Course paper

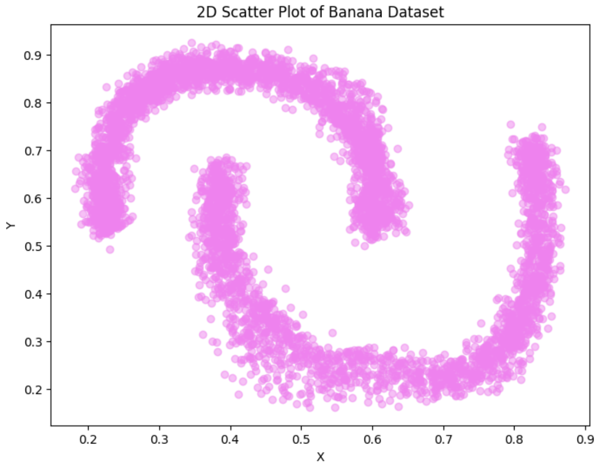
Cluster Analysis for Business Fall 2024

ELE 39091

Group Members:
ID 1109253
ID 1109249

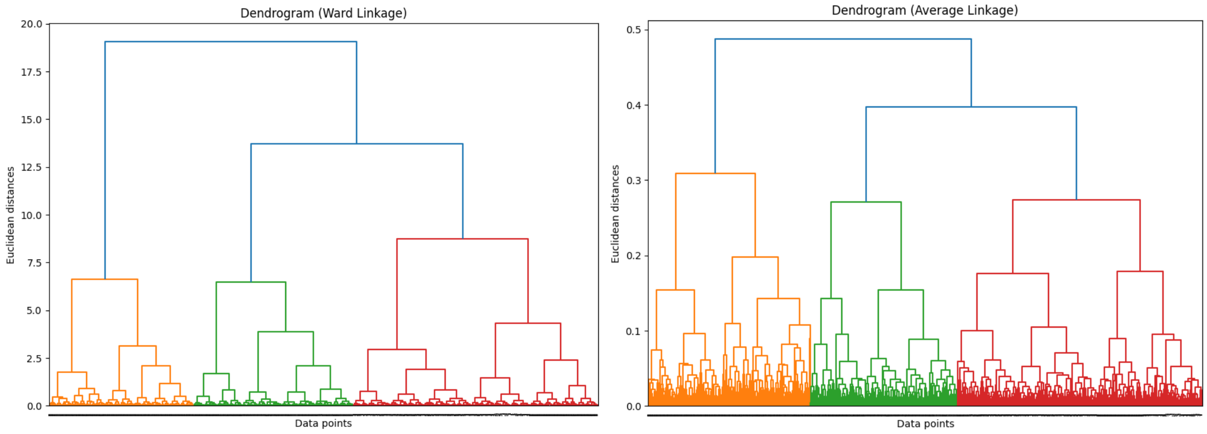
1) Task 1: Clustering and Representation Learning

a) After loading the toy "Banana Dataset", we visualise the data in a 2-dimensional scatter plot by displaying on the y-axis the "y feature" (as labelled in the dataset), and on the x-axis the "x feature".



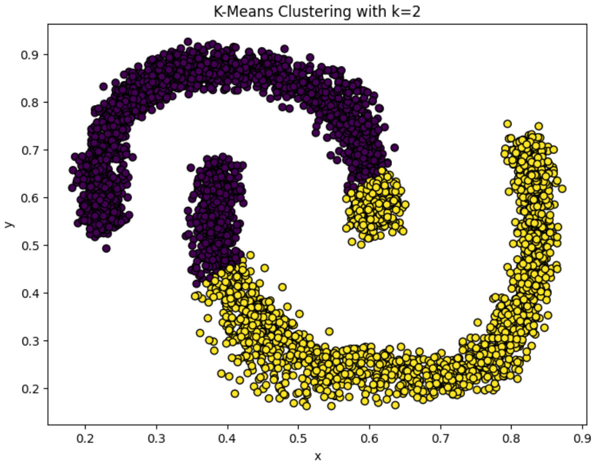
b) Based on our first visualisation of the data, we believe the optimal number of clusters to be two. Indeed, the data points seem to form two distinct, curved clusters, resembling a banana shape, with the space between these two regions relatively empty.

c) From the plots of the dendrograms shown below, it seems that the optimal number of clusters would be three for both methods. This conclusion is drawn based on the evaluation of the vertical distance (Euclidean distance) at which the merges take place; indeed, when there is a "large vertical jump", it signifies that the data points, assigned to the clusters being merged at that point, are far apart in terms of distance, and this is what happens after the data is merged into the three primary clusters (coloured orange, green, and red). However, if we compare it



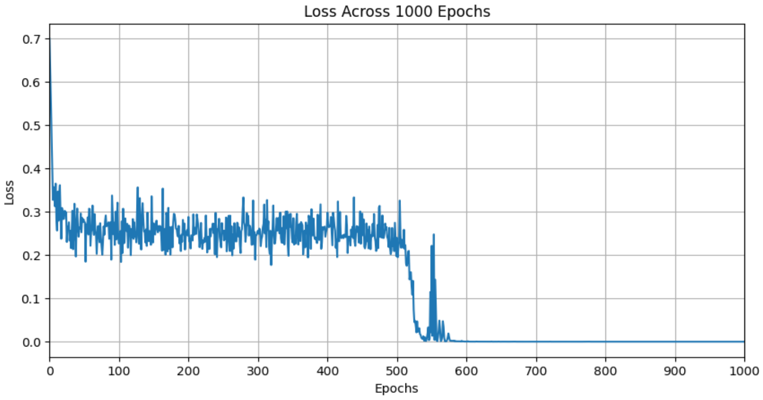
with our previous intuition drawn from the scatter plot visualisation, we can see that while the visual separation in the scatter plot points to two well-defined clusters, the dendrograms offer a more "granular" view, revealing that additional structure may exist within the data, leading to the identification of a third cluster. The disagreement between our two results reveals the need for an in-depth analysis, with the evaluation of metrics such as the Silhouette score and the Davies-Bouldin score (as reported in the appendix).

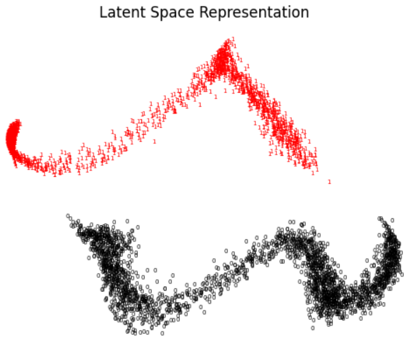
d) After selecting the K-parameter equal to two based on our intuition from point b), we run the K-means algorithm and visualise its result in a 2-dimensional scatter plot where the K-means cluster labels are used as colours in the scatters. While K-means with K = 2 does split the data into two groups, the clusters in this dataset follow a curved shape. However, K-means assumes "spherical" clusters, and because of this it doesn't capture the real shape of the clusters very well, especially in regions where the clusters are close together, resulting in points near the boundary being assigned to the wrong cluster.



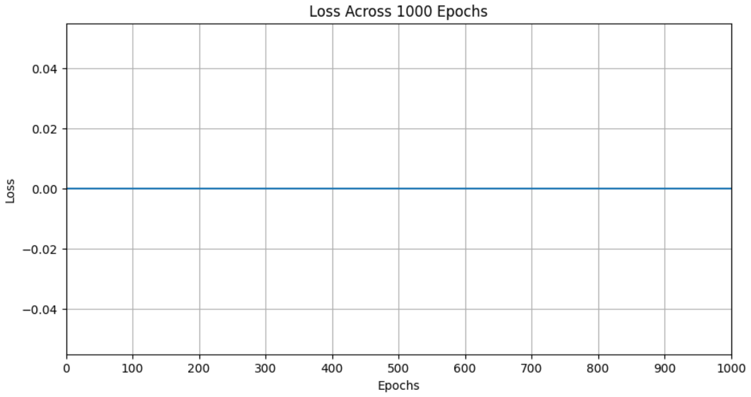
e) After defining and training the autoencoder, we plot the loss across the number of epochs and generate the latent representation for all data using the function "rep_learning" and visualize it in a 2-dimensional scatter plot.
As shown by the below plots, after the first training, the loss significantly decreases and the latent space representation separates the data into two distinct groups, while after a few trainings, the loss stagnates at zero and the latent space representation appears distorted, probably suggesting overfitting.

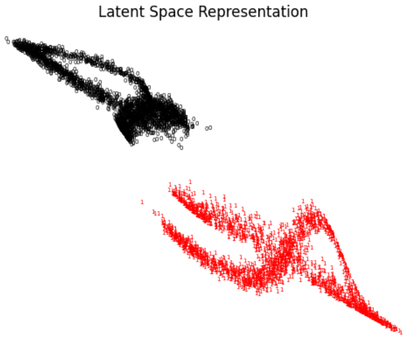
- results after first training:



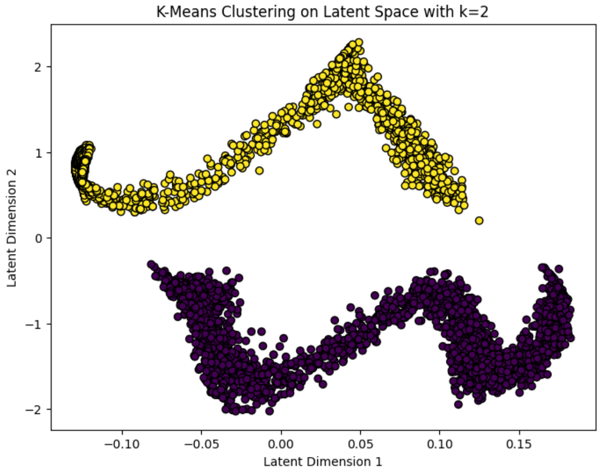


- results after few trainings:



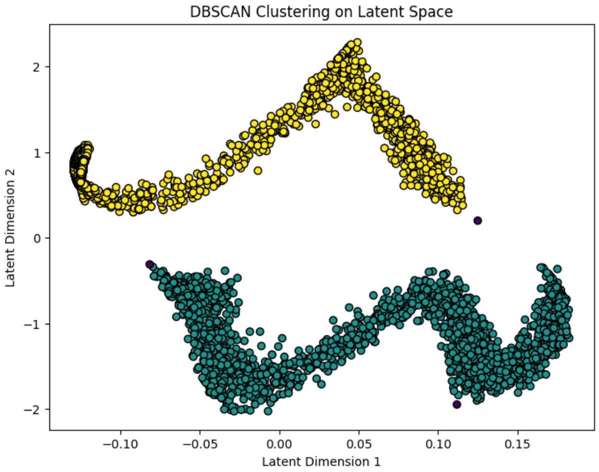


f) After clustering the obtained latent space in the first training with K-means, we obtain the plot shown below. If we compare it to the result obtained in point d), we can observe that clustering in the latent space produces a more accurate separation of the data into two groups. This is because the latent space, generated by the autoencoder, effectively captures the underlying structure of the data, allowing K-means to perform more effectively even with non-linear cluster shapes, while the original data in point d) presented significant overlap and misclassification due to the curved shape and proximity of the clusters.



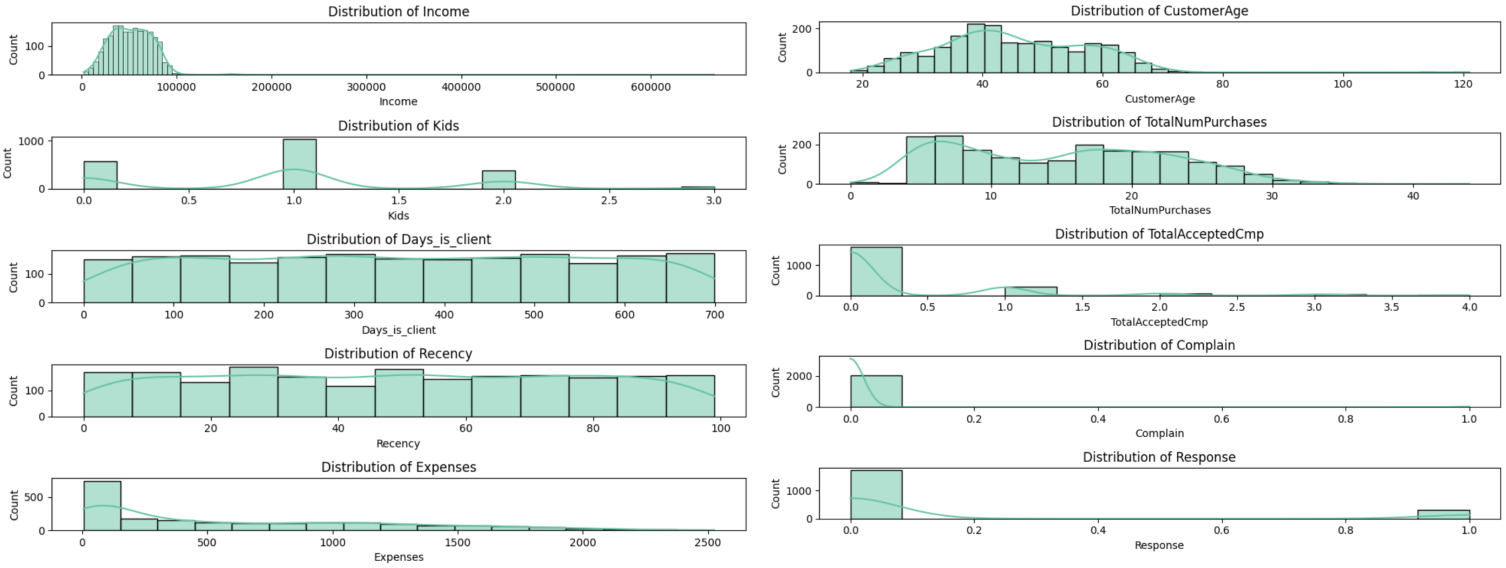
g) For this dataset, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) could be a more suitable clustering algorithm than K-Means. This is because DBSCAN is particularly effective in identifying clusters with non-linear shapes, such as the banana-shaped structures seen in this dataset, which, on the other hand, K-Means struggles with due to its assumption of spherical clusters. Additionally, unlike K-Means, DBSCAN does not require to know “a priori” the number of clusters, making it more flexible for datasets where the optimal number of clusters is unknown. Another advantage of DBSCAN is its ability to handle noise by identifying outliers as points that do not belong to any cluster, which is evident in the latent space clustering results, where a few points are correctly classified as noise (the violet points).

DBSCAN identifies clusters by iteratively examining each point’s local density. It starts with an unvisited point and checks how many neighbors fall within a specified radius (ε). If a point has enough neighbors (defined by min_samples), it becomes a core point and initiates a new cluster. The algorithm then expands this cluster by adding all points within the ε-radius of each core point, recursively including their neighbors until no more points can be added. Points that don’t meet the density requirement are labeled as noise. This process repeats with new unvisited points, allowing DBSCAN to detect arbitrarily shaped clusters, with ε and min_samples controlling the process. Finally, some tuning of the ε-radius s and min_samples parameters is often required to ensure that the algorithm correctly captures the cluster density (as carried out in the appendix).

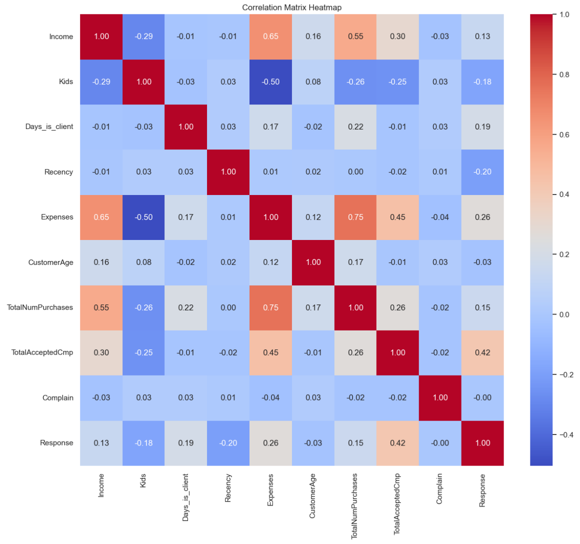


2) Task 2: Retail Store Business Case

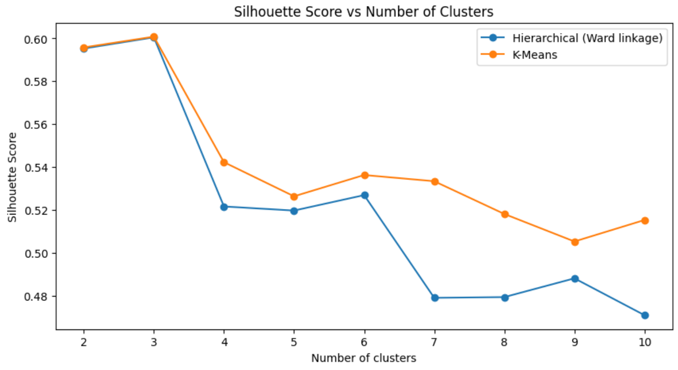
a) In our EDA, after checking for null values and dropping the duplicates, we proceeded with a univariate analysis of both categorical and numerical variables. The numerical variables reveal considerable diversity in the customer base: “Income” displays a wide range with a high standard deviation, indicating income inequality, while “Expenses” and “TotalNumPurchases” show significant variability (most customers seem to have low spending and purchase frequencies, while there’s a subset of “high-value” customers), “CustomerAge” and “Days_is_client” also span a broad range (reflecting the diversity in both age and loyalty), “Recency” shows that customer activity is unevenly distributed (some customers are highly engaged while others are inactive for a longer time), and finally “Complain” and “Response” have very low averages (indicating minimal customer complaints and low campaign engagement overall). Moreover, the categorical variables provide further insights: “Education” shows that most customers are graduates (customer base is relatively well-educated), and “Marital_Status” indicates that most customers are in a partnership. Together, these variables reveal a diverse customer base with different spending patterns, and engagement with the company.

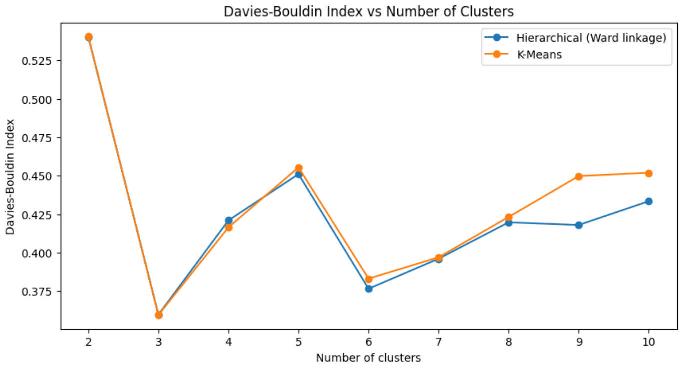


b) Instead of directly reducing the number of dimensions, PCA rotates the data so that the new axes (the principal components) align with the directions of maximum variance, preserving as much information as possible. These components are orthogonal and ordered by how much variance they capture, allowing for a more efficient representation of the data. Even though the number of components remains the same as the original variables, PCA eliminates redundancy and noise, making the data more suitable for further analysis without losing important variance. Moreover, the correlation matrix (before PCA) reveals that there is very little correlation, indicating that most variables capture different aspects of the data, so each dimension provides unique contributions. Finally, notice that although the variables were originally 12, after the one hot encoding they became 15 due to the to the categorical variables being split into multiple binary columns, and that's why we have 15 variables after PCA.



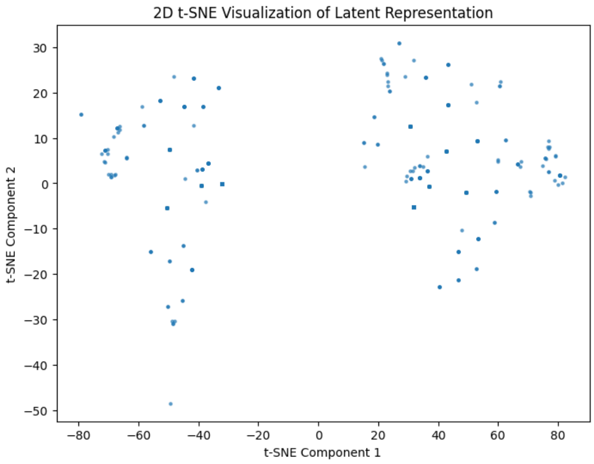
c) We decided to plot the Silhouette Score and the Davies-Bouldin Index in the same graph for both methods, to better display the outcomes.





d) The cross-validation suggests that 3 clusters are optimal, as indicated by both the Davies-Bouldin Index and the Silhouette Score. In the Davies-Bouldin Index (lower values are better), K-Means reaches its lowest value at 3 clusters, while the hierarchical (Ward linkage) method also performs well at 3 clusters. In the Silhouette Score (higher values are better), K-Means achieves a peak at 3 clusters, while the Ward linkage method also sees a similar pattern by reaching a similar score at 3 clusters. Hence, both metrics agree that 3 clusters represent the best fit for the data.

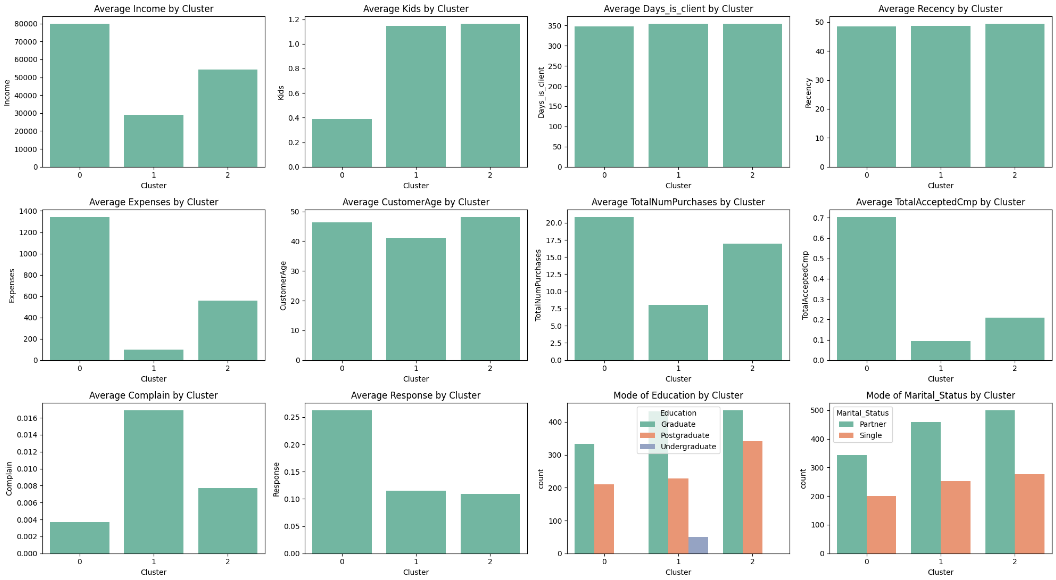
e) After generating the latent representation for all data using the function "rep_learning", we visualize it in a 2-dimensional scatter plot using the t-SNE dimensionality reduction method as displayed here.



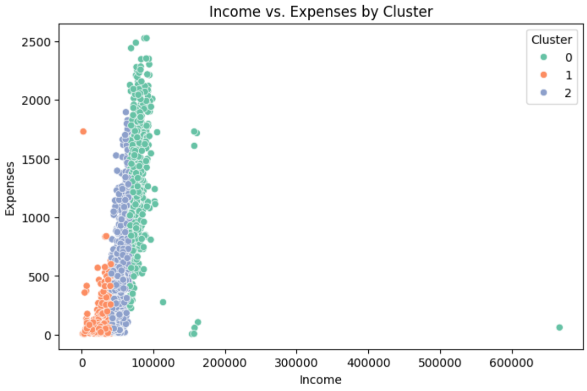
f) In the t-SNE scatter plot, we can see two distinct groups of data points, suggesting the presence of only two clusters. However, this is different from the result in exercise d), where the Davies-Bouldin Index and Silhouette Score suggested that three clusters are the best fit for the data. Notice that, when we computed the Davies-Bouldin Index and Silhouette Score, we used the complete dataset after applying PCA, which preserved more information about the data's structure, while the t-SNE plot reduces the data to two dimensions, so it might not show all the details of the data, making the third cluster harder to see.

g) To segment the customers according to their features, we applied the K-means algorithm with three clusters on the PCA transformation, and then continued our analysis with the original features of our dataset. This means that after clustering, we linked the assigned clusters back to the original dataset, preserving each observation's alignment with its original features. Importantly, PCA maintains the order of observations, so each cluster assignment corresponds directly to the initial data, allowing us to analyse clusters in terms of the original customer features.

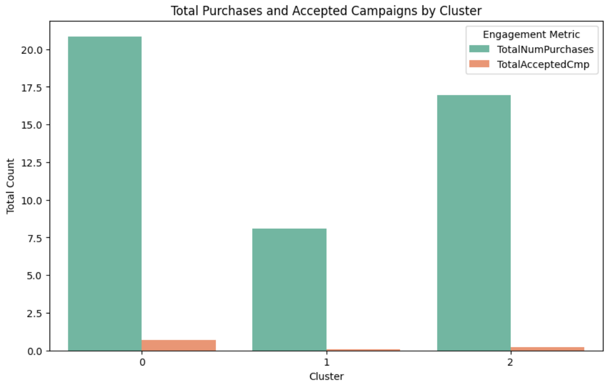
Our cluster analysis revealed three distinct customer profiles, each with specific spending habits, engagement levels, and preferences, as illustrated in the histograms below (reporting the mean value for each numerical features). The clusters are evenly distributed and capture substantial and diverse portions of the customer base, enabling the retail company to tailor marketing strategies effectively to meet their unique needs.



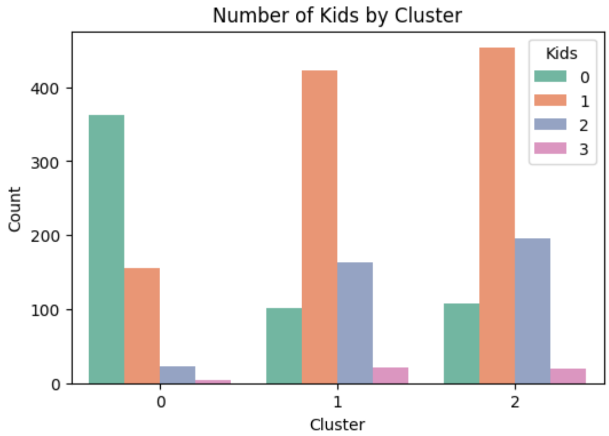
The first cluster (Cluster 0) represents a high-income, high-spending group. Customers in this cluster are more likely to have fewer children and exhibit a significantly higher level of total expenditures compared to others. They also tend to be more responsive to campaigns, accepting more offers than the others, indicating a lifestyle oriented toward personal or career pursuits. This cluster has a notable proportion of customers in higher education levels, and a relatively even mix of married and single individuals. This group stands out for its loyalty and willingness to spend, making it valuable for premium or exclusive offers.



The second cluster (Cluster 1) is, in contrast, characterized by lower-income, low-spending customers who are highly family-oriented, with a greater number of children. Members of this cluster have the lowest total spending and are generally the least engaged in terms of purchasing frequency and response to campaigns, suggesting a cautious approach to spending. They are predominantly graduates, with fewer customers in higher education levels, and most are married. Although they interact minimally with marketing offers, this group may be valuable for budget-friendly promotions and family-oriented products.



The third cluster (Cluster 2) represents a middle-income, family-oriented group that falls between high-spending and budget-conscious clusters. Primarily educated, married with children, these customers prioritize household stability. They demonstrate moderate spending habits and tend to be selective with promotions, engaging with campaigns only when offers align with their priorities of quality and affordability, making them reliable but cautious consumers. This group’s focus on practicality makes them ideal for family-oriented, value-driven products, with potential for steady loyalty when offered cost-effective solutions and occasional exclusive deals.



In conclusion, our analysis enables the retail company to identify the key features of each customer segment and target them effectively, reaching the wider audiences of Cluster 1 and Cluster 2 while offering premium options to high-value customers in Cluster 0. This approach ultimately boosts customer satisfaction, loyalty, and growth in each segment.