"Unlocking Customer Insights: Machine Learning Analysis of Supermarket Sales Data for Targeted Marketing and Sales Optimization"

DSC 478
Super Market Sales

Members: Aniket Surve, Nachiketh Reddy, Utkarsha Deshkar

Dataset Link – https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales/data

Executive Summary: -

Our project revolves around the analysis of supermarket sales data to gain insights into customer behaviour. Utilizing machine learning methods, we aim to categorize customers based on their purchasing patterns and forecast the specific group to which they belong. The dataset, sourced from Kaggle, encompasses comprehensive information on supermarket sales, encompassing details such as date, branch, customer profiles, product categories, and sales figures. This dataset serves as a valuable resource for enhancing our understanding of customer preferences, subsequently informing and optimizing marketing and sales strategies.

Methods:

Data Preprocessing, Exploratory Data Analysis(EDA), Correlation Matrix, Multicollinearity, K-means Clustering, PCA.

Data Analysis:

Correlation Matrix:

The dataset comprises various features, including 'cogs' (Cost of Goods Sold), 'Total' (Total sales amount), 'Rating' (Customer rating), 'gross income' (Gross income from sales), 'Tax 5%' (Sales tax amount), 'Quantity' (Number of items purchased), and 'Unit price' (Price of a single unit of a product). Notably, there exists a robust positive correlation (0.63 to 1.00) among 'cogs,' 'Total,' 'Rating,' 'gross income,' 'Tax 5%,' 'Quantity,' and 'Unit price.' However, 'Quantity' and 'Unit price' exhibit a weaker negative correlation (-0.01). This information is crucial for understanding how these features are interconnected, providing insights into potential patterns and dependencies within the supermarket sales dataset. Due to multicollinearity among the columns: 'cogs', 'gross income', 'Tax 5%' and ‘total’ we are removing columns_to_remove = ['cogs', 'gross income', 'Tax 5%'].

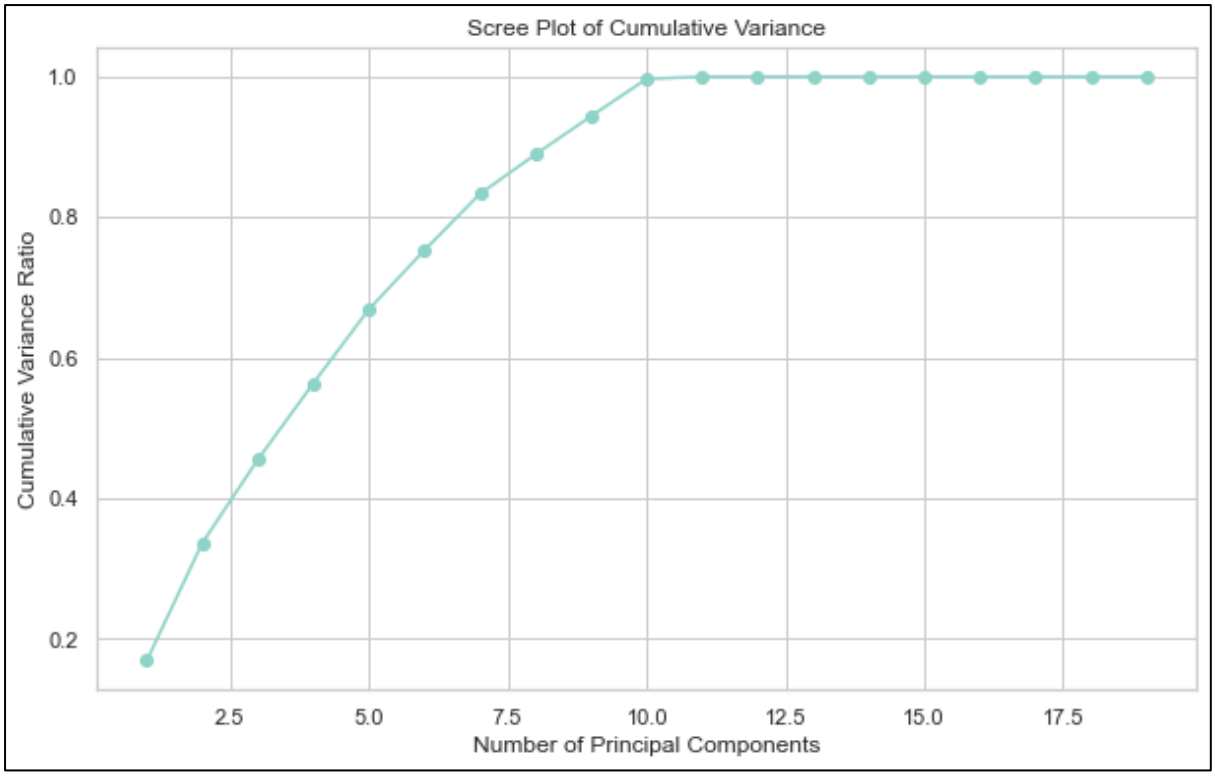
STANDARDIZATION:

Standardizing columns ‘Unit Price’, ‘Total’ and ‘Quantity’.

	Unit price	Quantity	Total
count	1.000000e+03	1.000000e+03	1.000000e+03
mean	-1.187939e-16	5.562217e-17	-2.420286e-17
std	1.000500e+00	1.000500e+00	1.000500e+00
min	-1.721668e+00	-1.543480e+00	-1.270692e+00
25%	-8.608740e-01	-8.590099e-01	-8.078714e-01
50%	-1.669588e-02	-1.745399e-01	-2.812422e-01
75%	8.406991e-01	8.521652e-01	6.037682e-01
max	1.672416e+00	1.536635e+00	2.928371e+00

PCA:

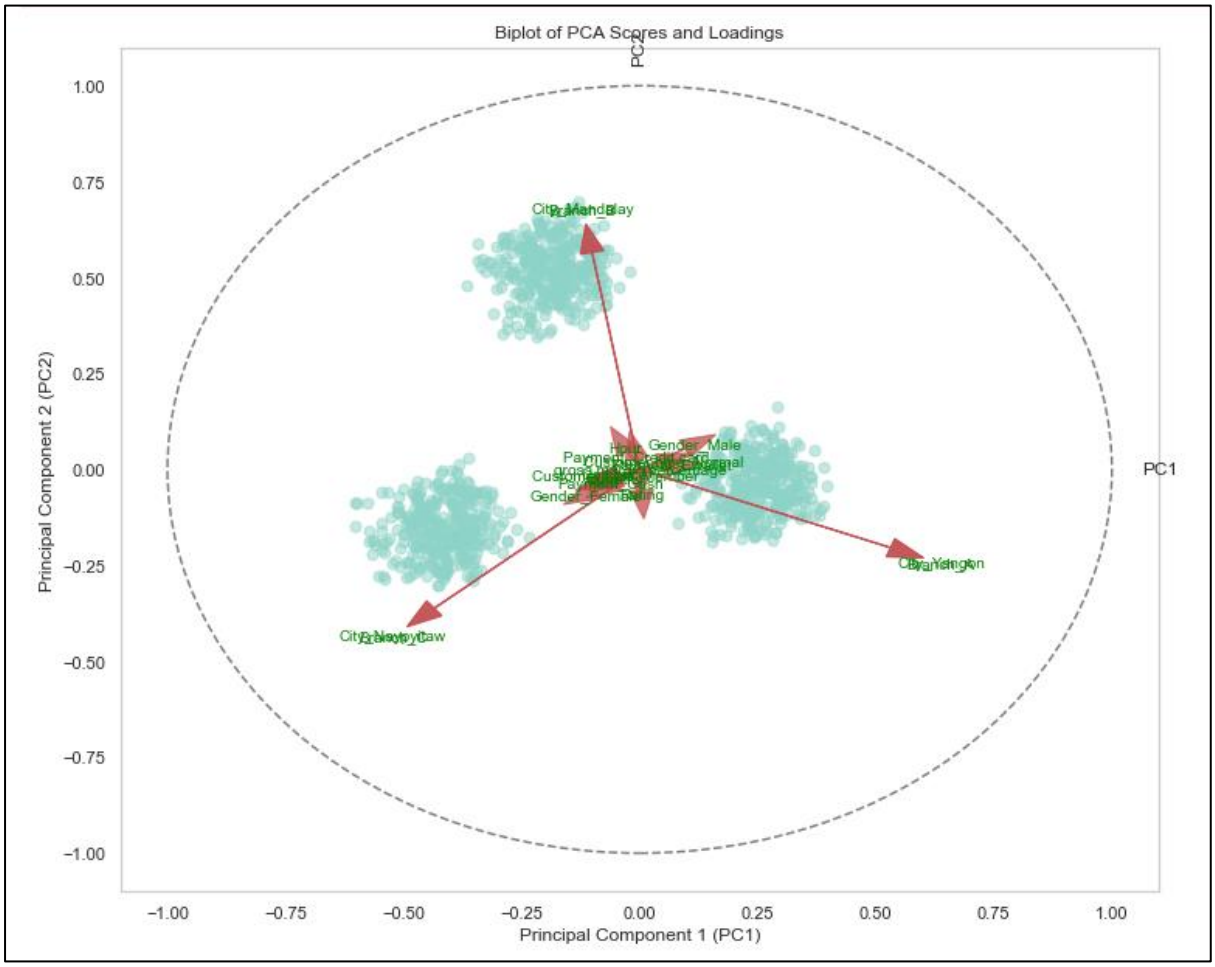
The output reveals a comprehensive representation of supermarket transactions for categorical variables. Product-related information, such as 'Product line,' 'Unit price,' 'Quantity,' and 'Total,' offers insights into the type, pricing, and quantity of items purchased, along with the total sales amount. Sales metrics, including 'gross margin percentage,' 'Rating,' and 'Hour,' provide additional context, detailing profit margins, customer ratings, and the timing of purchases. Branch, city, customer type, gender, and payment method details are encoded into binary columns, offering a structured numerical format for machine learning models. This preprocessing step enhances the interpretability and utilization of categorical information, enabling models to analyze and predict diverse aspects of each transaction. The resulting DataFrame encapsulates a wealth of information, combining both numerical and encoded binary columns for a holistic representation of supermarket sales data.



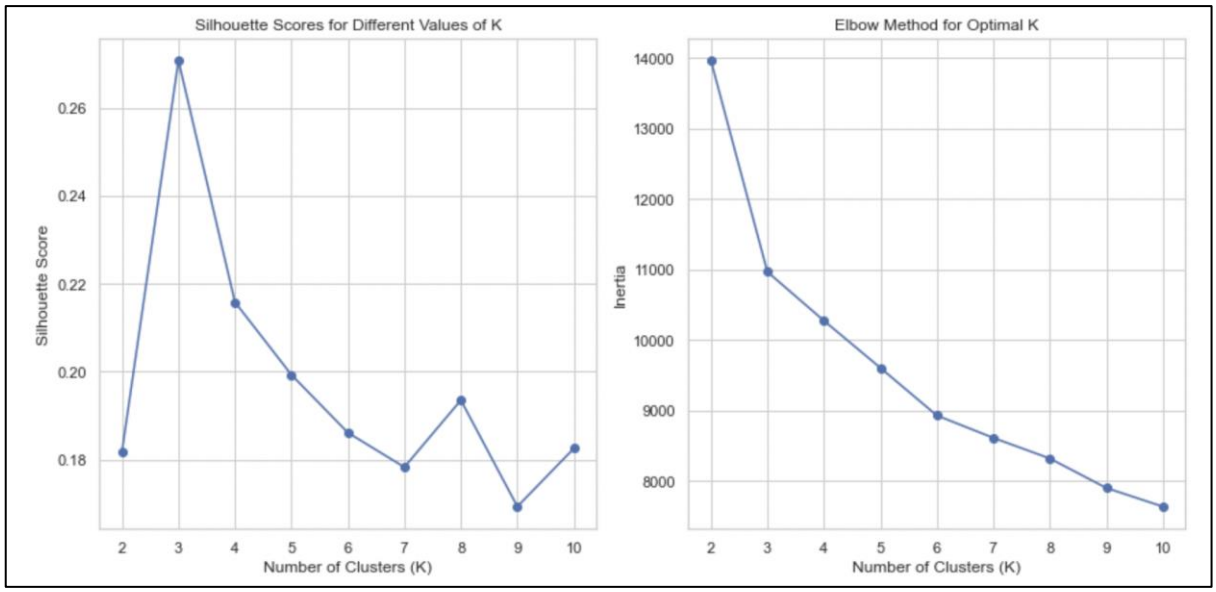
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	Product line
0	1.975556	-1.124873	-2.267212	0.304444	0.444272	-1.341706	-1.296979	1.398897	-0.228474	Health and beauty
1	-1.892746	-1.788977	0.886484	1.084775	-2.507078	1.349855	-0.329713	1.653474	1.159637	Electronic accessories
2	2.500356	-0.709763	1.021953	0.445429	0.536167	0.009330	2.227657	0.361821	0.664332	Home and lifestyle
3	2.331005	-0.777762	-0.267207	-1.139855	2.064104	-1.186608	-0.979353	-0.110550	0.322128	Health and beauty
4	2.502892	-0.780743	0.928978	1.562180	2.035795	-0.932981	-0.546534	0.223042	-0.077079	Sports and travel

Each row in the dataset corresponds to a unique observation, and the values in columns 'PC1' to 'PC9' indicate the impact of original features on these principal components. The 'Product line' column categorically identifies the type of product for each observation, providing insights into how principal components relate to different product categories. Positive and negative values in the principal component columns represent the direction and strength of the contribution of original features. For example, a positive PC1 value of 1.975556 indicates a significant positive influence of original features, with the corresponding 'Product line' being 'Health and beauty.' This condensed representation facilitates understanding relationships between principal components and product categories in the dataset.

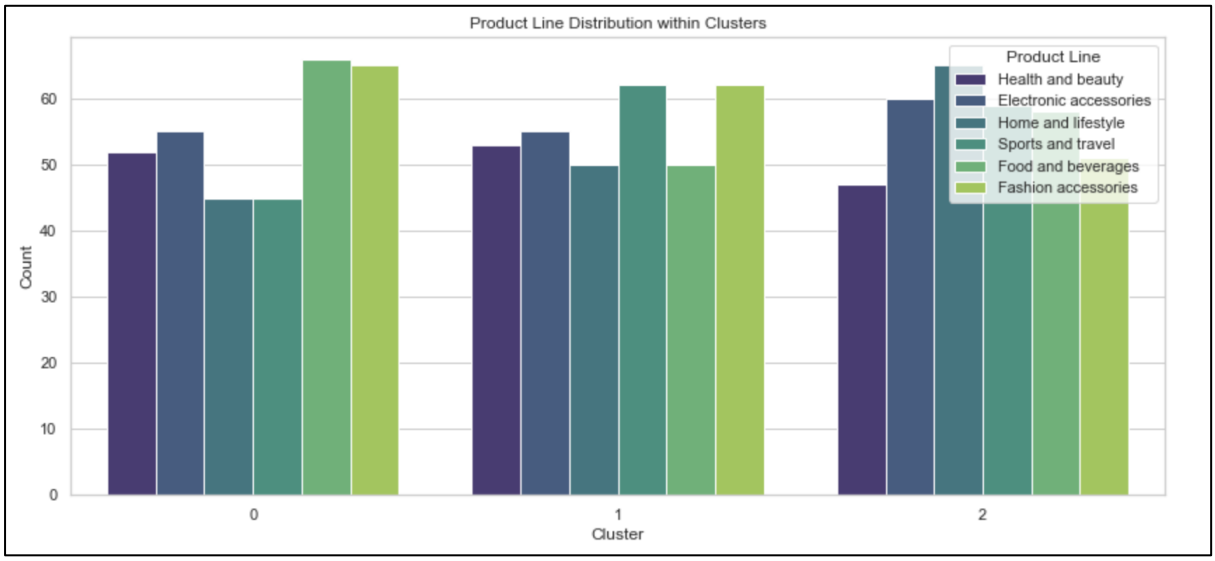
PCA Biplot shows 3 distinct groupings:

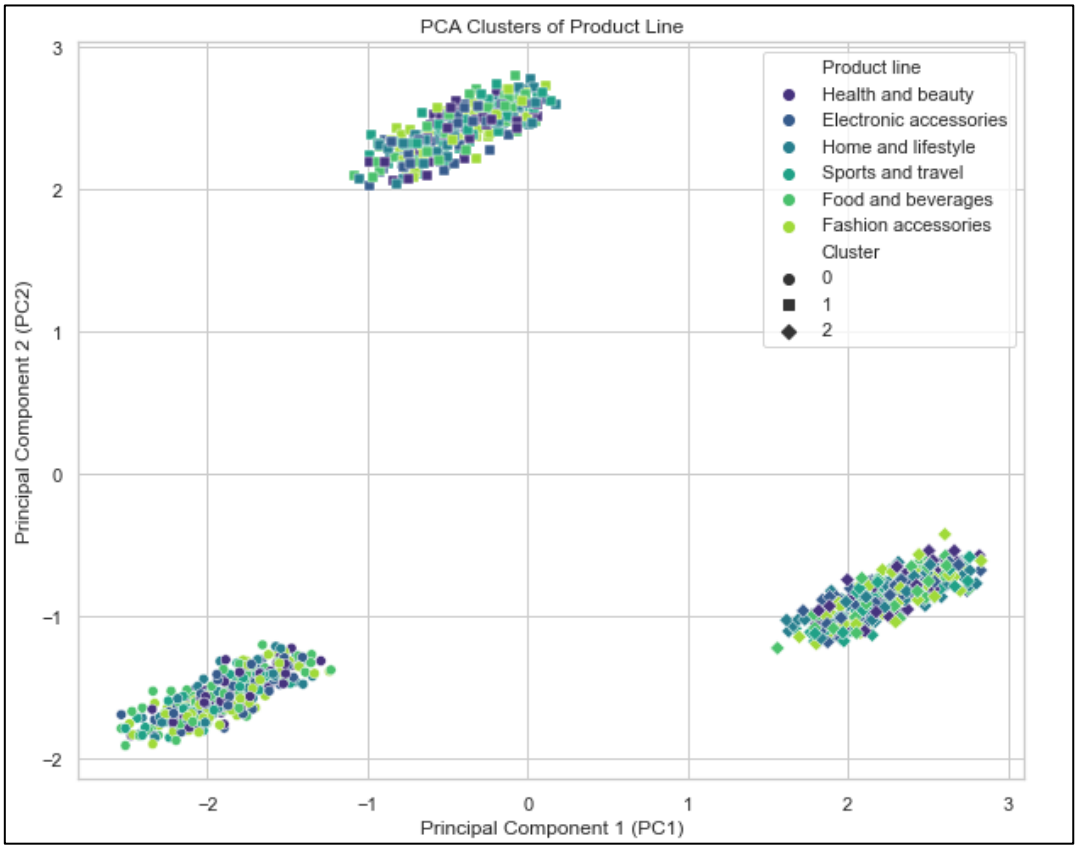


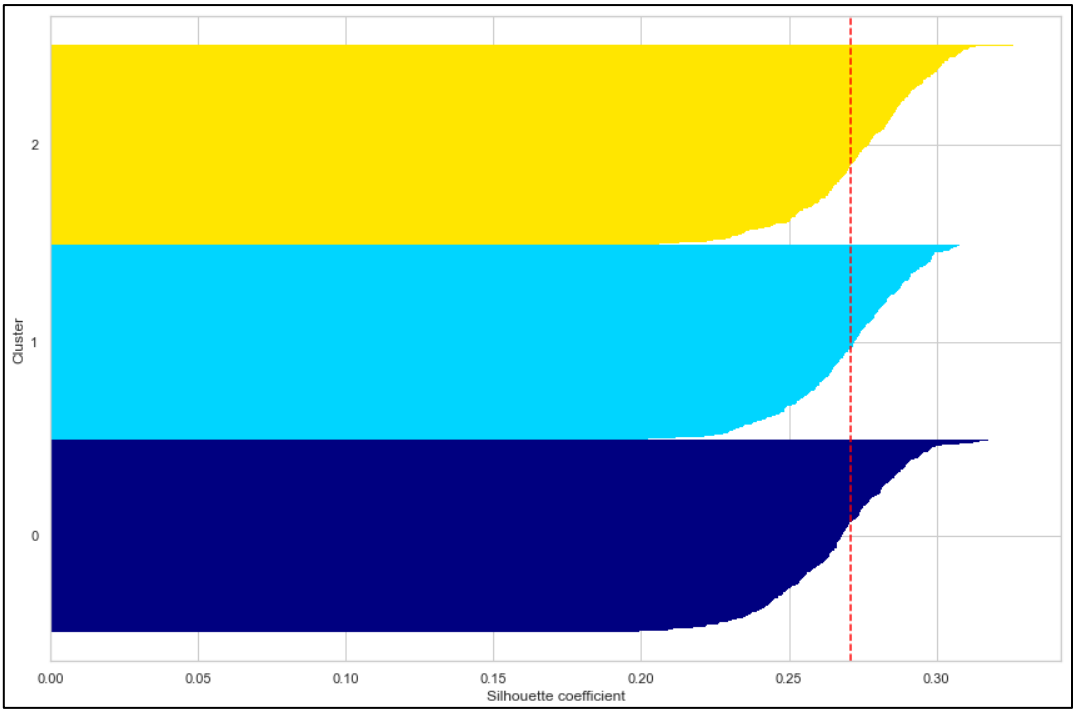
K-means Clustering:



Upon analyzing the Inertia and Silhouette Score, it's evident that the ideal cluster count (K) stands at three. As the number of clusters increases, Inertia declines, underscoring the need for a balanced approach. The Silhouette Score, pivotal for assessing cluster distinctness, peaks at K=3, signifying well-defined cluster boundaries. Hence, we strongly advocate employing K=3 in the K-Means algorithm for this dataset.







	Product Line	Cluster	PC1 Score	PC2 Score	PC3 Score	PC4 Score
0	Health and beauty	0	-1.830129	-1.522004	0.586099	-0.234588
1	Health and beauty	1	-0.411327	2.41419	0.003203	0.098579
2	Health and beauty	2	2.276398	-0.831217	-0.056881	-0.029805
3	Electronic accessories	0	-1.882849	-1.557648	0.60313	0.284783
4	Electronic accessories	1	-0.453767	2.402473	-0.072115	0.071589
5	Electronic accessories	2	2.256	-0.84724	-0.261621	-0.038846
6	Home and lifestyle	0	-1.895122	-1.558942	0.16858	-0.408942
7	Home and lifestyle	1	-0.392521	2.428948	0.078466	0.262339
8	Home and lifestyle	2	2.230242	-0.876696	-0.356261	0.126446
9	Sports and travel	0	-1.952728	-1.583925	-0.130751	-0.226029
10	Sports and travel	1	-0.447571	2.415609	-0.071641	0.110344
11	Sports and travel	2	2.241991	-0.874528	-0.278666	0.138358
12	Food and beverages	0	-1.939142	-1.572835	0.204786	-0.083303
13	Food and beverages	1	-0.465119	2.415129	-0.402416	-0.134191
14	Food and beverages	2	2.281366	-0.835555	-0.094782	-0.107977
15	Fashion accessories	0	-1.88188	-1.564125	0.394328	-0.089944
16	Fashion accessories	1	-0.436211	2.426771	-0.055397	-0.090202
17	Fashion accessories	2	2.24908	-0.859887	-0.258081	0.259754

	PC5 Score	PC6 Score	PC7 Score	PC8 Score	PC9 Score
0	0.095163	-0.218075	0.211091	-0.000752	-0.001147
1	0.373521	0.031603	-0.004211	0.306975	0.128507
2	-0.122156	0.139572	0.019784	-0.236387	0.086674
3	0.07519	0.243477	-0.125983	-0.167846	0.17622
4	-0.146333	0.237323	-0.130613	0.05204	0.19093
5	-0.004493	0.031976	0.100519	-0.077408	-0.041161
6	-0.061103	-0.228534	0.022629	0.007294	0.044254
7	0.142481	-0.186807	-0.088586	0.070714	0.170764
8	0.108478	0.102091	-0.193217	-0.010633	0.039616
9	-0.075942	-0.23333	0.243815	0.022078	0.10201
10	-0.007455	0.251933	-0.114875	-0.032515	-0.13648
11	0.043194	0.053001	-0.071293	0.094469	0.004315
12	0.063778	0.088237	-0.144303	-0.150339	-0.046159
13	-0.195728	-0.333507	0.281348	0.089507	-0.079813
14	0.088756	0.091872	0.178153	0.015691	-0.04694
15	0.016887	-0.166182	0.053239	0.173822	-0.154027
16	-0.298687	0.046036	-0.136972	-0.098305	-0.170164
17	-0.138541	-0.138145	0.056485	-0.048964	-0.151273

K = 3

```
Silhouette Score (PCA): 0.2707593873949745
Completeness (PCA): 0.0052854856892948976
Homogeneity (PCA): 0.0032429316797622162
```

Silhouette Score (PCA): 0.2707 The Silhouette Score measures how well-defined the clusters are in the data. A score close to 1 indicates well-separated clusters. In this case, the score of 0.2707 suggests moderate separation, indicating that the clusters are not as distinct as desired.

Completeness (PCA): 0.0053 Completeness measures whether all data points that are members of the same ground truth cluster are also members of the same cluster in the predicted set. A low value (close to 0) suggests that the clusters do not capture all the information about the ground truth clusters.

Homogeneity (PCA): 0.0032 Homogeneity measures how much each cluster contains only data points that are members of a single class. Similar to completeness, a low value (close to 0) indicates that the clusters do not align well with the true classes.

The obtained scores suggest that the clustering based on Principal Component Analysis did not result in well-defined, internally homogeneous, or complete clusters. This may indicate challenges in capturing meaningful patterns in the data using the chosen clustering method or the need for further exploration of optimal clustering parameters. It's advisable to experiment with different clustering algorithms, adjust hyperparameters, or explore additional preprocessing steps to improve the clustering performance.

Report Analysis:

The analysis provides a robust framework for customer segmentation and targeted marketing strategies. By leveraging the identified clusters, businesses can tailor their approaches to meet the diverse preferences of different customer groups. Moreover, understanding principal component contributions offers valuable insights into factors influencing customer preferences, enabling more informed decision-making for product development and marketing endeavours.

Cluster Separation: Clusters 0, 1, and 2 show distinctive patterns in the scores across all principal components for each product line. Health and beauty, Electronic accessories, Home and lifestyle, Sports and travel, Food and beverages, and Fashion accessories all have different positions within each cluster.

Principal Component Contributions: PC1 and PC2 have significant contributions to the separation of clusters.

Negative scores in PC1 and PC2 for Cluster 0 suggest a commonality among product lines in this cluster, while positive scores in PC1 and PC2 for Cluster 2 indicate a different set of preferences.

PC3 to PC9 contribute to additional nuances in the clustering, capturing more detailed variations in product line preferences.

Product Line Analysis: Health and beauty in Cluster 0 has negative scores in PC1 and PC2, suggesting a preference for products with lower scores in these components. Electronic accessories in Cluster 1 has positive scores in PC1 and PC2, indicating a preference for products with higher scores in these components. Home and lifestyle in Cluster 2 shows a mix of positive and negative scores across different principal components. Sports and travel, Food and beverages, and Fashion accessories exhibit similar patterns of differentiation across clusters.

Customer Segmentation: The clustering provides a basis for customer segmentation, grouping products and customers with similar preferences. Businesses can target marketing strategies based on the preferences identified in each cluster, tailoring their approach to meet the specific needs of different customer groups. Understanding the principal component contributions helps in interpreting the factors that influence customer preferences.

Recommendations for Business Strategy: Businesses can use this information to optimize product recommendations for each customer segment. Tailoring marketing campaigns and promotions based on the identified patterns can enhance customer engagement and satisfaction. Further analysis of specific product features corresponding to each principal component can guide product development and inventory management.