Applying Cluster Analysis in R for Business Segmentation

Burcu Gündüz Altay
Fresenius University of Applied Science

Data Analysis for Decision Making (WS 2025/26)
Prof. Dr. Stephan Huber
2025-12-11

Author Note

Correspondence concerning this article should be addressed to Burcu Gündüz Altay, Email: altay.burcu@stud.hs-fresenius.de

Abstract

Cluster analysis is a fundamental method in data-driven segmentation, enabling analysts to identify natural groupings in customer datasets without predefined labels. As firms increasingly rely on personalized marketing, CRM optimization, and targeted communication, statistically derived clusters offer a more nuanced understanding of consumer behavior than traditional demographic segmentation. This handout provides an academic overview of cluster analysis with a focus on K-Means and its application in business environments. A structured R workflow is presented to demonstrate how clustering can be implemented and evaluated using real data. Emphasis is placed on managerial interpretation, illustrating how analytically derived customer segments can support targeting, pricing, and retention strategies.The document has been prepared using an APA-compliant Quarto template to ensure proper academic formatting (Huber, 2024) and concludes with an exercise designed to reinforce the practical application of clustering in R.

Applying Cluster Analysis in R for Business Segmentation

1 Introduction

The increasing complexity of modern consumer markets has accelerated the shift toward data-driven segmentation strategies. Traditional demographic segmentation often fails to capture heterogeneous behavioral patterns, particularly in markets where customers engage with brands across multiple channels and touchpoints. Cluster analysis offers a systematic approach for identifying groups of individuals who exhibit similar characteristics, deriving structure directly from the data rather than from predefined assumptions (James et al., 2021) . This makes clustering an essential tool for personalization, customer lifetime value management, and resource allocation. In business analytics, clustering is frequently applied to variables such as spending behavior, engagement metrics, and purchasing frequency, providing firms with actionable insights for strategic decision-making.

2 Conceptual Foundations of Cluster Analysis

Cluster analysis is an unsupervised machine learning technique that groups observations based on similarity across selected variables (James et al., 2021). Unlike supervised methods, clustering does not rely on labeled data; instead, it reveals inherent patterns within datasets. The objective is to maximize within-cluster similarity while ensuring that clusters differ from one another. By providing structure where managerial intuition may fall short, clustering contributes to more evidence-based strategic decisions and marketing appliances.

3 Clustering Methods in Business Analytics

3.1 *K*-Means Clustering

K-means is one of the most widely used clustering methods in both academic and commercial settings (Hartigan & Wong, 1979). The algorithm partitions observations into *k* clusters by iteratively updating cluster centroids and assigning each data point to the nearest centroid. Its strengths include computational efficiency and intuitive interpretability. However, K-means assumes spherical clusters and similar cluster sizes, making it sensitive to outliers and variable scaling. Despite these limitations, its speed and simplicity make it highly effective for business segmentation.

3.2 Hierarchical Clustering

Hierarchical clustering groups observations and visualises how clusters merge or split using a dendrogram (Kassambara & Mundt, 2020). Although useful for exploratory analysis, it is less scalable for larger datasets. Therefore, this handout focuses on K-means due to its simplicity, speed, and practicality for business segmentation.

4 Determining the Number of Clusters

Two evaluation techniques are commonly used in business analytics:

4.1 Elbow Criterion

The Elbow Method examines the reduction in within-cluster variation as *k* increases. After a certain point, additional clusters yield diminishing improvements. The elbow in the curve indicates a balance between model complexity and explanatory power (James et al., 2021).

4.2 Silhouette Coefficient

The silhouette measure evaluates how well an observation fits within its assigned cluster compared to other potential clusters (Rousseeuw, 1987). Values close to +1 represent well-defined clusters, whereas negative values indicate poor separation. High average silhouette scores reflect strong cluster validity and are often used to justify the selection of *k* in marketing applications.

5 R Workflow for Cluster Analysis

5.1 Data Preparation

In this step, data is loaded, basic summary statistics are examined, and variables are standardised to ensure that all variables contribute equally to the analysis.

```r
df <- read.csv("Mall_Customers.csv")
summary(df)

df_scaled <- scale(df[, c("Age", "Annual.Income", "Spending.Score")])
```

5.2 Determining the Number of Clusters

These commands visually assess cluster quality using the Elbow and Silhouette methods for different *k* values.

```
library(factoextra)

fviz_nbclust(df_scaled, kmeans, method = "wss")        # Elbow method
fviz_nbclust(df_scaled, kmeans, method = "silhouette") # Silhouette method
```

5.3 Running *K*-Means Clustering

This step creates clusters by running the K-Means algorithm for the selected *k* value and assigns each observation to the relevant cluster.

```
set.seed(123)

k3 <- kmeans(df_scaled, centers = 3, nstart = 25)
k3$centers
k3$cluster
```

5.4 Visualizing Clusters

At this stage, the clusters are reduced to two dimensions using principal component analysis and presented visually, allowing the level of separation between segments to be examined (James et al., 2021).

6 Interpreting the Clusters: Business Insights

Cluster interpretation is central to generating managerial value. The following example illustrates how statistically derived clusters translate into actionable business segments:

Exercise :

Objective : To learn how to perform *K*-Means Clustering in R using retail data (Carseats dataset), evaluate the optimal number of clusters, visualize the results, and interpret cluster outcomes in a store performance segmentation context.

Task 1 Load the Dataset

Upload car seat data, select variables to be used for store segmentation, and scale these variables for *k*-means.

```
library(ISLR)
library(tidyverse)
library(factoextra)
```

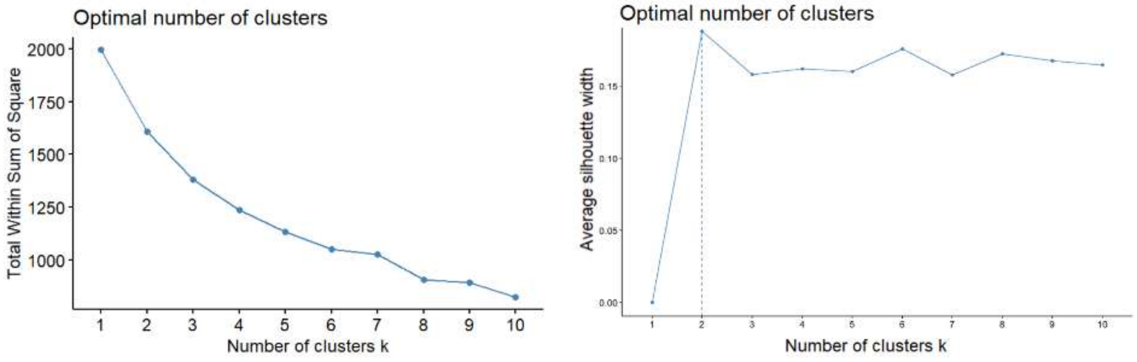
```r
data("Carseats")
df <- Carseats


df_scaled <- scale(df[, c("Sales", "CompPrice", "Income", "Advertising",
"Population")])
summary(df_scaled)
```

Task 2 How Many Clusters Make Sense? (Elbow + Silhouette)

Using elbow and silhouette plots to determine how many clusters *k* are appropriate for store segmentation.

```r
fviz_nbclust(df_scaled, kmeans, method = "wss")
fviz_nbclust(df_scaled, kmeans, method = "silhouette")
```

Figure 1



WSS decreases very rapidly from k=1 to k=3. After 3, the rate of decrease slows down significantly.This indicates that 3 clusters represent a good equilibrium point. Silhouette scores reach their highest value for k=2, but scores between k=3 and k=7 are also very close to each other and at an acceptable level.2 provides clearer differentiation, while 3 provides segmentation that offers greater insight. In terms of business decisions, three clusters make the most sense: both supported by Elbow and better reflect store diversity.

Task 3 Run *K*-Means with *k* = 3

Run the *k*-means algorithm with the selected number of clusters (e.g. *k* = 3) and examine the size and centres of each cluster.

```r
set.seed(123)
k3 <- kmeans(df_scaled, centers = 3, nstart = 25)


k3$size       # cluster sizes
k3$centers    # cluster profiles
```

Cluster 1: 129 shops, Cluster 2: 115 shops,Cluster 3: 156 shops Segments are of balanced size.

Task 4 Compare with *k* = 4 and Silhouette Scores

Compare the average silhouette scores for *k* = 3 and *k* = 4 and determine which model provides better cluster separation.

```r
library(cluster)


set.seed(123)
k4 <- kmeans(df_scaled, centers = 4, nstart = 25)


sil3 <- silhouette(k3$cluster, dist(df_scaled))
sil4 <- silhouette(k4$cluster, dist(df_scaled))


mean(sil3[, 3])
mean(sil4[, 3])
```

The average silhouette score is 0.1717 for *k* = 3 and 0.1750 for *k* = 4. This difference is very small and does not create a significant advantage in practice. On the other hand, *k* = 3 is better in terms of business decisions because: Segment sizes are more balanced, Easier to interpret, More suitable for strategy development, No unnecessary model complexity.

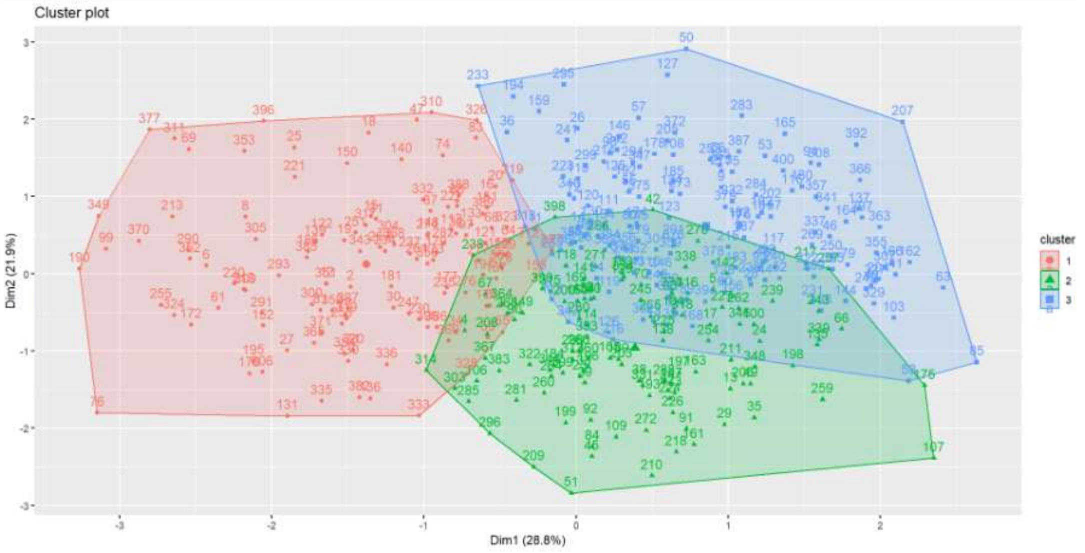
Task 5 Visualize the Clusters

Visually examine how stores are segmented according to selected variables.

```r
fviz_cluster(k3, data = df_scaled)
```

The graphic proves that the segments are truly distinct from one another,meaning that store segmentation is meaningful from a business perspective. Through this analysis, sales

Figure 2



growth and resource efficiency can be achieved by developing different marketing and investment strategies tailored to the needs of each store.

7 Managerial Significance of Cluster-Based Segmentation

Cluster analysis supports a wide range of strategic marketing decisions. By identifying latent behavioral groups, firms can improve targeting efficiency, tailor value propositions, design differentiated pricing strategies, and personalize CRM communications. Clustering also assists in identifying high-potential segments, detecting churn-prone users, and optimizing marketing spend. When combined with business expertise, analytically derived clusters substantially enhance a firm's ability to deploy precise and effective marketing interventions.

8 Conclusion

Cluster analysis provides a rigorous foundation for understanding complex customer behavior and developing actionable business segments. Through techniques such as K-Means and evaluation methods including the Elbow and Silhouette criteria, analysts can identify meaningful groups that inform strategic marketing decisions (Rousseeuw, 1987). R offers a transparent and replicable workflow for conducting clustering, visualizing results, and deriving insights. As firms increasingly rely on personalization and data-driven decision-making, mastery of clustering methods has become essential for effective business segmentation.

9 Affidavit

I hereby affirm that this submitted paper was authored unaided and solely by me. Additionally, no other sources than those in the reference list were used. Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and authorship. This paper, either in parts or in its entirety, in the same or similar form, has not been submitted to any other examination board and has not been published.

I acknowledge that the university may use plagiarism detection software to check my submission. I agree to cooperate with any investigation of suspected plagiarism and to provide any additional information requested by the university.

Burcu Gündüz Altay

Im Mediapark 4C, 50670 Cologne, Germany

11 December 2025

10 Checklist

- ☒ The handout contains 3-5 pages of text.
- ☒ The submission contains the Quarto file of the handout.
- ☒ The submission contains the Quarto file of the presentation.
- ☒ The submission contains the HTML file of the handout.
- ☒ The submission contains the HTML file of the presentation.
- ☒ The submission contains the PDF file of the handout.
- ☒ The submission contains the PDF file of the presentation.
- ☒ The title page of the presentation and the handout contain personal details (name, email, matriculation number).
- ☒ The handout contains a bibliography, created using BibTeX with an APA citation style.
- ☒ Either the handout or the presentation contains R code that proofs the expertise in coding.
- ☒ The filled out Affidavit.
- ☒ The link to the presentation and the handout published on GitHub.

Burcu Gündüz Altay, 11.12.2025, Cologne,Germany

11 References

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28*(1), 100–108.

Huber, S. (2024). *Apaquarto template for students*. https://github.com/hubchev/temp_apa_stu

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in r* (2nd ed.). Springer.

Kassambara, A., & Mundt, F. (2020). *Factoextra: Extract and visualize the results of multivariate data analyses*. https://CRAN.R-project.org/package=factoextra

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.