Group 29 Project Report on

Performance of Hybrid Clustering Techniques in Market Segmentation

Shreya Pandey (M24CSA030), Prateek (M24CSA022), Princu Singh (M24CSA024)

1 Introduction:

Market segmentation is a critical process that organizations undertake in an effort to understand and target different groups within their customer base. This means that market segmentation helps businesses identify customer groups who share similar characteristics, behaviors, or preferences and whose needs can be served through targeted marketing strategies, product development, and customer service efforts. Commonly used traditional approaches include clustering methods, such as k-means or Gaussian Mixture Models (GMM), in matching similar data objects together by grouping them for further targeted engagement with customer segments.

However, given that customer data becomes large in size and dimensionality, it is often not feasible for traditional clustering methods to deliver meaningful, accurate segments. In such cases, datasets, especially high-dimensional ones with diverse customer profiles, are challenging for these methods; this often leads to less reliable segments when behaviours and attributes associated with customers are wide in range. In such a scenario, conventional clustering models may oversimplify groupings or fail to capture some subtle patterns of behaviour, reducing the effectiveness of the segmentation process.

To overcome these, new hybrid clustering techniques have emerged as a strong alternative. Hybrid methods benefit from the strengths of several algorithms together: speed and simplicity of k-means; flexibility in overlapping clusters, as in GMM; and noise-handling capabilities, like DBSCAN. This implies that hybrid methods can improve both the accuracy and interpretability of segmentation results.

2 Problem Statement:

Hybrid clustering techniques, which combine multiple algorithms, offer a promising solution to these challenges. By leveraging the strengths of different clustering approaches, hybrid methods aim to improve the validity and interpretability of segmentation outcomes, especially in high-dimensional and varied data.

The primary goals of this project are to:

- **Enhance segmentation performance** by implementing hybrid clustering techniques, improving the accuracy and stability of customer segments.
- **Design and apply hybrid algorithms** that combine k-means, DBSCAN, and GMM to address the limitations of traditional methods.

By combining the simplicity of k-means for initial segmentation, the noise-handling capability of DBSCAN, and the flexibility of GMM for overlapping clusters, this project aims to develop a robust framework for market segmentation. The outcome will demonstrate how hybrid clustering can provide more precise, actionable customer insights, helping organizations make better-informed business decisions.

3 Literature Survey:

RESEARCH PAPER	NOVELTY	SHORTCOMINGS
K-means Clustering: A half-century synthesis	Includes the research done in this area in the last 50 yrs	The clusters are sensitive to outliers.
Unsupervised K-Means Clustering Algorithm	Doesn't require the initial number of clusters a priori	Handling noisy data, scalability
DBSCAN Clustering Algorithm Based on Density	Refining how density peak points are discovered	Lack of Parameter Tuning Discussion
Gaussian Mixture Model Clustering with Incomplete Data	Unified imputation and clustering process, Extensive experimentation	Computational complexity, assumes Known Number of Clusters
A Combination of K-Means and DBSCAN in Customer Segmentation	Hybrid Clustering Approach, Handling non-circular clusters	Heuristic Combination Method, Absence of ROI Evaluation

4 Dataset:

Dataset used: Customer Segmentation Dataset

Source: Kaggle (Link to Dataset)

The dataset used for this project is from Kaggle's **Customer Segmentation** Collection, containing detailed records of customer profiles and behaviors. It includes multiple attributes relevant for segmentation. Below is a breakdown of each feature:

- **CUST_ID**: A unique identifier for each credit card holder.
- **BALANCE**: The remaining account balance available for purchases, indicating the financial status of the cardholder at a given time.
- **BALANCE_FREQUENCY**: A measure between 0 and 1 that shows how often the balance is updated, where 1 indicates frequent updates and 0 indicates infrequent updates.
- **PURCHASES**: The total amount spent on purchases, reflecting overall spending habits.
- **ONEOFF_PURCHASES**: The maximum amount spent in a single transaction, useful for understanding large, one-time purchase behavior.
- **INSTALLMENTS_PURCHASES**: The amount spent on installment-based purchases, indicating a preference for payment over time.
- **CASH_ADVANCE**: The total amount of cash advances taken, which can be a sign of financial liquidity needs.

- **PURCHASES_FREQUENCY**: Frequency of purchases made on a regular basis, scaled between 0 and 1, with 1 being frequent and 0 being rare.
- **ONEOFF_PURCHASES_FREQUENCY**: Frequency of one-time purchase transactions, again scaled from 0 to 1.
- **PURCHASES_INSTALLMENTS_FREQUENCY**: Frequency of installment-based purchases, scaled from 0 to 1, showing the habit of using installments.
- **CASH_ADVANCE_FREQUENCY**: Frequency of taking cash advances, providing insight into the reliance on cash advances.
- **CASH_ADVANCE_TRX**: Total number of cash advance transactions, which reflects how often a user accesses quick cash.
- **PURCHASES_TRX**: Total number of purchase transactions, indicating purchase activity.
- **CREDIT_LIMIT**: The maximum credit card limit, showing the borrowing capacity allowed to each user.
- **PAYMENTS**: The total amount paid by the user, reflecting repayment behavior.
- **MINIMUM_PAYMENTS**: The minimum amount paid by the user, useful for analyzing compliance with minimum payment requirements.
- **PRC_FULL_PAYMENT**: The proportion of total charges paid off in full by the user, providing insight into financial discipline.
- **TENURE**: The length of time the user has been a credit card holder, which may correlate with loyalty or spending patterns.

5 Methodology:

5.1 Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is a crucial step in data analysis that helps to summarize the key characteristics of a dataset, often with visual methods. The following steps were performed:

- **Data Shape and Types:** The dataset consists of 8950 rows and 18 columns, with a mix of numerical and categorical data types. The columns related to customer spending, payment history, and credit details were identified as key features for clustering.
- **Missing Values:** We identified some missing values, specifically in CREDIT_LIMIT and MINIMUM_PAYMENTS. These were handled by either imputation or exclusion, based on the data's relevance to clustering.
- **Duplicate Checks:** The dataset was checked for duplicates, revealing no redundant rows.
- **Descriptive Statistics:** Summary statistics were computed for numerical columns to understand distributions and identify potential outliers which is shown in figure 1:
- **Uniqueness of Columns:** The number of unique values in each column was assessed using the apply method.

	mean	std	min	25%	50%	75%	max
BALANCE	1564.647593	2081.584016	0.000000	128.365782	873.680279	2054.372848	19043.138560
BALANCE_FREQUENCY	0.877350	0.236798	0.000000	0.888889	1.000000	1.000000	1.000000
PURCHASES	1003.316936	2136.727848	0.000000	39.800000	361.490000	1110.170000	49039.570000
ONEOFF_PURCHASES	592.503572	1659.968851	0.000000	0.000000	38.000000	577.830000	40761.250000
INSTALLMENTS_PURCHASES	411.113579	904.378205	0.000000	0.000000	89.000000	468.650000	22500.000000
CASH_ADVANCE	978.959616	2097.264344	0.000000	0.000000	0.000000	1113.868654	47137.211760
PURCHASES_FREQUENCY	0.490405	0.401360	0.000000	0.083333	0.500000	0.916667	1.000000
ONEOFF_PURCHASES_FREQUENCY	0.202480	0.298345	0.000000	0.000000	0.083333	0.300000	1.000000
PURCHASES_INSTALLMENTS_FREQUENCY	0.364478	0.397451	0.000000	0.000000	0.166667	0.750000	1.000000
CASH_ADVANCE_FREQUENCY	0.135141	0.200132	0.000000	0.000000	0.000000	0.222222	1.500000
CASH_ADVANCE_TRX	3.249078	6.824987	0.000000	0.000000	0.000000	4.000000	123.000000
PURCHASES_TRX	14.711476	24.858552	0.000000	1.000000	7.000000	17.000000	358.000000
CREDIT_LIMIT	4494.449450	3638.815725	50.000000	1600.000000	3000.000000	6500.000000	30000.000000
PAYMENTS	1733.336511	2895.168146	0.000000	383.282850	857.062706	1901.279320	50721.483360
MINIMUM_PAYMENTS	838.486229	2335.473910	0.000000	164.391437	295.779348	794.656428	76406.207520
PRC_FULL_PAYMENT	0.153732	0.292511	0.000000	0.000000	0.000000	0.142857	1.000000
TENURE	11.517935	1.337134	6.000000	12.000000	12.000000	12.000000	12.000000

Figure 1: Statistical description of each numerical column

5.2 Visualization:

Visualization plays a critical role in understanding and interpreting complex datasets by providing a clear and intuitive graphical representation of the data. It allows us to identify patterns, trends, and relationships among variables, facilitating deeper insights. Various visualization techniques such as scatter plots, distribution charts, and correlation matrices are employed to explore the dataset, identify outliers, and analyze feature distributions, ultimately helping to inform decision-making and model development.

- **Correlation Matrix**: A heatmap-style correlation matrix is used to understand the relationships between different features. By visualizing the correlation, it becomes easier to detect any redundancies or strong relationships that can be useful for feature engineering.

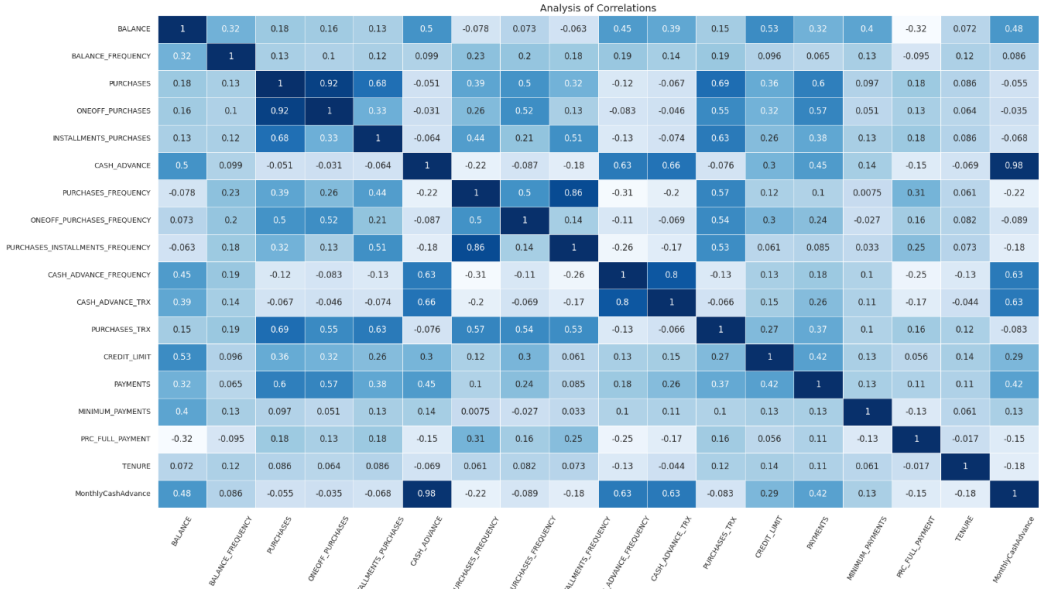

Figure 2: Correlation Matrix for the features

- **Scatter Plots**: A series of scatter plots were created to analyze the relationship between PURCHASES and various other features (like BALANCE, CREDIT_LIMIT, etc.), color-coded by TENURE to identify any trends or clusters within each tenure category.
- **Subplot Grid for BALANCE**: A set of scatter plots was generated to visualize BALANCE against each feature, helping to better understand the relationship between BALANCE and other financial attributes**.**
- Furthermore, the visualization also presents the relationship between CREDIT_LIMIT and BALANCE for different TENURE values. A main scatter plot displays the data with different colors for each TENURE category, showcasing how BALANCE varies with CREDIT_LIMIT across different tenure levels. Smaller scatter plots (subplots) are used to highlight individual tenure categories, with each subplot emphasizing the distribution of CREDIT_LIMIT and BALANCE for the selected tenure. This allows for a detailed comparison of how the two features interact across different tenure levels, with visual enhancements such as custom scatter plot styles and grid lines for clarity.
- **Comparison of Purchases Amount and Total Transaction Frequency by Tenure**:This section visualizes the comparison of the minimum, average, and maximum values for PURCHASES and PURCHASES_TRX across different tenure levels. The left plot represents the PURCHASES data, while the right plot compares PURCHASES_TRX. For each tenure group, data points are plotted for the minimum, maximum, and average values, with annotations for clarity.

Figure 3: purchase amount and Total Purchase Transaction Comparison

The distribution of key features was visualized using histograms and KDE plots, which provide an understanding of their spread and potential skewness. To summarize central tendencies, a bar chart depicting the average values of each feature was created. Outlier detection was carried out using the standard deviation method, where values that deviated beyond three standard deviations were identified and removed. Finally, a correlation matrix was generated to assess the relationships between different features, helping to reveal any underlying patterns and dependencies within the data.

5.3 K-Means:

To analyze the dataset further, K-means clustering was employed with hyperparameters configured to a random initialization method, ten initializations (n_init=10), a maximum of 300 iterations (max_iter=300), and a fixed random seed to ensure reproducibility.

To determine the optimal cluster count, the Elbow Method was applied by plotting the inertia across cluster numbers (1-10). A marked "elbow" in this plot indicated the optimal cluster count where adding more clusters provides minimal improvement in explaining variance. Additionally, silhouette scores and Calinski-Harabasz scores were plotted for clusters (k=2-10) to validate the cluster structure further. The silhouette coefficient, in particular, offered insights into the density and separation of clusters, while the Calinski-Harabasz score assessed the compactness and dispersion.

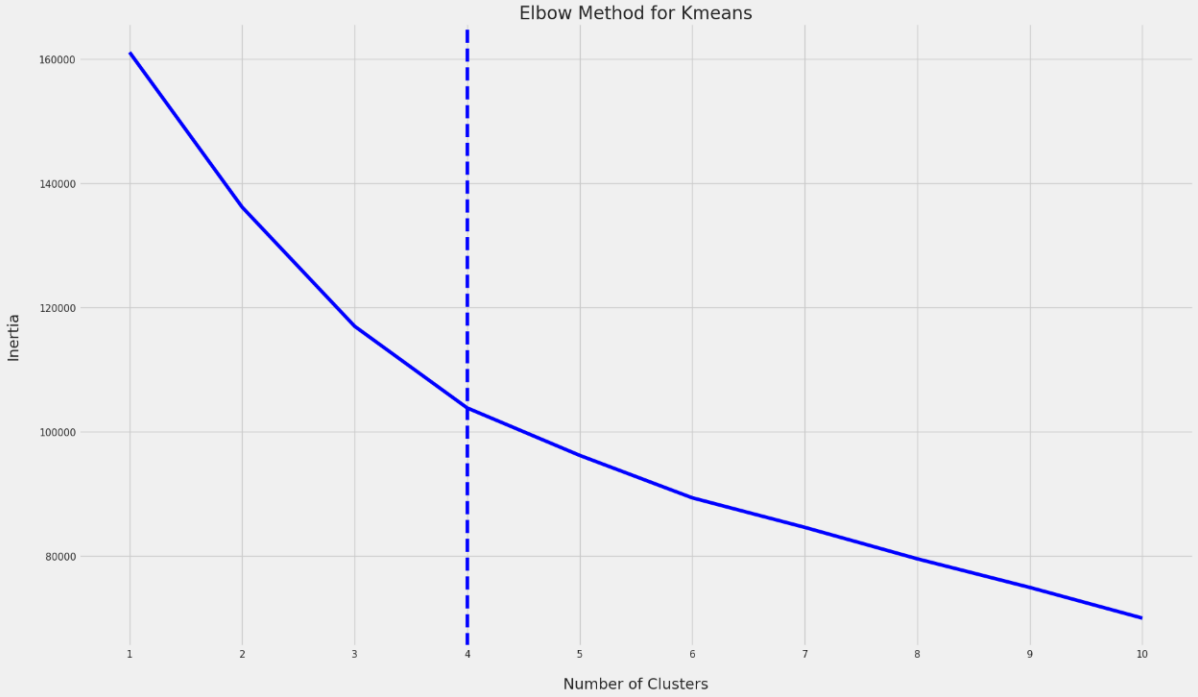

Figure 4: Graph of Elbow Method for K-Means

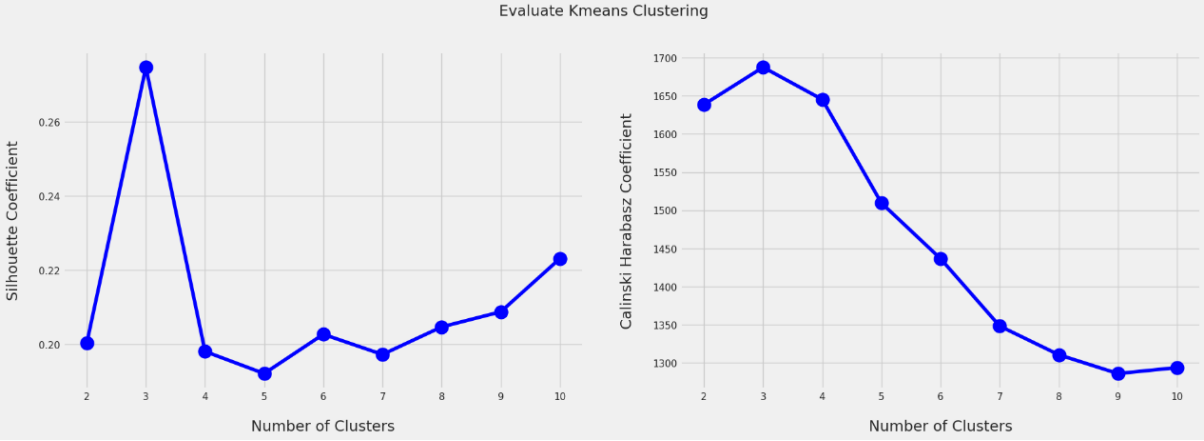

Figure 5: Plots for Silhouette Coefficient vs Number of Coefficient and Calinski-Harabasz Coefficient vs Number of Clusters for K-Means Clustering

After determining the optimal number of clusters, the final K-means model was configured to use three clusters, as the metrics collectively suggested this choice. The cluster labels from this model were added

to the original dataset to analyze each cluster's composition and count. A subsequent count plot and pie chart illustrated the distribution of data points across clusters, providing a quick visual summary of each cluster's size and proportion.

Further, Principal Component Analysis (PCA) was applied to reduce the high-dimensional data into a 2D space. A scatter plot of the two principal components, color-coded by clusters, was then generated. The centroids for each cluster were marked to highlight the central point of each group, with red 'X' markers designating these centroids. This 2D plot offered an effective visualization of the clusters, showing both separation and overlap within the data.

Figure 6: Scatter plot of the two principal components with K-Means Clusters

A more granular examination was carried out by creating scatter plots for each feature against the 'BALANCE' variable, with clusters represented by different colors. Each subplot includes black 'x' markers to represent cluster centroids, facilitating comparisons of feature distributions across clusters. This approach allowed the visual assessment of feature relationships within clusters, offering deeper insights into feature behavior.

Finally, the average silhouette score for the three-cluster solution was calculated, providing an overall metric of cluster separation quality. This score complements the visual analysis, further confirming the suitability of the chosen clusters.

5.4 DBSCAN:

To further analyze the dataset, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was utilized, a clustering method effective for identifying clusters of varying shapes and managing outliers. DBSCAN operates by defining clusters as areas of high data point density, separating them from regions of lower density, which are treated as noise or outliers.

To determine the optimal eps and min_samples values, a parameter grid was created with eps ranging from 3.0 to 5.0 in 0.25 increments, and min_samples ranging from 20 to 40 in increments of 10. Each parameter combination was evaluated using two metrics: the Silhouette score, assessing the density and separation quality of clusters, and the Calinski-Harabasz score, indicating compactness and separation between clusters. These metrics were plotted across the parameter grid to visually identify combinations yielding high clustering quality.

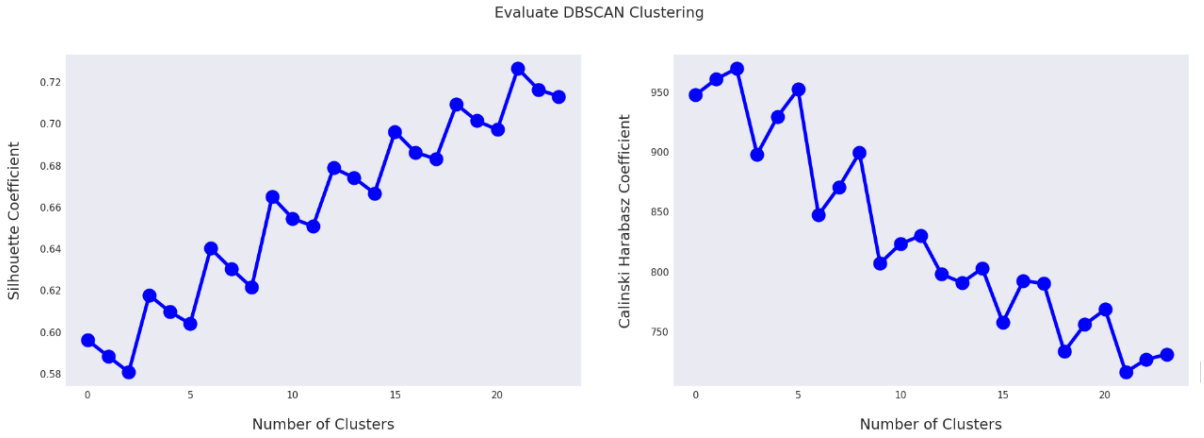

Figure 7: Plots for Silhouette Coefficient vs Number of Coefficient and Calinski-Harabasz Coefficient vs Number of Clusters for DBSCAN clustering

After examining the evaluation plots, the final DBSCAN model was configured with eps = 3.5 and min_samples = 40. This configuration balanced the cluster quality and the total number of clusters formed, with clear separation in the dataset and minimal outliers.

After reviewing these plots, the final DBSCAN model was set with eps = 3.5 and min_samples = 40, resulting in two primary clusters (labeled 0 and -1, where -1 indicates noise points). The final model's performance yielded a Silhouette Coefficient of 0.62 and a Calinski-Harabasz score of 899.06, confirming satisfactory clustering quality and separation.

The results were visualized by creating scatter plots of each feature against the 'BALANCE' variable, with different colors indicating clusters.

5.5 Gaussian Mixture Model (GMM):

For further analysis, the Gaussian Mixture Model (GMM) was employed to cluster the data into two components, using a full covariance matrix for flexibility in modeling the cluster shapes. After fitting the model, predictions were made, and the cluster means were calculated. These means were then transformed back to the original scale to provide more interpretable results.

The Silhouette Score for this model was found to be 0.11, which indicates that the clusters have weak separation. However, the Calinski-Harabasz score of 964.91 suggests that the clusters are compact and well-dispersed, indicating a reasonable fit.

The final cluster labels were appended to the dataset, and the data was visualized by plotting scatter plots of the 'BALANCE' variable against other key features. These plots were color-coded by the predicted clusters to illustrate how each feature behaves across the two clusters.

Furthermore, the scatter plots were generated to provide a visual understanding of how the features are distributed across the identified clusters.

5.6 Hybrid clustering technique:

- The hybrid clustering method was developed by combining the results of three distinct clustering algorithms: **K-Means**, **DBSCAN**, and **Gaussian Mixture Models (GMM)**. First, the cluster labels from each of the algorithms were merged into a new DataFrame. The next step involved the use of **majority voting** to determine the final cluster label for each data point, assigning the label that appeared most frequently across the three methods.
- The distance of each data point from the centroids of each method (K-Means, DBSCAN, and GMM) was computed. For DBSCAN, noise points were assigned a large distance value to distinguish them from other points. These distances served as new features, allowing for a feature-level hybridization of the clustering process.
- A secondary **K-Means clustering** algorithm was applied to this feature set, where the distances to the centroids of K-Means, DBSCAN, and GMM were used as the input for the clustering model. This step aimed to refine the hybrid clustering output by leveraging the information from the three clustering methods in a unified feature space.
- The hybrid clustering model was then evaluated using the **Silhouette Score**, which provided insight into the quality of the final clusters. This score measured how similar data points were to their own cluster compared to other clusters, with a higher score indicating better-defined clusters.

6 Results:

6.1 3D PCA Visualization:

PCA was applied to reduce the feature space to three dimensions for visualizing the hybrid clustering results. The resulting 3D scatter plot highlights the clusters in the reduced space, showing the distinct separation of data points based on the hybrid clustering approach.

Figure 8: 3D Visualisation of Hybrid Clustering

6.2 Scatterplot of Hybrid Clusters:

A scatterplot was created using two distinguishing features—distance to K-Means centroid and distance to GMM centroid. The plot provides a clear visualization of how the hybrid clusters are distributed across these two features, with centroid markers highlighted to demonstrate the central positions of the clusters.

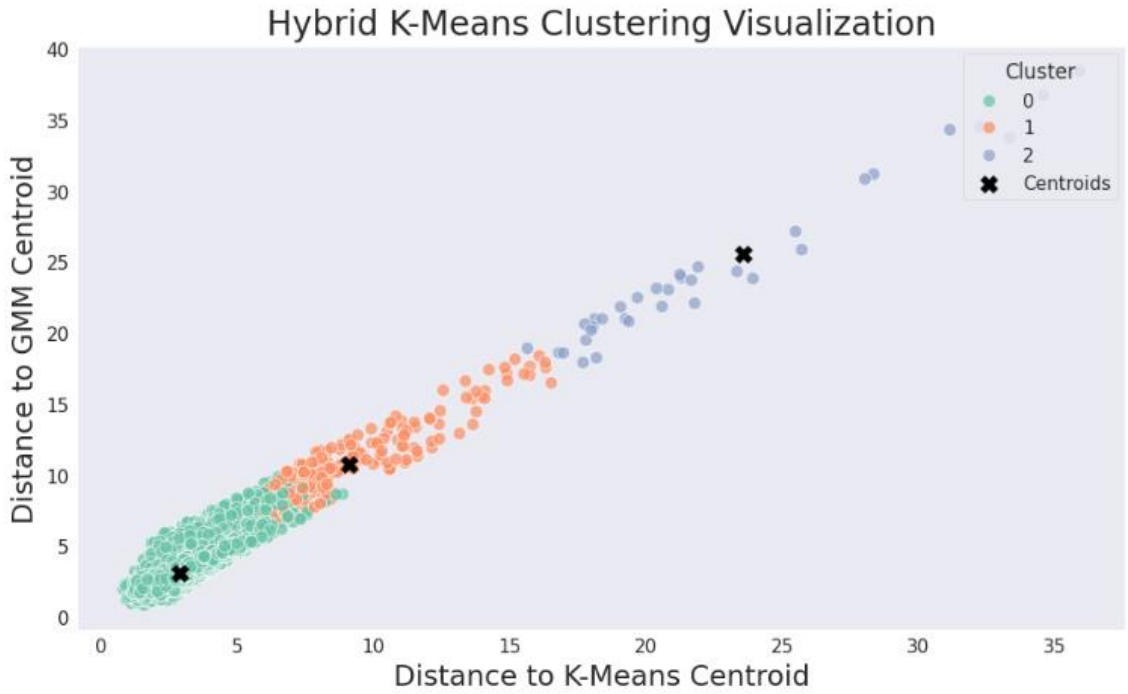

Figure 9: Scatterplot for Hybrid K-Means Clustering

6.3 Silhouette Score for Feature-Level Hybrid Clustering:

The silhouette score for the feature-level hybrid clustering approach was calculated to evaluate the clustering quality. The resulting score was 0.99 at random state 42, indicating that the feature set derived from the hybrid approach formed well-defined and distinct clusters.

6.4 Cluster Profiles Heatmap:

The cluster profiles were visualized using a heatmap as shown in Figure 10, where the mean values of the features for each final cluster were plotted. This heatmap revealed the distinct profiles for each cluster, offering insights into how the data points were grouped based on the hybrid clustering approach.

In the heatmap generated after applying hybrid clustering, we observe three distinct clusters labeled 0, 1, and 2, each representing different customer segments based on various purchasing behaviors and financial habits. Key insights from the heatmap include:

- Cluster 0 displays relatively low values across most features, indicating customers with minimal engagement in purchases, low cash advances, and smaller credit limits. These individuals may represent a segment with lower financial activity and spending.
- Cluster 1 is characterized by moderate levels across features like BALANCE, CASH_ADVANCE, and PAYMENTS. However, it shows a notable tendency for purchases and one-off purchases, suggesting that this segment includes customers who make periodic but meaningful purchases. The higher PURCHASES_TRX and CREDIT_LIMIT values indicate that this cluster may represent mid-level spenders with moderate credit utilization.

- Cluster 2 stands out with the highest values in features such as PURCHASES, ONEOFF_PURCHASES, and MINIMUM_PAYMENTS. This segment likely consists of high spenders with frequent large one-off purchases and higher minimum payment amounts, potentially indicating customers with high financial activity and dependency on credit facilities.

Thus, these clusters indicate a natural segmentation of customers into low, moderate, and high spenders, with unique patterns in credit usage and purchasing habits. This segmentation can guide targeted marketing strategies, helping to address the distinct needs of each customer group effectively.

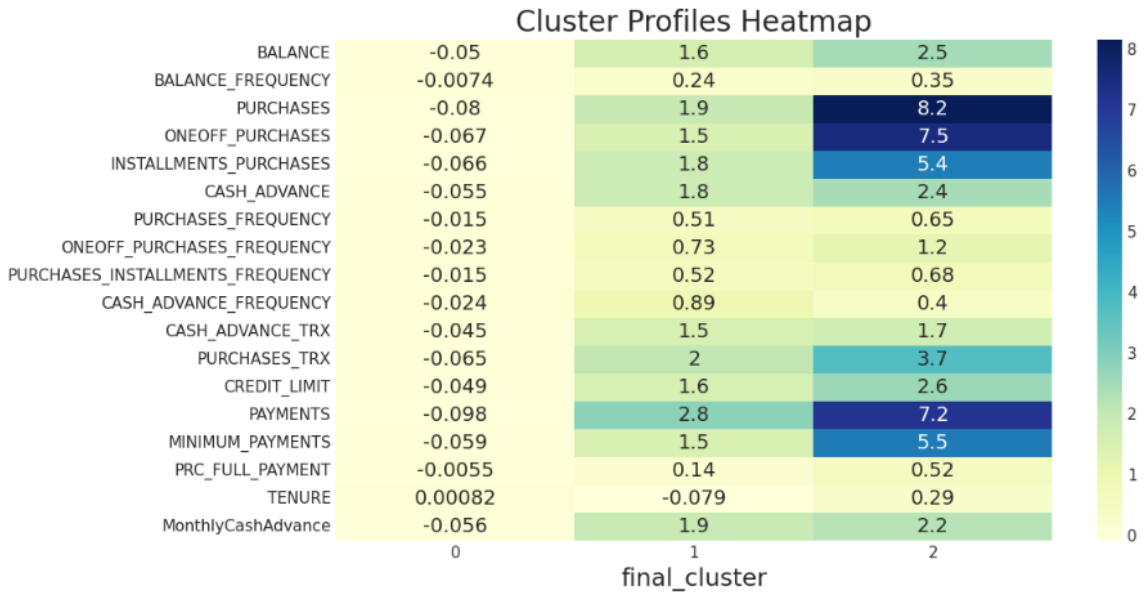

Figure 10: Cluster profiles Heatmap for Hybrid Clustering

7 Conclusion:

This project undertakes a hybrid approach toward clustering analysis. K-Means, DBSCAN, and Gaussian Mixture Models (GMM) thus provide the study with an all-round approach toward market segmentation based on their financial behaviors and purchasing habits. With the feature enhancements through clustering distances, DBSCAN noise identification, and labels from the individual algorithms, a richer dataset has been generated for further analysis.

Using K-Means, based on the hybrid set of features, three distinct customer segments did exist: Cluster 0 - low spenders, Cluster 1 - moderate spenders, and Cluster 2 - high spenders. Each cluster captures customers with peculiar patterns in purchases, credit usage, and payment behaviors: Cluster 0-low spenders with minimal engagements; Cluster 1-moderate spenders characterized by periodic purchases; and Cluster 2-high spenders who are high-value customers making frequent large purchases. These clusters are further validated using silhouette scores, the quality of the hybrid approach in customer base segmentation.

The cluster profiles from the heat map and feature importances suggest that customers have different financial activities that have to be approached in a targeted manner. For instance, low spenders in Cluster 0 need promotional offers, while high spenders in Cluster 2 need special credit offers or adjustments in their credit limits according to the spending behavior.

8 References:

1. S. Dolnicar and B. Grün, Market segmentation analysis: Understanding it, doing it, and making it useful, **Market Segmentation Analysis**, p. 324, 2018, doi: 10.1007/978-981-10-8818-6_2.

2. A. D. Chaturvedi, J. D. Carroll, P. Green, and J. A. Rotondo, "A feature-based approach to market segmentation via overlapping K-centroids clustering," **Journal of Marketing Research**, vol. 34, pp. 370-377, 1997.

3. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the E-M algorithm," **Journal of the Royal Statistical Society** B, vol. 39, pp. 1-38, 1977.

4. A. K. Jain and R. C. Dubes, **Algorithms for Clustering Data**, Englewood Cliffs, NJ: Prentice Hall, 1988.

5. A. K. Jain, "Data clustering: 50 years beyond K-means," **Pattern Recognition Letters**, vol. 31, pp. 651-666, 2010.

6. M. S. Yang, C. Y. Lai, and C. Y. Lin, "A robust EM clustering algorithm for Gaussian mixture models," **Pattern Recognition**, vol. 45, pp. 3950-3961, 2012.

7. S. S. Li, "An improved DBSCAN algorithm based on the neighbor similarity and fast nearest neighbor query," **IEEE Access**, pp. 99, 2020.

8. B. E. Cahyana, U. Nimran, H. N. Utami, and M. Iqbal, "Hybrid cluster analysis of customer segmentation of sea transportation users," **Journal of Economics, Finance and Administrative Science**, vol. 25, no. 50, pp. 321-337, 2020.

9. P. Barjatiya, "Customer segmentation using K-means clustering with PySpark: Unveiling insights for business success," **Medium**, 2023.

10. D. Dey, "DBSCAN clustering in ML - Density based clustering," **GeeksForGeeks**, 2023.