Marketing Campaign Customer Segmentation Analysis and Report.

Data-Driven Insights and Recommendations.

February 23th, 2025.

Contents / Agenda

Executive Summary

In today's competitive marketplace, businesses face the challenge of not only attracting but also retaining customers. With a diverse customer base, each with unique needs and preferences, generic marketing strategies are often inefficient and fail to engage customers effectively. Marketing represents a major expense for businesses, and inefficient marketing can result in considerable waste of resources. By employing customer segmentation, companies can focus their marketing efforts on the most receptive segments, which significantly reduces costs and enhances the effectiveness of their campaigns. This understanding and addressing of specific customer needs allows businesses to craft personalized strategies that resonate well with each segment, thereby improving customer loyalty and increasing the return on marketing expenditures.

Solution Design

Marketing segmentation offers the advantage of predicting future buying behaviors and preferences, which is instrumental in guiding product development and inventory management. This approach also facilitates dynamic adaptation, enabling businesses to swiftly adjust their marketing strategies in response to changes in customer behavior or market conditions. The goal of this segmentation analysis is to apply advanced unsupervised learning techniques such as K-Means, K-medoids, DBSCAN, Hierarchical Clustering, and Gaussian Mixture Models, to extract relevant data-driven insights and help the business leverage customer segmentation to craft personalized and impactful marketing strategies. In this report, we will attempt to answer the following relevant questions: How can customers be segmented based on their purchasing behavior and demographic characteristics? What are the characteristics of each customer segment? How can the marketing strategies be improved based on these segments?

Data Overview

The dataset structure is composed of 2240 entries with 27 customer attributes with demographic information that includes factors such as age, income, education, marital status, and household size provide insights into the customer's background. Additionally, the dataset also provides transactional data that includes details about purchases across different product categories (like wines, meats, fruits), total spending, and frequency of purchases. The dataset also provides engagement metrics, particularly data on how customers interact with marketing campaigns, such as responses to promotions, participation in surveys, or clicks on email campaigns.

Summary Statistics

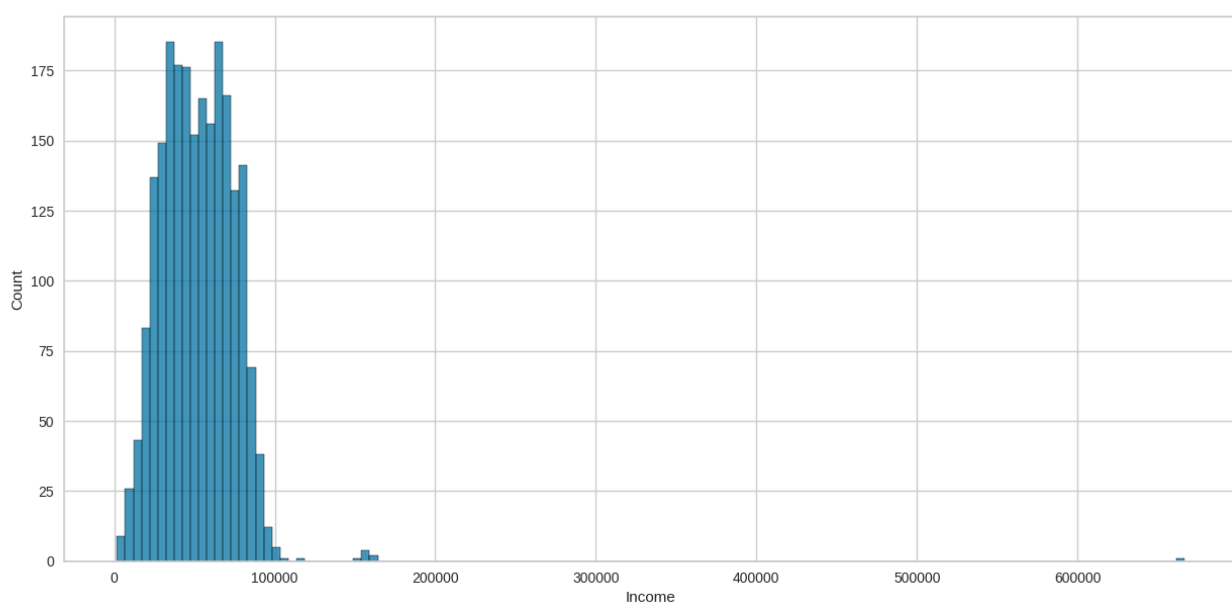
Column	Unique	Top	Freq	Mean	Std	Min	Median	Max
ID	-	-	-	5592.16	3246.66	0.0	5458.5	11191.0
Year_Birth	-	-	-	1968.81	11.98	1893	1970.0	1996
Education	5	Graduation	1127	-	-	-	-	-
Marital_Status	8	Married	864	-	-	-	-	-
Income	-	-	-	52247.25	25173.08	1730	51381.5	666666.0
Kidhome	-	-	-	0.44	0.54	0	0.0	2
Teenhome	-	-	-	0.51	0.54	0	0.0	2
Dt_Customer	663	31-08-2012	12	-	-	-	-	-
Recency	-	-	-	49.11	28.96	0	49.0	99
MntWines	-	-	-	303.94	336.60	0	173.5	1493
MntFruits	-	-	-	26.30	39.77	0	8.0	199
MntMeatProducts	-	-	-	166.95	225.72	0	67.0	1725
MntFishProducts	-	-	-	37.53	54.63	0	12.0	259
MntSweetProducts	-	-	-	27.06	41.28	0	8.0	263
MntGoldProds	-	-	-	44.02	52.17	0	24.0	362
NumDealsPurchases	-	-	-	2.33	1.93	0	2.0	15
NumWebPurchases	-	-	-	4.08	2.78	0	4.0	27
NumCatalogPurchases	-	-	-	2.66	2.92	0	2.0	28
NumStorePurchases	-	-	-	5.79	3.25	0	5.0	13
NumWebVisitsMonth	-	-	-	5.32	2.43	0	6.0	20
AcceptedCmp1-5, Response	-	-	-	-	-	-	-	-
Complain	-	-	-	0.01	0.10	0	0.0	1

Data Analysis and Key Insights

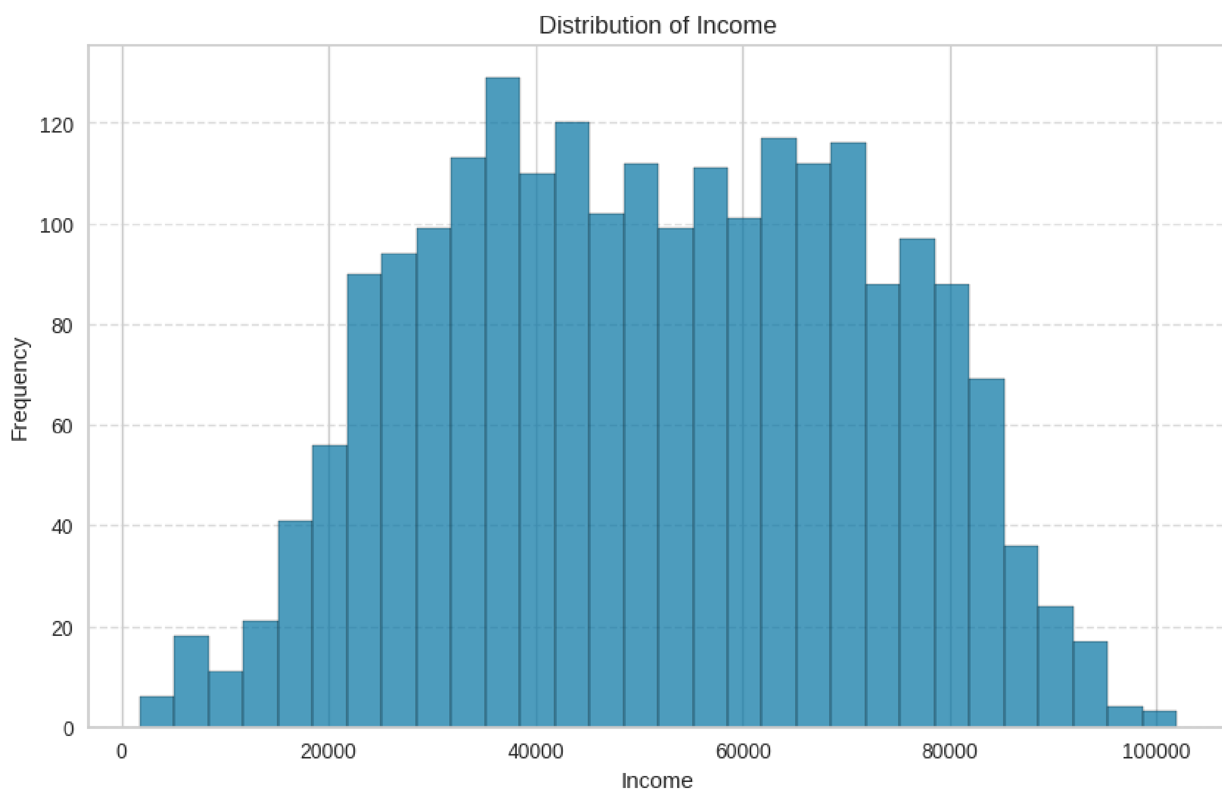
Univariate Analysis

The dataset provides a comprehensive view of customer demographics and purchasing behaviors. The customers' birth years span over a century, from 1893 to 1996, highlighting a wide age range. Educational backgrounds vary significantly, encompassing levels from Graduation to PhD, which may influence purchasing patterns and marketing strategies. Marital status is categorized into groups such as Single, Married, and Together, offering insights into different household dynamics. Household incomes range broadly from $1,730 to $666,666, with a median income of $51,381.5 as shown in the histogram below, suggesting diverse economic statuses.

Income Distribution Histogram



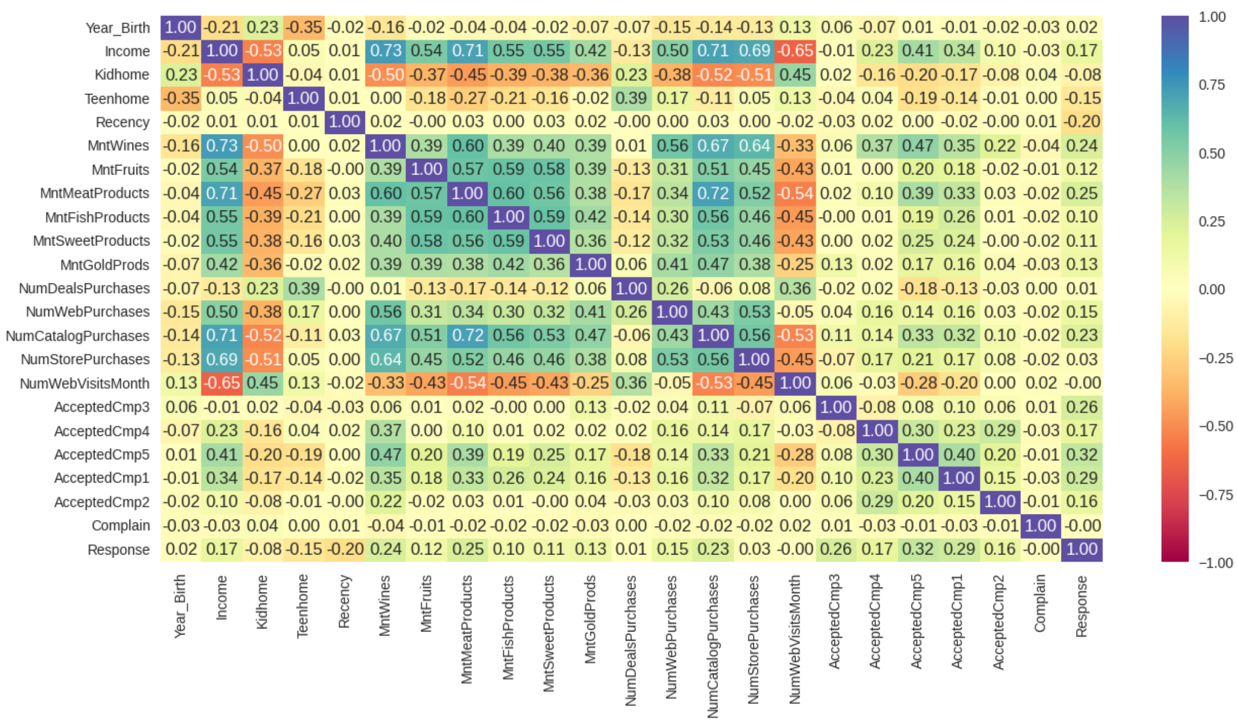
There are some outliers in the data. For instance, the income distribution histogram, as well as its boxplot equivalent, clearly highlight the presence of extreme values on the right end of the distribution. These outliers significantly exceed the general range of the data and must be handled properly. Since these outliers represent a vary small number of customers, they should be dropped from the analysis for the purpose of this report. After dropping the outliers from the income distribution, we get a better picture of the income distribution of the company's customers.



Bivariate Analysis

For this portion of the analysis, the dataset reveales several key patterns essential for understanding customer behavior and tailoring marketing strategies. A prominent pattern is the relationship between income and spending, where higher income levels correlate with increased expenditure across various product categories. This suggests that income is a significant predictor of purchasing behavior, potentially guiding targeted marketing efforts. Marital status as well as education level also seem to be positively correlated with income. Married couples and individuals with higher degree levels (e.g. Master, and PhDs) appear to have higher incomes when compared to single, divorced, or widows. Additionally, the age distribution within the customer base shows that different age groups have distinct spending habits, which could influence marketing strategies tailored to specific demographics.

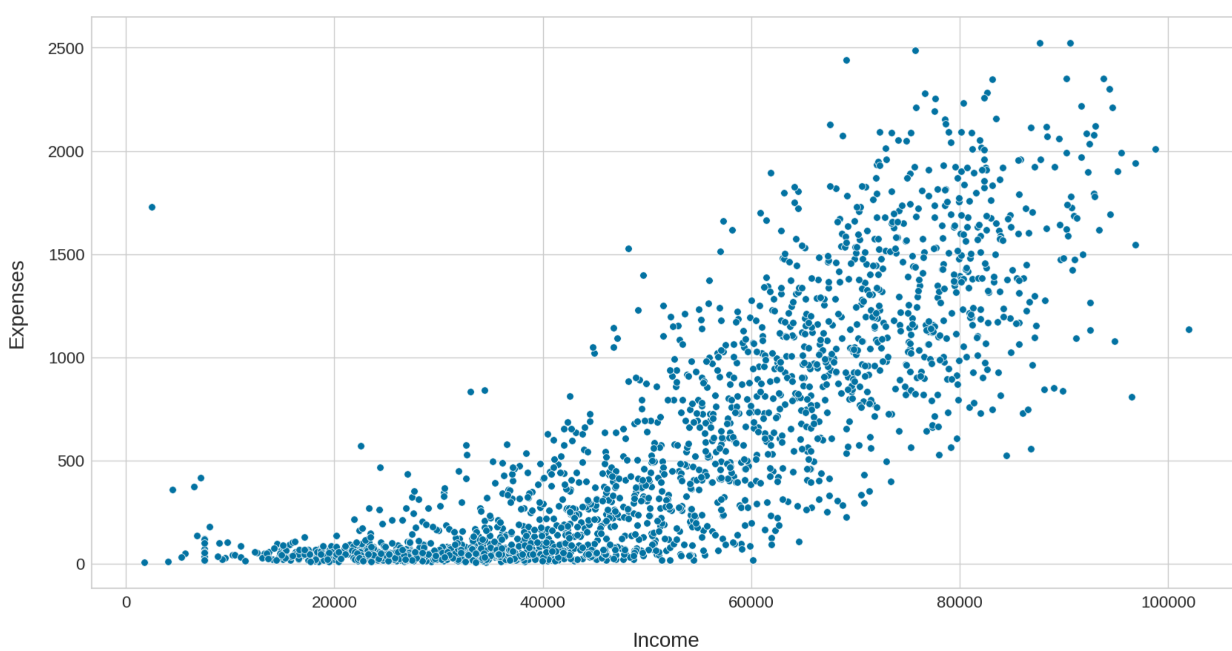
Variable Correlation Heatmap



The data also provides an overview of customer family compositions, with the number of children and teenagers in the household ranging from 0 to 2. There is a negative correlation between the number of children in the home (Kidhome) and household income, with a correlation coefficient of approximately -0.53. This indicates that as the number of children in the home increases, there tends to be a decrease in household income. Additionally, customer engagement is detailed through metrics such as the date of joining, ranging from 2012 to 2014, and recency of purchases, which varies from 0 to 99 days. Spending behavior is tracked across various categories with significant variations; for instance, expenditures on wine over the last two years range up to $1,493, while spending on fruits and meat products reaches up to $199 and $1,725, respectively. This rich dataset can be instrumental in tailoring marketing campaigns and understanding consumer needs across different segments.

When comparing income vs expenses, the data shows a positive correlation between income and expenses, suggesting that as income increases, expenses also tend to increase. A large number of data points are clustered at the lower income levels with relatively lower expenses. This suggests that a significant portion of the customer base has modest incomes and correspondingly modest spending habits. Most data points are concentrated below an income of $60,000 and expenses of $500, suggesting that the bulk of the customer base falls within this economic bracket as seen in the Income vs Expense scatterplot below.

Income vs Expenses Scatterplot



Another critical observation is the varied levels of engagement with marketing campaigns among different segments. Some customer segments are more responsive than others, indicating varying degrees of brand loyalty or interest in specific products. This insight is particularly valuable for optimizing marketing campaigns to enhance customer response and drive sales.

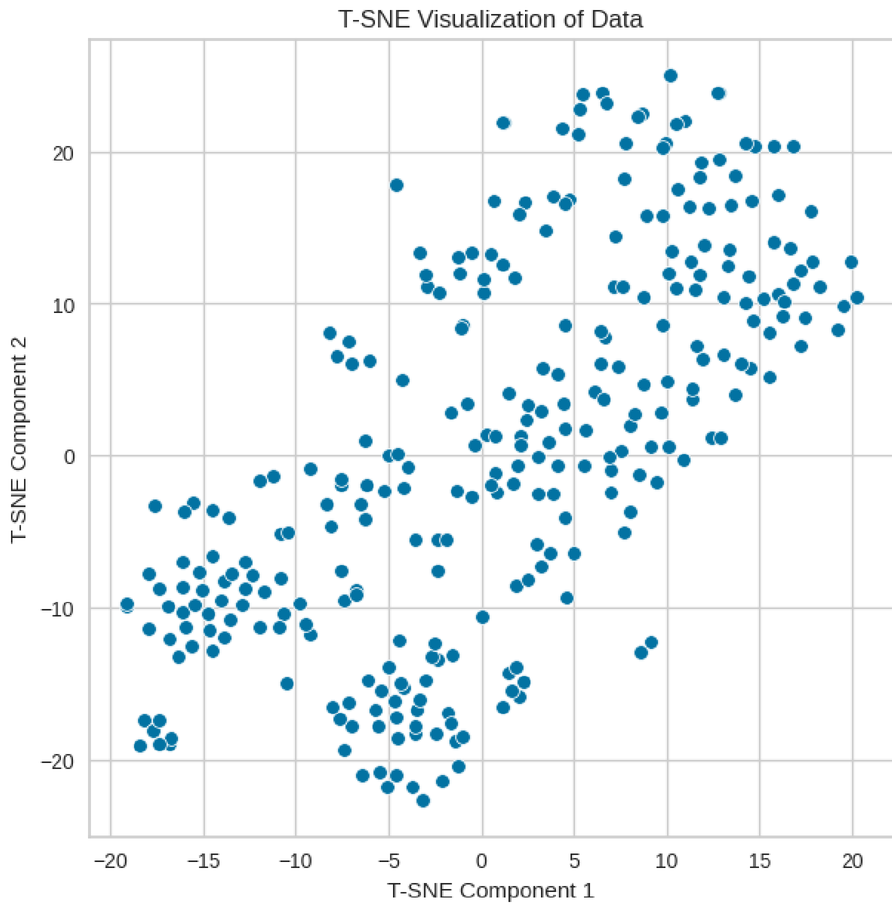
Scaling the Data

Prior to applying unsupervised learning models like K-Means, K-medoids, DBSCAN, Hierarchical Clustering, and Gaussian Mixture Models, the data must first be cleaned up and organized. To do this, the data was divided between numerical and categorical data. Numerical data includes variables such as income (about 1.071429% of total income), total spent, and number of purchases that were analyzed for mean, median, range, and standard deviation to understand their central tendency and dispersion. Categorical data includes Education level, marital status, and responses to marketing campaigns (accepted or declined offers) were summarized with frequency counts to identify the most common categories. There were some missing values (24 total) in the Income column. These values were addressed in the analysis by imputing with the median to help maintain the data's integrity without introducing bias.

Standardization was applied to numerical features by scaling the data to normalize such data which is essential for effective clustering and Principal Component Analysis (PCA). Dimensionally reduction techniques such as PCA were applied to the data to reduce the dimensionality of the dataset while retaining the characteristics that account for the most variance in the data. This was helpful in visualizing high-dimensional data and improved the performance of clustering algorithms. Moreover, t-Distributed Stochastic Neighbor Embedding (t-SNE) was used to explore data structure through t-SNE, which is particularly good at visualizing clusters or groups in data to further understand the data.

t-SNE Clustering

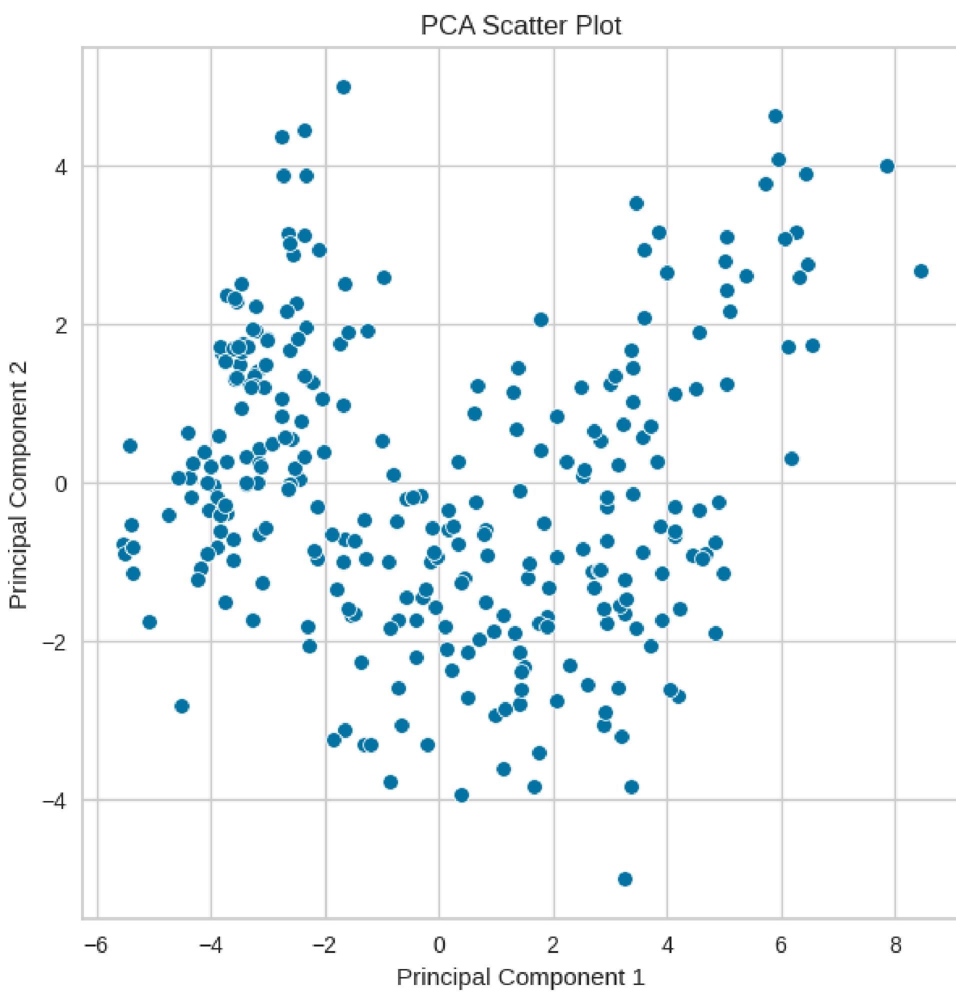
t-SNE was applied to the scaled data resulting in a scatterplot with visually distinct clusters as shown below.



The distinct clusters suggests that there are groups within the dataset that share similar characteristics. These clusters vary in density and separation, with some being densely packed, suggesting a high similarity within those groups, while others are more spread out or isolated, representing unique customer segments. This clustering can be instrumental in devising targeted marketing strategies, where customers in tight clusters may respond similarly to marketing efforts, whereas those in sparser groups might require tailored approaches. Points that stand alone could be considered outliers, representing atypical customer behaviors.

PCA Clustering

Since numerous variables in the dataset are highly correlated, this causes multicollinearity, impacting the clustering method and leading to ineffective cluster profiling biased towards a few variables. To mitigate this issue, PCA was employed to reduce the multicollinearity among the variables, thereby improving the clustering results as shown below.



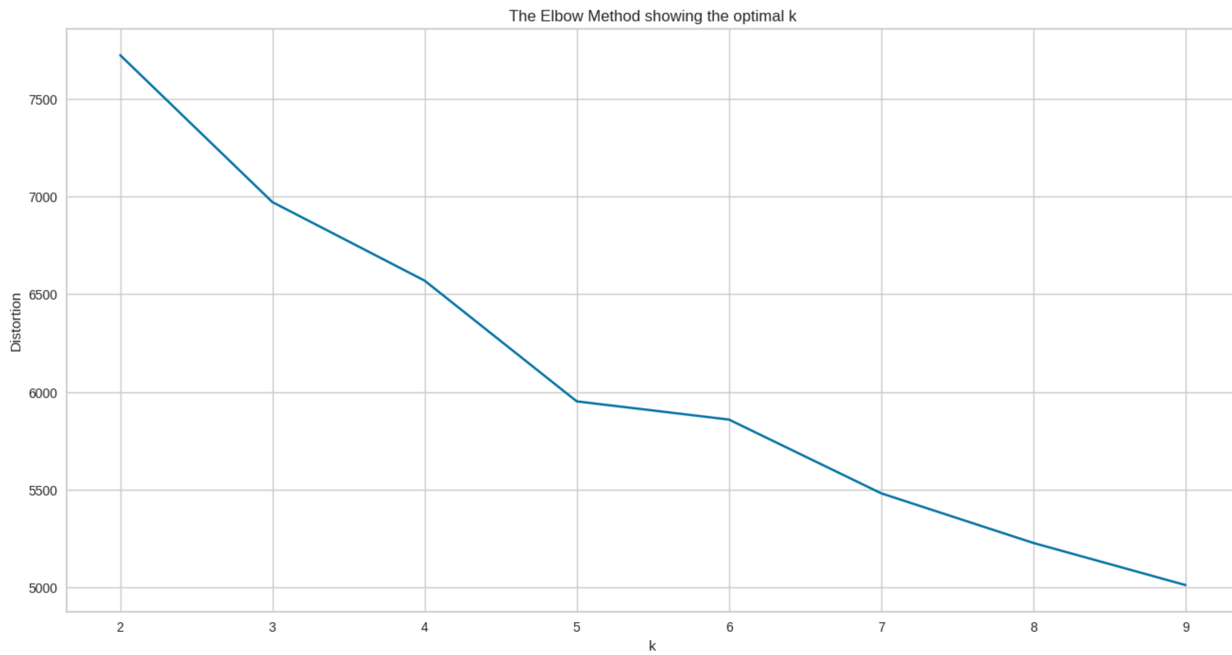
The PCA scatter plot shows how the data spreads out across two main components (x and y axes), hinting at some patterns but not showing clear groups. The first component is particularly important as it captures most of the variation in the data, which means it's a key factor in understanding the overall structure of the dataset. The second component, while not as significant, still shows some variation and helps to reveal more about the relationships in the data.

From the plot, we can also spot a few potential outliers, especially noticeable on the first component. These are points that stand out because they're quite different from the rest, suggesting they might have unique characteristics. While the PCA has simplified the data by reducing its dimensions, the overlapping and clustering of points suggest that the data relationships are complex. To get a clearer picture, we might need to use additional analysis methods or clustering techniques to identify more distinct groups and understand the data better.

Unsupervised Learning and Models

K-Means Clustering

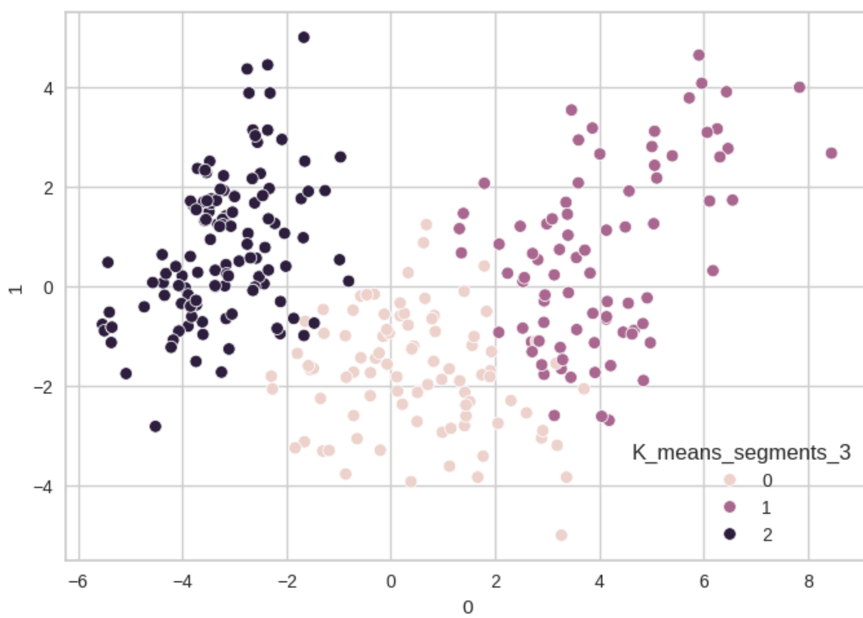
For this analysis, K-Means clustering was used to partition the data into k distinct clusters. To determine the ideal number of k clusters to use, the elbow method was applied. In the graph shown below, the elbow is seen for K=3 and K=5 where some drop in distortion is observed. Since the elbow method provided ambiguous results, the Silhouette Coefficient Score was calculated to measure of how similar an object is to its cluster compared to other clusters.



The result was n_clusters = 3 was closest to 1, and thus, 3 was found to be the best value for k. When running K-Means with k = 3 on the PCA data, resulted in the plot shown below where three clusters are clearly identified and differentiated by three distinct colors (purple, pink, and navy). These colors represent the different segments into which the data has been grouped. As expected, the principal components capture the most significant variance in the dataset, with the horizontal axis (x-values) representing the most variation.

Each cluster corresponds to a unique customer segment, differentiated by key features such as demographic traits such as age and income spending behaviors, and engagement levels. In this case, navy blue represents the highest income and high spending individuals, purple represents older, high income and moderate engagement and spending, and pink represents younger, with lower to mid-range incomes individuals.

K-Means Scatterplot with K = 3.

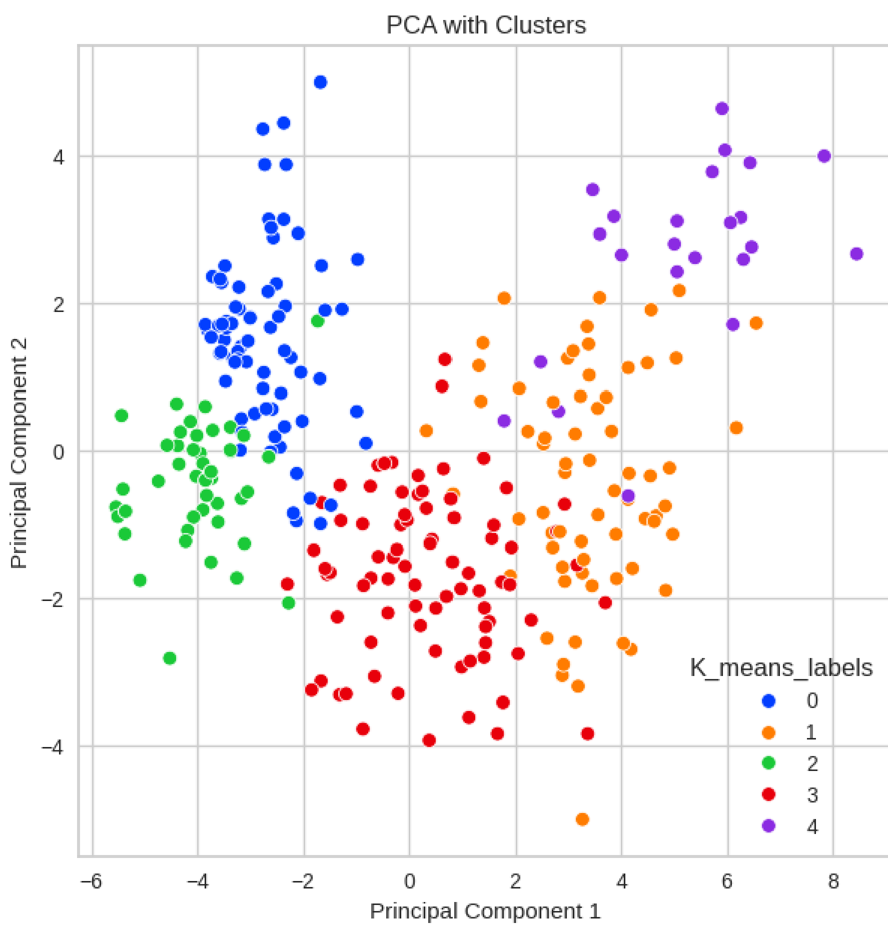


Analysis of K-Means Cluster Profiles

K-Means clustering has delineated several clear and distinctive customer segments, each exhibiting unique spending habits and their own demographic traits. Cluster 0 emerges as the most affluent, with the highest expenditures across various product categories like wines, meats, fruits, fish, and sweets, suggesting a preference for high-end goods. This cluster also shows the highest number of total purchases and deals, making them highly active shoppers. Cluster 1, had the most recent birth years and highest website visits, likely tech-savvy consumers who are actively engaging online but spending less overall.

Cluster 3 stands out for having the highest income but does not lead in most spending categories. This suggests a selective spending pattern likely focused on specific high-value purchases, evident in these customers' lead in web purchases. On the other hand, cluster 4 displays a strong sensitivity to marketing campaigns, with the highest acceptance rates for several promotions, which demonstrates their responsiveness to direct marketing efforts. This cluster also has the highest number of complaints and oldest average age. This may indicate a more demanding or critical customer base.

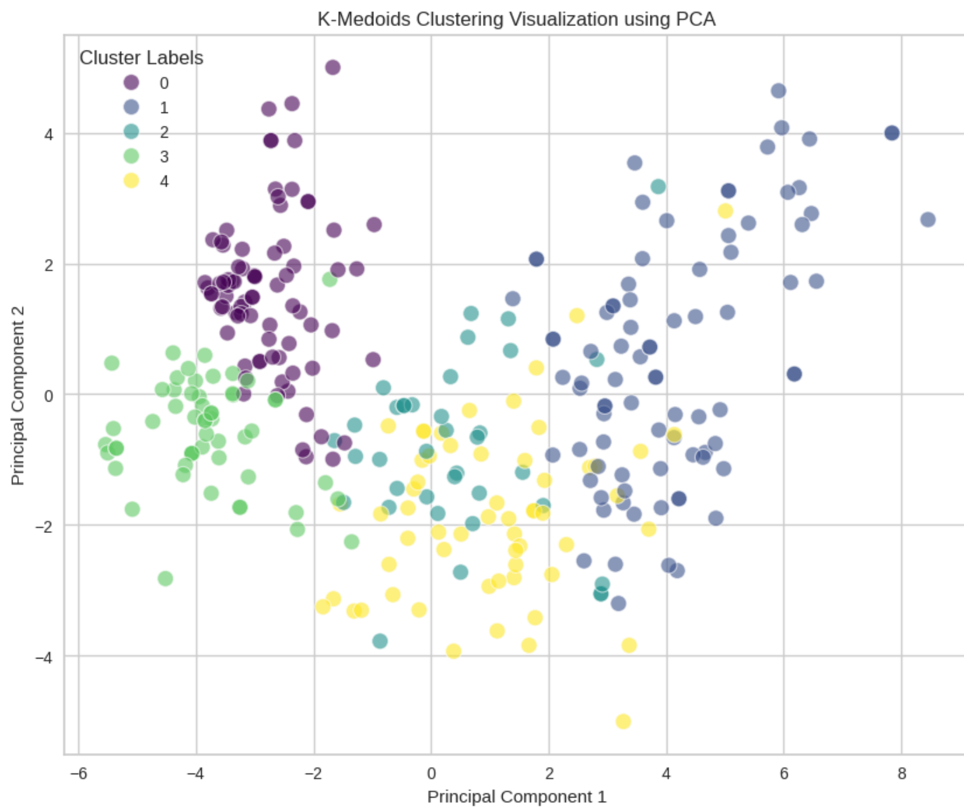
These insights can aid in crafting targeted marketing strategies that cater to the specific preferences and behaviors of each segment, optimizing resource allocation and enhancing customer satisfaction and retention. However, the clustering profiles with K=3 primarily distinguish customers based on high, medium, and low-income levels, but they don't provide deeper insights into the diverse types of customers within the dataset. To further deepen into the different customer segments, let's consider increasing the number of clusters to K=5 as shown in the scatterplot below. This approach may yield more detailed and actionable cluster profiles, potentially revealing unique customer behaviors or preferences.



The scatterplot reveals detailed insights and observations. For instance, cluster 0 (blue) suggests high-income, selective spenders focusing on high-value transactions while cluster 1 (orange) indicates average earners with moderate, well-distributed spending habits. Cluster 2 (green) represents more price-sensitive, infrequent shoppers, hinting there is an opportunity for targeted promotions in this segment. Cluster 3 (red) on the other hand, includes customers with higher incomes, that are likely loyal to specific products, showing focused but substantial spending. Cluster 4 (purple) consists of premium customers with significant purchasing power across multiple categories. This more detailed segmentation is preferred because it allows for tailored marketing strategies to enhance customer satisfaction and optimize business performance by addressing the unique needs and spending behaviors of each group.

K-Medoids Clustering

The next technique is K-Medoids. This technique is more robust to noise and outliers compared to K-Means, which helps refine and understand the data. In this case, K-Medoids was initialized with 5 as the number of clusters (K=5) yielding a silhouette score of approximately 0.1282. When comparing the distribution of the K-Medoids scatterplot with the summary statistics deeper insights can be observed. The scatterplot reveals five customer segments, each exhibiting unique characteristics and behaviors. The resulting scatterplot is shown below.



Analysis of K-Medoids Cluster Profiles

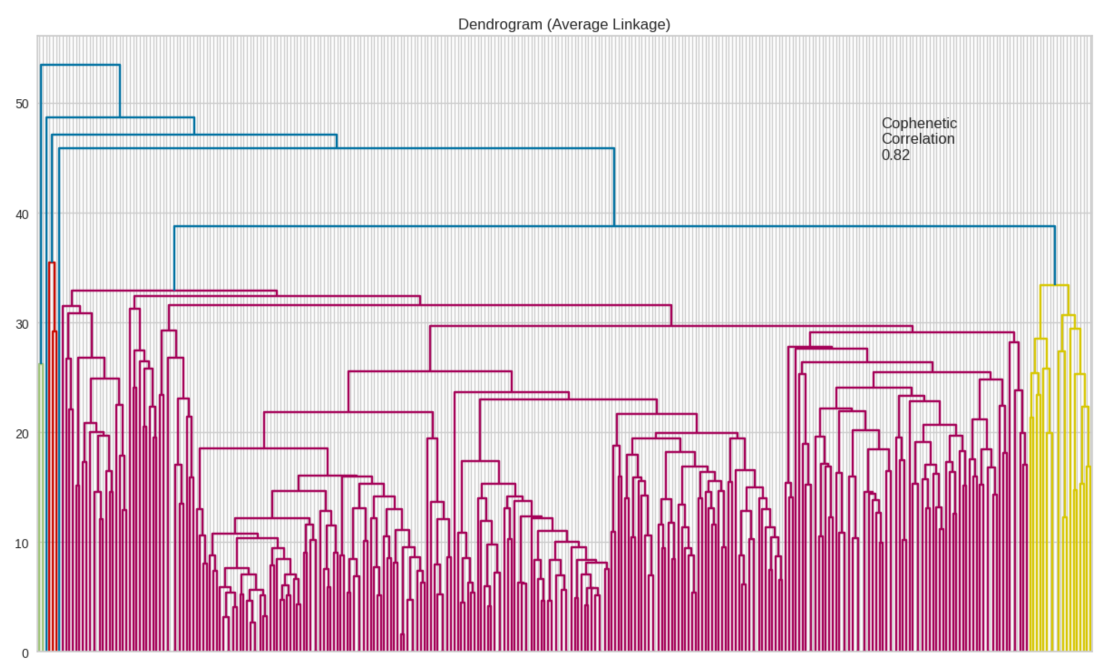
The K-Medoids clustering reveals five distinct customer profiles based on demographics, spending behaviors, and brand engagement. Cluster 0 consists of younger individuals with a high income, engaging significantly with the brand and spending the most, particularly on wines and meats. Their high engagement and expenditure make them prime targets for premium product offerings and loyalty programs. Cluster 1, with slightly older demographics and lower income levels, shows cautious spending habits, with fewer total purchases and lower expenses. They require more personalized marketing strategies to enhance their engagement and conversion rates.

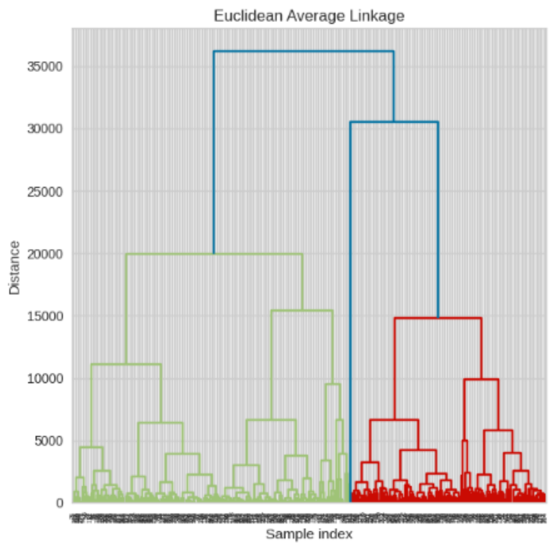
Clusters 2 and 4 are characterized by high incomes and receptiveness to marketing campaigns, with cluster 2 showing the highest campaign acceptance. This suggests a strong potential for upselling and cross-selling high-end products. Cluster 3, contains the oldest participants and largest families. This cluster shows conservative spending on luxury items and a moderate response to marketing initiatives, indicating there is a need for value-driven promotions in this segment.

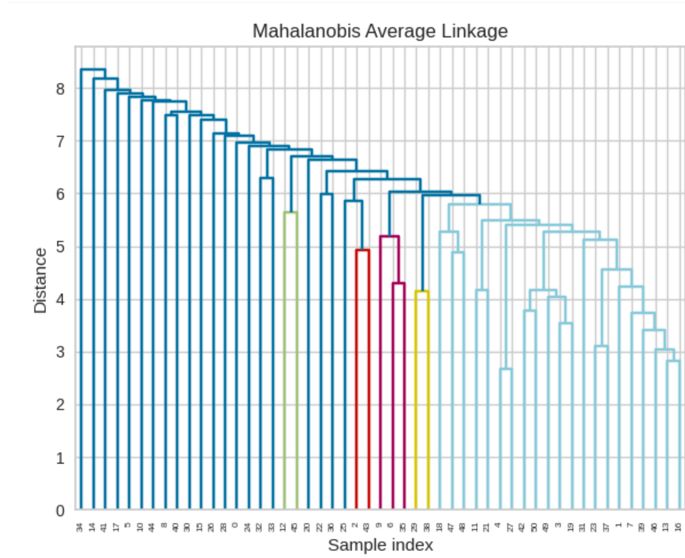
Hierarchical Clustering

Hierarchical clustering was used to build a hierarchy tree of clusters and observe whether there are natural groupings in the data. Unlike partitioning clustering methods like K-Means or K-Medoids, hierarchical clustering does not require a pre-specified number of clusters. Instead, it is initialized with the number of data points (N), where each data point is its own cluster. The highest cophenetic correlation (an essential metric for assessing the effectiveness of hierarchical clustering) was found to be approximately 0.8393 in this case. The average linkage dendrograms using Cityblock, Chebyshev and Mahalanobis distance are shown in the graphs below.

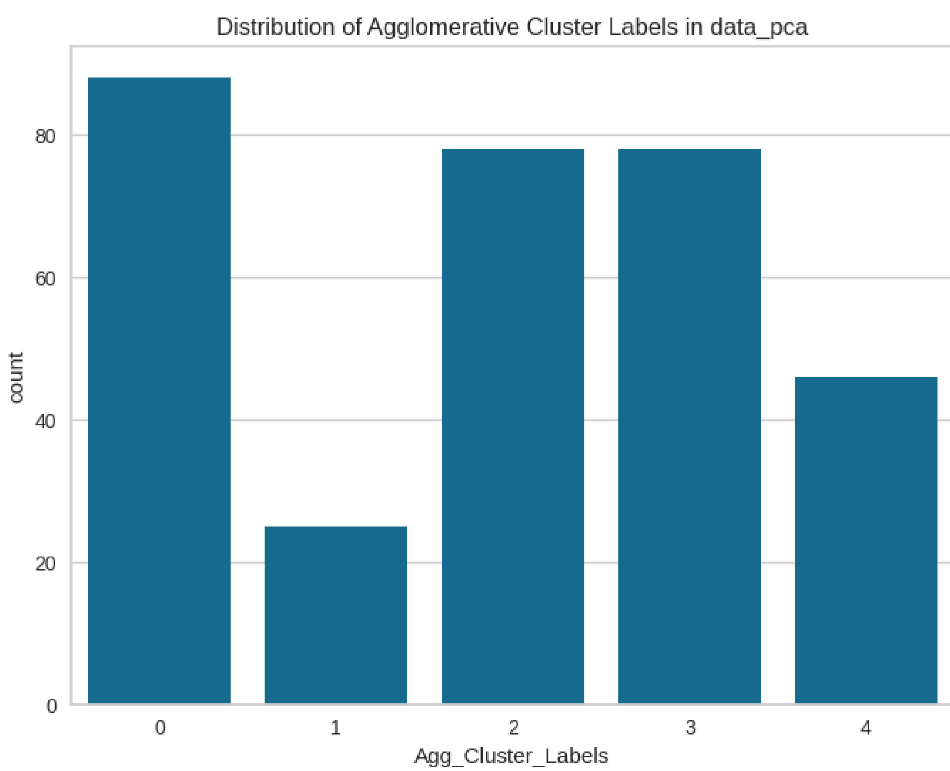
Cityblock Average Linkage Dendrogram





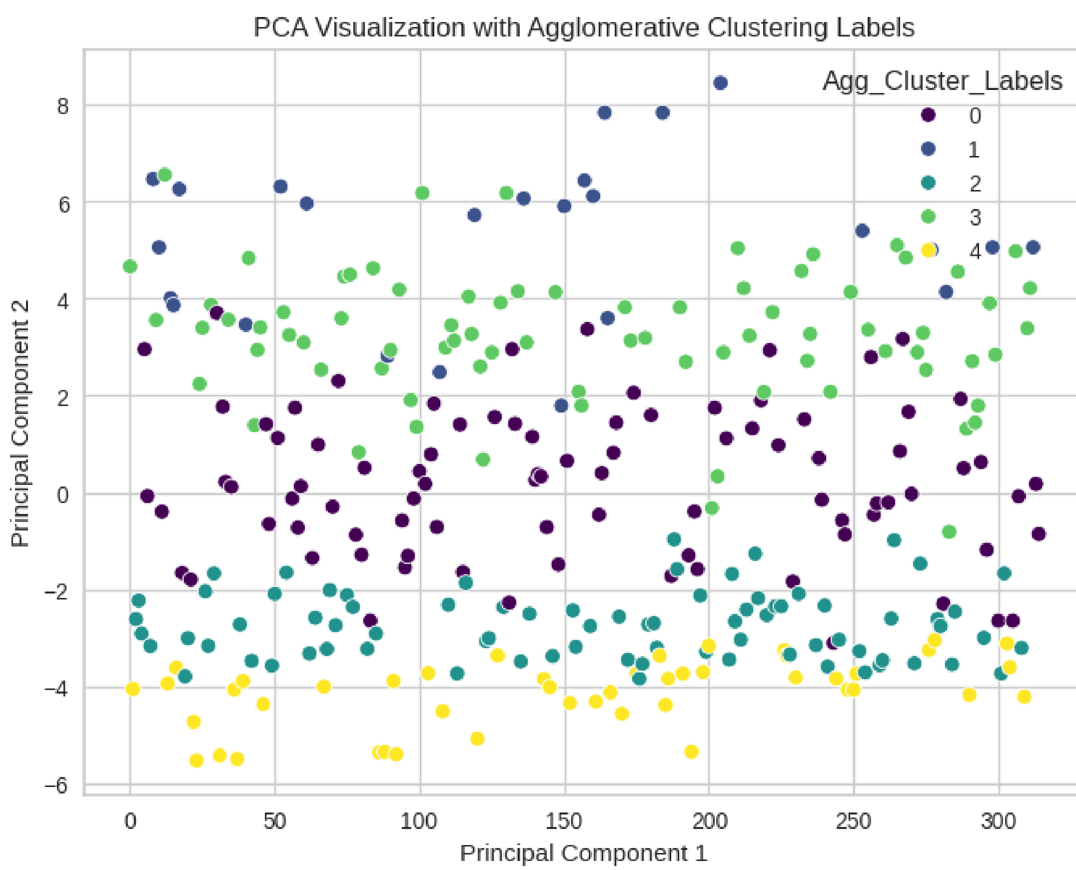


Below is the bar chart for the distribution of agglomerative cluster labels assigned to our dataset. Each bar corresponds to one of the five clusters (labelled from 0 to 4), and the height of each bar indicates the number of data points assigned to each cluster.



Next, let's visualize the clusters using PCA. The resulting hierarchical scatterplot shown below. The scatterplot reveals distinct customer behaviors across five clusters. Cluster 0 (purple) consists of conservative spenders or possibly newer customers with lower overall engagement and spending across product categories. Strategies for this group should focus on value deals and entry-level products to enhance their engagement. Cluster 1 (dark blue) represents customers with moderate purchasing power who engage in frequent transactions across a variety of products, making them ideal targets for up-selling and cross-selling promotions.

Agglomerative Scatterplot



Clusters 2 (light blue), 3 (green), and 4 (yellow) exhibit more unique purchasing behaviors that require tailored approaches. Cluster 2 customers, showing high activity in specific purchasing areas, might benefit from targeted marketing that focuses on specialty products or bundles tailored to their interests. Cluster 3, is characterized by infrequent but significant purchases, and could be engaged through personalized marketing and loyalty programs aimed at increasing transaction frequency. Finally, cluster 4 includes high-spending customers likely to respond to premium offerings and exclusive deals. This suggests the focus here should be on high-value products and services to cater to their diverse but significant investment interests.

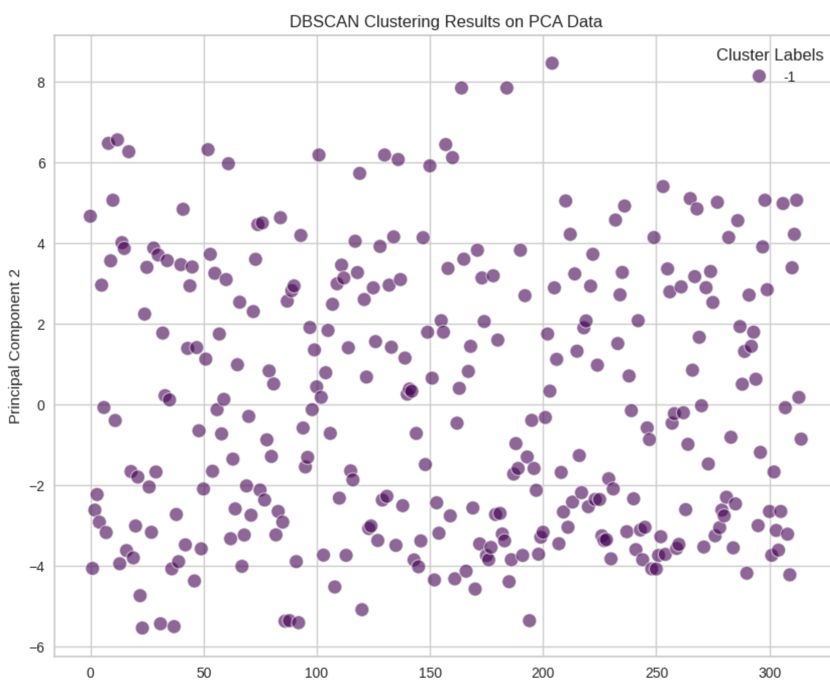
Analysis of Hierarchical Cluster Profiles

The customer segmentation within the dataset showcases varying demographics and spending behaviors across five clusters. Cluster 0 comprises slightly younger customers, born on average in 1968, with a higher income around $53,747. They exhibit recent and frequent interactions with the brand, reflected in their lower recency figures and significant expenditure, especially on wines and meats, totaling an average of $601. Although their engagement over days is robust, they show a lower propensity to respond to marketing campaigns compared to other groups. Cluster 1's older demographic, born in 1963, earns slightly less, with an income of $49,376, and has the most children at home. This demographic's cautious spending is evidenced by lower outlays on luxury items and a moderate number of total purchases, though they show the highest acceptance rate of marketing campaigns, indicating a potential openness to promotions.

On the other hand, cluster 2's younger patrons, with a birth year of 1970 and higher incomes, displays a clear preference for quality in their purchases with the highest spending across all segments, particularly on wines, fruits, and meats. This group shows a strong brand loyalty and extensive engagement over days and significant expending, yet their acceptance of marketing campaigns is conservative. Cluster 3, averaging a birth year of 1970 but with lower incomes, tends toward more conservative spending habits, particularly on meats and sweets, which might reflect budgetary constraints or different priorities. They engage reasonably well with the brand but are the least responsive to marketing campaigns. Lastly, cluster 4 includes the oldest average birth year of 1965 and is similar to cluster 3 in income but has the largest families and the lowest spending per category. This suggests a focus on basic needs over luxury purchases, with an average brand relationship and campaign responsiveness. These insights enable tailored marketing strategies to enhance customer experience and brand loyalty across diverse customer profiles.

DBSCAN

DBSCAN groups densely packed data points and identifies outliers in sparse areas, using parameters eps and min samples. In this case, the distribution was used to optimize the hyperparameters, initializing eps and min samples to 5, and using the Euclidean metric which resulted in the scatterplot shown below.

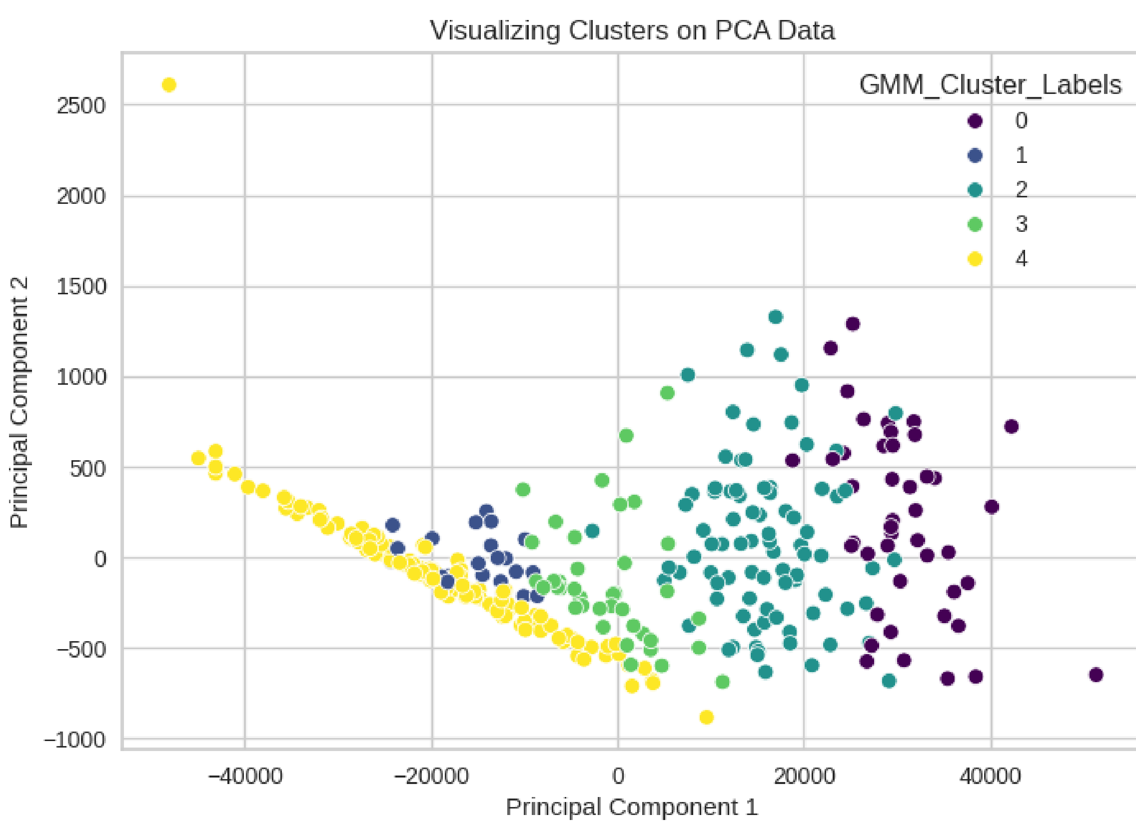


Unfortunately, the scatterplot produced by DBSCAN did not provide any meaningful insights for the purpose of this report. No apparent differentiated clusters were formed because the DBSCAN failed to segment the data meaningfully due to most clusters being labeled as noise (-1). Therefore, DBSCAN could not be obtained and this clustering technique prove not to be useful or adequate for analysis for the purpose of this report.

Gaussian Mixture Model (GMM)

The last model considered for this analysis was the Gaussian Mixture Model. It assumes that each cluster follows a Gaussian distribution for which the parameters are estimated from the data making GMM ideal in anomaly detection. After calculating the silhouette score and applying GMM to the PCA data, the model produced the scatterplot shown below.

GMM Scatterplot



The plot highlights five distinct clusters, each represented by a unique color. However, the tight grouping of some clusters suggests similar profiles among their points, whereas the dispersion and overlap in others reflect a less distinct separation, which is further emphasized by a low silhouette score of about -0.2549. This score indicates poor clustering performance, with significant overlap complicating the practical use of these clusters for segment analysis. However, GMM cluster profiles provided useful insights for targeted marketing strategies.

Analysis of GMM Cluster Profiles

The Gaussian Mixture Model (GMM) clusters for the dataset delineate five distinct customer profiles, each marked by specific demographics, spending habits, and engagement levels. Cluster 0 customers, with the highest average income and an older demographic, show significant expenditure across various product categories and have the highest interaction with marketing campaigns. This suggests a potential for premium product targeting. Cluster 1, the youngest and lowest income group on the other hand, displays very cautious spending behavior, especially in non-essential goods, indicating a segment that may be more responsive to cost-saving offers and essential product promotions.

Moreover, Cluster 2's customers, while having a relatively high income, show significant spending on luxury items like wines and meats, coupled with high brand loyalty, making them ideal candidates for exclusive offers and loyalty rewards. Cluster 3 and 4 contain middle-aged to younger individuals with moderate to upper-middle income brackets but differ in family composition and spending, with cluster 3 focusing less on luxury and more on necessity and cluster 4 showing balanced spending. These insights can help tailor marketing strategies that resonate with each cluster's lifestyle and purchasing power to enhance customer satisfaction and brand engagement.

Comparison of Modeling Techniques

In comparing various clustering techniques used (K-Means, Hierarchical Clustering, DBSCAN, Gaussian Mixture Models (GMM), and K-Medoids), we find distinct advantages and limitations. For instance, K-Means Clustering provided a clear and straightforward segmentation, making it easy to interpret and apply in marketing strategies. It excelled in distinguishing between different income and expenditure groups. However, it struggled to capture more nuanced groupings as well as with outliers and noise.

In the case of Hierarchical Clustering, this technique offered a detailed view of data structure, which helped in understanding the data at various levels of aggregation. It was particularly useful for small datasets but less effective for large datasets The choice of linkage (average linkage in our case) showed better cophenetic correlation, suggesting that it preserved the original distances well. K-Medoids was more robust with noise and outliers than K-Means, showing a balanced performance in identifying clusters based on real-world data points.

DBSCAN's performance in identifying arbitrary-shaped clusters and handling outliers was mixed. While it effectively managed noise and outliers, tuning its eps and min samples parameters proved challenging, resulting in many data points being labeled as noise. Gaussian Mixture Models (GMM) was adept at handling overlapping clusters. It provided a more nuanced understanding of the dataset, especially where clusters were not clearly separable. The negative silhouette score indicated some issues with cluster separation.

Solution Proposal

Considering the detailed insights and limitations observed across various models, the proposed solution design in this case is the K-means clustering model. K-means tends to produce well-separated clusters if the data conforms reasonably well to spherical shapes, as seen in some of the scatterplots. Even with its limitations when handling some of the noise and outliers, the distinct clusters identified by K-means can provide actionable insights for customer segmentation. K-means is not only straightforward to implement but also scales efficiently to large datasets, making it suitable for practical applications where interpretability and computational efficiency are crucial.

While K-means requires careful selection of the number of clusters, the use of methods like the elbow method or silhouette scores provides a systematic approach to determine an optimal k-value, enhancing the model's effectiveness. K-means performed relatively well in terms of silhouette score compared to other models like DBSCAN and GMM, which had issues with either identifying too many outliers or overlapping clusters, respectively. Given these factors, K-means offers a balanced approach, providing clear and meaningful customer segments that can directly inform targeted marketing strategies and customer engagement initiatives.

Conclusion and Recommendations

This analysis provided a clear segmentation of the customer base which can significantly enhance the effectiveness of targeted marketing campaigns. By aligning marketing strategies with the identified segments, businesses can better meet customer needs and improve both sales and customer satisfaction. Models such as K-means and K-medoids are effective in identifying customer clusters based on spending habits, allowing for targeted strategies such as upselling premium products to high spenders or offering discounts to encourage more spending among low spenders. DBSCAN is adept at distinguishing between highly engaged and minimally engaged customers, which can inform loyalty programs and re-engagement campaigns. Meanwhile, Gaussian Mixture Models (GMM) excel in detecting customers with overlapping preferences, which is advantageous for designing effective cross-selling strategies. Each of these models provides insights that can be leveraged to refine marketing approaches and enhance customer engagement.

The clustering analysis of customer data provided by the above-mentioned techniques, revealed critical insights for targeted marketing strategies. The analysis identified distinct spending habits across product categories, showing clear divisions between high, moderate, and low spenders, which can inform targeted promotions. Engagement levels varied, with some clusters displaying higher brand loyalty, suggesting these groups might benefit from personalized loyalty programs. Additionally, demographic differences such as age, income, and family size offer opportunities for tailored marketing messages, with potential strategies ranging from premium offers for affluent older customers to cost-effective promotions for younger families. These insights are invaluable for enhancing customer satisfaction, improving retention, and optimizing marketing expenditure.

The next step for the business is to optimize the effectiveness of customer segmentation. It is recommended to initiate pilot marketing campaigns targeted at identified segments to test and refine strategy recommendations. Concurrently, it is also recommended to implement a continuous feedback loop to collect and analyze data from these campaigns, updating the segmentation model regularly to adapt to new customer behaviors and shifting preferences, ensuring the approach remains dynamic and responsive to market changes.

Appendix

Code Snippets for Data Preparation and Clustering Implementation:

```
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import DBSCAN
from sklearn.mixture import GaussianMixture
```

 # Load and preprocess data

```
data = pd.read_csv("marketing_campaign.csv")
```

 # Scaling data

```
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_model)
```

 # PCA Transformation

```
pca = PCA(n_components=n)
data_pca = pd.DataFrame(pca.fit_transform(df_scaled_cleaned))
```

 # K-Means Clustering

```
kmeans = KMeans(n_clusters=5, random_state=0)
kmeans.fit(data_pca)
```

 # DBSCAN Clustering

```
dbs = DBSCAN(eps=0.5, min_samples=5, metric='euclidean')
dbs.fit(data_pca)
```

 # Gaussian Mixture Model

```
gmm = GaussianMixture(n_components=5, random_state=1)
gmm_labels = gmm.fit_predict(data_pca)
```

Technical Diagrams:

- PCA scatterplots for visualizing clusters.
- Elbow method graph for determining optimal number of clusters in K-Means.

Cluster Profiles:

K-Means and K-Medoids: Identified through scatter plots and statistical means of clusters, focusing on spending habits, engagement levels, and demographic characteristics.

DBSCAN: Used to highlight clusters with various engagement levels and resistance to specific marketing strategies, despite issues with parameter tuning.

GMM: Provided insights despite low silhouette scores, indicating overlapping preferences among customers.

Performance Metrics: Best silhouette score mentioned for GMM was -0.2549, indicating poor performance.

Additional Details: This report provides details on how the clustering techniques explored in the analysis can be utilized for marketing strategies, including tailored marketing strategies, personalized communications, and efficient resource allocation.