Comparison of K-Means and HDBSCAN Clustering Approaches to Enhance Marketing Strategies

RAPHAEL IBRAIMOH[1], ADETUNJI ADEROBA[2]
[1, 2] *School of Science, Engineering and Environment, University of Salford*

Abstract- ***Maintaining and growing market share is non-negotiable for businesses regardless of economic swings and the instability of many sectors. Many companies set aside large funds for marketing and advertising their goods and services meant to support their corporate objectives. But a Proxima (2023) analysis indicates that 60% of this spending may be better used, primarily because of poor targeting of the appropriate audience depending on their capacity and purchase patterns. Personalising marketing and sales communication and targeting will help one to maximise client satisfaction and optimise return on investment. This work investigates the application of machine learning models to examine a real-world dataset of 3,900 distinct consumers who regularly buy accessories, outerwear, shoes, and clothes. Customer clusters were segmented and understood using the Recency, Frequency, and Monetary (RFM) Model, K-Means and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) methods. High, medium, and low-paying clients were found by means of RFM scores. The results resulted in the creation of focused marketing plans emphasising on the 7Ps (Product, Place, Price, Promotion, People, Process, and Physical Evidence), therefore creating business prospects for favourable changes in several client categories. A reusable Python tool was also developed to examine big databases going forward.***

Indexed Terms- ***Customer Segmentation, HDBSCAN Clustering, KMeans Clustering, and The Recency Frequency Monetary Model.***

I. INTRODUCTION

A thorough awareness of customer behaviour is vital for creating and applying effective marketing plans in the present corporate climate. Customer segmentation, where a client base is split into several categories based on shared traits, has become essential for raising consumer loyalty and satisfaction (Wedel & Kamakura, 2000). The advent of big data and sophisticated analytics has significantly enhanced the importance of consumer segmentation by providing more precise and actionable insights (Daelemans & Goethals, 2008).

The Zendesk Customer Experience Trends Report (2023) claims that 60% of consumers make purchases based on the service they expect to meet their demands and wants (Lahey, 2023). Ignoring potential consumer dissatisfaction can significantly impact companies, leading to lost revenue and reduced profitability. According to an Adobe survey, 54% of consumers have severed ties with a company due to a breach of trust, often caused by excessive commercial messaging (Pozo, 2021). The American Customer Satisfaction Index (ACSI) indicates that businesses with low satisfaction ratings, such as Comcast's Xfinity Mobile, may experience declining customer loyalty and increased turnover rates.

Several examples illustrate the consequences of poor marketing strategies:

- Coca-Cola's 1980s "New Coke" failed due to an underappreciated emotional link to the original recipe, leading to a rebranding as "Coca-Cola Classic" (Thompson, 2024).
- Gap Inc. faced widespread criticism and wasted marketing resources when it changed its recognisable logo without proper research or testing in 2010 (Williams, 2021).
- J.C. Penney's 2011 "fair and square" pricing strategy, which eliminated sales and discounts, failed to appeal to consumers, resulting in a significant drop in sales (Dennis, 2024).
- Pepsi's 2017 campaign with Kendall Jenner was criticised for trivialising social justice movements and was quickly withdrawn (Solon, 2020).

- Google Glass faced privacy concerns, usability issues, and a lack of compelling use cases, leading to its failure to gain widespread acceptance (Weidner, 2024).

These examples underscore the importance of understanding consumer behaviours and tailoring marketing plans accordingly. In a highly competitive and dynamic business environment, developing effective marketing strategies depends on a deep understanding of consumer behaviour. K-Means and HDBSCAN are two clustering methods that offer powerful tools for consumer segmentation based on buying behaviour. This paper evaluates these clustering techniques using the Kaggle Consumer Behaviour and Shopping Habits Dataset to improve marketing strategies. The dataset, comprising 3,900 entries, provides insights into consumer purchase patterns and preferences across various demographic, behavioural, and transactional variables (Aslam, 2023).

Addressing the challenges companies face in understanding and meeting the diverse needs of their clients drives this research. Many segmentation techniques fall short in fully capturing the complex and varied nature of consumer behaviour. By employing K-Means and HDBSCAN clustering, companies can better understand their customer base and tailor their marketing strategies to target specific customer segments.

A. Problem Statement

Businesses need help properly classifying their consumers and using these groups to develop exact marketing strategies to overcome the challenges of the availability of consumer data.

1. How can K-Means and HDBSCAN clustering methods practically segment customers in the given dataset in a way that closely mirrors patterns of genuine consumer behaviour?
2. From the marketing perspective, what are the advantages and disadvantages of any clustering technique?
3. How may the consumers' behaviour model and market indicators be combined with the knowledge from these clustering methods to improve marketing plans?

B. Goals and Objective of Study

To enhance marketing strategies, the goal is to compare K-Means and HDBSCAN clustering approaches to consumer shopping behaviour and purchase patterns datasets.

Objectives:

- Apply K-Means and HDBSCAN clustering methods on the given dataset.
- To assess every clustering technique's performance in line with suitable metrics.
- To combine improved consumer segmentation, the clustering algorithms with the consumer's behaviour.
- Using the knowledge acquired from the clustering analysis, build marketing plans based on key market indicators like Price, Product, Promotion, Place, etc.

C. Significance of Study

By focusing on consumer groups, this study seeks to enhance marketing strategies to optimise the resource economy and raise conversion rates. It also emphasises improving consumer experience by customising goods to fit personal tastes, raising customer loyalty and pleasure. Knowing customer segmentation can also help businesses maximise their product lines by enabling them to decide which popular items, areas, or tastes to focus on so lowering waste and enhancing inventory control.

The findings of the study can support strategic decisions at different corporate levels by guaranteeing that operations fit consumer preferences and needs. By proving the use of K-Means and HDBSCAN clustering in practical situations and the possibility of sophisticated analytical techniques to expose important data from challenging datasets, it also promotes data science.

In the data-driven environment of today, ethical issues are also quite important since data security and compliance are necessary to keep customer integrity and confidence. The study intends to give thorough knowledge of client behaviour by means of data-driven segmentation, so helping businesses and consumers.

II. METHODOLOGY

The approach analyses consumer behaviour data using structured clustering techniques, helping companies improve customer satisfaction and modify their marketing plans. This approach guarantees the results' validity and correctness, guiding strategic marketing decisions.

A. Dataset Structure and Description

The shopping behaviour dataset is a tabular data format published in CSV format, allowing Jupiter Notebook accessibility. It covers numerous rows and columns, each indicating a single transaction or customer entry and a specific attribute or variable linked to the transaction or customer. The dataset includes attributes such as customer ID, age, gender, item purchased, category, purchase amount (USD), location, size, color, season, review rating, subscription status, shipping type, discount applied, promo code used, previous purchases, payment method, and frequency of purchases.

B. Data Preprocessing

To facilitate data manipulation, visualisation, preprocessing, and clustering, Python libraries such as Hdbscan, Pandas, Matplotlib.pyplot, Seaborn, Numpy, os, sklearn.preprocessing, sklearn.metrics, and sklearn.cluster are employed. Data cleansing was implemented to guarantee the dataset's accuracy, consistency, and value. This process involved the identification and management of missing values, the correction of inconsistencies, the treatment of outliers, and the mapping of specified values to standardised terms. The `OrdinalEncoder` was imported and columns were defined to encode in order to convert categorical variables into numerical variables using the sklearn.preprocessing module. The RFM model and the 7Ps marketing mix indicators were used to guide feature engineering, which resulted in the creation of additional columns and the assignment of rankings. The `StandardScaler` from sklearn.preprocessing was employed to standardise and encode numerical features. The final DataFrame then displayed the combined selected features, confirming that the dataset is prepared for machine learning models.

C. Feature Selection

Adopting the concept of the relationship between Consumer Buying Behaviour (RFM Model) and Market Mix (7Ps) based on the mathematical function below (Azzam & Ali, 2019):

$Y=f(X)$

$Y=X_1+X_2+X_3\ldots+X_n$

RFM = Product + Price + Place + Promotion + People + Physical Evidence + Process

Features Selected	Redefined Feature
Category	Product
Purchase Amount (USD)	Price
Shipping Type	Place
Discount Applied	Promotion
Age	People
Review Rating	Physical Evidence
Payment Method	Process
Frequency of Purchases	Recency
Previous Purchases	Frequency
Purchase Amount (USD) * Previous Purchases	Monetary

Table 1: Feature Selection

D. Exploratory Data Analysis

Understanding data structure, spotting trends, and exposing relationships between variables all depend on EDA. Methods comprise descriptive statistics, data visualisation, missing value analysis, and categorical data analysis. Descriptive statistics find the mean, median, mode, standard deviation, and range of numerical variables; data visualisation shows numerical variable distribution, data dispersion, and relationships using histograms, box plots, scatter plots, and heatmaps. Missing value analysis handles missing values either by deletion or imputation. Bar charts and cross-tabulation for variables like gender and category help categorical data analysis visualise frequency distributions.

E. KMean Clustering Algorithm

K-Means clustering is an unsupervised machine learning method that divides a dataset into K non-overlapping groups (Bock, 2007). Each data point falls into the cluster with the closest mean. The process involves choosing K random centroids, assigning each

data point to the closest centroid based on Euclidean distance,

$$d(x_i, c_j) = \sqrt{\sum_{m=1}^{n} (x_{im} - c_{jm})^2}$$

Where (x_i) is a data point, (c_j) is a centroid, and (n) is the number of features.

and recalculating the centroids as the mean of all data points allocated to each cluster.

$$c_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

Where (C_j) is the set of cluster (j) points.
The process was repeated until the centroids no longer showed significant change or a maximum number of iterations was reached. The Elbow Method computes the sum of squared distances between data points and their designated centroids using the Within-Cluster Sum of Squares (WCSS) approach.

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

where (C_i) is the (i)-th cluster, (x) is a data point, and (\mu_i) is the centroid of (C_i).
The optimal number of clusters is found at the "elbow point," where the decline in WCSS slows. This method produces meaningful consumer segments from a trade-off between cluster compactness and computing efficiency, enabling customised marketing plans and individual recommendations.

F. HDBSCAN Clustering Algorithm

Extending DBSCAN, HDBSCAN is a hierarchical clustering method allowing the identification of clusters with different densities and forms. It does not call for a set number of clusters and is more flexible than K-Means. Built on data point density, HDBSCAN identifies flat clusters using a stability metric to create a hierarchy of clusters. Particularly in situations when a small minimum cluster size is required but must prevent too many micro-clusters in highly populated areas, this hybrid technique can show the advantages of varied densities of data clustering. One can use the technique for HDBSCAN cluster candidates without requiring hierarchy adjustment. Defined as the distance between two places, key mathematical components of HDBSCAN consist in the core distance and the mutual reachability distance (Malzer & Baum 2020).

$$\text{mreach}(p, q) = \max(\text{core\_dist}(p), \text{core\_dist}(q), \text{dist}(p, q))$$

where (core_dist(p)) is the core distance of (p) and (dist(p, q)) is the Euclidean distance between (p) and (q).

G. Silhouette Score

The quality of the two machine learning cluster methods is evaluated using the Silhouette Score. It gauged an object's cohesiveness; that is, how similar it was to its own cluster, against other clusters' separability. The score indicates a well-matched data object and a great distance from surrounding clusters since it ranges between -1 and +1. It also shows whether the sample crosses the decision line between two consecutive clusters (Strobl et al., 2020). The calculation is

$$s = \frac{b - a}{\max(a, b)}$$

Where:
(a) is the mean distance between the sample and all other points in the same cluster (intra-cluster distance).
(b) is the mean distance between the sample and all points in the nearest cluster that the sample is not a part of (nearest-cluster distance).

H. Recency Frequency Monetary (RFM) Score

The customers were segmented based on their purchasing behaviour using the RFM model and analysed the average RFM scores for each cluster (Theng & Bhoyar, 2023).

RFM Preprocessing

This involved the preprocessing of customer transaction data to calculate recency, frequency, and monetary values for each customer (“RFM Analysis Analysis Using Python,” 2024).

Recency (R): indicates the number of days since the customer’s last shopped.
Calculation:
Recency = Current Date−Last Purchase Date
From the dataset, 'R' is 'Frequency of Purchases'

Frequency (F): indicates the total number of purchases made by the customer within a specific period. Calculation:
Frequency = Total Number of Purchases
From the dataset 'F' is 'Previous Purchases'

Monetary (M): indicates total amount of money spent by the customer within a specific period. Calculation:
Monetary = Sum of All Purchase Amounts
From the dataset 'M' is the sum of 'Purchase Amount (USD)' * 'Previous Purchases'

Ranking and Scoring
After values for Recency, Frequency, and Monetary were derived, customer were ranked based on these values ("RFM Analysis Analysis Using Python," 2024).

Recency Rank: Customers with the most recent purchases get the highest scores. Using this python code on the value .rank(ascending=False)

Frequency Rank: Customers with the highest number of purchases get the highest scores. Using this pyhton code on the value .rank(ascending=True)

Monetary Rank: Customers who spent the most get the highest scores. Using this pyhton code on the value .rank(ascending=True)

Overall RFM Score
The overall RFM score was the sum of the individual ranks:

RFM Score=Recency Rank+Frequency Rank+Monetary Rank

The clusters of customers was assigned and defined based on the calculated RFM scoring and ranking as denoted below:

RFM Score	Customer Cluster Label
0.0 and below	Low-Valued
0.1 to 0.9	Medium-Valued
1.0 and Above	High-Valued

Table 2: RFM Score Rule

I. Marketing Strategy Formulation
For both K-Means and HDBSCAN clustering systems, the RFM scores and ranks function was applied to the dataset to create scores for every customer group and ranks for every customer individually. Calculating the average RFM score for every cluster helped one to grasp the traits of every consumer group. This approach enables the identification of low-value, medium-value, and high-value consumers as well as appropriate modification of marketing plans.

III. RESULTS

The outcome of the study will focus on major key areas to buttress the application of KMean and HDBSCAN clustering algorithms to influence effective marketing decisions. An overview of the descriptive statistics, handling of outliers and noisy dataset for both machine learning algorithms, correlation heat matrix, visualisation of the clustering output, tabular representation of clusters, silhouette and RFM score, recommended marketing strategies based on the 7Ps.

A. Descriptive Statistics
The numerical data analysis below shows unique customers with an average age of 44, an average purchase amount of $59.76, a review score of 3.75 out of 5, past purchases of 25.35, and a frequent purchasing period of 89 days.

```
# Exploratory Data Analysis (EDA)
#descriptive statistics of numerical columns
df.describe()
```

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases	Frequency of Purchases
count	3900.000000	3900.000000	3900.000000	3900.000000	3900.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.749949	25.351538	89.133077
std	1125.977353	15.207589	23.685392	0.716223	14.447125	119.037566
min	1.000000	18.000000	20.000000	2.500000	1.000000	7.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000	14.000000
50%	1950.500000	44.000000	60.000000	3.700000	25.000000	30.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000	90.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000	365.000000

Table 3: Descriptive Statistics

B. Identified Outlier
Two plots showing purchase frequency are shown in Figure 1. With a peak of almost 1400 on the count axis, the first chart, which depicts the frequency range of 0–50, indices the range of most transactions. Half as tall, the second bar, which represents the frequency range of 50–100, reaches about 700 on the count axis.

Shorter, signifying less counts in those areas, the third and fourth bars show greater frequency ranges.

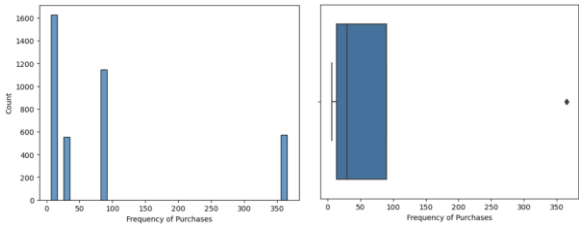

Figure 1: Frequency of Purchases

Representing the interquartile range (IQR), the box plot displays a central rectangle on the frequency axis ranging from just below 50 to just above 100. The median line inside the rectangle represents the frequency of buying median value. Representing variability outside the upper and lower quartiles, whiskers stretch from either end of the box. A dot placed further down the frequency axis indicates an outlier, that is, a much greater purchase frequency than the rest of the data.

C. Handling Outliers

KMeans Machine Learning algorithm is Sensitive to outliers and noisy data points, which can skew clustering outcomes and produce erroneous clusters and poor performance. The code in Figure 3.16 caps outliers at a maximum value of 90 days to help to reduce their impact without sacrificing important data points. This guarantees that, even with extreme values brought into a tolerable range, the most of the data stays unaltered.

The cap of ninety days is based on the upper IQR point of the "Frequency of Purchases," therefore assuring that most data points stay whole. This method is less extreme than eliminating outliers, which can greatly shrink the dataset and maybe compromise the validity of the research. The KMeans method can run better by limiting outliers since it lessens the impact of extreme data, hence producing more accurate and significant clusters.

Any "Frequency of Purchases" number higher than 90 in the given sample code in Figure 3.15, which caps at 90, effectively manages outliers and maintains all 3900 dataset items.

D. Correlation Heat Map

Using two techniques of the KMeans Algorithm in which outliers have been treated and the HDBSCAN Algorithm, the picture shows two heatmaps comparing correlation matrices of variables.

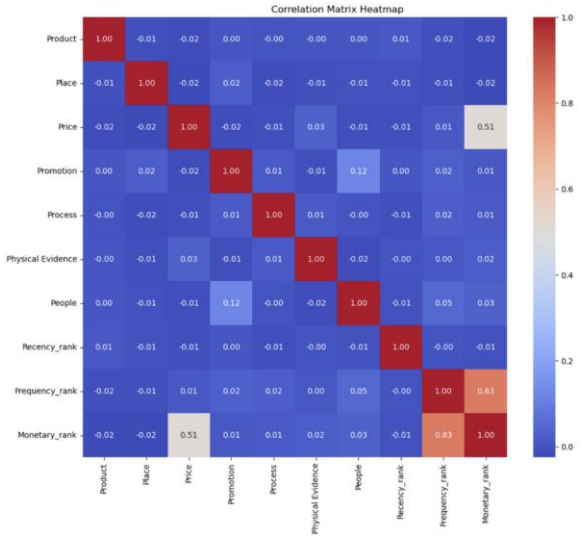

Figure 2: Correlation Heatmap Matrix

With positive correlations between variables like "Monetary_rank" and "Price," positive correlations between variables like "Frequency_ rank" and "Monetary_ rank," and little to no correlation with variables like "Physical Evidence" and "People" the heatmaps exhibit similar trends. A very slight difference exists in the “Recency_ rank” in both plots. The main variations are methodology sensitivity; HDBSCAN is a density-based method while KMeans is a centroid-based method. These variations can help one understand how every method views relationships between variables, hence enhancing data analysis and clustering outcomes.

E. Optimal Clusters for KMean and HDBSCAN

Optimal Numbers of Cluster for Features of 7Ps and RFM Models		
	HDBSCAN	KMEAN
Cluster -1	657	-
Cluster 0	10	1366
Cluster 1	10	1826
Cluster 2	3223	708
Cluster 3	-	-
	4	3

Table 4: Optimal Numbers of Clusters

F. KMean Clustering Visualisation

Three separate data point clusters, with each symbolised by a different colour (red, green, and blue) shown on the KMeans clustering analysis graph in Figure 3. The data score was derived from the means of the mean of the DataFrame final_df's "Recency_rank," "Frequency," and "Monetary" columns. This allows one to classify consumer groups according to their purchase patterns. With consideration for the mean of "Product," "Place," "Price," "Promotion," "Process," and "Physical Evidence" columns, the 7Ps score assesses many facets of the marketing mix, which was both computed to visualise the scattered plot in Figure 3.

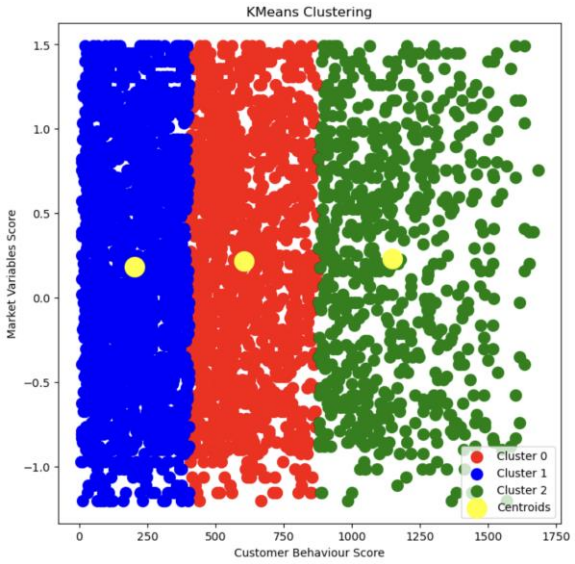

Figure 3: KMean Clustering

These clusters go under Cluster 0, Cluster 1, and Cluster 2 labels. Showing the y_kmeans array's unique value count, the cluster_counts of Cluster; 0 shows 1366 times; 1 shows 1826 times; 2 shows 708 times. Each cluster's centroids is at the core points around which the data points are arranged, are shown by the yellow points. The market variables score is shown on the y-axis; the customer behaviour score is shown on the x-axis. The well-separated groups show effective, based on similarity grouping. The average position of the data points inside every cluster is indicated by the centroids at their centre.

G. HDBSCAN Clustering Visualisation

The HDBSACN algorithm helps to concentrate on core clusters (Cluster 0, 1 and 2). Cluster densities help discover essential clusters and grasp their features, improving the general cluster structure. The challenge encountered was handling noise and outliers in the data, which the implementation of HDBSCAN identified outliers as separate clusters in Cluster -1, depicted below in Figure 4.

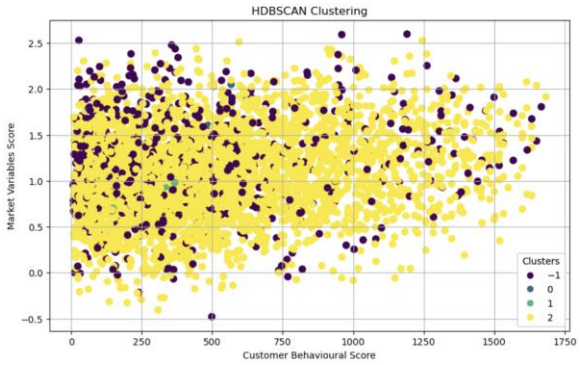

Figure 4: HDBSCAN Clustering

Density was computed using the probability attribute, facilitating the understanding of the density of every cluster as shown in Figure 5. Dense clusters and core points, including densities, were highlighted in another scatter plot. This visual aid clarifies the clustering findings and improves the grasp of clustering.

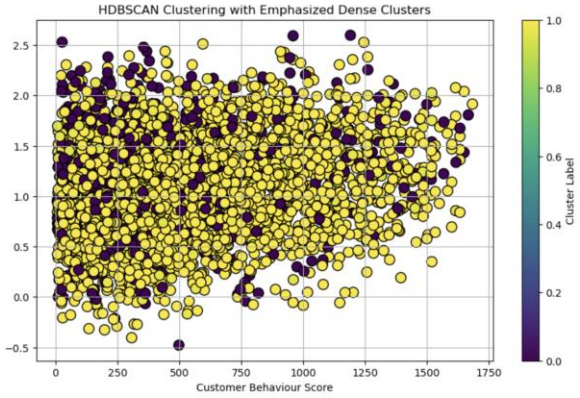

Figure 5: HDBSCAN Dense Clustering

H. Silhouette Clusters Evaluation

The performance of K-Means and HDBSCAN was evaluated using Silhouette Score as shown in table 6. K-Means provided clear and distinct clusters but struggled with noise and varying densities. HDBSCAN, on the other hand, excelled in identifying clusters with varying shapes and densities and

managing noise, though it required careful parameter tuning.

Silhouette Clusters Evaluation		
	Silhouette Score	Silhouette Level
HDBSCAN	0.20	Low
KMEAN	0.58	Moderate

Table 5: Silhouette Score

KMean's clustering appears to perform way better than the silhouette score of a moderate degree of '0.58', while the silhouette score of HDBSCAN being close to '0' at '0.28' degree portrays the decision boundary of the clusters is very close and indicates that the clusters may not be well-defined.

The complexity of the dataset, as eminent in the noisiness of the data, portrays a real-world scenario of how complex human behaviour could be. KMean Approach could have performed better in the silhouette score due to the data manipulation in handling outliers, in which, in the real world, consumer behaviours cannot be directly influenced, only through the use of an effective marketing strategy. Which levels of the HDBSCAN model are more appropriate to adopt in spite of its lag in its silhouette score, which that realistically treats the dataset wholistically, without compromise in the quality of data.

I. RFM Score Evaluation

RFM scores were calculated for each customer and integrated into the clustering process, shown in Table 7. This enhanced the segmentation by identifying high-value customers, allowing for more targeted marketing strategies.

RFM Score Evaluation				
Cluster / Customer Segment	RFM Score			
	HDBSCAN	Customer Label	KMEAN	Customer Label
Cluster -1	-0.42	Low Value	-	-
Cluster 0	-2.09	Low Value	0.92	Medium Value
Cluster 1	-2.12	Low Value	-1.67	Low Value
Cluster 2	0.098	Medium Value	2.53	High Value

Table 6: RFM Score

HDBSCAN's low RFM score for Cluster - 1 suggests that these clients are not particularly valuable. Their monetary values, frequency, and recency may be low. Based on HDBSCAN, Cluster 0 has a low RFM score; but, based on KMEAN, this cluster has a medium RFM score. This implies that whilst HDBSCAN regards some consumers as less desirable, KMEAN finds them to be rather valuable. Both HDBSCAN and KMEAN rank Cluster 1 as having low value, meaning that these consumers are not regular or heavy spenders.

Based on HDBSCAN, Cluster 2 boasts a medium RFM score; based on KMEAN, it boasts a high RFM score. Given their large spending and regular purchases, this points to these consumers as valued.

J. Marketing Strategy Development

Insights from the clustering analysis were used to tailor marketing strategies according to the 7Ps model. This comprehensive approach ensured that all aspects of the marketing environment were addressed, leading to more effective consumer engagement:

Product Strategy

High Valued Customers: Ensure high quality and offer customised and personalised products based on their purchasing history and preferences.

Medium Valued Customers: Ensure high quality and offer relevant and personalised products based on their needs and preferences.

Low Valued Customers: Targeted promotions to increase the monetary value of purchases. Eg complementary products

Price Strategy

High Valued Customers: Loyalty/exclusive discounts, and bundle sale offer.

Medium Valued Customers: Loyalty/exclusive discounts to encourage repeated purchases

Low Valued Customers: Offering discounts on repeat purchases

Place Strategy
High Valued Customers: Products to be accessible on multiple channels and offer early access to new products.

Medium Valued Customers: Products to be accessible on multiple channels.

Low Valued Customers: Products to be accessible on multiple channels. Offer free home delivery

Promotion Strategy
High Valued Customers: Personalised marketing on special offers.
Medium Valued Customers: Targeted marketing campaigns to highlight products that match their purchase history.
Low Valued Customers: Focus on retention campaigns: personalised follow-ups, special offers, or loyalty rewards.

People Strategy
High Valued Customers: Offer exceptional customer service
Medium Valued Customers: Offer exceptional customer service
Low Valued Customers: Gather feedback to understand their needs and preferences better.

Process Strategy
High Valued Customers: Simplified the purchasing process and gathered feedback to continuously improve their experience.
Medium Valued Customers: Simplified the purchasing process and gathered feedback to continuously improve their experience.
Low Valued Customers: Simplified the purchasing process and gathered feedback on best mode of service

Physical Evidence Strategy
Branding consistence and highly quality packaging for all customers.

CONCLUSION

Using the RFM model and the 7Ps of Marketing Mix, this paper examined K-Means and HDBSCAN clustering techniques to improve marketing tactics. The given dataset comprised several factors connected to consumer behaviour and demography. Important results showed that although K-Means Clustering was successful in producing separate clusters depending on purchase behaviour, it needed the prior specification of the number of clusters. With careful parameter adjustment for best results, HDBSCAN Clustering effectively detected clusters of varied forms and densities handling noise.

Every customer received an RFM score, which improved segmentation by pointing up valuable consumers. More successful consumer engagement resulted by customising marketing plans based on the 7Ps model using the findings from the clustering analysis, so improving customer interaction. Physical evidence was used in this regard.

The results coincide with body of knowledge already in publication on the efficiency of K-Means and HDBSCAN in use in clustering tasks. Combining clustering findings with the RFM model and the 7Ps framework offers a fresh method that offers a more whole perspective on consumer segmentation and marketing plan building.

Enhanced data quality, sophisticated feature engineering, algorithm optimisation, real-time clustering, cross-channel analysis, and longitudinal research are among the possible product improvements. Legal, social, ethical, and professional concerns like data privacy, intellectual property rights, inclusivity, consumer trust, bias in data, transparency, industry standards, best practices, and ongoing learning were taken under consideration.

All things considered, this study offers insightful analysis of how well K-Means and HDBSCAN clustering methods fit marketing plans. Improving data quality, boosting feature engineering,

investigating hybrid clustering techniques, and following industry standards and best practices should all take front stage in future efforts.

REFERENCES

[1] Aslam, S. (2023, October 19). Consumer Behaviour and Shopping Habits Dataset: Kaggle. Retrieved August 7, 2024, from https://www.kaggle.com/datasets/zeesolver/consumer-behaviour-and-shopping-habits-dataset/data

[2] Azzam, Z. A., & Ali, N. N. (2019). The Relationship between Product Mix Elements and Consumer Buying Behaviour– A Case of Jordan. Global Journal of Economic and Business, 6(2), 375–384. https://doi.org/10.31559/gjeb2019.6.2.10

[3] Bock, H. H. (2007). Clustering Methods: A History of k-Means Algorithms. Studies in classification, data analysis, and knowledge organization (pp. 161–172). https://doi.org/10.1007/978-3-540-73560-1_15

[4] Daelemans, W., & Goethals, B. (2008). Machine Learning and Knowledge Discovery in Databases. Springer Science & Business Media.

[5] Dennis, A. (2024, July 2). 13 Customer Experience Challenges to Overcome (2024). The Whatfix Blog | Drive Digital Adoption. https://whatfix.com/blog/customer-experience-challenges/

[6] Lahey, S. (2023, December 27). Customer dissatisfaction: A guide to handling unhappy customers. Zendesk.

[7] Malzer, C., & Baum, M. (2020). A Hybrid Approach To Hierarchical Density-based Cluster Selection. https://doi.org/10.1109/mfi49285.2020.9235263

[8] Pozo, D. (2021, November 5). How losing trust costs brands customers – and what marketers can do to prevent it. Marketing Week. Retrieved August 7, 2024, from https://www.marketingweek.com/losing-trust-costs-brands-customers/

[9] RFM Analysis Analysis Using Python. (2024). GeeksforGeeks. https://www.geeksforgeeks.org/rfm-analysis-analysis-using-python/

[10] Sam. (2024, July 10). Eliminate waste and complexity in your digital advertising budget. Proxima. https://proximagroup.com/proxima-perspectives/eliminate-waste-and-complexity-in-your-digital-advertising-budget/#:~:text=Proxima's%20research%20into%20the%20state,traffic%20and%20poor%20viewability%2Fplacement

[11] Solon, O. (2020, April 16). Kendall Jenner's Pepsi ad was criticized for co-opting protest movements for profit. The Guardian. https://www.theguardian.com/fashion/2017/apr/04/kendall-jenner-pepsi-ad-protest-black-lives-matter

[12] Stewart, G., & Al-Khassaweneh, M. (2022). An Implementation of the HDBSCAN* Clustering Algorithm. Applied Sciences, 12(5), 2405. https://doi.org/10.3390/app12052405

[13] Strobl, M., Sander, J., Campello, R. J. G. B., & Zaïane, O. (2020). Model-based Clustering with HDBSCAN*. In University of Alberta.

[14] Theng, D., & Bhoyar, K. K. (2023). Feature selection techniques for machine learning: a survey of more than two decades of research. Knowledge and Information Systems, 66(3), 1575–1637. https://doi.org/10.1007/s10115-023-02010-5

[15] Thompson, J. (2024, July 18). Big Brands That Lost Customers' Satisfaction in 2023 [Where CX Went Wrong + Data]. https://blog.hubspot.com/service/companies-that-lost-customers

[16] Wedel, M., & Kamakura, W. A. (2000). Market Segmentation. In International series in quantitative marketing. https://doi.org/10.1007/978-1-4615-4651-1

[17] Weidner, J. B. (2024, July 3). Why Google Glass Failed. Investopedia. https://www.investopedia.com/articles/investing/052115/how-why-google-glass-failed.asp#:~:text=Google%20Glass%20was%20marketed%20as,Week%20and%20in%20relevant%20advertisements.

[18] Williams, A. (2021, December 8). Learning from the Gap Logo Redesign Fail. The Branding

Journal.
https://www.thebrandingjournal.com/2021/04/le
arnings-gap-logo-redesign-fail/