Multivariate Statistical Analysis – 2

MINI PROJECT

By:

Amirta V:21060641004

Samruddhi Hindlekar:21060641041

Shruti Deshmukh:21060641047

Vamsikrishna A:21060641055

Introduction

Living organisms are classified into groups or sets on the basis of likenesses. Such a systematic manner of classifications aids in simplifying the study of the wide variety of organisms. In 1996, a five kingdom classification was proposed by R. H. Whittaker. This type of kingdom classification includes five kingdoms namely, Monera, Protista, Fungi, Plantae and Animalia. Different classifications of living beings have different codon frequency bias. Codon usage frequency refers to the frequency of occurrence of synonymous codons in coding DNA. With the help of Clustering, an unsupervised ML algorithm, data on codon frequencies (occurrences) can be analyzed for patterns.

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters. The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

Principal Component Analysis is an unsupervised learning algorithm that is used for 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. These new transformed features are called the Principal Components. It is one of the popular tools that is used for exploratory data analysis and predictive modeling. It is a technique to draw strong patterns from the given dataset by reducing the variances. PCA generally tries to find the lower-dimensional surface to project the high-dimensional data.

Dataset:

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. There are total 69 columns out of which column 1 gives the kingdom, column 2 states the DNA type of the organism, column 3 is the species ID, column 4 states the total number of codons in the organism's DNA, column 5 is the species name and the rest 63 columns represent the frequencies of codon usage in the DNA with each column representing one type of codon. There are a total of 13028 entries in the dataset.

A codon is a series of three nucleotides or a triplet that encodes a specific amino acid residue in a polypeptide chain. For example, UUU, UUA, UUC are codons. Codon usage frequency refers to the frequency of occurrence of synonymous codons in coding DNA. There are a total 64 different types of codons in DNA. Codon frequency is important for determining the gene expression and cellular function as it influences processes such as RNA processing, protein translation and protein folding.

Originally in the data, the 'kingdom' column represents sub kingdoms which falls under the broader category of kingdom's that are, Monera, Virus, Plant, Animalia. Protista is also one of the kingdoms but this dataset doesn't have observations on it. For easy reference and for the purpose of performing external validation, an extra column is created such that Monera represents archaea, bacteria and plasmids, Virus represents Bacteriophages, Plant contains only plant and animalia represents invertebrates, vertebrates, mammals, rodents and primates.

Methodology

Data manipulation: The data needs to prepared as such to perform the necessary analysis. Initially, the column "Kingdom", "DNAtype", "SpeciesID", "Ncodons" and "SpeciesName" were removed as they are needed to perform the PCA or clustering. Since the columns "UUU" and "UUC" are not in numeric format, these columns are converted to numeric format. All entries with missing values are dropped.

```
> colnames(c)
 [1] "Kingdom"     "DNAtype"     "SpeciesID"   "Ncodons"     "SpeciesName" "UUU"         "UUC"         "UUA"         "UUG"
[10] "CUU"         "CUC"         "CUA"         "CUG"         "AUU"         "AUC"         "AUA"         "AUG"         "GUU"
[19] "GUC"         "GUA"         "GUG"         "GCU"         "GCC"         "GCA"         "GCG"         "CCU"         "CCC"
[28] "CCA"         "CCG"         "UGG"         "GGU"         "GGC"         "GGA"         "GGG"         "UCU"         "UCC"
[37] "UCA"         "UCG"         "AGU"         "AGC"         "ACU"         "ACC"         "ACA"         "ACG"         "UAU"
[46] "UAC"         "CAA"         "CAG"         "AAU"         "AAC"         "UGU"         "UGC"         "CAU"         "CAC"
[55] "AAA"         "AAG"         "CGU"         "CGC"         "CGA"         "CGG"         "AGA"         "AGG"         "GAU"
[64] "GAC"         "GAA"         "GAG"         "UAA"         "UAG"         "UGA"
```

Fig 1: Column names in the Data

Question 1: The optimal clusters obtained are four as seen from the elbow curve in Figure 2. Also, there are majorly four kingdoms and hence four optimal clusters are obtained from the elbow curve.

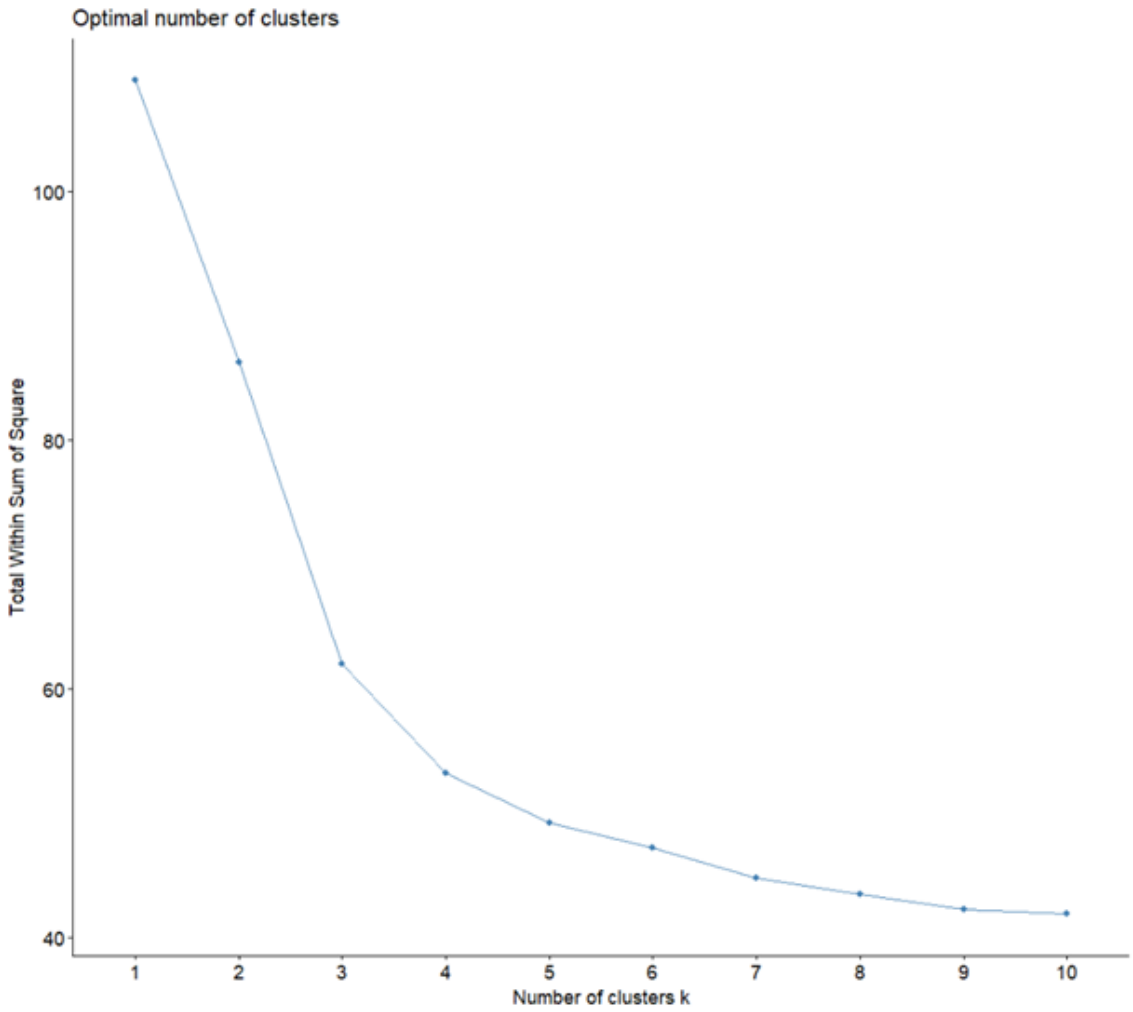

Fig 2: Elbow curve

Question 2: K-means clustering is performed with optimal clusters as 4. Code for the same in Fig 3 while the clusters can be visualized in figure 4.

```
> #Taking k=4
> set.seed(123)
> km.sol_1 <- kmeans(codon, 4, nstart = 30)
>
```

Fig 3: K-means clustering with optimal clusters

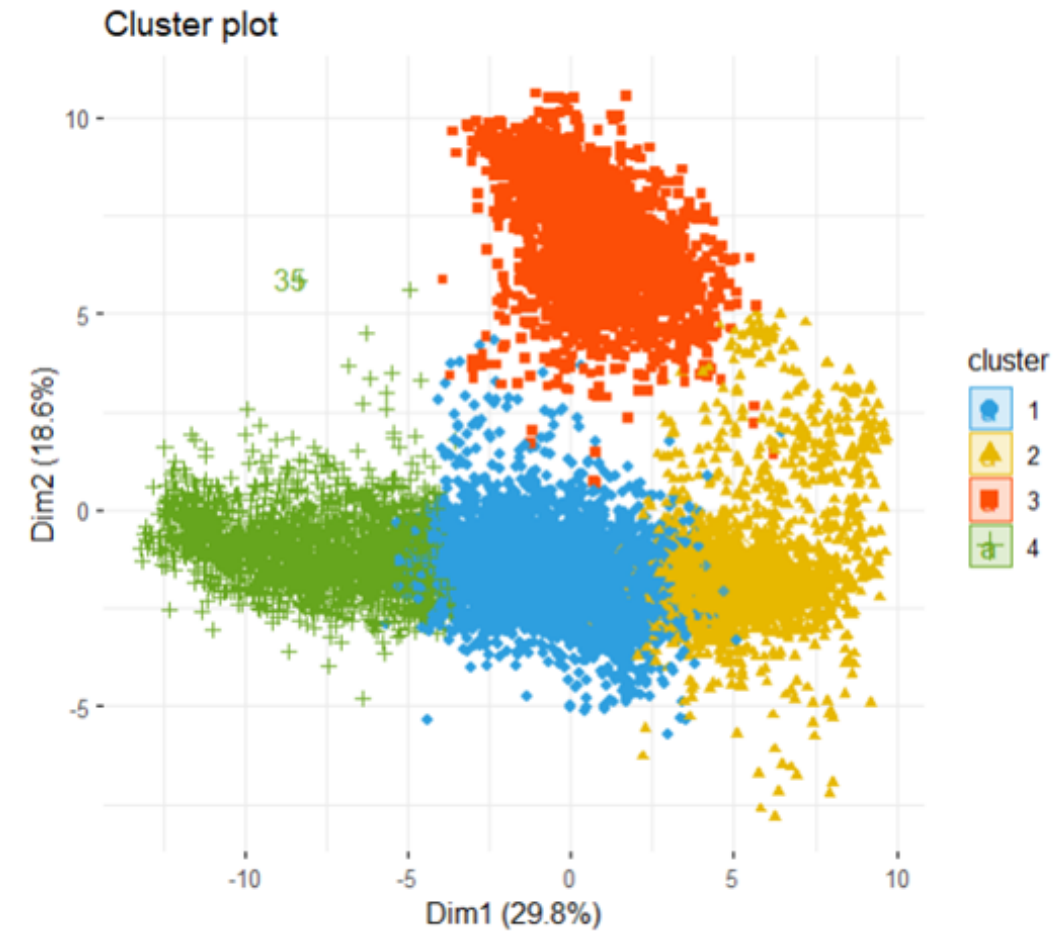

Fig 4: Visualization of the four clusters

Question 3:

One way ANOVA is applied to find which of the variables are significantly contributing for the clustering.

Idea behind ANOVA: To check whether the cluster means corresponding to each variable are significantly different from each other or not. If they are significantly different then the variable is significantly contributing for the clustering, otherwise the variable does not significantly contribute for clustering.

Here ANOVA is applied for each variable but considering the clusters as blocks.

Hypothesis:

Ho: There is no significant difference in the cluster means.

H1: There is a significant difference in the cluster means.

From the ANOVA, it is found that out of 64 features, only 61 features are significantly contributing for the clustering. Considering only these 61 features, the clustering is performed. Clustering is performed after removing the features that are not significant and it can be visualized in figure 5.

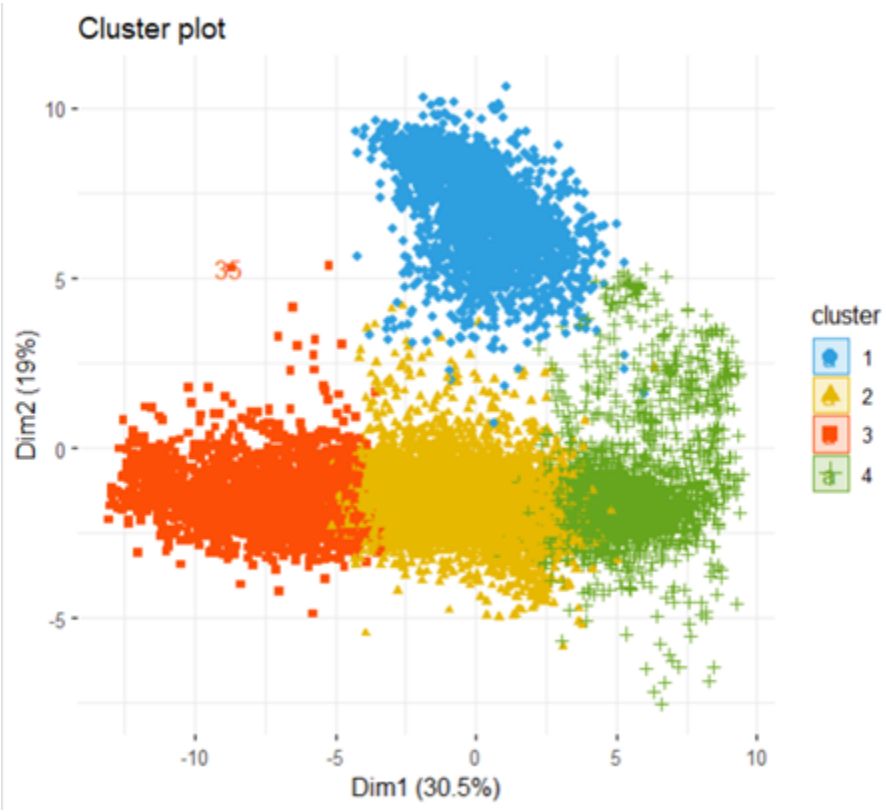

Fig 5: Visualization of the cluster

Question 4:

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

Misclassification Table: This table is formed between the original class of the observation form the data and the classes assigned to the observation by k means clustering. The rows represent the original class and the columns represent the assigned class.

- **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

Table 4.1

R output:

```
            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 non significant variables using ANOVA

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

Table 4.2

R Output:

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

Clustering Method comparison Table:

	1	2	3	4
1	0	5406	0	14
2	0	16	0	3182
3	2318	1	0	0
4	0	37	2052	0

Table 4.3

R Output:

```
      1    2    3    4
1     0 5406    0   14
2     0   16    0 3182
3  2318    1    0    0
4     0   37 2052    0
```

From the above tables it can be seen that removing the variables which are not significantly contributing to the clustering and applying k means cluster again does not reduce the misclassification. So, for the given data the k means clustering can be applied to the whole data.

Question 5: The principal component analysis is carried out for the complete data. The Scree plot and summary of the analysis is as observed in the figure 6.

From the scree plot and the summary of the analysis, it is clear that the first 7 principal components contribute to almost 74% of the variance in the data. Also, the scree plot is observed to start flattening after the 7th principal component and hence the first 7 principal components may be considered for further analysis.

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.05496 0.04333 0.02296 0.01779 0.01553 0.01365 0.01332 0.01227 0.01119 0.01060 0.009931 0.009698 0.009134 0.009014
Proportion of Variance 0.36127 0.22461 0.06308 0.03787 0.02884 0.02230 0.02122 0.01801 0.01499 0.01343 0.011800 0.011250 0.009980 0.009720
Cumulative Proportion  0.36127 0.58588 0.64896 0.68684 0.71568 0.73798 0.75920 0.77721 0.79220 0.80563 0.817430 0.828670 0.838650 0.848370
                          PC15     PC16     PC17     PC18     PC19     PC20     PC21     PC22     PC23     PC24     PC25     PC26     PC27
Standard deviation     0.008053 0.007737 0.007716 0.007464 0.007244 0.007075 0.006898 0.006817 0.00654 0.006184 0.006154 0.005994 0.005768
Proportion of Variance 0.007760 0.007160 0.007120 0.006660 0.006280 0.005990 0.005690 0.005560 0.00512 0.004570 0.004530 0.004300 0.003980
Cumulative Proportion  0.856130 0.863290 0.870410 0.877070 0.883350 0.889340 0.895030 0.900580 0.90570 0.910270 0.914800 0.919100 0.923080
```

Fig 6a: Summary of Analysis

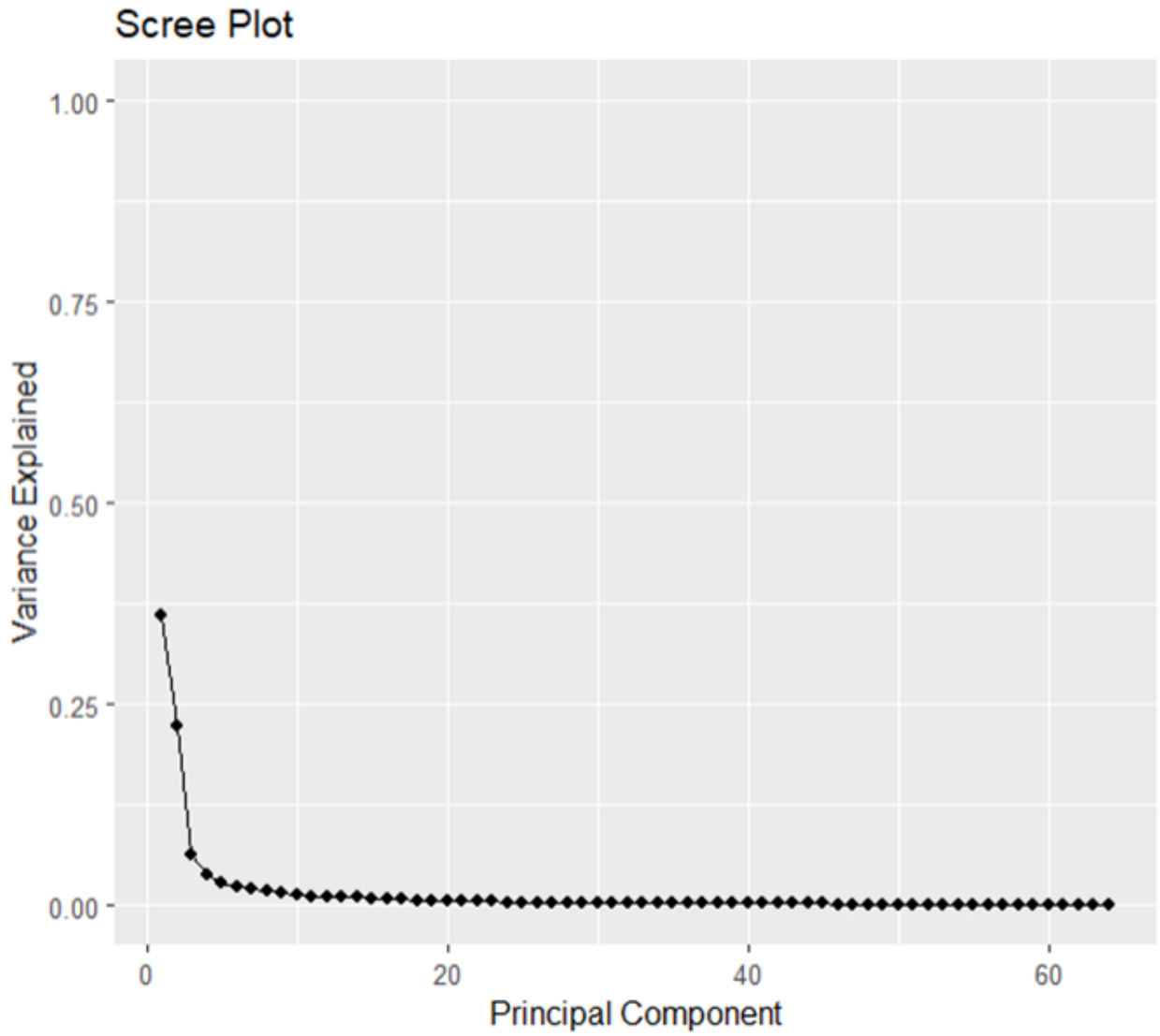

Fig 6b: Scree Plot

Question 6: The kmeans clustering is applied to PCA for various values of k and nstart. The within sum of squares are calculated each time and the elbow curve is plotted between various values of k and within sum of squares represented in figure 7. From the elbow curve it is very evident that the optimal number of clusters is 4. Also, broadly there are four kingdoms in the given data. Hence, there are naturally four groups leading to 4 optimal clusters in the data.

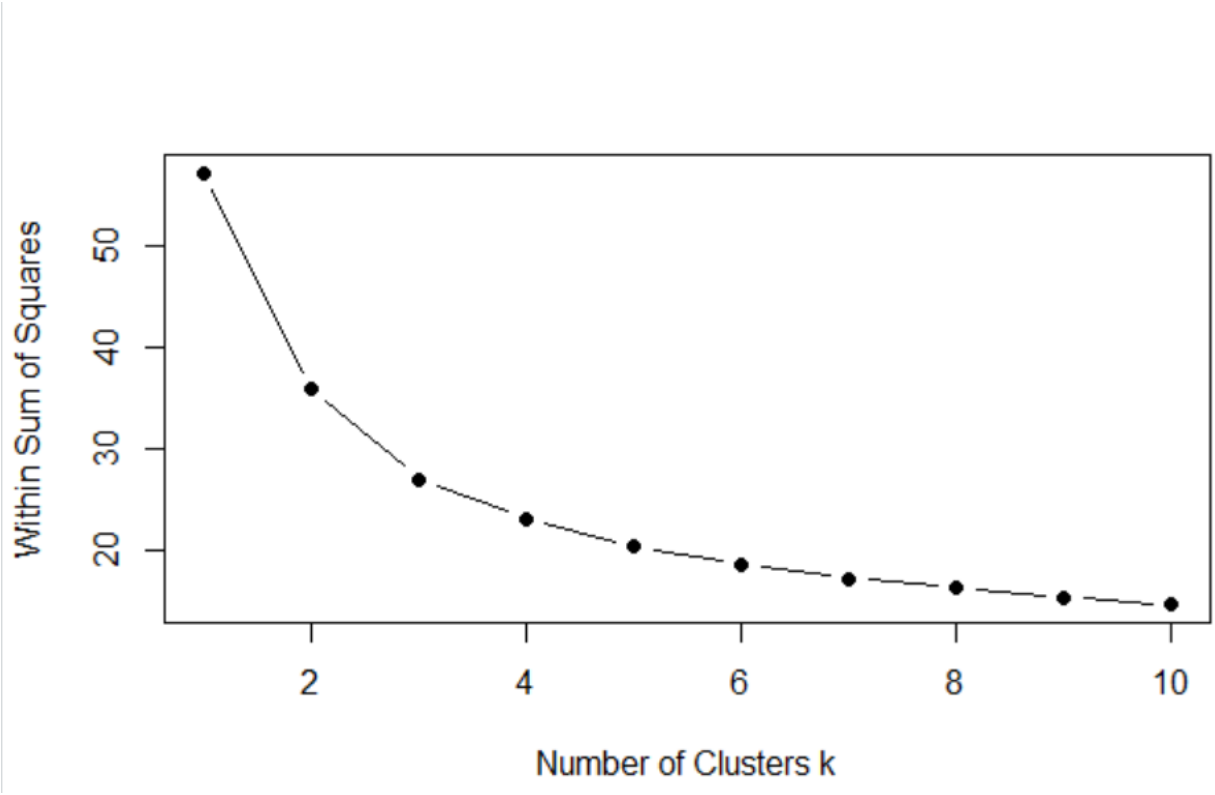

Fig 7: Elbow curve (x-axis is number of clusters and y-axis is within sum of squares)

Question 7: kmeans() function in R was used to perform the same. Number of clusters (k) was chosen as 4 from the previous question. figure 8 shows the clusters to which the observation belongs.

```
> k_m<-kmeans(pca1,4,iter.max=100,nstart=10)
> k_m$cluster
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28
  2   2   1   1   1   1   4   1   2   4   4   4   4   4   4   4   1   2   2   4   4   2   2   2   2   2   2   1
 29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56
  1   1   2   1   1   1   2   1   1   4   1   1   1   1   1   1   1   1   1   1   1   2   2   2   1   1   1   2
 57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84
  2   1   1   1   4   1   1   4   1   1   1   4   2   1   1   1   1   1   1   1   1   2   1   1   1   2   1   1
 85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112
  2   1   4   4   4   1   1   1   1   1   1   1   4   1   1   1   1   4   1   4   4   4   4   4   4   4   1   4
113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
  4   4   1   4   4   4   4   1   4   4   4   4   4   1   1   1   4   4   4   4   4   1   1   1   1   4   4   4
141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168
  4   1   1   1   1   1   4   4   4   1   1   4   1   1   1   1   1   1   4   1   1   1   4   4   1   4   1   1
169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196
  4   1   1   1   1   1   1   1   1   1   1   1   2   1   1   4   4   4   4   4   4   4   4   4   4   4   4   4
197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224
```

Fig 8: Cluster outputs of kmeans

Question 8: ANOVA is applied to each PCA considering clusters as blocks. This is to find which PCA is significantly contributing for the clustering. The null and alternative hypothesis is as follows.

H0: Means of all clusters are equal

H1: Means of all clusters are not equal

The obtained p-values for each principal component are as shown in figure 9.

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

Table 1: P-values of ANOVA for each PC

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

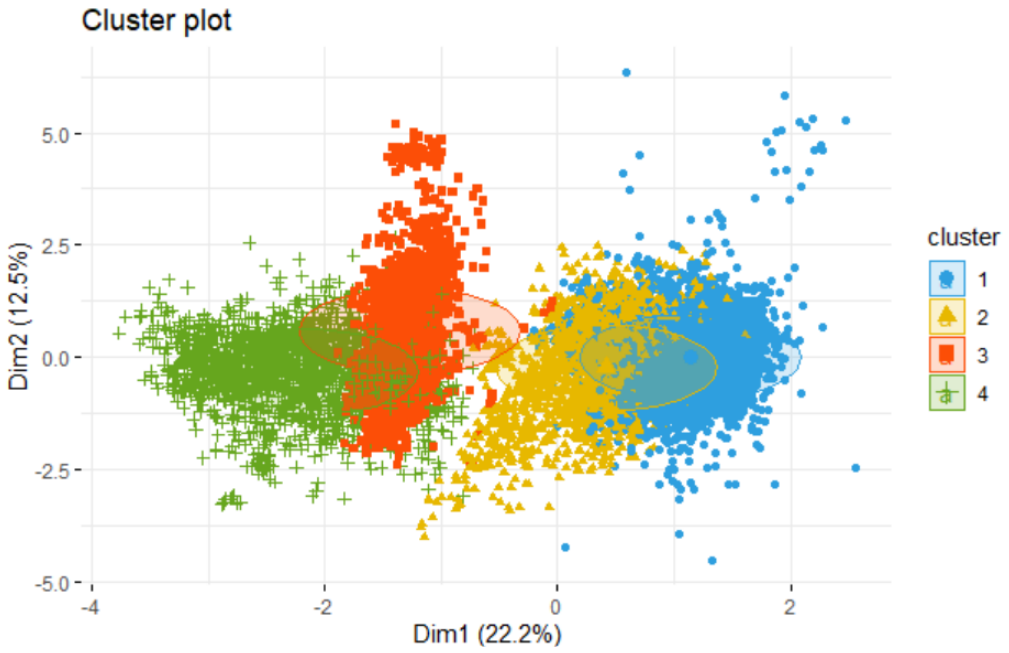

Fig. 9

Question 9: Since all the PCs are significant (from the results of ANOVA). There was no need for dropping out of any PCA. As there were no dropouts of PCA the result of K means clustering performed in the previous question will be the same.

Question 10: (Conclusions)

Comparison of all 3 clusters:

- **Misclassification Table**

  Cluster Analysis 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

Table 10.1

Cluster Analysis 2: After removing non significant variables using ANOVA

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

Table 10.2

Cluster Analysis 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

Table 10.3

The miscalculation percentage was calculated for each of the cluster analysis results and for all of its 4 clusters. The intuition used to calculate it was that, for each cluster the sum of the misclassifications were calculated and divided by the total elements in the cluster. The results obtained were as follows:

Cluster	Cluster Analysis 1	Cluster Analysis 2	Cluster Analysis 3
Animalia	47%	47%	47%
Monera	57%	69%	56%
Plant	59%	60%	59%
Virus	26%	26%	26%

Table 10.4

Here it has been observed that the misclassifications increased in cluster analysis 2 as compared to cluster analysis 1. Thus there was no significant difference in the clustering results after dropping the non significant variables obtained on performing ANOVA. When PCA was applied on the entire dataset, 7 principal components were selected. After clustering this data, the misclassification was found almost similar to that of the 1st clustering analysis results. Thus the suggested method would be to go for PCA first and then perform the clustering. This grants us the benefits of dimension reduction and may save time and costs for larger datasets.

References:

1. "Kingdom (biology) - Wikipedia" https://en.m.wikipedia.org/wiki/Kingdom_(biology)
2. https://www.geeksforgeeks.org/ml-principal-component-analysispca/

Appendix:

R Codes

```
#MSA 2 - MINI PROJECT

library(factoextra) #Visualizations

#Load the dataset
c<-read.csv(file.choose(),header=TRUE)
c<-c[-c(487,5064),] #These rows have absurd values
head(c)

#Creating the main_kingdom column by clubbing the sub kingdoms
c$main_kingdom[c$Kingdom=='arc'] <-'Monera'
c$main_kingdom[c$Kingdom=='bct'] <-'Monera'
c$main_kingdom[c$Kingdom=='plm'] <-'Monera'
c$main_kingdom[c$Kingdom=='phg'] <-'Virus'
c$main_kingdom[c$Kingdom=='vrl'] <-'Virus'
c$main_kingdom[c$Kingdom=='pln'] <-'Plant'
c$main_kingdom[c$Kingdom=='inv'] <-'Animalia'
c$main_kingdom[c$Kingdom=='vrt'] <-'Animalia'
c$main_kingdom[c$Kingdom=='mam'] <-'Animalia'
c$main_kingdom[c$Kingdom=='rod'] <-'Animalia'
c$main_kingdom[c$Kingdom=='pri'] <-'Animalia'
View(c)

codon<-c[,-c(1,2,3,4,5,70)]
head(codon)
str(codon)
codon$UUU=as.numeric(codon$UUU)
codon$UUC=as.numeric(codon$UUC)
sum(is.na(codon)) # No NAs
```

```
#####Q.1 Value of k
#Elbow bent chart from factoextra library
fviz_nbclust(codon, kmeans, method = "wss")
#fviz_nbclust(codon, kmeans, method = "silhouette")
#We select k=4

#Taking k=4
set.seed(123)
km.sol_1 <- kmeans(codon, 4, nstart = 30)


fviz_cluster(km.sol_1, data = codon,
             palette = c("#2E9FDF", "#E7B800", "#FC4E07","#66A61E"),
             ellipse.type = "euclid", # Concentration ellipse
             repel = TRUE, # Avoid label overplotting (slow)
             ggtheme = theme_minimal())

#Adding the cluster numbers to the datapoints
codon <- cbind(codon, cluster1=km.sol_1$cluster)
head(codon)


#Getting column names
cols=colnames(codon)[-65]


signi=c()  #Important Variables
not_signi=c()   #Nonsignificant variables

for(i in cols){
  result.aov=aov(codon[,i]~cluster1,data=codon)
  p=summary(result.aov)[[1]][["Pr(>F)"]][1]
  if(p<0.05){signi=c(signi,i)}
  else{not_signi=c(not_signi,i)}
}

#Removing the non significant variables
codon1=codon[,signi]
head(codon1)
```

```
#We will cluster the new dataset now
set.seed(123)
km.sol_2 <- kmeans(codon1, 4, nstart = 30)

fviz_cluster(km.sol_2, data = codon1,
             palette = c("#2E9FDF", "#E7B800", "#FC4E07","#66A61E"),
             ellipse.type = "euclid", # Concentration ellipse
             repel = TRUE, # Avoid label overplotting (slow)
             ggtheme = theme_minimal())

#Adding the cluster numbers to the datapoints
codon1<- cbind(codon1, cluster2 = km.sol_2$cluster)
head(codon1)


# Question 4
#Getting the miss classification number
table(c[,70],codon[,65])

#Getting the miss classification number after removing the columns
table(c[,70],codon1[,62])

#Table between without removing variables and with removing variables.
table(codon[,65],codon1[,62])


#Conducting PCA

data<-codon[,-c(65)]
head(data)
pca<-prcomp(data,center = TRUE, scale.= FALSE)
summary(pca)#74% for 7 principle components
#calculate total variance explained by each principal component
varexplained = pca$sdev^2 / sum(pca$sdev^2)

#create scree plot
library(ggplot2)

qplot(c(1:10), varexplained[1:10]) +
```

```
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot") +

  ylim(0, 1)

#7 components
pca1<-pca$x[,1:7]
head(pca1)

#Converting into data frame
pca1<-as.data.frame(pca1)

#Question 6
set.seed(123)
L1=c()
L2=c()

for (i in (1:50)){
  c1<-kmeans(pca1,i,iter.max = 100,nstart=10)
  L1[i-1]<-c1$tot.withinss
  L2[i-1]<-c1$betweenss
}
library(ggplot2)

ggplot(x=c(1:10), y=L1[1:10]) +
  geom_line() +
  xlab("Number of Clusters") +
  ylab("Within Sum of Squares") +
  ggtitle("Elbow Curve") +

  ylim(0, 1)

plot(1:10,L1[1:10],type="b", pch = 19,xlab="Number of Clusters k",ylab="Within Sum of
Squares")
#lines(2:10,L2[1:9],type="l") #This is for between

#optimal K value is 4 ie k=4
```

```
"""
L3=c()
L4=c()

for (i in (2:50)){
  c1<-kmeans(pca1,i,iter.max = 100,nstart=15)
  L3[i-1]<-c1$tot.withinss
  L4[i-1]<-c1$betweenss
}
plot(2:10,L3[1:9]/L4[1:9],type="l")# optinmal k =4
plot(2:10,L3[1:9],type="l")
lines(2:10,L4[1:9],type="l")

"""

#Question 7
set.seed(123)
k_m<-kmeans(pca1,4,iter.max=100,nstart=10)
k_m$cluster
#Question 8

#Binding the PCA dataset and the clusters
d_c=cbind(pca1,k_m$cluster)
head(d_c)
colnames(d_c)
##Visualizing the clusters
fviz_cluster(k_m, data = d_c,
             palette = c("#2E9FDF", "#E7B800", "#FC4E07","#66A61E"),
             ellipse.type = "euclid", # Concentration ellipse
             repel = TRUE, # Avoid label overplotting (slow)
             ggtheme = theme_minimal())


#Doing ANOVA by considering each clusters as the blocks
pval=c()
for (i in (1:6)){
  a<-aov(d_c[,i]~d_c[,7])
  pval[i]<-summary(a)[[1]][1,5]

}
```

```
a<-aov(d_c[,1]~d_c[,7])
summary(a)[[1]][1,5]
#Since all p values are near zero, all hypothesis are being rejected. Hence the block means are significantly differnt
#for each PCA. Hence No PCA need to be removed. All PCA are significantly contributing for the clusters

# Question 9
#We need not do this again because as per result we need not remove any PCA value. So the kmeans cluster that we did
#previously will remain the same.

#Question 10
#Getting the miss classification number
table(c[,70],codon[,65])
#Getting the miss classification number after removing the columns
table(c[,70],codon1[,62])
#Table between without removing variables and with removing variables.
table(codon[,65],codon1[,62])
#Table
table(c[,70],d_c[,8])
```