SRI RAMACHANDRA

INSTITUTE OF HIGHER EDUCATION AND RESEARCH

(Category - I Deemed to be University) Porur, Chennai

SRI RAMACHANDRA ENGINEERING AND TECHNOLOGY

CSE 320 DATA MINING

PROJECT REPORT CA4

Name : D SANTHOSH

Unique ID : E0120009

Year : III

Quarter : Q1

Department : B.Tech CSE (AI & ML)

Faculty Name : Prof. Shiyamala

Academic Year : 2022

PROJECT REPORT

Abstract:

In the iris Dataset, the main aim is to cluster the flowers based on their petal length, petal width, sepal length and sepal width. Taking into factor many key-points, the appropriate solution can be derived.

Introduction:

Iris can be a powerful means to identify the species of the flowers. This technique can be used by data scientists to find the species of the flowers using data mining models.

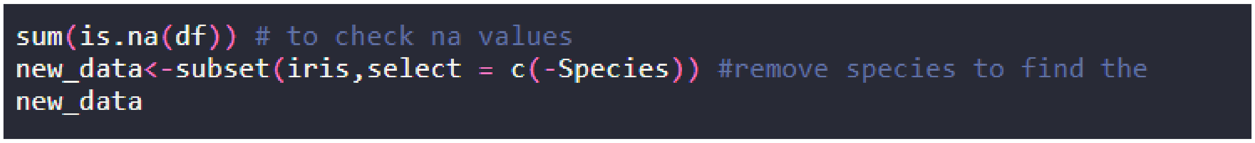
About Dataset:

Data Description:

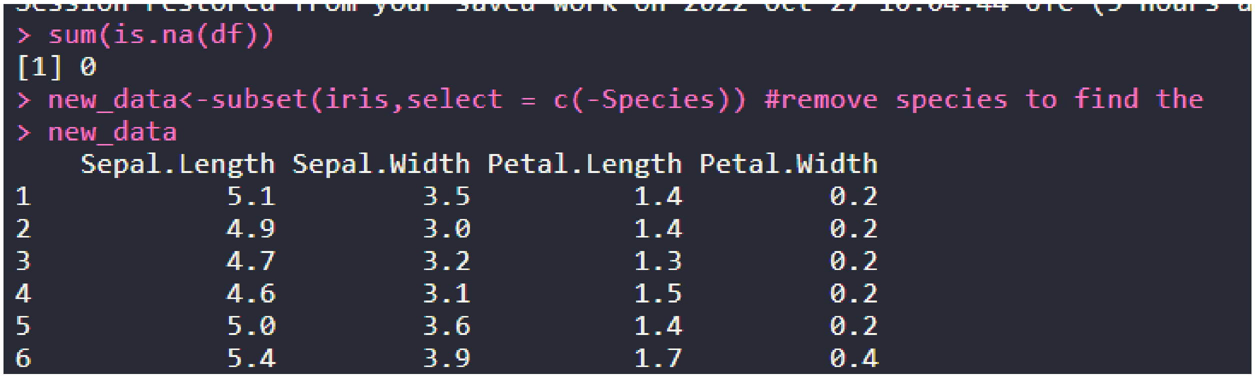
We will take into consideration the Iris dataset which contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris setosa, Iris virginica and Iris versicolor).

Data PreProcessing:

Procuring the data was not that difficult ,the cleaning process did not take much time either.Invalid data such as negative values, misprinted values and outliers were removed along with duplicate values.



```
sum(is.na(df)) # to check na values
new_data<-subset(iris,select = c(-Species)) #remove species to find the
new_data
```



```
Session restored from your saved work on 2022 Oct 27 10:04:44 UTC (5 hours a
> sum(is.na(df))
[1] 0
> new_data<-subset(iris,select = c(-Species)) #remove species to find the
> new_data
    Sepal.Length Sepal.Width Petal.Length Petal.Width
1            5.1         3.5          1.4         0.2
2            4.9         3.0          1.4         0.2
3            4.7         3.2          1.3         0.2
4            4.6         3.1          1.5         0.2
5            5.0         3.6          1.4         0.2
6            5.4         3.9          1.7         0.4
```

Machine Learning Model:

K means Clustering:

The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.

- Specify number of clusters K.
- Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters is not changing.

Elbow Method:

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.

Hierarchical Clustering:

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA.
In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.

Complete Linkage Clustering:

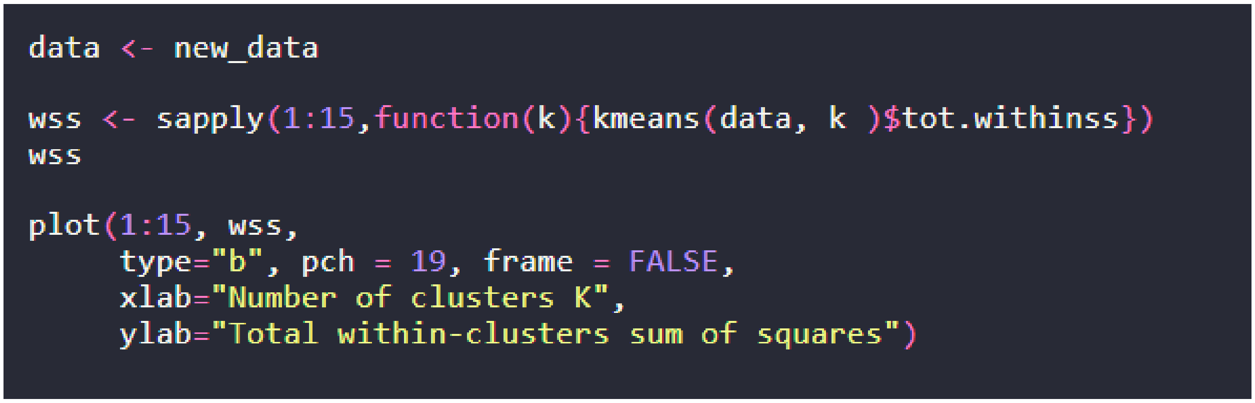
It is the farthest distance between the two points of two different clusters. It is one of the popular linkage methods as it forms tighter clusters than single-linkage.

Average Linkage Clustering:

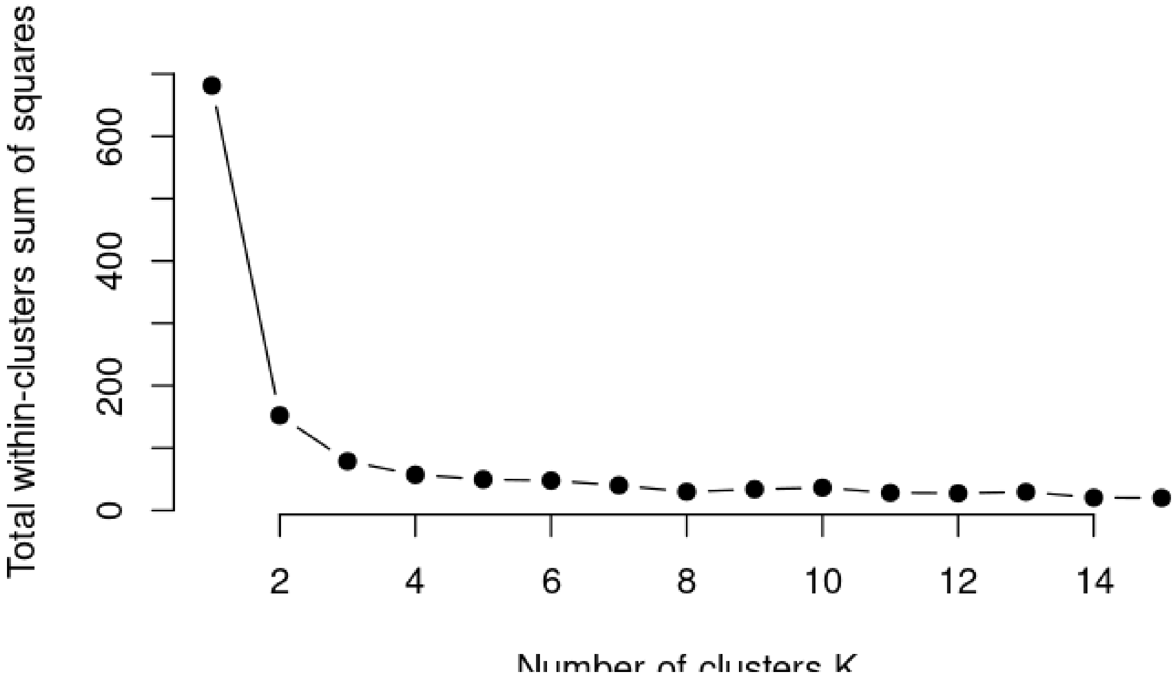
It is the linkage method in which the distance between each pair of datasets is added up and then divided by the total number of datasets to calculate the average distance between two clusters. It is also one of the most popular linkage methods.

To find the species of the flowers (K Means Algorithm):

Elbow Method:

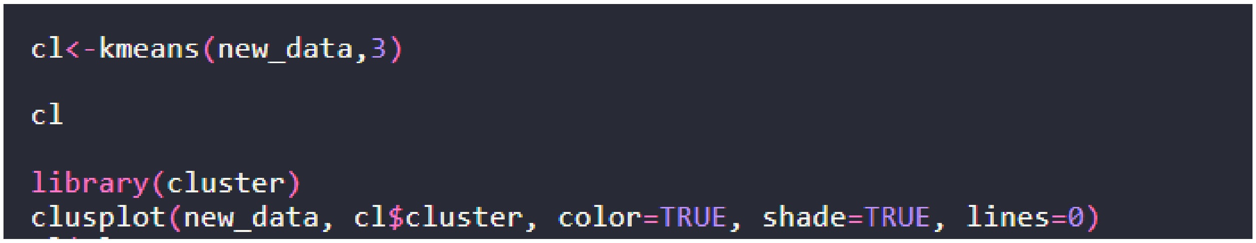


```
data <- new_data

wss <- sapply(1:15,function(k){kmeans(data, k )$tot.withinss})
wss

plot(1:15, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```

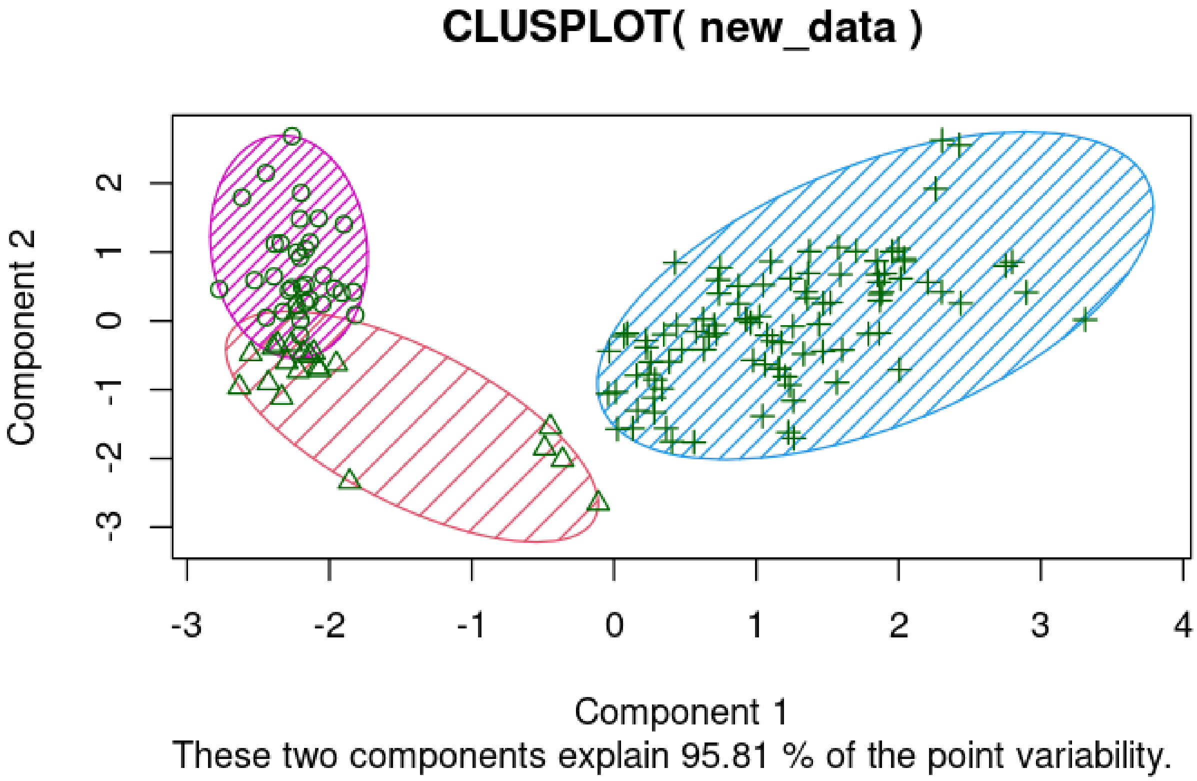


To determine the optimal number of clusters, we have to select the value of k at the “elbow” ie the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is 3.

KMeans Algorithm:



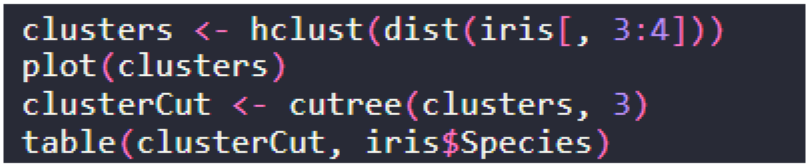
```
cl<-kmeans(new_data,3)

cl

library(cluster)
clusplot(new_data, cl$cluster, color=TRUE, shade=TRUE, lines=0)
```



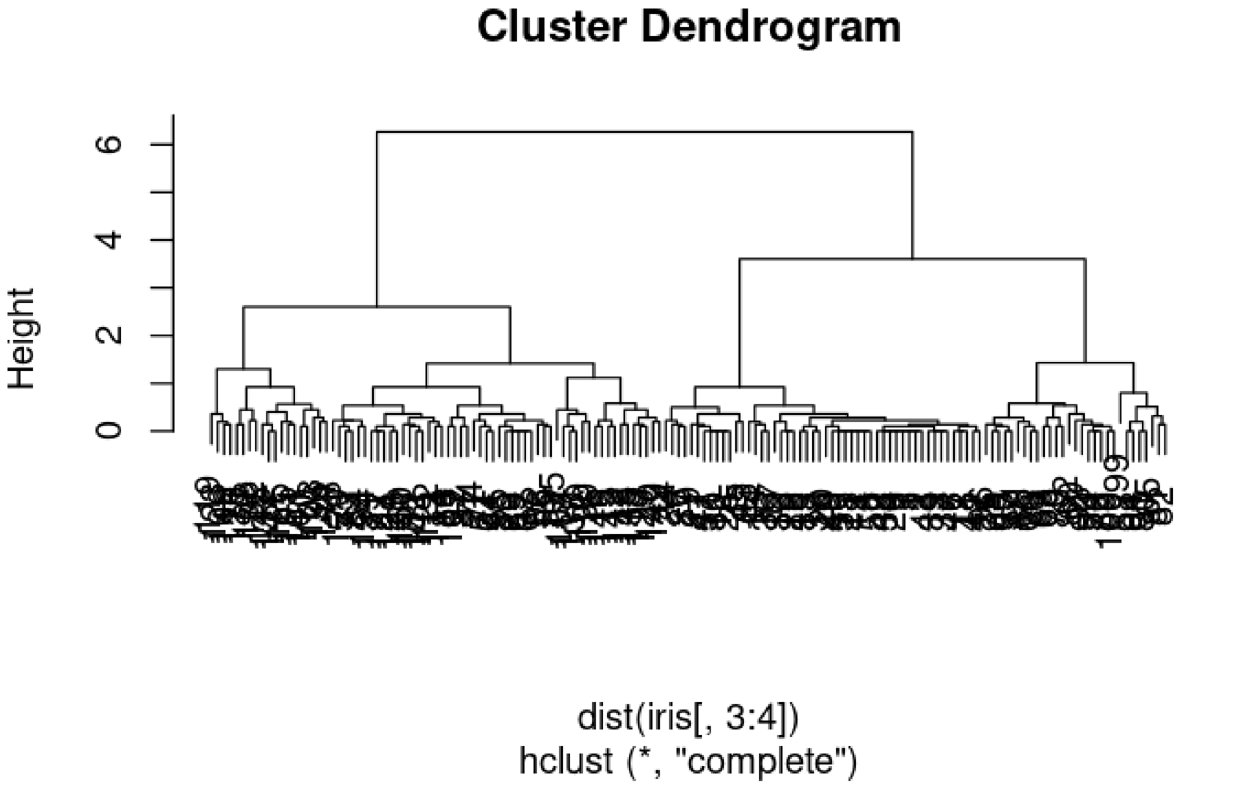
Clustering Analysis gives us a noticeably clear insight about the different segments of the customers, namely setosa ,versicolor, verginica . We can segregate them based on their petal length, petal width. Here, once we have identified the clusters we can make sure that they belong to the species .The setosa species can be avoided, and we must not be catering to it. The rest can be promotionally catered to according to the category. There are also some outliers which should be taken into consideration as they are the ones with similar attributes.

To find the species of the flowers (Hierarchical Algorithm):

Complete Linkage Method:



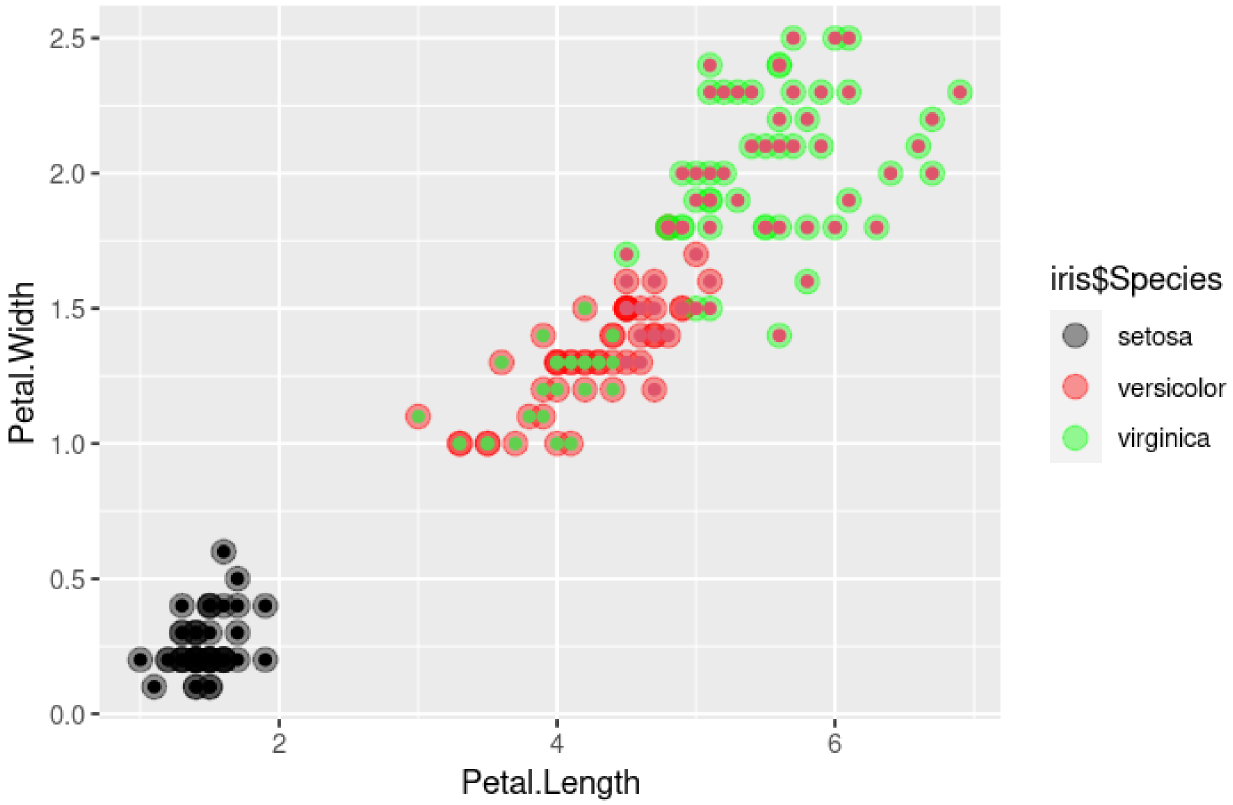
```
clusters <- hclust(dist(iris[, 3:4]))
plot(clusters)
clusterCut <- cutree(clusters, 3)
table(clusterCut, iris$Species)
```



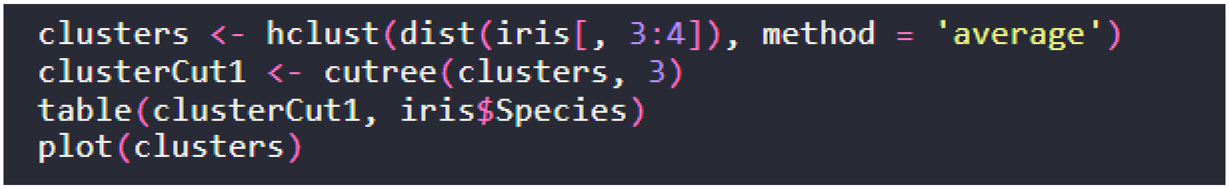
In this case, two clusters do not seem to work as a good separation between the three species.

Using three clusters separates all iris-setosa specimens in one cluster and all but one specimen of Iris-virginica in another one.

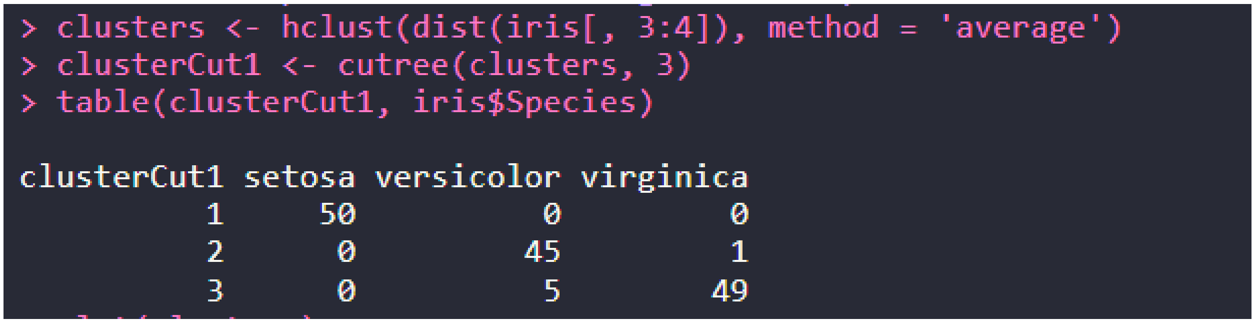
In this case, using "Complete" method for hierarchical clustering gives better results. However, Iris-versicolor data is still problematic for the algorithm.



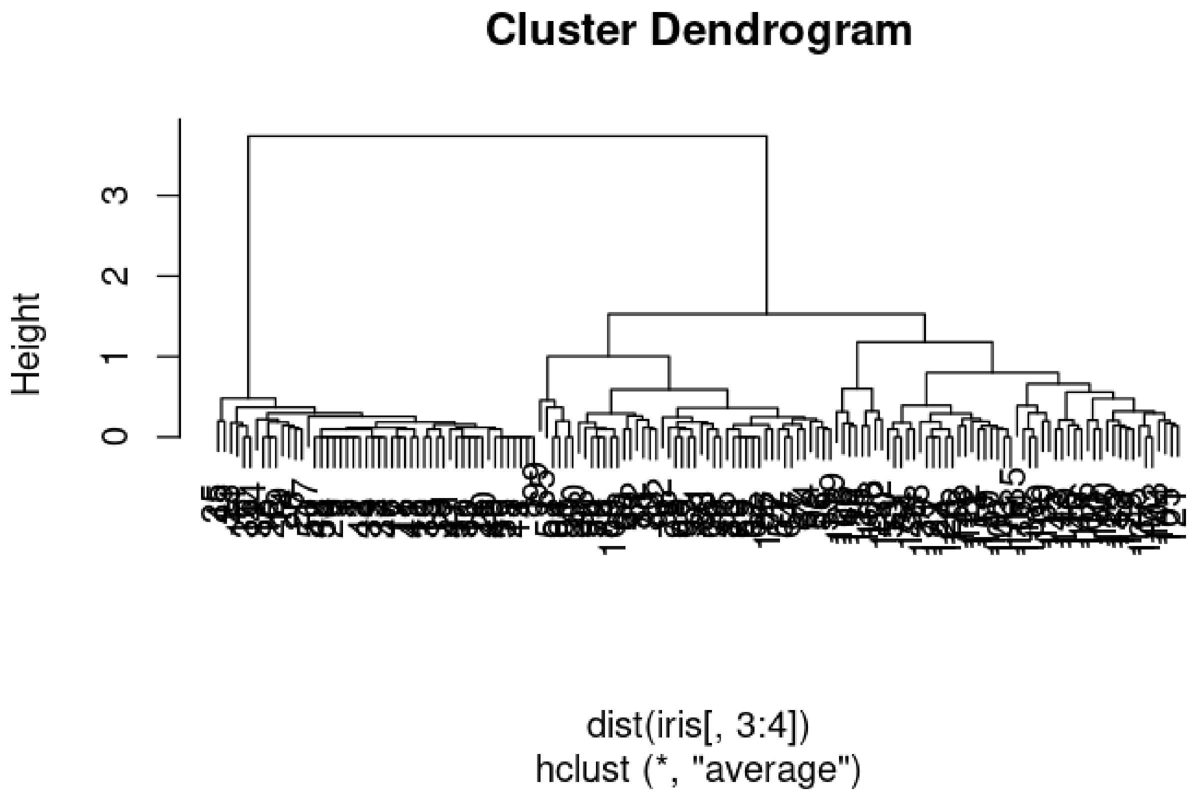
Average Linkage Method:

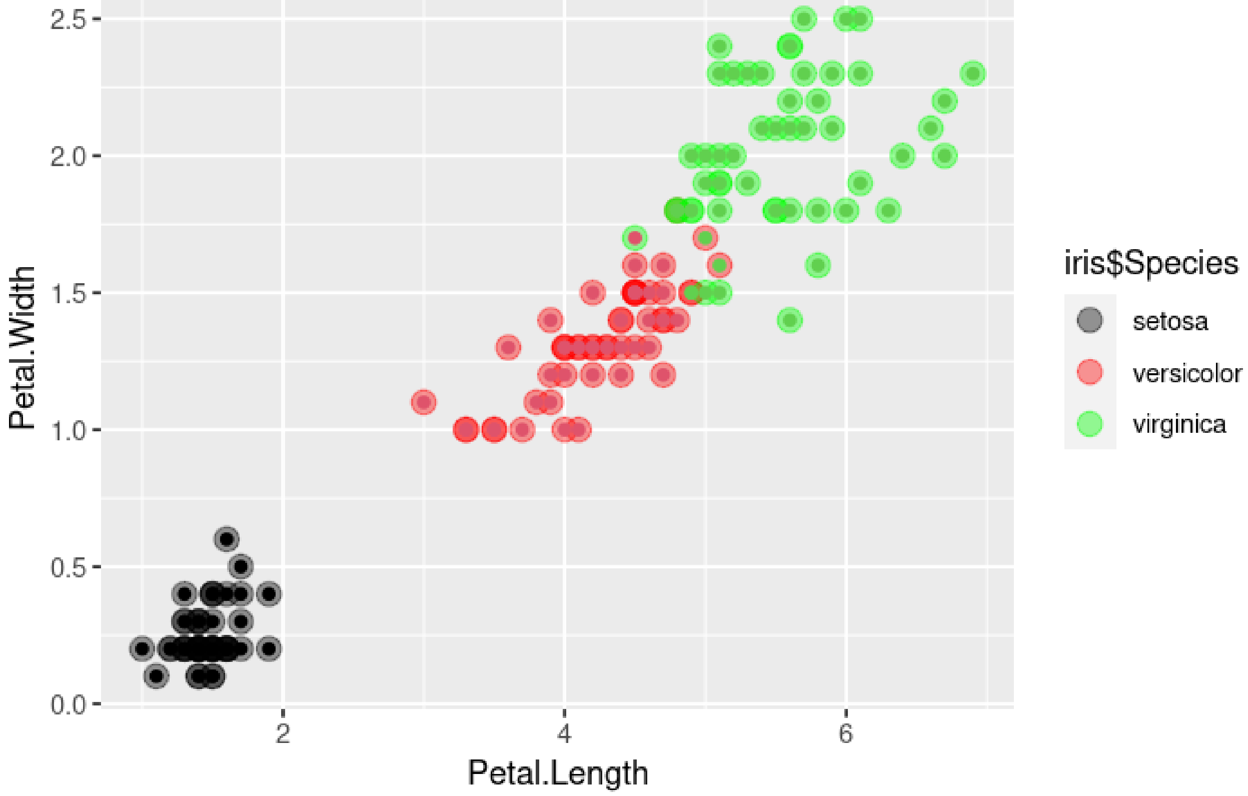


```
clusters <- hclust(dist(iris[, 3:4]), method = 'average')
clusterCut1 <- cutree(clusters, 3)
table(clusterCut1, iris$Species)
plot(clusters)
```



```
> clusters <- hclust(dist(iris[, 3:4]), method = 'average')
> clusterCut1 <- cutree(clusters, 3)
> table(clusterCut1, iris$Species)

clusterCut1 setosa versicolor virginica
          1     50          0         0
          2      0         45         1
          3      0          5        49
```


Cluster Dendrogram
Height
0
1
2
3
dist(iris[, 3:4])
hclust (*, "average")


Petal.Width
0.0
0.5
1.0
1.5
2.0
2.5
2
4
6
Petal.Length
iris$Species
setosa
versicolor
virginica

Using the average linkage clustering algorithm with 3 clusters maintains the accuracy for the classification of Iris-Setosa and improves the accuracy of the classification of Iris-Versicolor. This method seems more accurate than the Complete method, in this case with three clusters there are 16 specimens outside of their species that is almost 90% of the specimens are properly classified by this algorithm and with the Complete method this ratio was around 81%.

Conclusion:

By analyzing the given data into appropriate graphs, we have optimized strategies and ensured that efficiency is maintained.
It could separates the data in three groups:

- Definitely Iris-Setosa
- Basically all Iris-Virginica
- Basically all Iris-Versicolor