DOI: http://dx.doi.org/10.26483/ijarcs.v16i6.7368

Volume 16, No. 6, November-December 2025

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

UNSUPERVISED LEARNING APPROACHES FOR CUSTOMER SEGMENTATION: A PRACTICAL FRAMEWORK WITH VALIDATION METRICS

Aiysha Siddiqui
Assistant Professor, School of Computer Science, Engineering and Technology
ITM SLS Baroda University
Vadodara, Gujarat, India

Abstract: A dataset for retail marketing that includes demographics, purchase recency, and category-level expenditure is used in this study to investigate data-driven costumer segmentation. Building a repeatable preprocessing pipeline that encodes categorical, scales numeric features, and imputes missing values is the next step after we have completed the systematic cleaning and feature engineering. Using internal validation measures (Silhouette, Calinski–Harabasz, and Davies–Bouldin), we carry out a comparison of K-Means, Agglomerative Clustering, Gaussian Mixture Models, and DBSCAN across a whole spectrum of cluster counts. Visual diagnostics consist of distribution graphs, a heatmap illustrating the association between spending and income, and principal component analysis forecasts. A limited number of coherent and behaviourally unique segments are revealed by the study. These segments have significant distinctions in terms of income, spend composition, and household characteristics. These segments provide practical recommendations for targeting and personalization endeavours.

Keywords: Customer segmentation; clustering; K-Means; Agglomerative Clustering; Gaussian Mixture Model; DBSCAN; Silhouette Score; Calinski–Harabasz Index; Davies–Bouldin Index; marketing analytics; PCA visualization; feature engineering; profiling

I. INTRODUCTION

One of the most important aspects of contemporary retail marketing is customer segmentation. Instead of addressing all of its customers in the same manner, businesses partition their clientele into distinct groups based on the relevant behavioural or demographic qualities they share. Traditional segmentation is frequently based on rules and is founded on the intuition of business professionals. While these methods are transparent, they have the potential to overlook hidden structures that are the result of multivariate interactions including income, product mix, and engagement history amongst variables. A supplementary approach is provided by unsupervised learning, which enables the data to show natural groups that are compact within clusters and distinct between clusters.

This paper places a strong emphasis on both the methodological rigour and the practical utility of the findings. The segmentation process is meant to be resistant to problems with the quality of the data, repeatable over several iterations, and aligned with interpretability that is independent of the model to be used. To prevent overfitting or spurious separation, visual diagnostics and internal metrics are used as a sanity check. The result is an intellectual exercise in clustering; and it is a collection of segments that can be identified, explained, and operationalized in marketing efforts.

II. LITERATURE REVIEW

A basic study of cluster analysis in marketing is provided by Ufeli et al. (2025). They place an emphasis on a disciplined workflow that begins with the formulation of the issue and the selection of variables and continues through preprocessing (standardization, transformations), the selection of similarity measures and algorithms, and, most importantly, validation on the entire process. They provide a list of common pitfalls, such as arbitrary variable sets, unscaled features, unexamined outliers, and an excessive reliance on a single algorithm or k, while also advocating for triangulation across methods (such as K-Means and hierarchical), stability checks, and external/managerial validation to guarantee that segments are interpretable and actionable. The actual segmentation process is directly influenced by their direction, which includes the following: engineering significant features, scaling mixed units, testing different algorithms and k values, evaluating stability and internal validity, and profiling segments in relation to business objectives.

Clustering approaches are surveyed by Hu et al. (2024) through the use of a unified taxonomy. These methods include partitioning, hierarchical, density-based, grid-based, and model-based clustering. Considerations of distance metrics, scalability, and high-dimensional effects are included. Other than highlighting parameter sensitivity (for example, eps/min_samples in DBSCAN), the need for careful preprocessing, and the role of dimensionality reduction for diagnostics rather than fitting, they provide a rationale for comparing multiple families, sweeping k, tuning density parameters, and using PCA plots as qualitative checks. They also detail algorithmic behaviors and trade-offs, such as the tendency of K-Means to favor spherical clusters, the hierarchical granularity of agglomerative, the ability of DBSCAN to find arbitrary shapes and noise, and the flexibility of the model-based flexibility.

Mixture modelling is a probabilistic framework for clustering that was established by Murthy et al. (2025). It is centred on the EM method, which is used for maximum likelihood estimation and model selection using information criteria (AIC/BIC). They explore covariance parameterizations (spherical, diagonal, and complete) for Gaussian mixtures, as well as identifiability, initialisation, and local maxima, as well as concerns such as singularities and label switching. They motivate soft assignments (posterior probabilities) for uncertainty-aware profiling. Their study sheds light on the connections between K-Means, which is a limited example, the benefits of elliptical clusters, and expansions to robust mixes, such as t-mixtures. For the purpose of customer segmentation, the book recommends use GMM in conjunction with K-Means, taking into account covariance restrictions for stability, using BIC as an alternative k-selection criterion, and using posterior probabilities to quantify segment membership and direct downstream targeting.

Within the context of e-commerce, Tabianan, Velu, and Ravi (2022) investigate customer segmentation with a particular focus on behavioural signals and use K-Means to separate clients according to their buying behaviour. Their goal is to maximise long-term value by identifying customer categories that are lucrative and tailoring their offerings to meet those segments. Behavioural data is the sole data that is operationalised in this work, which focusses on within-cluster similarity and between-cluster dissimilarity. The article includes demographic, psychographic, behavioural, and geographic segmentation criteria. They investigate the connections between different sorts of events, groups of items, and categories, and they argue that clusters enable vendors to prioritise high-profit cohorts and maximise exposure and promotions. In terms of methodology, the study is in line with the standard practice of feature engineering of behavioural interactions and K-Means due to its simplicity and scalability. However, it inherits the assumptions and sensitivities of K-Means, which are roughly spherical clusters in standardised space, as well as scaling, initialisation, and k preference. The contribution is a blueprint that is application-oriented and emphasises practical advantages such as targeting and retention. However, it also leaves potential for deeper validation, such as stability checks, alternative algorithms, and temporal robustness, as well as richer feature sets, such as explicit rules of engagement and channel engagement.

Tripathi et al. (2025) uses a system that blends simulated and surrogate data to analyse the behaviour of the approach under controlled settings to assess PCA-based clustering for market segmentation. The study emphasises three pillars: rigorous dimensionality reduction (selecting an adequate number of main components), principled clustering (they choose K-Means as the baseline), and extensive validation using both internal and external measurements. Using principal component analysis (PCA) in a systematic manner can increase clustering stability and interpretability, as demonstrated by their findings; nevertheless, improper PC selection might mask significant structure or magnify noise. The framework enable repeatable segmentation operations that are resilient to high dimensionality, and it also makes model diagnostics clear, which includes stability, validation curves, and explainability of segments in PC space. This framework's utility is twofold. To prevent artefacts, the research suggests that component selection should be coupled with validation and sensitivity studies. The study warns against using principal component analysis as a general pre-step.

Implications for clinical treatment and a synthesis: A systematic pipeline for segmentation is shown to be supported by these articles collectively. According to Tabianan et al., using behavior-centric characteristics and K-Means as a solid baseline is recommended; nonetheless, it is possible to confirm segment actionability and compare it to alternatives. Use internal indices (Silhouette, Calinski–Harabasz, Davies–Bouldin), stability tests, and exterior criteria wherever possible. According to Tripathi et al., dimensionality reduction should be seen as a design option that should be proven rather than assumed. This indicates that, in practical situations: to assure interpretability and marketing relevance, standardise and engineer behaviour characteristics; compare K-Means, hierarchical, and model-based approaches; if principal component analysis is used prior to clustering, tweak the number of principal components using validation; and profile segments based on original features (Bombina et al., 2024).

III. EXPERIMENT AND RESULTS

Dataset

Kaggle's marketing_campaign.csv (tab-separated) dataset is a retail marketing collection that incorporates demographic characteristics with purchase and interaction statistics. The dataset was created by Kaggle. The year of birth, which is used to calculate age, the level of education, the marital status, the makeup of the family (children and teenagers living at home), and income are all examples of demographic fields. Behavioural fields include category-level spending amounts, which are typically prefixed with Mnt (for example, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProds, and MntGoldProds), as well as recency and customer enrolment date (Dt_Customer), which together summarise how recently a customer has been active and how long they have been active. The segmentation is able to capture both the ability to spend and the choice mix since it takes into account both absolute amounts and proportions, other than the demographic background.

A first examination of the data reveals the typical peculiarities that are associated with real-world marketing data. These peculiarities include inconsistent date formats across different locations, occasional missing revenue entries, and long-tailed expenditure distributions with a small number of really high-performing consumers. The dataset is sufficiently extensive to allow for value-driven as well as preference-driven segmentation capabilities, while at the same time being sufficiently compact to allow for interactive exploration.

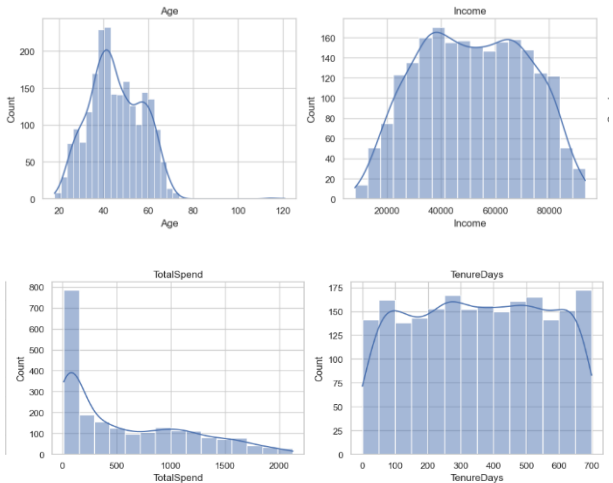

Figure – 1 : EDA [Own development]

Data Preprocessing

When it comes to data preparation, the method favours conservative modifications that improve robustness without imposing significant modelling assumptions. When dates are processed, protections are included to prevent mixed formats. An alternative pass that does not use day-first is attempted when day-first parsing results in a significant number of missing values. The option that has a smaller number of missing entries is chosen to be preserved. It is possible to determine the length of the customer's connection by computing the tenure in days based on the Dt_Customer. To prevent problems caused by unexpected strings further down the line, Year_Birth is converted into a numeric value and then used to calculate Age.

The category-level expenditure variables, which are all columns that begin with Mnt, are aggregated into TotalSpend to obtain an accurate representation of purchasing power and product mix. To differentiate between preferences and purchasing power, the proportion of each category is calculated by dividing the total expenditure by the total amount spent, and the undefined situations are filled with zero. A Kids feature that includes Kidhome and Teenhome, as well as a straightforward IsPartnered flag that is derived from marital status ("Married" or "Together" mapped to 1), are the components that best summarise the structure of households. For the purpose of avoiding the creation of artificial separation, numerical IDs and dataset constants like ID, Z_CostContact, and Z_Revenue are not included in the modelling process.

Outliers have the potential to disproportionately impact the outcomes of clustering because it is dependent on distance geometry. It is necessary to trim Income and TotalSpend at the first and 99th percentiles to minimize leverage without distorting the majority distribution. This is done while retaining missing values in their original state so that they may be imputed in a systematic manner. After that, a ColumnTransformer is used to generate the modelling matrix. The numerical variables are imputed with the median and scaled using StandardScaler. On the other hand, the categorical variables are imputed with the category that occurs the most frequently and one-hot encoded with unknown handling enabled. The transformation is made repeatable across environments thanks to this pipeline, which also works to prevent information from leaking out from future data (Jahanian, et al, 2025).

The association between the various characteristics is depicted in the heatmap that can be seen below. There is a significant relationship between income and the overall amount of money spent, as well as the quantity of items of each respective category. Given that a bigger income enables a greater number of things to be purchased, this makes perfect sense.

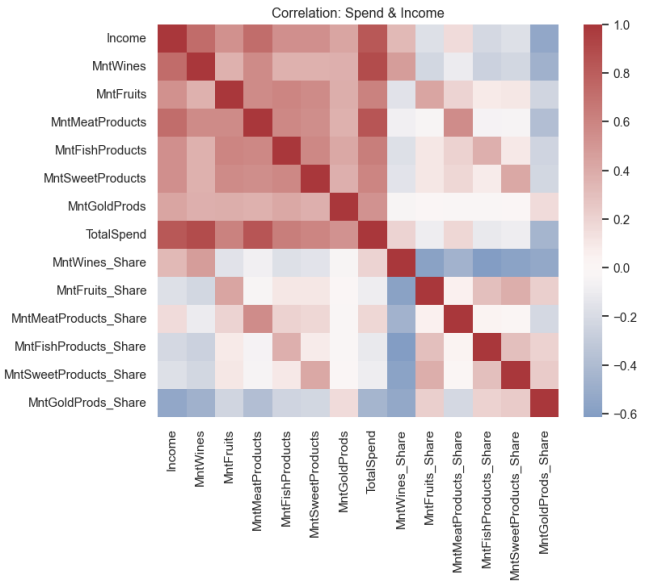

Figure – 2 : Correlation Matrix [Own development]

Model Training

In the modelling section, complementary clustering families are compared to reduce the possibility that the findings reflect the bias of an algorithm rather than the signal obtained from the data. K-Means is a robust baseline that gives preference to spherical clusters in the converted feature space. It is also quick, stable with many initialisations, and simple to comprehend through the use of centroids. Agglomerative Clustering is a method that constructs clusters in a hierarchical fashion. They are able to reveal structure at different levels of granularity and are less susceptible to the initialization of the centroid. Gaussian Mixture Models are a generalization of K-Means that allow clusters to have elliptical covariance. These models also provide soft assignments (probabilities) that represent uncertainty. Last but not least, DBSCAN is an option that is based on density and has the ability to identify non-convex shapes and label outliers as noise. This technique eliminates the requirement of pre-specifying the number of clusters (Amorim et al., 2025).

Throughout the training process, every parametric technique is trained on $k \in \{2, 3, 4, 5, 6, 7\}$. K-Means is able to lessen the sensitivity of local minima by performing repeated restarts (n_init=10). The agglomerative For the purpose of forming compact groups, clustering makes use of linkage on the preprocessed space. Random seeds are used to initialise GMMs for the sake of repeatability. This allows for complete covariance within clusters, provided that they are supported. DBSCAN is executed with a significant epsilon (eps=0.8) and min_samples=10 as an untuned baseline. This is done in recognition of the fact that high-dimensional standardized data typically need thorough calibration of density thresholds. Visualization is the only purpose of principal component analysis (PCA), while clustering is carried out on the entire preprocessed feature matrix.

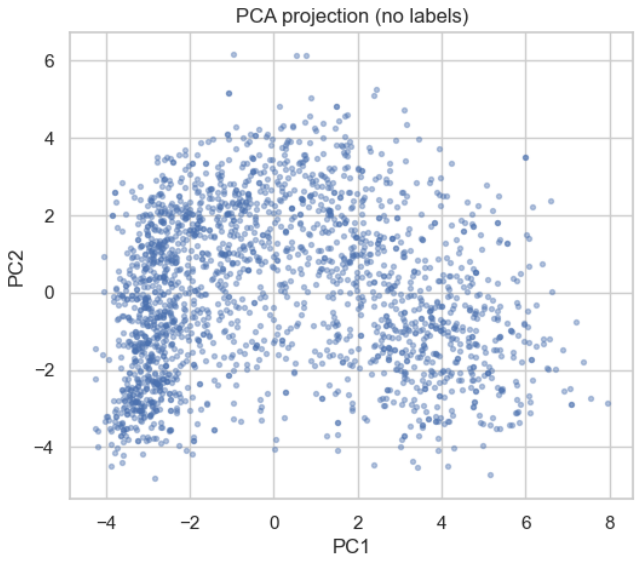

Figure – 3 : PCA Projection [Own development]

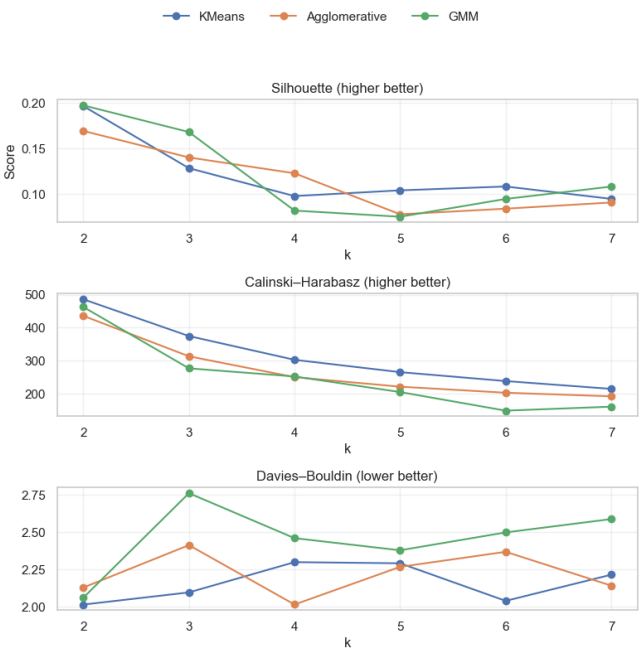

Figure – 4 : Clustering Methods [Own development]

IV. Comparison and Results

Internal validation measures are used to proceed with model comparison. These metrics strike a compromise between within-cluster compactness and between-cluster separation. A normalised measure of how well each point fits inside its allocated cluster in comparison to the next best possibility is provided by the silhouette coefficient. Higher values indicate a more distinct separation between the points involved. The Calinski–Harabasz algorithm prefers clusters that are compact and well-separated in terms of their dispersion ratings. By taking into account the average similarity between each cluster and its most comparable peer, Davies–Bouldin provides a penalty for overlapping clusters; lower values are preferable. Consistency across various measurements increases confidence that the identified groups are not artefacts of a single criterion, despite the fact that these metrics do not constitute certain "truths."

Over the whole range of k that was evaluated, K-Means and Gaussian Mixtures often provide the greatest silhouette scores at small to moderate k (generally between 2 and 3). This indicates that the natural structure of the dataset favours a limited number of wide segments rather than a large number of fine-grained micro-clusters. Agglomerative Clustering is competitive for values of k that are comparable, and the opportunity to study the dendrogram (conceptually) adds interpretative value; but, for the purpose of our research, we depend on the cluster labels rather than the whole tree. DBSCAN has a tendency to categorise a significant portion of observations as noise in the standardised, somewhat high-dimensional space when the default parameters are used. This is a behaviour that is expected and does not necessarily indicate bad performance; the only thing that has to be adjusted is the density parameters to correspond with the data magnitude.

#	Silhoutte Score	Calinski-Harabasz	Davies Bouldin	k	Model
1	0.197	462.65	2.06	2	GMM
2	0.196	486.90	2.01	2	KMEANS
3	0.169	435.77	2.12	2	AGGLOMERATIVE
4	0.167	276.91	2.76	3	GMM
5	0.140	313.39	2.41	3	AGGLOMERATIVE
6	0.128	374.61	2.09	3	KMEANS
7	0.122	250.33	2.01	4	AGGLOMERATIVE
8	0.108	161.02	2.58	7	GMM
9	0.104	265.44	2.29	5	KMEANS
10	0.097	302.81	2.29	4	KMEANS
11	0.094	214.82	2.21	7	KMEANS
12	0.094	148.94	2.49	6	GMM
13	0.090	192.32	2.14	7	AGGLOMERATIVE
14	0.084	203.17	2.36	6	AGGLOMERATIVE
15	0.082	252.41	2.45	4	GMM
16	0.078	221.48	2.26	5	AGGLOMERATIVE
17	0.075	205.18	2.37	5	GMM

To achieve success in operations, interpretability is absolutely necessary. When cluster assignments are projected onto the principal component analysis space, it is possible to see clearly distinguishable regions for the configurations that perform the best. These regions have little overlap and gradients that correspond to the directions with the greatest variation. More crucially, segment profiling on the original variables that are accessible by humans reveals significant behavioural differences. There is a recurrent sector that combines above-average income and total spending with greater proportions dedicated to wine and gold items, which is consistent with premium-oriented purchasing. This group frequently has a distinct family composition, which may include a greater number of children or teenagers living at home, which connects with various product mixtures such as sweets and fruits. Another section demonstrates lower income and total spending, with significantly higher proportions in daily categories. consumers who have been with the company for a longer period of time and are older tend to congregate together and exhibit more consistent spending habits. On the other hand, consumers who have had the company for a shorter period of time or who are younger

tend to form groups that have more varied activity and lower total expenditure.

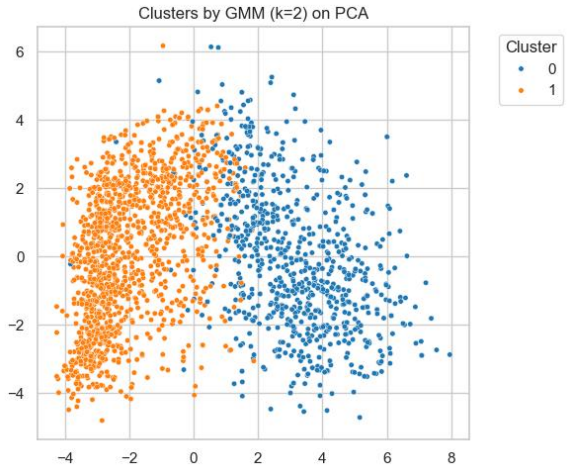

Figure – 5 : GMM Clusters [Own development]

Category	Cluster 0	Cluster 1
Age	47.07	44.17
Income	70717	39774
TotalSpend	1221.39	200.23
TenureDays	377.96	339.73
Kids	0.52	1.23
IsPartnered	0.62	0.66
MntWines	604.14	110.25
MntFruits	55.62	7.51
MntMeatProducts	351.64	39.97
MntFishProducts	78.23	11.07
MntSweetProducts	57.21	7.29
MntGoldProds	74.55	24.14

These profiles may be translated into basic activities from a marketing point of view. The most effective way to attract premium-oriented segments is to provide them with unique discounts, early access, and curated packages that place an emphasis on high-end categories. Multi-pack discounts, seasonal deals, and cross-category bundles that coincide with the purchase habits of family-oriented value shoppers are examples of promotions that are customised to household needs and provide benefits to these shoppers. Smaller but more regular nudges, reminders connected to recency, and cross-sells into adjacent categories that extend the customer's basket without needing huge increases in expenditure are all examples of types of re-engagement methods that are suitable candidates for segments who have lower activity or budget constraints.

It is important to keep in mind that internal measures can only approximate the validity of external metrics. It is encouraging that top-performing k is consistent across both K-Means and GMM; however, future validation could include holdout-based stability checks (for example, adjusted Rand index across resamples), temporal robustness (for example, retraining on earlier versus later periods), or business-grounded evaluation such as an increase in campaign response rates when segments are targeted with tailored messages.

V. Conclusion

The purpose of this study is to demonstrate a pipeline for customer segmentation that is both practical and reproducible, and that can be used in real-world analytics environments. A method that places an emphasis on solid data preparation, thorough feature engineering that differentiates between preference and capacity to spend, and a systematic evaluation of clustering models is being used. The findings suggest that a limited number of segments, which are consistently discovered by K-Means and Gaussian Mixtures across all internal measures, are capable of capturing important behavioural characteristics without compromising interpretability when compared to other segments. It has been demonstrated using visual diagnostics and profiling that there are differences in segments along income, total expenditure, product mix, and home context that are predicted and should be acted upon.

It is important to take note of a number of restrictions and extensions. Exploring a grid for DBSCAN (eps, min_samples) or adopting OPTICS might make it possible to find non-convex structures that centroid-based algorithms have overlooked. Density-based clustering can be strong when it is customised to the scale of the data. The use of explicit RFM measurements (Recency, Frequency, and Monetary) and channel interaction elements (online versus shop versus catalogue) has the potential to further differentiate between high-value consumers and clients that are at risk. In conclusion, the most effective method for determining the value of segmentation for a firm is to conduct controlled experiments. These studies involve assigning customized offers to each group and assessing the additional lift in conversion, average order value, or retention over an acceptable time horizon.

As a whole, this paper reveals that the application of data science techniques, in conjunction with the implementation of disciplined preprocessing and transparent assessment, results in the generation of segments that conform to marketing intuition and also make significant statistical sense. These segments serve as a basis for coordinated targeting, personalization, and resource allocation, therefore assisting organizations in transitioning from providing outreach that is universally applicable to providing customer engagement that is evidence-based.

VI. References

[1] Ufeli, C. P., Sattar, M. U., Hasan, R., & Mahmood, S. (2025). Enhancing customer segmentation through factor analysis of mixed data (FAMD)-based approach using K-means and hierarchical clustering algorithms. Information, 16(6), Article 441

[2] Hu, L., Jiang, M., Dong, J., Liu, X., & He, Z. (2024). Interpretable clustering: A survey. arXiv. https://doi.org/10.48550/arXiv.2409.00743

[3] Murthy, S. A., Rao, K. N., & Rao, K. S. (2025). Market segmentation through finite mixture regression models with generalized normal distribution and hierarchical clustering. Journal of Information Systems Engineering & Management, 10(35s), 116–125

[4] Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. Sustainability, 14(12), 7243. https://doi.org/10.3390/

[5] Shailesh Tripathi, Nadine Bachmann, Manuel Brunner, Alican Tuezuen, Ann-Kristin Thienemann, Sebastian Pöchtrager, Herbert Jodlbauer, Evaluation of Clustering

with PCA for Market Segmentation: A Study Using Simulated and Surrogate Data, Procedia Computer Science, Volume 253, 2025, Pages 2063-2075, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2025.01.267.

[6] Bombina, P., Tally, D., Abrams, Z. B., & Coombes, K. R. (2024). SillyPutty: Improved clustering by optimizing the silhouette width. PLoS ONE, 19(6), e0300358. https://doi.org/10.1371/journal.pone.0300358

[7] de Amorim, R. C., & Makarenkov, V. (2025). Improving clustering quality evaluation in noisy Gaussian mixtures. arXiv preprint arXiv:2503.00379. https://doi.org/10.48550/arXiv.2503.00379

[8] Jahanian, M., Karimi, A., Eraghi, N. O., & Zarafshan, F. (2025). Adaptive clustering for medical image analysis using the Improved Separation Index (ISI). Scientific Reports, 15, Article 28191

https://www.kaggle.com/datasets/vishakhdapat/customer-segmentation-clustering

