BIG DATA ANALYTICS COURSEWORK

Introduction:

In this study we are analyzing EUROPEAN EMPLOYMENT data during 1979 using unsupervised learnings called clustering. Clustering is used widely to group similar data together based on their characteristics without labeled data. The main objective of this analysis is to identify patterns of employment structures across European countries using the two clustering techniques, K-means clustering and hierarchical clustering.

Here are the following steps for our analysis:

Step-1:

Load and preprocess data-

The analysis was conducted in R studio with the packages tidyverse for data manipulation, clustering for clustering algorithms, factoextra for visualizations and fpc for projection. We load and read the data into R. The dataset consists of employment percentage across eight different sectors with 26 European countries. We removed the country variable as it is not numeric identifier. The data is complete as there are no missing values in it.

Scale the data-

To remove the effect of differing scales, all variables in the data are standardized or normalized as clustering algorithm K-means is sensitive. In this we preferred standardized method.

Step-2:

Apply clustering-

K means- It helps to check the optimal number of clusters in advance.

We applied the two widely used methods, elbow and silhouette approach to find the optimal number of clusters. In elbow method it determines the k by analyzing within-Cluster sum of squares(wcss) whereas in silhouette method it evaluates the cluster quality for different k. In our analysis both methods suggested the cluster number k as 3.

The below figure shows us the methods

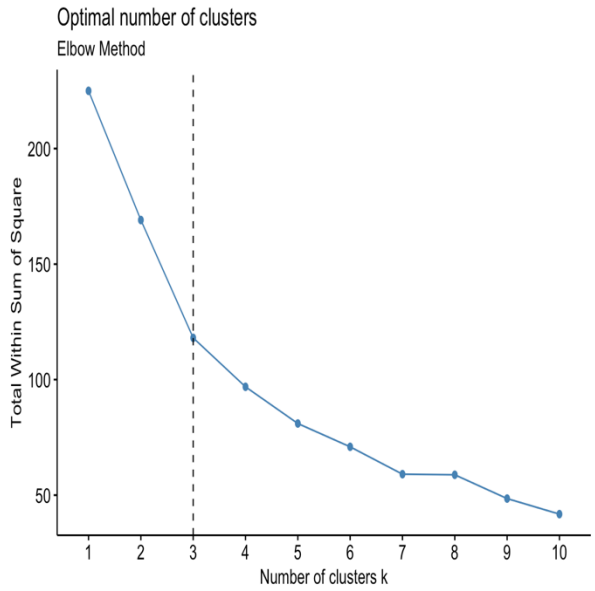


fig- 1: elbow method

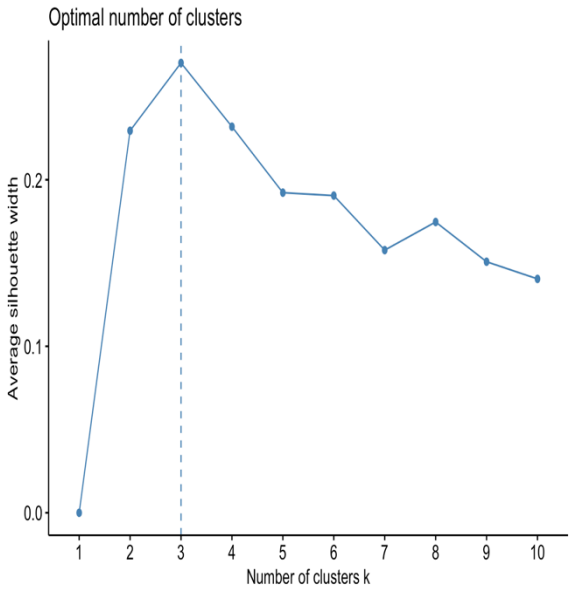


fig- 2: silhouette method

To perform K-means clustering, we applied three centers with 25 different configurations. Then the cluster size is 2,8,16.

```
> kmeans_result <- kmeans(scaled, centers = 3, nstart = 25)
> kmeans_result
K-means clustering with 3 clusters of sizes 2, 8, 16

Cluster means:
         Agr        Min         Man          PS        Con         SI        Fin        SPS
1  2.4840999 -0.1585972 -2.09163650 -0.81786082 -2.6223997 -1.5644365  0.7838767 -1.6725978
2  0.3526007  0.9496004  0.38775681  0.14568146  0.1121882 -0.9524484 -1.0332921 -0.4097020
3 -0.4868128 -0.4549756  0.06757616  0.02939187  0.2717058  0.6717788  0.4186614  0.4139257
           TC
1 -2.11729825
2  0.50582974
3  0.01174741

Clustering vector:
 [1] 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 1 2 2 2 2 2 2 2 1

Within cluster sum of squares by cluster:
[1] 13.44198 39.14856 65.48018
 (between_SS / total_SS =  47.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

from this figure we can say the cluster 1 is very small size it has 2 countries with unique employment patterns, where there is a high employment in agriculture, moderate in finances

and low employment in manufacturing, construction and services. So, we can say that these two countries rely on agriculture employment.

The cluster 2 has 8 countries with high mining employment, manufacturing, transportation or communication and low in services and finances employment.

The cluster 3 has 16 countries with strong employment in construction, finances, services and poor employment in agriculture and mining.

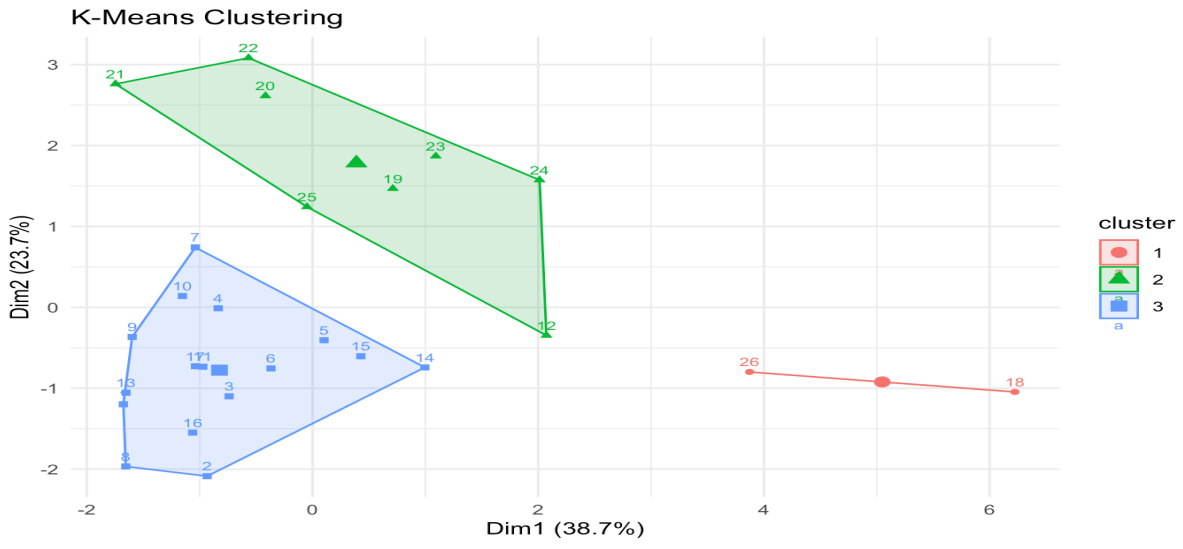
```
> split(europeanJobs$Country, europeanJobs$Cluster)
$`1`
[1] "Turkey"     "Yugoslavia"

$`2`
[1] "Greece"         "Bulgaria"       "Czechoslovakia" "EGermany"       "Hungary"
[6] "Poland"         "Rumania"        "USSR"

$`3`
 [1] "Belgium"     "Denmark"     "France"      "WGermany"    "Ireland"     "Italy"
 [7] "Luxembourg"  "Netherlands" "UK"          "Austria"     "Finland"     "Norway"
[13] "Portugal"    "Spain"       "Sweden"      "Switzerland"
```

This figure tells us that in cluster 1 the 2 countries are Turkey and Yugoslavia. In cluster 2 the countries are Greece, Bulgaria, Czechoslovakia, EGermany, Hungary, Poland, Rumania and USSR. In cluster 3 the 16 countries are Belgium, Denmark, France, WGermany, Ireland, Italy, Luxembourg, Netherlands, Uk, Austria, Finland, Norway, Portugal, Spain, Sweden and Switzerland.

Using (fviz cluster) function with scaled data and k-means result we plot the k means clustering. The below figure represents the PCA (principal component analysis) plot



The two principal component 38.7% and 23.7% shows the variation in data. The red colored is cluster 1, green is cluster 2 and blue is cluster 3. This plot shows that cluster 1 has very different employment pattern compared to cluster 2 and cluster 3. Whereas cluster 2 and cluster 3 has similar employment patterns.

Step-3:

K-means summary:

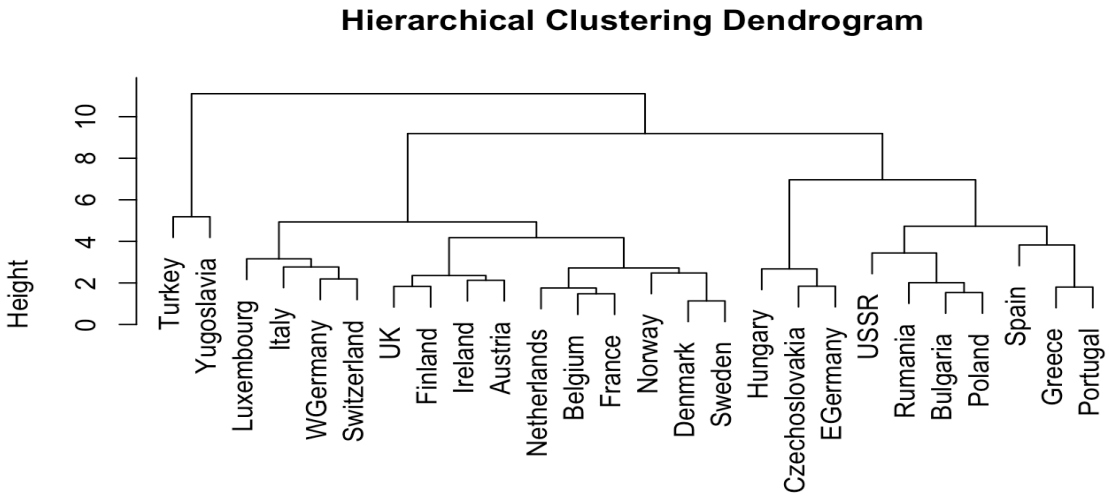
```
> kmeans_summary <- aggregate(. ~ kmeans_result$cluster, data = europeanJobs[, -1], FUN = mean)
> kmeans_summary
  kmeans_result$cluster     Agr    Min      Man      PS    Con       SI   Fin    SPS     TC
1                     1 57.7500 1.1000 12.35000 0.60000 3.8500  5.80000 6.200  8.600 3.6000
2                     2 24.6125 2.1750 29.72500 0.96250 8.3500  8.60000 1.100 17.225 7.2500
3                     3 11.5625 0.8125 27.48125 0.91875 8.6125 16.03125 5.175 22.850 6.5625
```

From this we can say that cluster 1 countries economy is dependable on agriculture as there is a 57.75% in farming. The cluster 2 countries mostly relying on manufacturing sector and the cluster 3 countries focusing on service sectors.

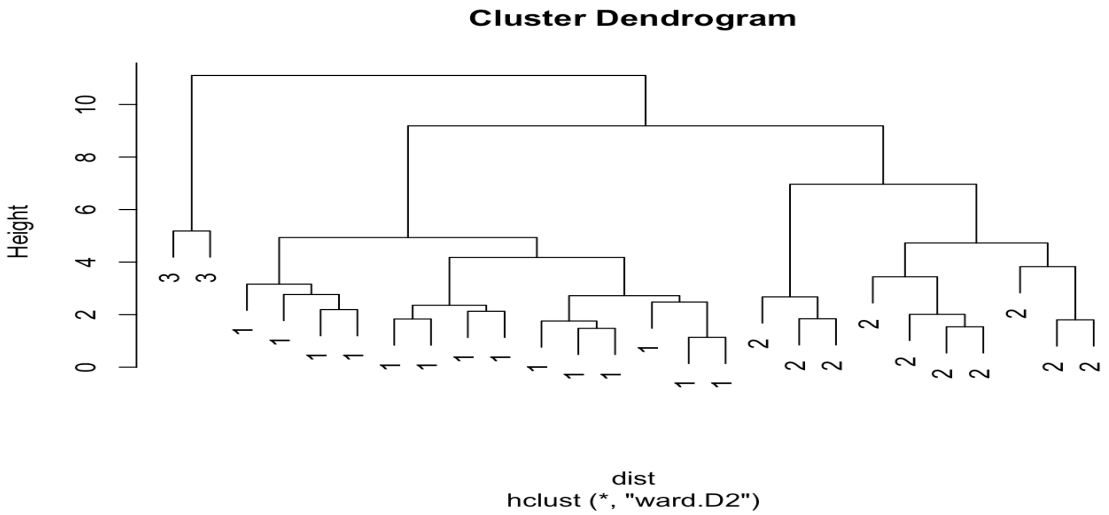
Step-4:

Hierarchical clustering: an analysis where it merges based on their similarity. In this there is no need to predefine the cluster number.

This clustering grouped the countries based on their employment patterns. It builds a tree like structure called dendrogram using Euclidean distance and wards method within the clusters.



The dendrogram cut at 3 clusters as K means analysis.



The left small branch with [3,3] shows our cluster 1 countries, the branch with 8 countries [2] shows the cluster 2 and the branch with 16 countries [1] shows the cluster 3.