BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name) 1. Mankame, Piyusha 2. Rutva, Dharmendra Joshi 3.	4. 5. 6. 7. **GROUP NUMBER:** 5a
MSc in: Business Analytics	
Module Code: SMM636	
Module Title: Machine Learning	
Lecturer: Dr Rui Zhu	**Submission Date: 5th April 2023**

Declaration:

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission:

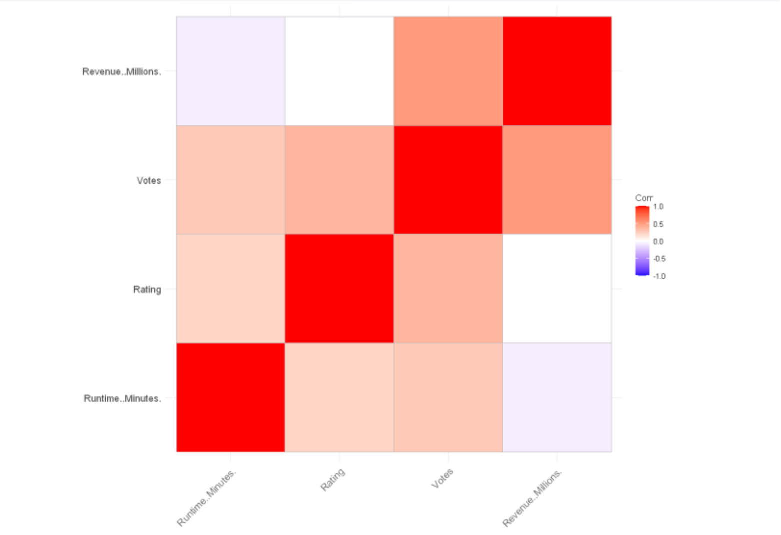
Final Mark: %

Final Report

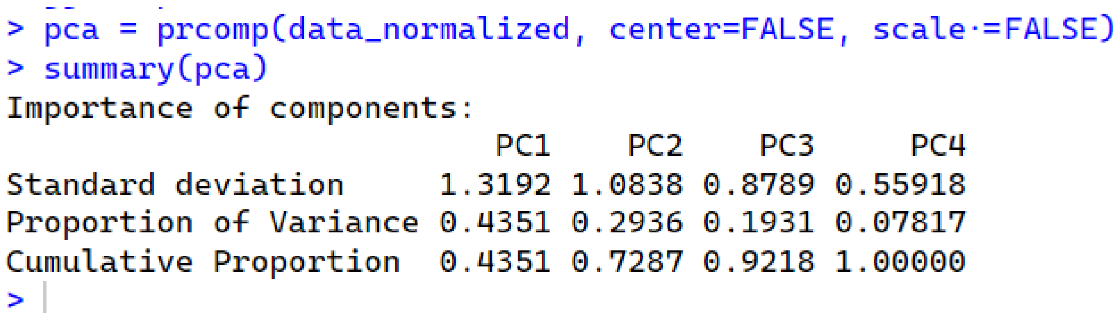
Question 1

***Part 1 - Apply PCA on the numerical variables excluding Year and report your findings*.**

After cleaning the dataset and replacing the null values in the last column with the column mean, we have plotted a correlation matrix to find the pairwise correlation between the variables of the dataset.

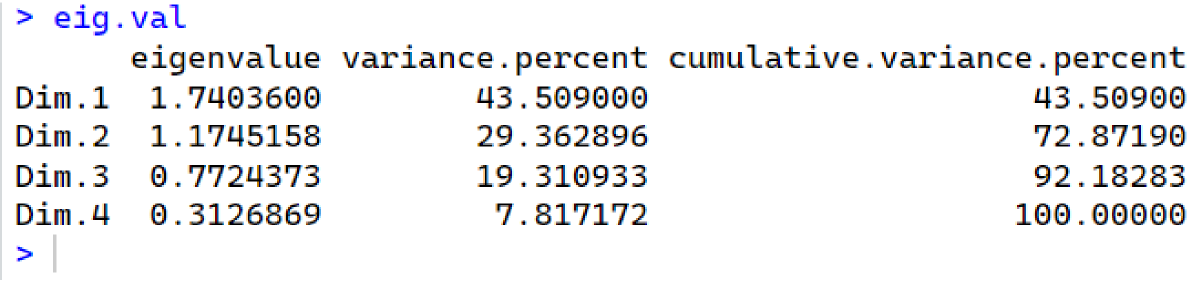


According to the plot, ‘Votes’ and ‘Revenue..Millions.’ are highly correlated followed by ‘Votes’ and ‘Rating’. It means that ‘Votes’ and ‘Revenue..Millions.’ have a direct relationship and the both increase or decrease together. While ‘Rating’ and ‘Runtime..Minutes.’ are not very highly correlated and hence don’t influence each other much.



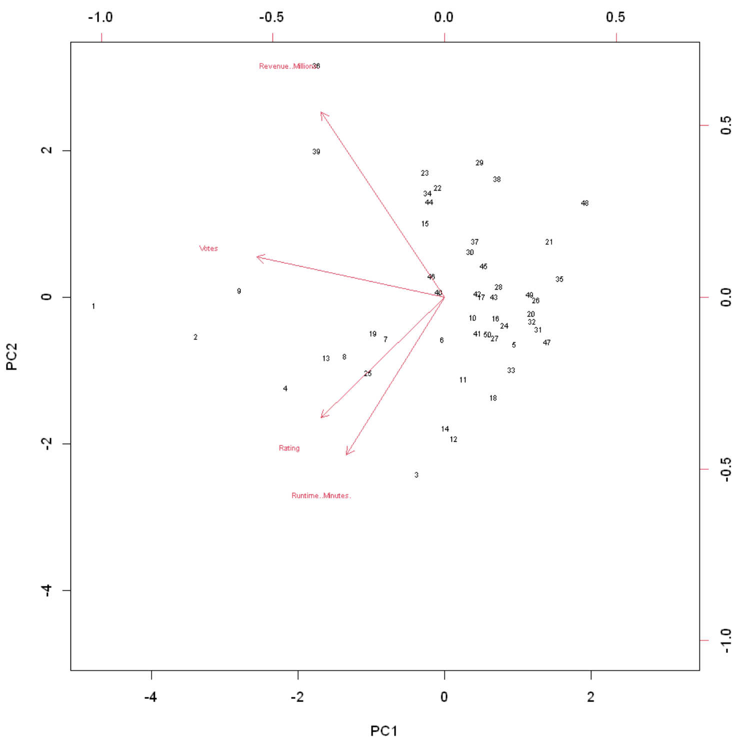
```
> pca = prcomp(data_normalized, center=FALSE, scale =FALSE)
> summary(pca)
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     1.3192 1.0838 0.8789 0.55918
Proportion of Variance 0.4351 0.2936 0.1931 0.07817
Cumulative Proportion  0.4351 0.7287 0.9218 1.00000
> 
```

The Standard deviation of PC1 captures the most variability of the dataset. PC1 explains 43.51% of the total variance and the PC2 explains 29.36% of the dataset. Together, they explain 72.87% of the variability of the dataset and hence are important for summarizing and identifying pattern of variations.

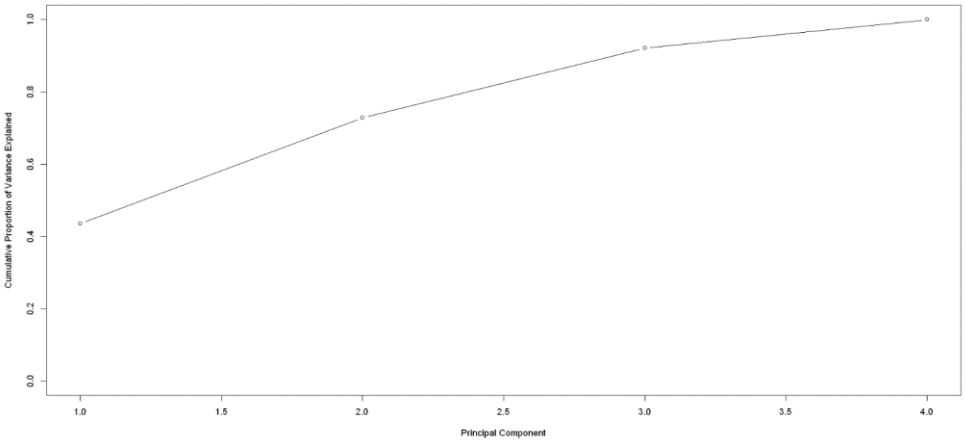


```
> eig.val
      eigenvalue variance.percent cumulative.variance.percent
Dim.1  1.7403600        43.509000                    43.50900
Dim.2  1.1745158        29.362896                    72.87190
Dim.3  0.7724373        19.310933                    92.18283
Dim.4  0.3126869         7.817172                   100.00000
> 
```

The first PC has the highest eigenvalue of 1.74 followed by the second PC with an eigenvalue of 1.17. According to the 'Kaiser criterion' we want to retain 2 principal components as they have eigenvalues above 1.



This the biplot depicting the two principal components and the direction of the variables. As per correlation matrix, ‘Votes’ and ‘Revenue..Millions.’ are positively correlated to each other and have a greater impact on the observations of the dataset. The plot also indicates that we have certain outliers such as 1, 2, 9, 36, 39, 48, etc.

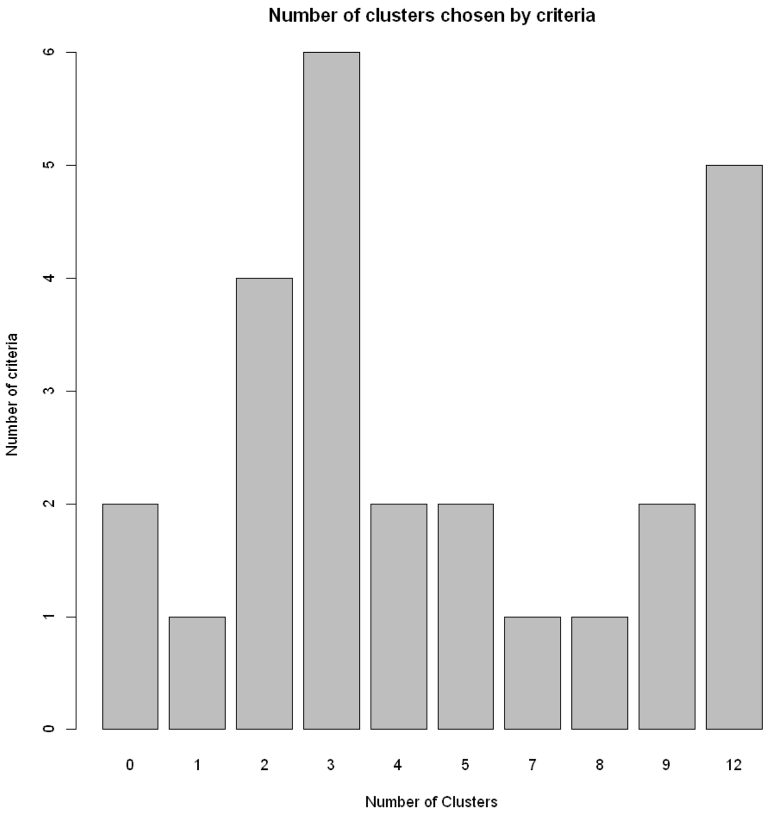


The Cumulative Variance plot shows that about 70% of the variance has been captured if we retain two principal components.

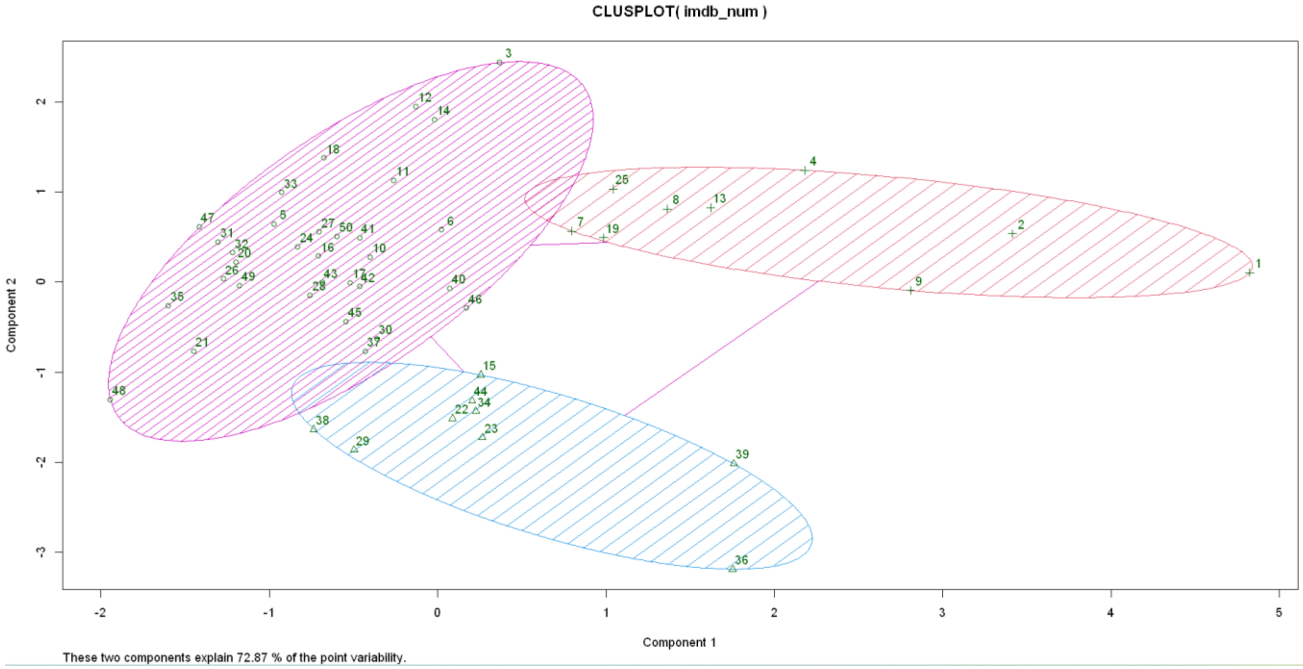
Part 2

A. K-Means Clustering

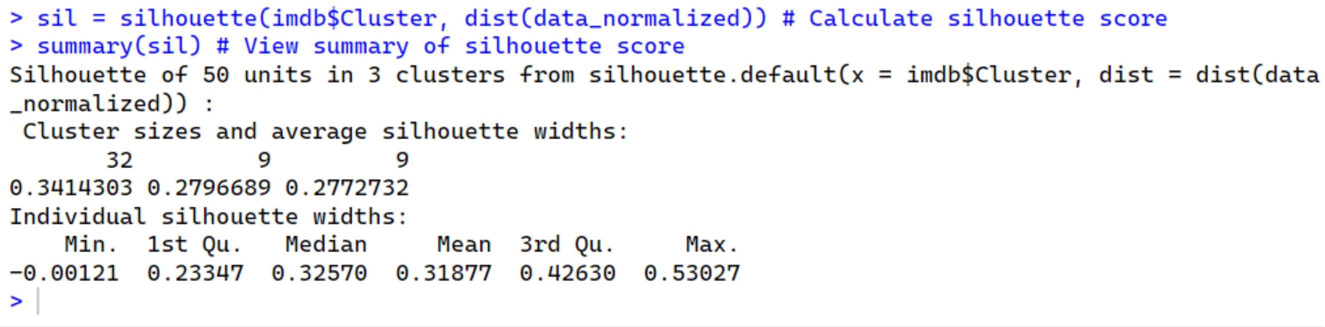
We used k-means clustering algorithm to cluster the movies based on their numerical variables excluding the Year variable. We chose k-means clustering because it is a widely used and efficient clustering algorithm that can handle large datasets. Additionally, k-means clustering is a centroid-based algorithm, which assigns each observation to the nearest centroid or cluster centre. This makes it easy to interpret and visualize the resulting clusters.



Here, we have used the 'NbClust()' function to determine the optimal number of clusters for the dataset. It perform the cluster analysis with k-means clustering using different values of k(number of clusters) and calculates different internal clustering validation indices to evaluate the cluster quality. I have also specified Euclidean's Distance as the argument to be used in the clustering analysis. According the the plot, 3 clusters would be the optimal number of clusters to represent the dataset.



According to the plot, the 3 clusters are able to represent the data logical manner.



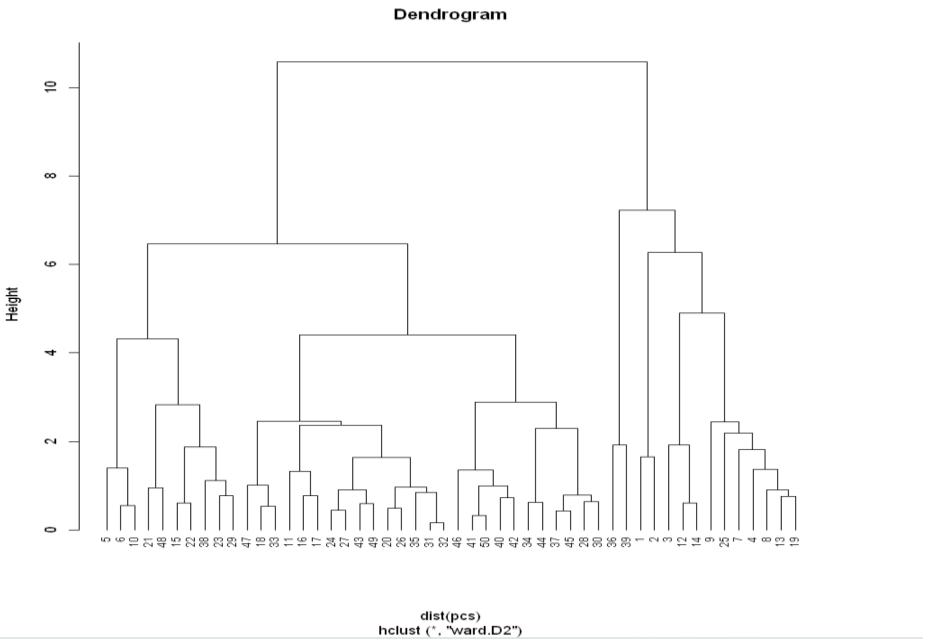
```
> sil = silhouette(imdb$Cluster, dist(data_normalized)) # Calculate silhouette score
> summary(sil) # View summary of silhouette score
Silhouette of 50 units in 3 clusters from silhouette.default(x = imdb$Cluster, dist = dist(data
_normalized)) :
 Cluster sizes and average silhouette widths:
       32         9         9 
0.3414303 0.2796689 0.2772732 
Individual silhouette widths:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-0.00121  0.23347  0.32570  0.31877  0.42630  0.53027 
> 
```

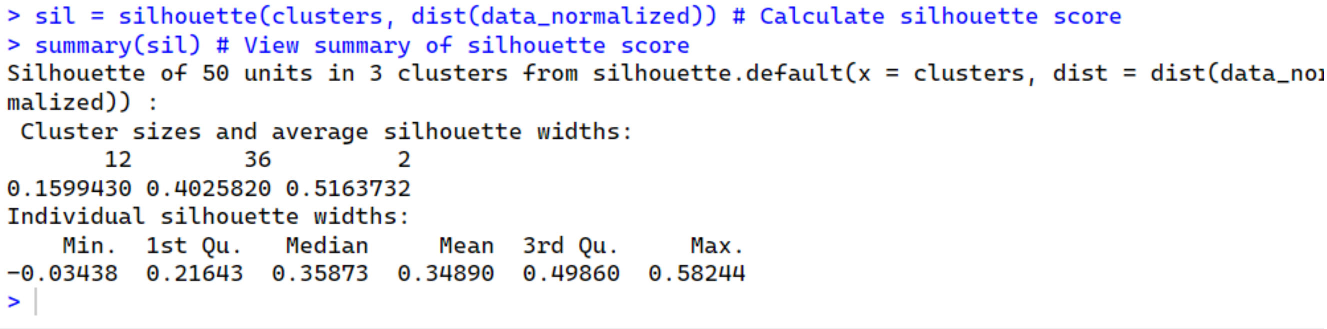
The summary shows that the cluster sizes and the average Silhouette widths for each cluster. The 1st cluster contain 32 datapoints(observations) while the 2nd and 3rd cluster contain 9 observations each. Also, Cluster 1 has the highest average Silhouette width of 0.341, indicating that its data points are well-clustered and have high similarity to other data points within the same cluster.

B. *Hierarchical Clustering*

We have also use Hierarchical Clustering to cluster the datapoints. One of the main advantages of hierarchical clustering over k-means clustering is that it does not require the number of clusters to be specified beforehand. This is because hierarchical clustering produces a tree-like structure that can be cut at different levels to produce different numbers of clusters, whereas k-means clustering requires the number of clusters to be specified before the analysis is performed.

Furthermore, hierarchical clustering can be more robust to noise and outliers in the data, as it considers all data points in the clustering process, whereas k-means clustering can be sensitive to outliers.

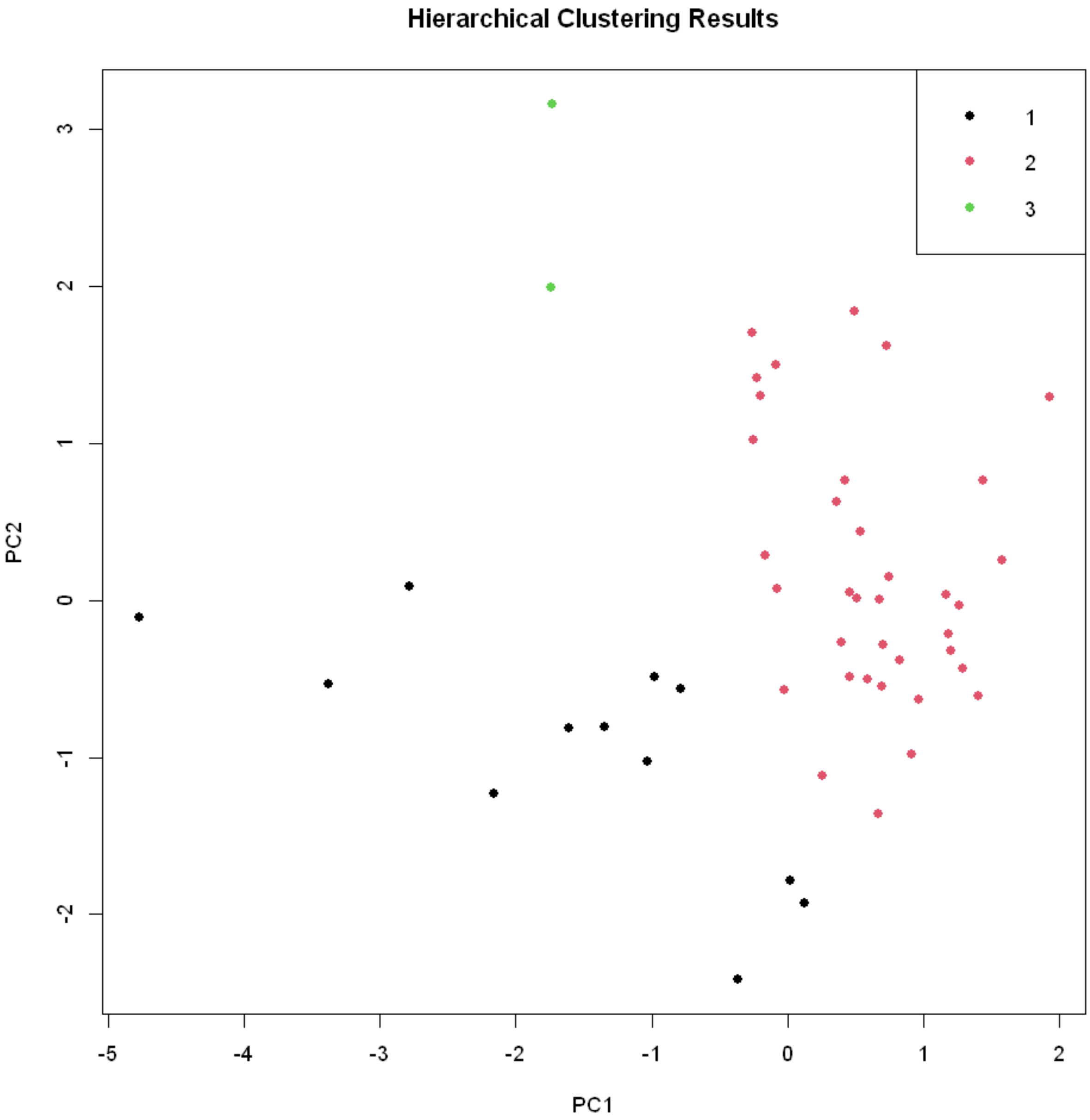




```
> sil = silhouette(clusters, dist(data_normalized)) # Calculate silhouette score
> summary(sil) # View summary of silhouette score
Silhouette of 50 units in 3 clusters from silhouette.default(x = clusters, dist = dist(data_no
malized)) :
 Cluster sizes and average silhouette widths:
       12        36         2 
0.1599430 0.4025820 0.5163732 
Individual silhouette widths:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-0.03438  0.21643  0.35873  0.34890  0.49860  0.58244 
> 
```

In the given output, we have performed hierarchical clustering with 3 clusters, and the summary output shows that the average silhouette score for all the clusters is 0.34 as compared against the silhouette score of k-means clustering which was 0.31. The cluster sizes are 12, 36 and 2. Cluster 2 has the highest average silhouette score of 0.40, indicating well-separated and dense clusters. Cluster 1 has the lowest average silhouette score of 0.16, indicating that some of the points may belong to other clusters. Cluster 3 has only two points, so its average silhouette score may not be reliable.

A higher mean silhouette score is generally considered an indicator of better clustering performance. Here, the mean silhouette score of hierarchical clustering is 0.34 while that of k-means is 0.31, it suggests that the hierarchical clustering algorithm has produced better-defined clusters than k-means. However, it's important to note that silhouette score is just one measure of cluster quality, and you should also consider other metrics and criteria when deciding which clustering algorithm to use.



Total Word Count – 746 words