Multivariate Data Analysis Using R

Cluster Analysis

Author: Opeyemi Ayanwale

Contents

1 Introduction

Cluster analysis is a multivariate tool for grouping or clustering items or observations based on how closely they are related. It can be used in market research to identify distinct groups of customers, city planning to group houses based on house type and geographical location, weather classification, bio-informatics, and so on. There are different clustering methods such as hierarchical clustering and non-hierarchical method (k-means clustering) (Brian et al., 2011).

The primary goal of this report is to conduct cluster analysis with the chatterjee-price attitude dataset which can be found in datasets package in R. The data is aggregated from a survey of clerical employees which contain 35 employees for 30 randomly selected departments (R Core Team and contributors worldwide, 2021). Non-hierarchical (k-means) method is used in this report to determine the distinct group in the Chatterjee-price attitude data. The following sections explain the details and findings of this analysis.

2 Problem Statement

The chatterjee-price attitude dataset examined in this report includes 30 observations and 7 variables from a questionnaire administered to approximately 35 employees of a financial organization. The percentages represent the proportion of positive responses to 7 questions from each of the 30 randomly selected departments. The variables in dataset includes overall rating, handling of employees complaints, special privileges, learning opportunities, raises based on performance, critical and advancement. They are numerical variables and there are no missing data in the dataset. The chatterjee-price attitude dataset can be found inside datasets package in R (R Development Core Team, 2021).

3 Statistical Methods

Non-hierarchical (k-means) method is used in this report to examine any existing cluster in the chatterjee-price attitude data.

3.1 Non-hierarchical (K-means method)

The non-hierarchical method is also known as k-means clustering method. K-means is an unsupervised classification algorithm that group objects into several 'k' clusters where k is the number of groups specified. The algorithms are implemented by the R function kmeans(), and these algorithms can compute the distance or dissimilarity between each pair of observations using the squared Euclidean distance measure. K-means aim is to minimize the within cluster of sum of squares (Brian et al., 2011).

The Euclidean distance

$$Dist_{XY} = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2}$$

where X and Y are two vector of length m

3.2 Cluster Analysis in R using K-means

Clustering for K = 2, 3, 4, 5 is carried out using the kmeans function and graphically represented using the fviz gap statistics generated by the clusgap function in the cluster and factoextra package in R (Maechler et al., 2021; Kassambara and Mundt, 2020).

K-means

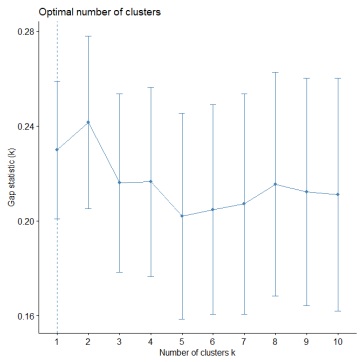
$$K(n) = kmeans\,(df,\, centers = m,\, nstart = 25)$$

where n specifies the number of clustering, df is the dataset, m specifies the cluster centers and nstart = 25 will generate the initial 25 configuration.

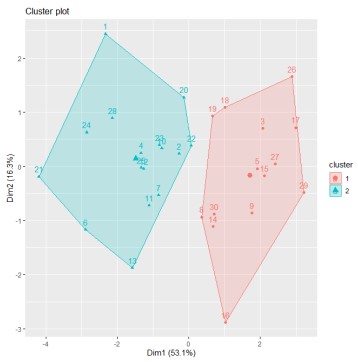
The statistical software R (R Development Core Team, 2021), version 4.1.1 was used for all analysis and visualizations.

4 Statistical Analysis

The distance matrix is calculated using the Euclidean method and the total number of objects equals to 30. Figure 1 shows gap statistic plot and Table 1 shows the within sum of squares (WSS) values. From the plot the gap statistics is highest at k = 2 clusters. K-means clustering with optimal K (K=2) is performed with 2 cluster sizes 14, 16.



(a) Optimal Cluster



(b) Cluster Plot

Figure 1: Cluster analysis plot.

Table 1: Within Sum of Squares (WSS).

	Within sum of squares	between$_{ss}/total_{ss}$
k=2	63.51206 62.63001	37.9 %
k=3	63.512055 7.616089 33.846393	48.3 %
k=4	27.309576 20.244250 7.616089 31.468155	57.3 %
k=5	7.616089 3.143101 37.925660 6.350570 20.244250	62.9 %

The WSS is calculated for each k value. The value of k with the least amount of WSS (37.9 %) is chosen as the optimal value, which is k = 2, which corresponds to the gap statistic plot result.

5 Summary

This report examined the chatterjee-price attitude of 35 employees of a financial organization. The aim of the report is to perform a cluster analysis to identify if there are natural groupings in the price attitude of the employees. K-means cluster analysis was used to identify the clusters and gap statistics was used to select the optimal cluster in the dataset. The optimal value of k is chosen at k = 2 because it has the least amount of within sum of squares (37.9 %). Gap statistics also show that the optimal k is 2 and from the plots, we have two properly partitioned clusters.

Bibliography

Brian, S. E., Sabine, L., Morven, L., and Daniel, s. (2011). *Cluster Analysis, 5th Edition*. JohnWiley Sons Ltd.

Kassambara, A. and Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2021). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.2 — For new features, see the 'Changelog' file (in the package source).

R Core Team and contributors worldwide (2021). *R: The R Datasets Package*. R Foundation for Statistical Computing, Vienna, Austria.

R Development Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.