, for a level of significance of alpha=0.05, we reject the null hypothesis and conclude that all the 6 principal components are significant as p-value is less than 0.05.

Question 10:

- Comparison of all three clusters:

Misclassifiaction Table:

For cluster 1

	1	2	3	4
Animalia	1218	566	2314	291
Monera	764	991	0	1308
Plant	1188	1026	3	306
virus	2250	615	2	184

For cluster 2

	1	2	3	4
Animalia	2313	1248	267	561
Monera	0	765	1304	994
Plant	3	1199	301	1020
virus	2	2248	180	621

Question 10:

Misclassifiaction Table:

For cluster 3 : After performing PCA on the entire dataset.

	1	2	3	4
Animalia	1214	568	2313	294
Monera	753	990	0	1320
Plant	1186	1026	3	308
Virus	2253	610	2	186

Refrences:

- "Kingdom (biology) - Wikipedia" https://en.m.wikipedia.org/wiki/Kingdom_(biology)

- https://www.geeksforgeeks.org/ml-principal-component-analysispca/

Thank You!