K-means Cluster Analysis

Gergely A. Marias

In this report, I explore a dataset of 2,000 users of a social networking app, aiming to segment them into meaningful groups using k-means clustering. Understanding user segmentation in a social networking app is significant as it enables personalized and targeted marketing strategies. By dividing the userbase into distinct groups based on their characteristics, the app can tailor content, features, and advertisements to better meet the specific needs and preferences of each segment. The central challenge is determining the optimal number of clusters for meaningful segmentation. After clustering, I interpret each cluster to understand the distinctive characteristics of its users. These insights will be vital for tailoring marketing strategies and making data-driven decisions to cater to the diverse needs of the userbase. K-means is a popular unsupervised machine learning technique used for clustering and data segmentation. The method is iterative and revolves around partitioning a dataset into K distinct clusters. The process begins with selecting K initial centroids (representative points) randomly within the data space. These centroids serve as the initial cluster centers. Each data point is then assigned to the nearest centroid based on a chosen distance metric, typically Euclidean distance. This step groups data points into clusters based on their similarity to the centroids. After data points are assigned to clusters, the centroids of these clusters are recalculated as the mean of all data points within each cluster. Steps 2 and 3 are repeated iteratively until convergence, meaning that the centroids no longer change or a predetermined number of iterations is reached. K-means clustering aims to minimize the Within-Cluster Sum of Squares (WCSS), which quantifies the compactness of clusters. The challenge lies in selecting the optimal number of clusters, K, as too few or too many clusters can hinder the meaningful interpretation of the data. Various techniques, such as the elbow method or silhouette analysis, are employed to determine the appropriate value for K.

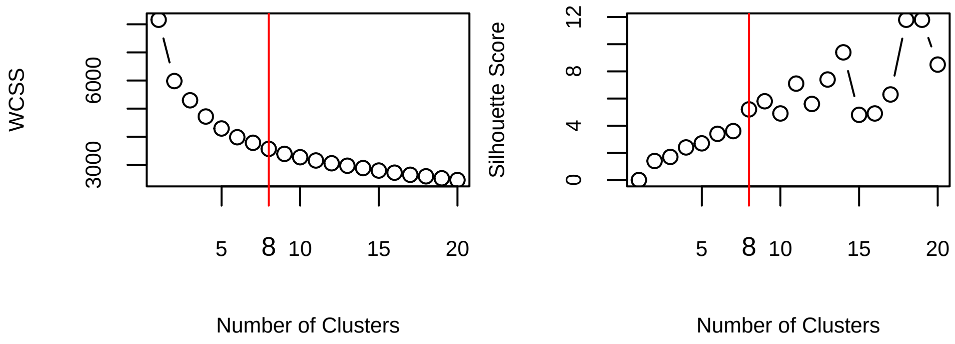

Figure 1: Optimizing Cluster Number with WCSS and Silhouette

Points with a higher silhouette coefficient indicate that the point is well matched to its own cluster and poorly matched to neighboring clusters, suggesting a good clustering but points with a lower silhouette coefficient indicate overlapping clusters. My data pre-processing started with the fact that I converted the “children” variable into a factor since it represents the number of children. However Categorical variables like “education,” “occupation,” “townsize,” and “children” cannot be directly used in many machine learning algorithms. Converting them into factors and then creating dummy variables ensures that these categorical attributes are appropriately represented in a numerical format. To ensure that variables on different scales have equal importance during clustering, I scaled the numerical variables “age” and “income.” Scaling the numerical prevents variables with larger ranges from dominating the clustering process.

I conducted k-means clustering, exploring a range of cluster numbers (1–20) and different random seeds (1–10) to assess the WCSS for each configuration. This involved running the k-means algorithm 200 times. Furthermore, I calculated silhouette scores for different cluster numbers with varying seeds to assess cluster quality. The results of the WCSS and Silhouette can be seen in Figure 1. However, the metrics indicated only that I need to choose a cluster number that strikes a balance between good evaluation metrics and can still be interpreted easily. Therefore,

Table 1: Stability of the cluster number 8 (left) and the most customers from each category (right)

WCSS value for K8	Category	Numbers	Percentage
Min. :3483	education1	1386	69%
1st Qu.:3507	children0	1133	57%
Median :3545	occupation1	1113	56%
Mean :3551	married	993	50%
3rd Qu.:3568	townsize0	989	49%
Max. :3944	female	914	46%

I decided to go with cluster number 8 (K8). - One reason for this decision was that I initially had to work with 8 features. I considered it as a kind of arbitrary threshold, but unfortunately, the analysis did not suggest a more optimal number. Therefore, I selected it based on interpretability. - To assess the stability of K8, I executed the algorithm 200 times with 200 distinct initial centroids, then examined the WCSS for each model. The findings are presented in Table 1 on the left side. If I had chosen K2, the WCSS would have been approximately 6,000. However, with K8, the average WCSS is nearly half of that at 3551. For further analysis, I can proceed with the model having the minimum WCSS value, which is 3483, by using set.seed(121).

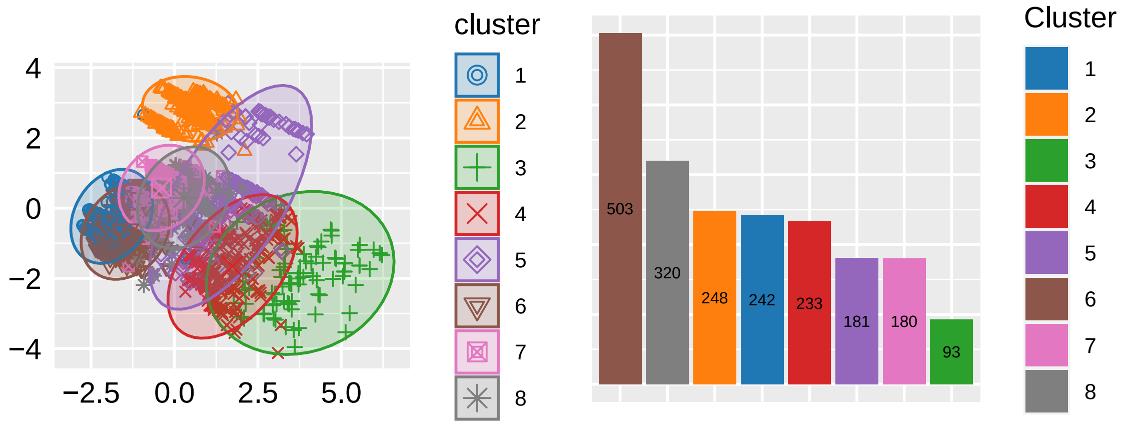

Figure 2: Cluster Map and Cluster Size Distribution

I created a visual representation of the clusters and additionally created bar plots to visualize the size of each cluster, showing how many users belong to each cluster (Figure 2). The biggest segment is cluster number 6 (C6), with 503 customers that exhibit the following characteristics: an average age of 31, an average income of $62,000, 99% with a high school education, 95% employed at a medium skill level, a 94% marriage rate, 79% female representation, 50% residing in small towns, and no children among half of its members. For further details, I refer to Appendix A, where Figures 3–4 can be found. The difference between the first (C6) and the second (C8) largest segments is that C8 is 92% unmarried and 92% male. The third largest segment is C2, with 248 customers, and its main characteristics are that their education level is unknown and they have no children. Moreover, it's quite clear that the richest segment is C3, even though it's the smallest with just 93 customers. This segment also stands out for having the highest average age at 62 and is unique in that it consists of highly skilled employees with a graduate school education. In a nutshell, over half of the customers in this analysis share some common characteristics: they have a high school education, no children, and are employed at a medium skill level, as shown in Table 1 on the right side.

In summary, I have gathered valuable insights into the customer base and I have reached a point where the limitations of the k-means method become evident. It suggests that perhaps using fewer than 8 clusters might be enough for interpretability, though this is a challenging decision and largely depends on the specific goals of the analysis, the desired level of precision and the context of future marketing interventions.

References:

Dolnicar, S., Grün, B., & Leisch, F. (2018). Market Segmentation Analysis (pp. 89-99).

Chapman, C., & Feit, E. M. (2015). R for Marketing Research and Analytics (pp. 302-303).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning (pp. 386-390).

Appendix A

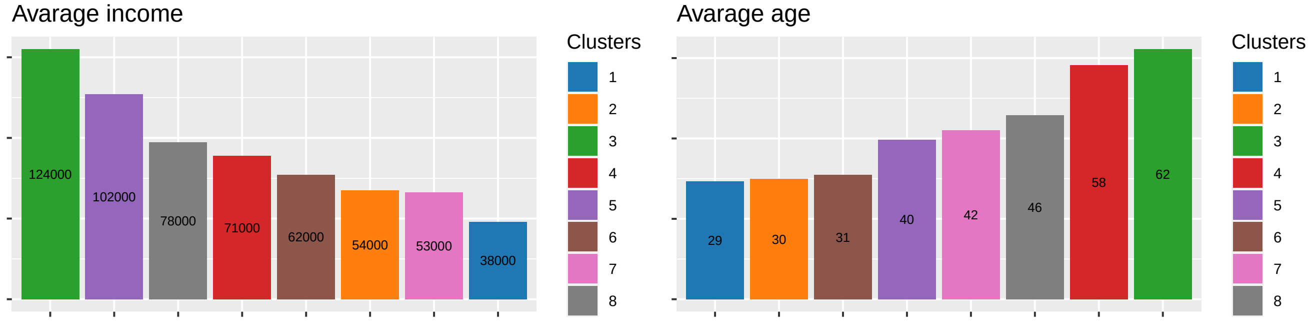

Figure 3: Income Distribution and Age Diversity by Clusters

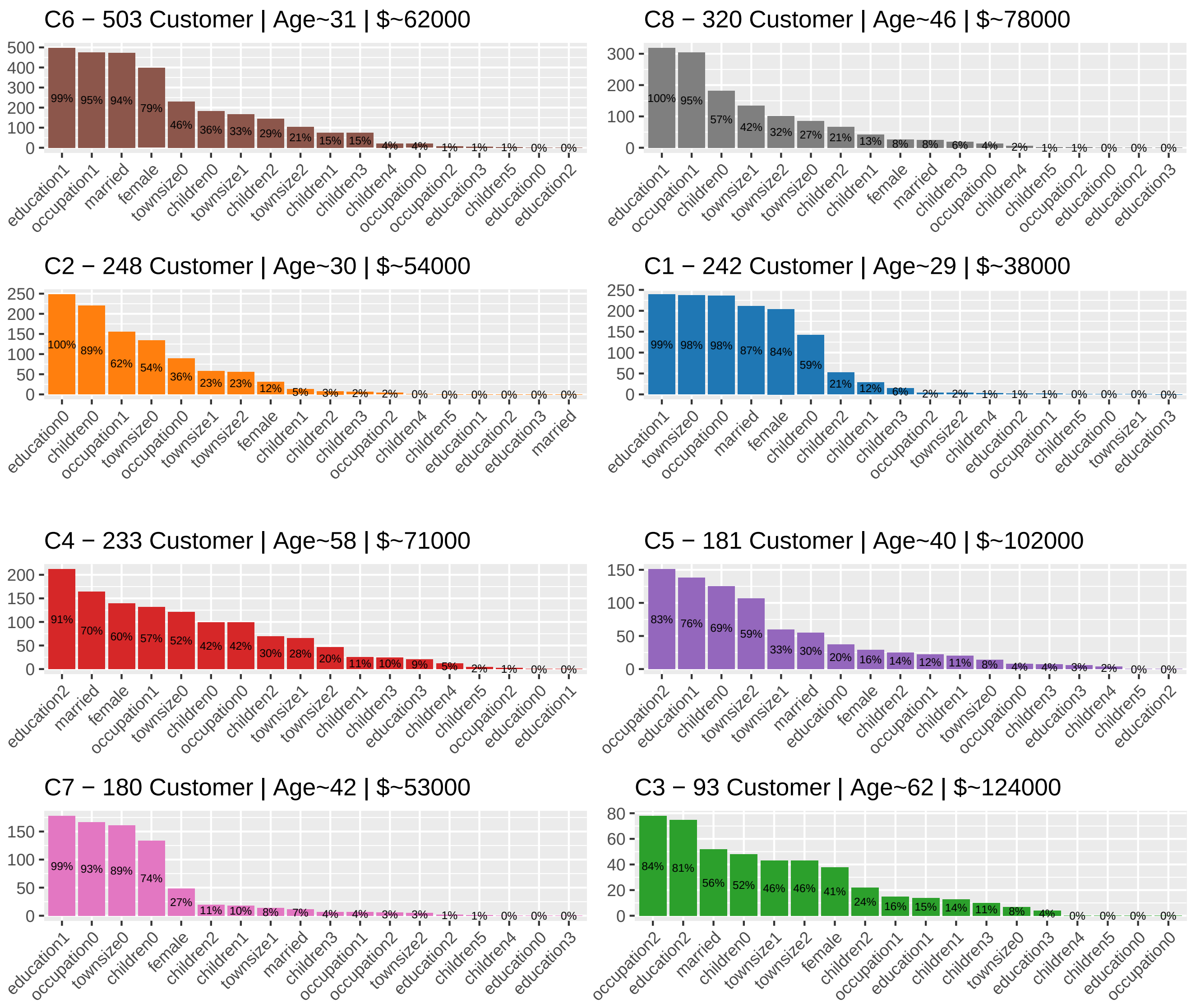

Figure 4: Customer Segments Analysis, segments by segments in descending order of size