**E-commerce Customer Segmentation and Recommendation System**

# I. Farzhana

*Assistant Professor Computer Science and Engineering*

Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology

Avadi, Tamil Nadu, India [ifarzhana@veltech.edu.in](mailto:ifarzhana@veltech.edu.in)

# T.Sudharsan

*Artificial Intelligence and Data Science*

Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology

*Avadi, Tamil Nadu, India*

[vtu22719@veltech.edu.in](mailto:vtu22719@veltech.edu.in)

# S.Ishtiyaq Marzuq

*Artificial Intelligence and Data Science*

Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology

*Avadi, Tamil Nadu, India*

[vtu22815@veltech.edu.in](mailto:vtu22815@veltech.edu.in)

# J.Suriya Prakash

*Artificial Intelligence and Data Science*

Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology

*Avadi, Tamil Nadu, India*

[vtu23199@veltech.edu.in](mailto:vtu23199@veltech.edu.in)

***Abstract*—Customer segmentation has become essential for personalized marketing tactics in today's fiercely competitive corporate environment. This research introduces an unsupervised machine learning approach, K-Means clustering, for consumer segmentation in order to boost the success rate of e-commerce marketing. We cluster customers based on common traits by utilizing customer data, including demographics and previous purchases. By implementing focused marketing efforts, firms are able to increase consumer happiness and engagement through segmentation. Additionally, the framework emphasizes how crucial real-time data processing is to dynamic segmentation and recommendation systems. Sales, customer personalization, and overall marketing efficiency all show a considerable improvement, according to the results.**

***Index Terms:* Customer Segmentation, Unsupervised Machine learning, Marketing, Recommendation Systems**

1. **INTRODUCTION**

Businesses are facing more and more difficulties in meeting the various wants of their customers as a result of the development of new industries and the growth of consumers. Customer segmentation is one of the latest marketing tactics that companies must use in order to remain competitive. By using customer segmentation, businesses can group clients according to shared traits, which leads to more focused and successful marketing The process starts with preparing customer data, which includes handling duplicate and missing entries. Next, RFM (Recency, Frequency, Monetary) metrics are used in feature engineering. After that, the K-Means method is used to divide the clientele into various clusters.

1. **EASE OF USE**

*A. Ensuring Customer Segmentation Is Effective*

K-Means clustering consumer segmentation systems need to be flexible, precise, and offer a user-friendly interface. Here are five key phrases to take into account:

* **User-Friendly Interface:** Marketing and sales teams can investigate the customer cluster without having technical skills due to the segmentation
* **Optimized Data Processing**: Reliability and scalability of the model are increased by integrating effective data pretreatment procedures, such as handling duplicates, outliers, and missing values, to guarantee the accuracy of the clustering results.
* **practical Insights:** By distinctly identifying different customer segments, the technology offers marketing campaigns practical insights. This enables companies to develop tailored marketing plans that improve client loyalty and engagement.
* **Real-Time Dynamic Segmentation:** The system can dynamically re-segment client groups in real time since it receives constant feedback on consumer behaviour. This guarantees that marketing plans are always in line with the most recent patterns of consumer behaviour.
* **Scalable Clustering method:** The K-Means method is designed to be as scalable as possible, meaning that it can effectively manage datasets that are getting bigger and bigger. This makes the system appropriate for companies of all sizes.

1. **TECHNICAL CHALLENGES AND SOLUTIONS**
2. ***Challenges***  
   There were various technical difficulties in creating and deploying the K-Means clustering customer segmentation system. These included managing sizable and complicated datasets, resolving issues with missing or duplicate values in the data, streamlining the procedure for scalability, and ensuring that the model could adapt to modifying, real-time consumer data.
3. ***Proposed Solutions***

Advanced data preprocessing methods were used to ensure data quality in order to overcome these obstacles, including imputation for missing data and duplication removal

Using effective clustering evaluation metrics and adjusting the number of clusters, the K-Means algorithm was made more efficient for handling large-scale datasets.

1. **Related Works**

**Customer Segmentation and Recommendation Using K-Means**

This section highlights key contributions and methodologies in customer segmentation using machine learning, specifically the K-means clustering algorithm, from various research studies.

Kansal's study on "Customer Segmentation using K-means Clustering" [1] demonstrates its effectiveness in identifying distinct customer segments based on purchasing behaviours. The study emphasizes the practical implementation of clustering techniques for businesses, paving the way for targeted marketing strategies and further improvements in customer relationship management.

Maheswari, [2] in her paper “Finding Best Possible Number of Clusters using KMeans Algorithm”, The Silhouette score and the Elbow method are utilized in the clustering process to enhance the accuracy of the algorithm. These techniques, combined with the Elbow method, contribute to the growing research on optimizing K-means performance, crucial for creating well-defined customer segments.

Wu's empirical study [3],on customer segmentation using the Recency, Frequency, and Monetary (RFM Model) and K-means clustering demonstrates the effectiveness of RFM features in segmenting customers based on purchasing patterns. This hybrid approach, combining domain-specific features with machine learning algorithms, offers valuable insights for e-commerce businesses, making it a valuable resource for segmentation outcomes.

Bhardwaj's article [4], on cluster validation techniques, particularly the Silhouette Coefficient, provides a quantitative measure of customer fit within a cluster. The Silhouette score validates the quality of K-means clustering results, ensuring robustness of customer segmentation models and preventing ineffective marketing efforts due to poor clustering.

Kumar's work "Implementing Customer Segmentation Using Machine Learning" [5], on the implementation side of customer segmentation emphasizes the importance of data preprocessing, feature engineering, and validation in setting up a K-means clustering model. This serves as a blueprint for practitioners to develop customer segmentation models, integrating machine learning into business processes.

Vijilesh's research on "customer segmentation using machine learning"[6], provides a comprehensive analysis of various clustering algorithms, including hierarchical clustering, which offers a nuanced understanding of their performance under different conditions. This comparative analysis is valuable for organizations seeking to enhance their customer data management.

Karaman's article “Segmentation by RFM Clustering” [7],   
highlights the effectiveness of combining RFM analysis with clustering techniques, emphasizing the importance of RFM features in identifying valuable customer segments. This approach, like Wu et al.'s, enhances customer retention and satisfaction by combining traditional business metrics with modern machine learning techniques.

**K-Means vs Other Clustering Methods**

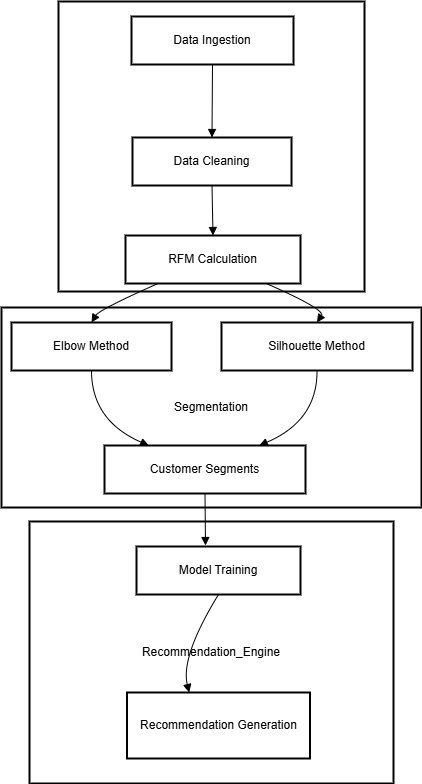
K-means is a popular choice for customer segmentation in e-commerce due to its efficiency and ease of use highlighted by, **Kansal et al. (2019)** [1] and **Wu et al. (2020)** [3].Different clustering methods like hierarchical clustering, DBSCAN, and GMMs can provide advantages depending on data nature and business needs, such as detecting outliers and handling non-convex shapes.Hierarchical clustering and GMMs offer versatile cluster shapes and scalability as alternative solutions.

**V. General architecture diagram**

The architecture starts with data collection from customer transactions and behavior. RFM analysis (Recency, Frequency, Monetary) is performed to score each customer. These RFM values are then used for K-Means clustering to segment customers into distinct groups. A recommendation engine provides product suggestions tailored to each segment based on their purchasing patterns. Finally, a feedback loop continuously refines the model for improved segmentation and recommendations over time**.** Additionally, continuous feedback loops and performance monitoring ensure the system adapts to evolving customer patterns, driving improved engagement and sales over time

The E-commerce Customer Segmentation and Recommendation System uses customer data from various sources, including transaction records, web behaviour, and CRM systems. The data is cleaned through standardization, outlier detection, and handling of null values to ensure quality. RFM analysis is performed, scoring each customer based on recent purchases, frequency, and spending. The Elbow Method and Silhouette Method are used to determine the optimal number of clusters for segmentation, selecting the best number of distinct customer segments. K-Means clustering is then applied to divide customers into segments like loyal, at-risk, or potential loyalists based on their RFM scores. These segments provide valuable insights into customer behaviour, guiding the model training phase.

The machine learning model is trained using segmentation results to predict future purchasing behavior. It is integrated into a Recommendation Engine, which generates personalized product suggestions for each customer segment. The system adapts based on customer engagement and purchasing patterns, ensuring relevance and timely suggestions. A continuous feedback loop refines both segmentation and recommendation models over time. Customer responses are monitored to fine-tune the system, ensuring it adapts to evolving customer behaviour. Performance monitoring ensures the system remains effective in driving customer engagement and increasing sales. This dynamic approach to customer segmentation and personalized recommendations fosters customer loyalty and business growth.



**Figure 1 Architecture Diagram**

**VI.PROPOSED SYSTEM**

The proposed system introduces an advanced customer segmentation solution utilizing K-Means clustering, designed to optimize marketing strategies for e-commerce businesses. This system enhances the way businesses interact with their customers by identifying distinct groups based on shared behaviors and characteristics. By leveraging clustering techniques, the proposed system ensures more personalized marketing efforts, resulting in improved customer engagement and satisfaction.

There are three major sections:

• Data Pre-processing.

• RFM score calculation.

• Cluster Creation.

By bringing in the requisite libraries listed below:

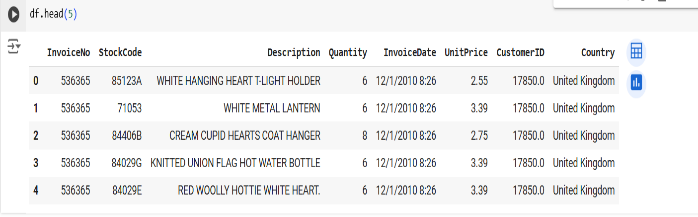
**Pandas –** Loading & Pre-processing of data.

**NumPy** **–** numeric data calculation.

**Seaborn, matplotlib** **–** Visualization.

**Scikit-**learn Machine Learning library

1. **Data Pre-Processing:** Involves preparing the dataset for analysis by standardizing features, handling missing values, and reducing dimensionality. This step ensures that the data is clean, well-scaled, and optimized for the clustering process, improving the accuracy and efficiency of the model.



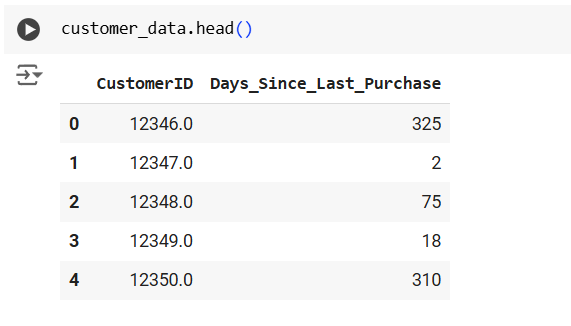
**Figure 2 features of the dataset**

**B.RFM SCORE**

RFM is a technique for dividing up the client base and evaluating customer value. It is an acronym for the following:

**Recency (R):** This measure shows how recent a customer's purchase was. A lower recency value denotes a more recent purchase, which suggests a higher level of brand engagement from the customer.

Days Since Last Purchas- This feature represents the number of days that have passed since the customer's last purchase.

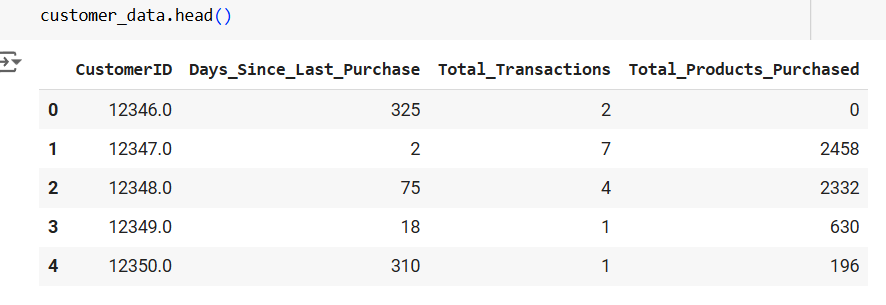


**Figure 3 Features of Recency (R)**

**Frequency (F):** This measure shows how frequently a consumer buys something within a given time frame. A customer with a higher frequency value is one who engages with the company more frequently, which may indicate greater satisfaction

Total Transactions**-**This feature represents the total number of transactions made by a customer**.**

Total Products Purchased: This feature indicates the total number of products (sum of quantities) purchased by a across customers all transactions

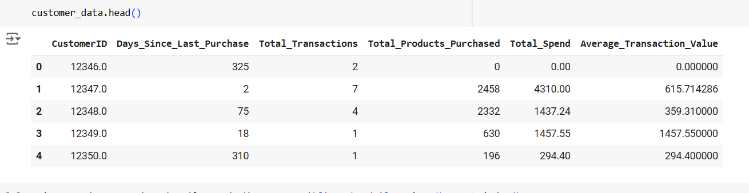


**Figure 4 Features of Frequency (F)**

**Monetary (M):** This measure shows how much money a client has spent overall over a specific time frame. Higher-valued customers have made more contributions to the company, suggesting that they may have a high lifetime value.

Total Spend: This feature represents the total amount of money spent by each customer

Average Transaction Value: This feature is calculated as the **Total Spend** divided by the **Total Transactions** for each customer.



**Figure 5 Features of Monetary (M)**

**C. Clustering**

K-Means is an unsupervised machine learning algorithm that works well with complex datasets for clustering applications. The dataset is divided into "K" non- overlapping, pre-defined clusters, with each data point belonging to a single group. By minimizing the within-cluster sum-of-squares (WCSS), sometimes referred to as inertia, the algorithm operates. Every time, data points are assigned to the closest centroid, and the centroids are recalculated using the average of the assigned points. This iterative process ensures effective and significant data clustering by continuing until convergence or a predetermined stopping criterion is satisfied.

**Steps to Implement K-means Clustering**:

**Try Various K Values:** To figure out which K value (the number of clusters) best fits your data, experiment with a range of values.

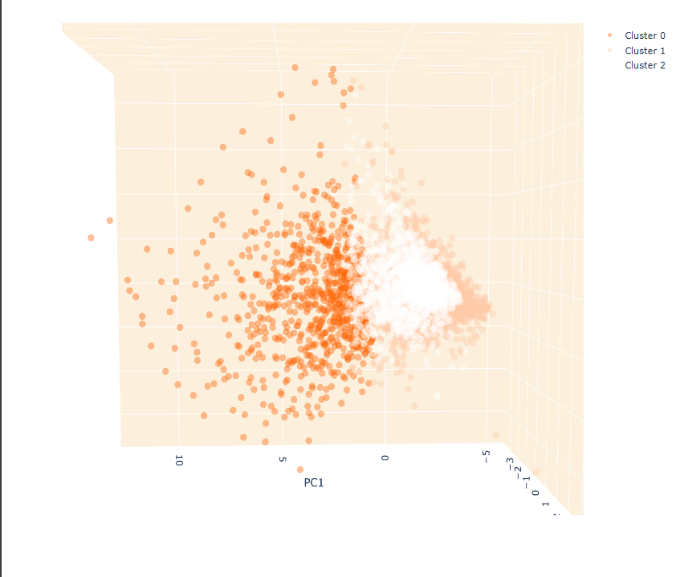
**Initialize K Centroids**: Randomly select K points from the dataset to serve as the initial cluster centroids.

**Assign Data Points to the Nearest Centroid**:

Calculate the distance between each data point and each centroid, then assign each data point to the closest centroid to form K clusters.

**Update Centroids**: Determine the average of all the points allocated to every centroid, then realign the centroid to reflect the updated mean position.

**Repeat:** Continue with steps 3 and 4 until there is no more significant change or convergence in the centroids.

****

**Figure 6 Clustering in 3D space**

**VII. Determine the Optimal K Using the Elbow Method**:

Plot, for a range of K values, the sum of squared distances between data points and the corresponding cluster centroids (inertia). The ideal value of K is suggested by the point where the rate of decrease slows down and forms a "elbow".

**Elbow method:**

The elbow method is a popular technique used to determine the optimal number of clusters (k) in K-means clustering. The idea behind the method is to run the K-means algorithm for different values of k and calculate the within-cluster sum of squares (WCSS), also known as inertia, for each k

**Steps:**

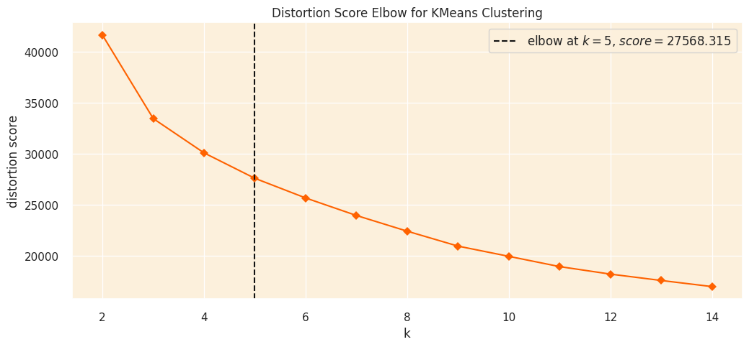
**Run K-means:** The K-means algorithm is used to fit a range of values, such as 1 to 10, into a dataset**.**

**Calculate WCSS:** The WCSS is calculated for each k points, calculating the sum of squared distances between each point and the centroid of its assigned cluster.

**Plot the graph**: The graph displays the number of clusters 𝑘 on the x-axis and the WCSS on the y-axis.

**Look for the” elbow”:** As k increases, the WCSS decreases, but the marginal gain diminishes, forming a noticeable bend in the graph.

**Optimal k:** The optimal number of clusters is determined by the point where the "elbow" occurs, which balances WCSS minimization with model complexity by fewer clusters.

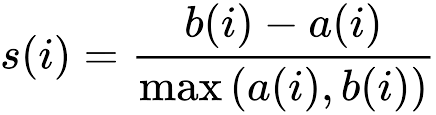


**Figure 8 Elbow Curve**

**VIII. Silhouette Method:**

The Silhouette Method is a statistical technique used to determine the optimal number of clusters in a dataset by assessing the consistency and separation of clusters within each data point, thereby determining its similarity to other clusters.

**Determining the silhouette coefficient for a point:  
• Calculate a(i):** Average distance between point i and other cluster points.  
**• Calculate b(i):** Average distance between point i and nearest cluster points.  
**• Calculate s(i):** using a formula.



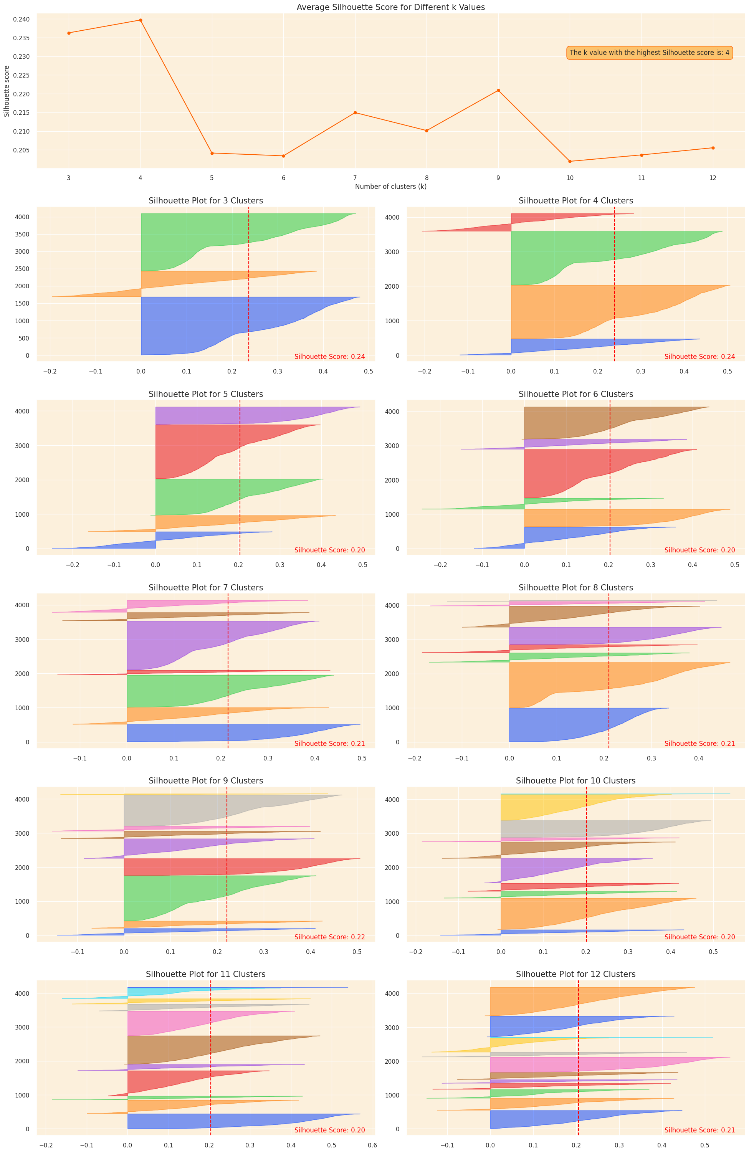
The Silhouette Method is a clustering technique that assesses cluster quality by considering cohesion and separation from other clusters, providing a more comprehensive measure of clustering performance than the Elbow Method, which only considers inertia. It produces a silhouette score, making it easier to compare different values of k. The Silhouette Method also generates a visual representation of silhouette coefficients for each data point, enabling easier identification of fluctuations and outliers within clusters, thereby determining the optimal number of clusters with higher confidence.

**XI. Elbow vs Silhouette**

The Elbow Method and Silhouette Method are crucial techniques in determining the optimal number of clusters when applying K-Means clustering in customer segmentation projects, such as e-commerce platforms. The Elbow Method is based on the concept of inertia, which is the sum of squared distances between each data point and the centroid of its assigned cluster. It involves running the K-Means algorithm for a range of cluster numbers and plotting the inertia as a function of the number of clusters. The key idea is to look for the "elbow" point on this plot, where the decrease in inertia becomes significantly smaller, indicating diminishing returns from adding more clusters. This point represents a balance between minimizing within-cluster variance and avoiding overfitting with too many clusters.

The Silhouette Method evaluates the quality of the clustering by measuring how similar each data point is to its own cluster compared to other clusters. The goal is to maximize the overall silhouette score for the dataset, indicating that the chosen number of clusters results in distinct, well-separated groups. In the context of customer segmentation, this method ensures that customer groups are distinct from each other, making it easier to design targeted marketing strategies for each segment.

Both methods can be complementary in a project, providing a more straightforward approach that helps identify a point of diminishing returns.



**Figure 9 Silhouette Score**

**X. Conclusion**

The output for customer segmentation and product recommendation using \*K-Means, \*\*RFM analysis, and the \*\*Elbow and Silhouette methods\* involves first calculating RFM scores for each customer to represent their purchase behavior. The \*Elbow method\* is used to determine the optimal number of clusters by finding the point where adding more

clusters yields diminishing returns, while the \*Silhouette method\* ensures the clusters are well-separated and distinct. After determining the optimal clusters, \*K-Means clustering\* groups customers into segments like high-value or dormant customers. Tailored product recommendations are then generated for each segment, such as premium offers for frequent buyers or discounts for low-engagement customers, enabling businesses to create personalized marketing strategies

The e-commerce customer segmentation and recommendation system could be improved by using advanced techniques like deep learning-based clustering, real-time data processing, AI-powered personalized marketing strategies, multi-channel customer data, and reinforcement learning. These techniques could enhance customer identification, provide dynamic segmentation updates, refine product recommendations, and optimize recommendations based on customer feedback and engagement.

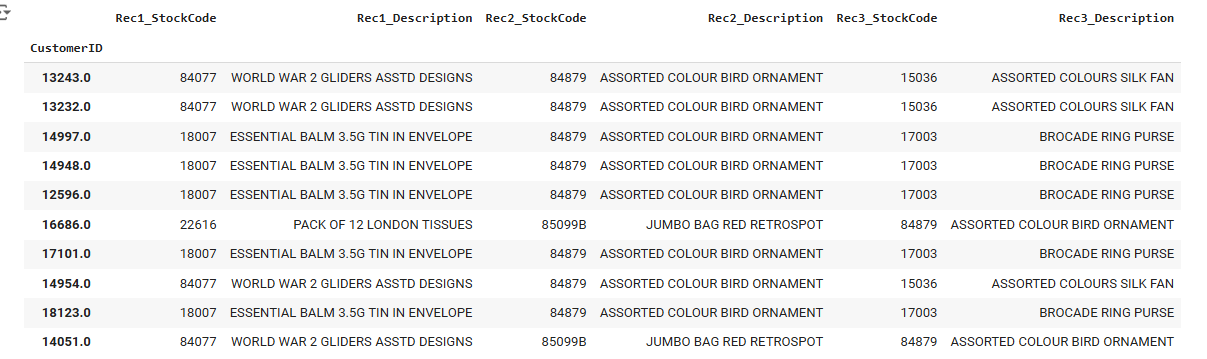


Figure 9 Output

**Reference**

[1]Tushar Kansal; Suraj Bahuguna; Vishal Singh; Tanupriya Choudhury, “Customer Segmentation using K-means Clustering,” IEEE, Jul. 2019.

[2]K. Maheswari, “Finding Best Possible Number of Clusters using KMeans

Algorithm,” International Journal of Engineering and

Advanced Technology (IJEAT), vol. 9, no. 1S4, Dec. 2019

[3] Wu, J., Shi, L., Lin, W. P., Tsai, S. B., Li, Y., Yang, L., & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviours Using a RFM Model and K -Means Algorithm. Mathematical Problems in Engineering, 2020

[4]Ashutosh Bhardwaj, “Silhouette Coefficient Validating clustering

techniques,” Towards Data Science, May 26, 2020.

[5] Dhiraj Kumar, “Implementing Customer Segmentation Using

Machine Learning”, July 10th 2021.

[6]V.Vijilesh, “CUSTOMER SEGMENTATION USING MACHINE

LEARNING,” International Research Journal of Engineering and

Technology (IRJET), vol. 08, no. 05, May 2021.

[7] Customer Segmentation. Segmentation by RFM clustering | by Baris Karaman Towards Data Science. (n.d.). Retrieved May 11, 2022,