INDIAN INSTITUTE OF TECHNOLOGY, KANPUR
DEPARTMENT OF MATHEMATICS AND STATISTICS

Statistical & AI Techniques in Data Mining

MTH443

Project Report

Customer Personality Analysis

Under the Guidance of: Prof. Amit Mitra
Department of Mathematics and Statistics, IIT Kanpur
amitra@iitk.ac.in

Submitted By: Dasari Charithambika **210302**
Divya Gupta **210353**
Soni Verma **211051**
Chakravartula Vinay Kumar **231080029**
Kajipally Sai Nihal **231080049**

IIT KANPUR, NOVEMBER 2024

ABSTRACT

Customer Personality Analysis plays a crucial role in understanding and refining a company's ideal customer profiles, offering insights into customer behaviors, preferences, and needs. This project focused on developing a data-driven approach to customer segmentation, allowing businesses to tailor products and marketing strategies to specific customer segments. By concentrating efforts on targeted customer types, companies can enhance marketing efficiency, reduce costs associated with broad outreach, and improve product relevance by focusing on segments most likely to engage with new offerings.

To conduct this analysis, we utilized various data science techniques on a customer segmentation dataset, employing methods that support both in-depth analysis and visualization. First, we applied Dimension Reduction techniques, such as Principal Component Analysis (PCA), to project customer segments into a lower-dimensional space, enhancing visualization and revealing the underlying structure of customer groups. Next, we calculated similarity and dissimilarity measures, including Euclidean and cosine distances, to evaluate how closely related or distinct different customers are based on their attributes.

Cluster Analysis was then performed using both hierarchical (agglomerative clustering) and non-hierarchical (k-means) methods to group customers into distinct clusters, enabling us to compare results and identify natural customer groupings. To further understand the distribution of important variables like income and spending, we applied Kernel Density Estimation. This provided valuable insights into how these variables are distributed within the customer base, contributing to a clearer understanding of customer financial behaviors.

Additionally, we employed Association Rule Mining, specifically through market basket analysis, to uncover common associations between products that customers frequently purchase together. This analysis helped identify product combinations with high co-occurrence, supporting product bundling strategies and targeted cross-selling initiatives.

Overall, this project highlights the significance of using sophisticated data science techniques to enhance customer segmentation. The insights gained offer practical implications for optimizing marketing strategies and personalizing product offerings, supporting companies in making informed, data-driven decisions that resonate with customer needs.

Contents

1 Introduction

In the Customer Personality Analysis project, we used Python and R to drive data-driven decisions, make predictions, perform analyses, and create visualizations. Our initial focus was on data cleaning, where we refined our dataset with additional features to deepen our understanding of customer profiles. We derived key features to enhance customer insights while removing redundant features for a streamlined dataset.

In Python, we implemented advanced techniques like PCA to reduce dimensionality, agglomerative clustering to form customer segments, and density estimation using joint plots. We also calculated similarity and dissimilarity matrices to assess how closely customers resembled one another in specific characteristics. In R, we applied Association Market Analysis to uncover frequent product pairings and analyze customer purchasing patterns.

This project emphasizes the importance of leveraging both Python and R for comprehensive customer personality analysis, empowering businesses with insights that allow for targeted marketing and personalized customer experiences.

2 About Dataset

We sourced the data from Kaggle https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

We cleaned the data from the **marketing_campaign.csv** dataset and preprocessed

- **Age** from Year_Birth for customer age.
- **Spent** feature for total spending on items like wines, fruits, meat & fish, sweets, and gold.
- Derive **Living_With** from Marital_Status to identify the living situation of couples.
- Add **Children feature** to show the total children in a household (both kids & teenagers).
- Create **Family_Size** feature for a clearer understanding of household composition.
- Add **Is_Parent** feature to indicate parenthood status.
- Simplify **Education** by creating three broader categories.
- Drop redundant features to streamline the dataset and Remove the outliers

After preprocessed,the final dataset is **marketing_campagin.csv** and dimension is **2202 x 30**

Education	Income	Kidhome	Teenhome	Recency	Wines	Fruits	Meat	Fish	Sweets	Gold	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Graduate	58138	0	0	58	635	88	546	172	88	88	3	8	10
Graduate	46344	1	1	38	11	1	6	2	1	6	2	1	1
Graduate	71613	0	0	26	426	49	127	111	21	42	1	8	2
Graduate	26646	1	0	26	11	4	20	10	3	5	2	2	0
Postgraduate	58293	1	0	94	173	43	118	46	27	15	5	5	3
Postgraduate	62513	0	1	16	520	42	98	0	42	14	2	6	4

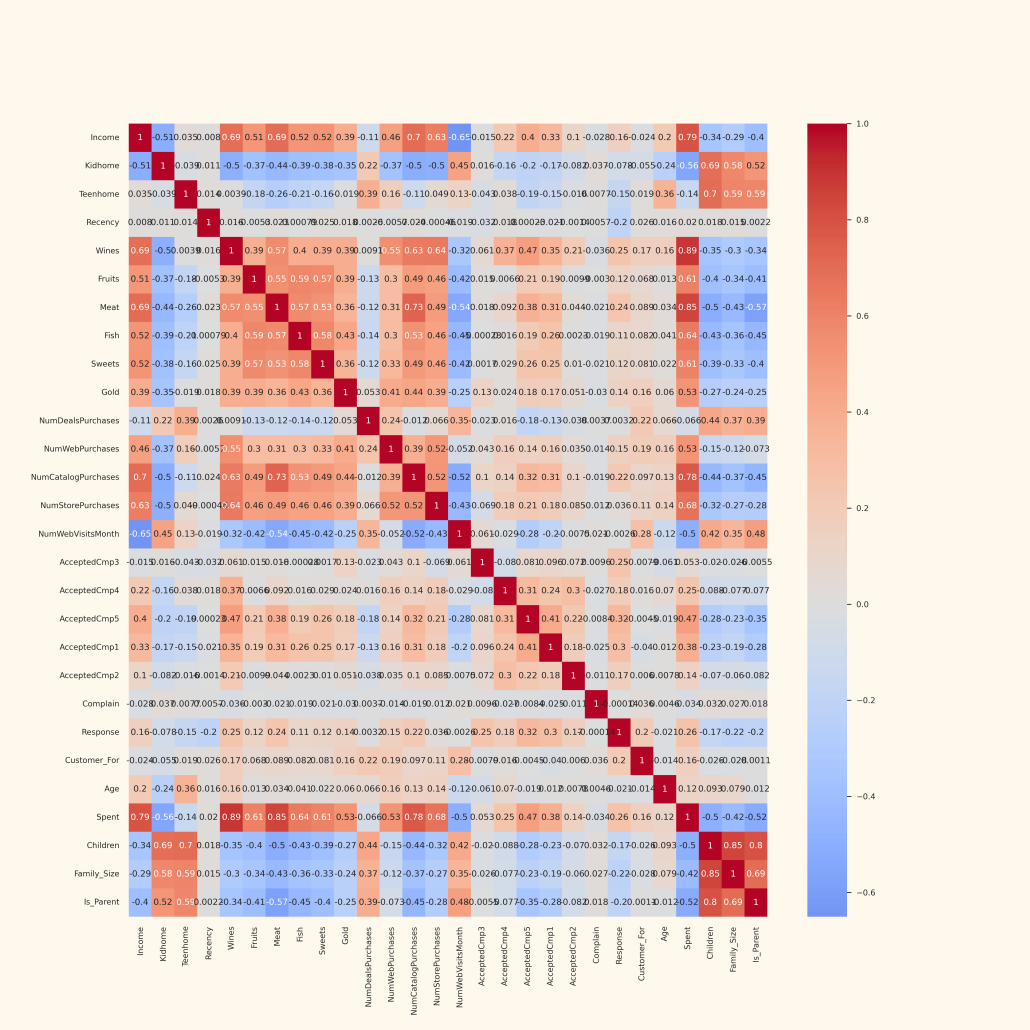
NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
10	4	7	0	0	0	0	0
1	2	5	0	0	0	0	0
2	10	4	0	0	0	0	0
0	4	6	0	0	0	0	0
3	6	5	0	0	0	0	0
4	10	6	0	0	0	0	0

AcceptedCmp2	Complain	Response	Customer_For	Age	Spent	Living_With	Children	Family_Size	Is_Parent
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>
0	0	1	663	64	1617	Alone	0	1	0
0	0	0	113	67	27	Alone	2	3	1
0	0	0	312	56	776	Partner	0	2	0
0	0	0	139	37	53	Partner	1	3	1
0	0	0	161	40	422	Partner	1	3	1
0	0	0	293	54	716	Partner	1	3	1

- Education: Education as factors (Undergraduate, Graduate, Postgraduate)
- Income: Amount earned
- Kidhome: no of kids in home & Teenhome: no of teenagers in home
- Recency: Number of days since customer's last purchase
- Wines, Fruits, Meat, Fish, Sweets, Gold: Amount spent on those items
- NumDealsPurchases: Number of purchases made with a discount
- NumWebPurchases: Number of purchases made through the company's website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to the company's website in the last month
- AcceptedCmp i: 1 if customer accepted the offer in the i^{th} campaign, 0 otherwise for all i = 1,2,3,4,5
- Response: 1 if the customer accepted the offer in the last campaign, 0 otherwise
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise
- Customer_For: number of days the person has became customer
- Age: age of customer
- Spent: Amount spent on buying those items
- Living_With: living with as factor of (Alone, Partner)
- Children: no of children(Kids + Teenagers)
- Family_Size: No of family members
- Is_Parent: 1 if parent, else 0

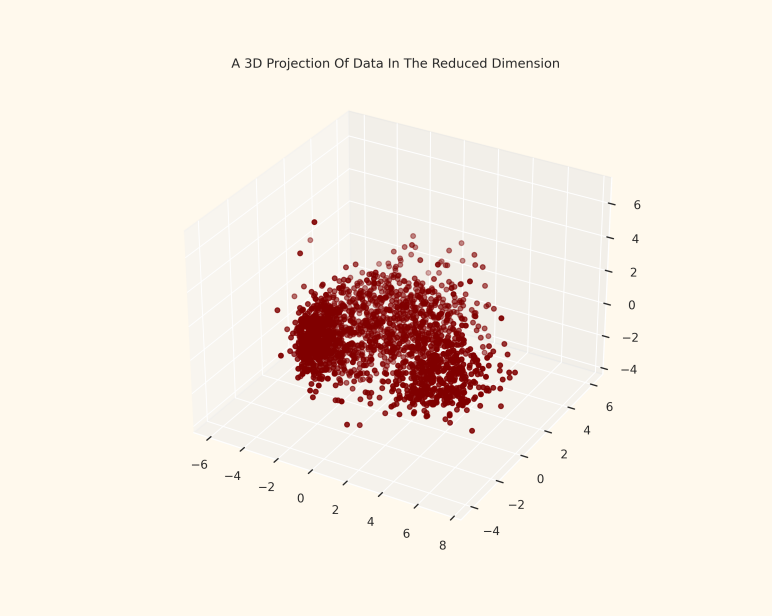
3 Correlation

To examine the correlation between features in the dataset, we used the `corr()` function to compute the correlation matrix.



4 Dimensionality Reduction

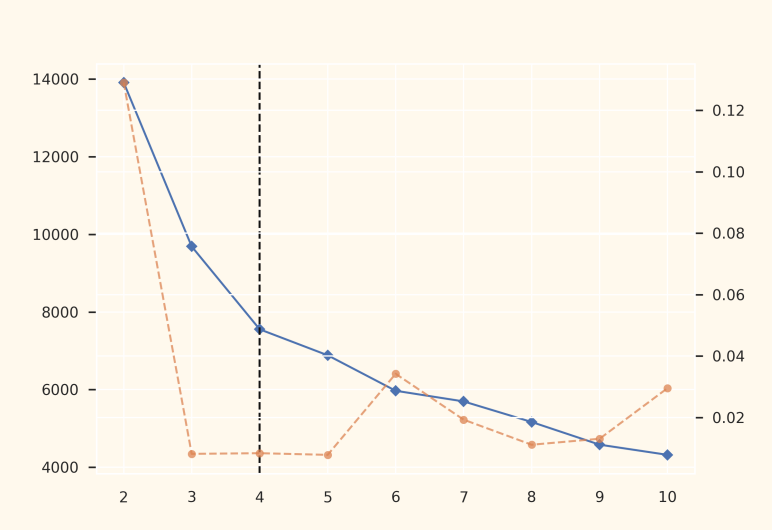
Before dimensionality reduction, we performed label encoding on categorical features using `LabelEncoder()` and scaled all features using `StandardScaler()` for uniformity. We then applied Principal Component Analysis using the python function `PCA()` to reduce the dataset to three principal components.



5 Clustering

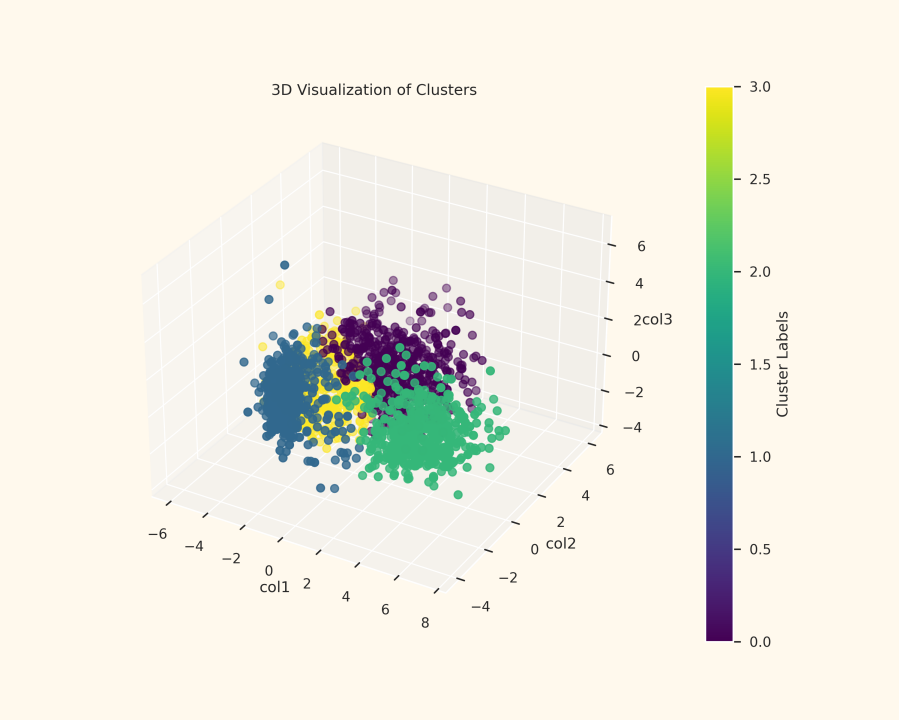
Reduced the dataset to three dimensions and prepared it for clustering using Agglomerative Clustering, a hierarchical method that iteratively merges data points until the desired cluster count is reached.

Applied the Elbow Method with `KElbowVisualizer()` to determine the optimal number of clusters.



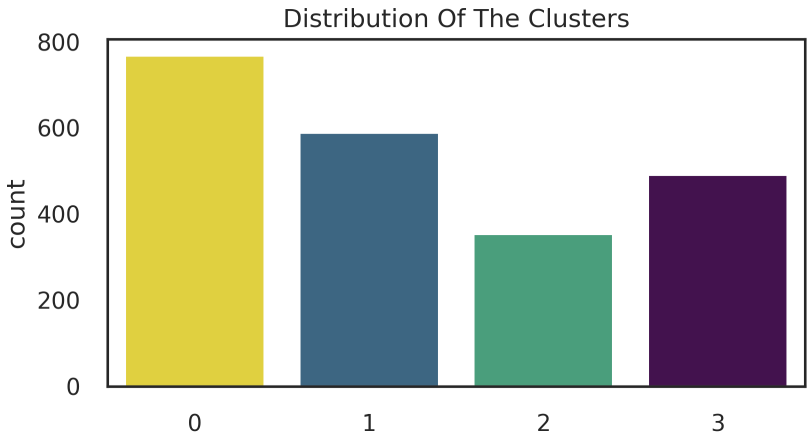
Based on the Elbow plot analysis, we identified $k = 4$ as the optimal number of clusters and proceeded with 4 clusters for further analysis. Performed clustering with `AgglomerativeClustering()` to group data points based on similarity.

Visualized the resulting clusters using a scatter plot for interpretation and analysis.



5.1 Distribution of Clusters

We identified 4 clusters in our dataset, and here is the distribution of data points across these clusters. The plot visually represents the count of data points in each cluster, allowing us to see the relative sizes of the clusters and identify any imbalances or predominant groupings within our segmented dataset.



- **Within and Between clusters distance**

  Average Within-cluster Distance: 6.263213199153256

  Average Between-cluster Distance: 8.229446699627276

  The higher between-cluster distance (8.22) compared to the within-cluster distance (6.26) indicates distinct clusters with cohesive groupings. The moderate gap between these distances suggests reasonably well-separated clusters, though further optimization may enhance distinctiveness if needed.

- **Average Cosine Similarity within each Cluster**

Cluster	**Average cosine similarity**
Cluster 0	0.3001421908443696
Cluster 1	0.4298392860970552
Cluster 2	0.2877551887118949
Cluster 3	0.44391531443574445

  Cluster 0: The average cosine similarity is 0.3001, indicating a relatively low similarity among its members. This suggests that the cluster may consist of diverse or less related items.

  Cluster 1: With an average cosine similarity of 0.4298, this cluster shows a moderate level of similarity, indicating that its members are more closely related than those in Cluster 0.

  Cluster 2: The average cosine similarity of 0.2878 reflects a lower level of cohesion, suggesting that the members are relatively diverse and less tightly grouped than those in Clusters 1 and 3.

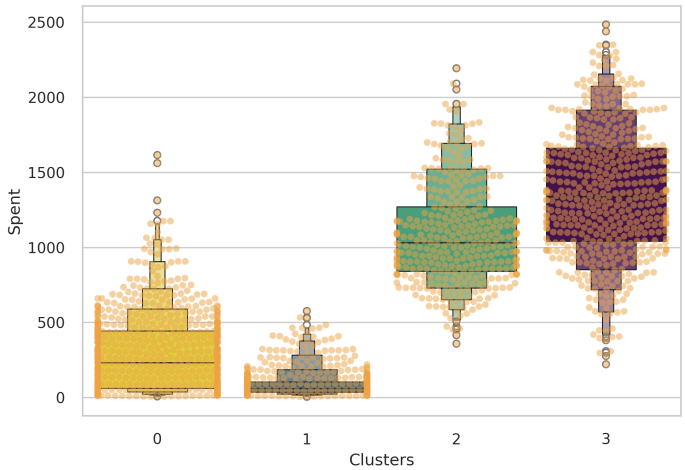
  Cluster 3: Exhibiting the highest average cosine similarity of 0.4439, this cluster shows the strongest relationship among its points, indicating that its members share more characteristics or features.

5.2 Scatter Plot between Income and Spent from different clusters

- Group 0: Average income & Low Spending
- Group 1: Low income & Low Spending
- Group 2: Average income & Average Spending
- Group 3: High income & High Spending

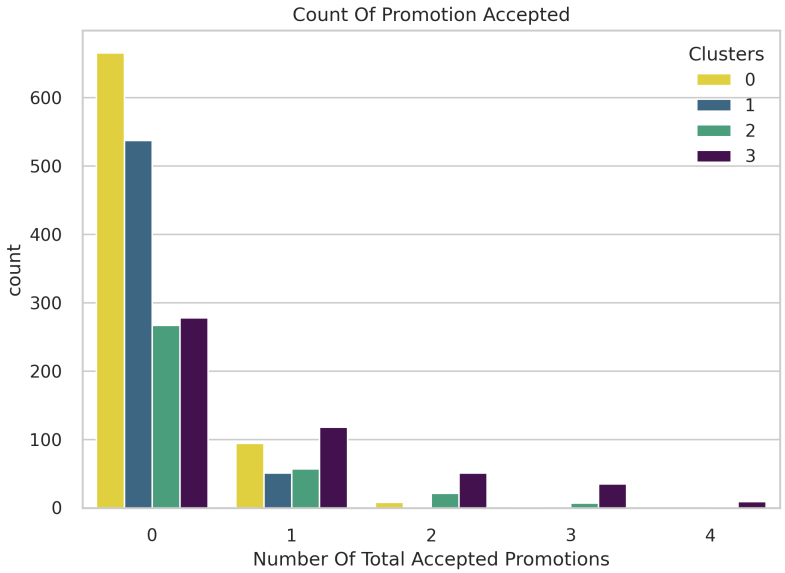
5.3 Distribution of clusters as per the products

Distribution of clusters as per the various products in the data. Namely: Wines, Fruits, Meat, Fish, Sweets, and Gold using the box and swarm plots.



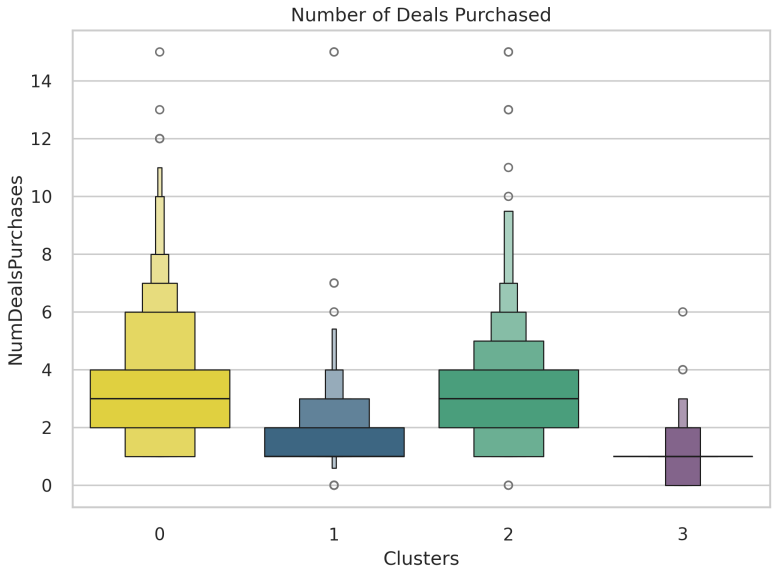
The plot shows that Cluster 3 has the largest group of customers, closely followed by Cluster 2. Analyzing the spending patterns of each cluster can help refine targeted marketing strategies.

5.4 Accepted offers in different campaign according to clusters



The response to the campaigns has been limited, with only a small number of participants overall. Additionally, no customers participated in all five campaigns. To enhance engagement and drive sales, more targeted and strategically planned campaigns may be necessary.

5.5 Deals Purchased according to clusters



The deals offered performed significantly better than the campaigns, with the most favorable outcomes observed in Cluster 0 and Cluster 2. However, Cluster 2, comprising our star customers, showed limited engagement with the deals. Additionally, no specific deal appeared to strongly appeal to Cluster 3.

6 Density Estimation

By creating individual joint plots between Spending and each variable in Places **(WebPurchases, CatalogPurchases, StorePurchases, and WebVisitsMonth)**, we can analyze how spending varies with each type of purchase behavior.

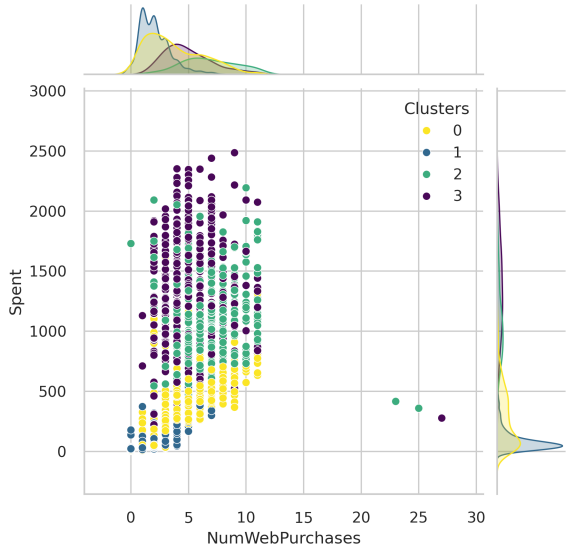


Figure 6.1: Joint plot btw Spent & Web Purchases

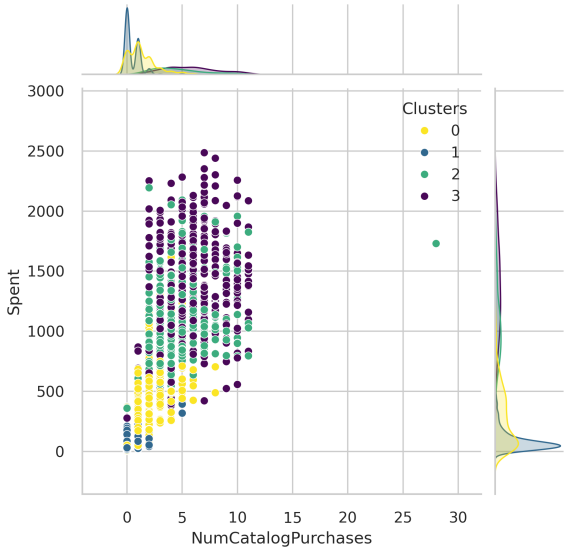


Figure 6.2: Joint plot btw Spent & Catalog Purchases

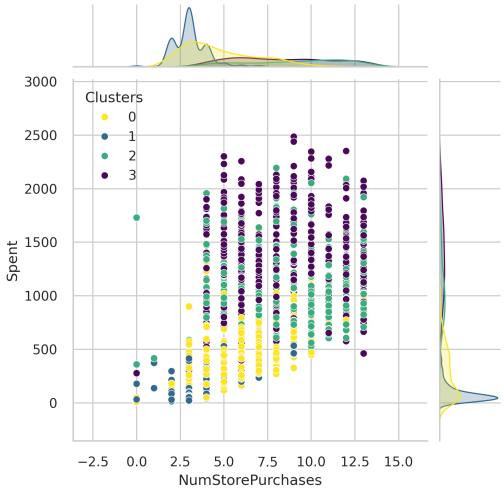


Figure 6.3: Joint plot btw Spent & Store Purchases

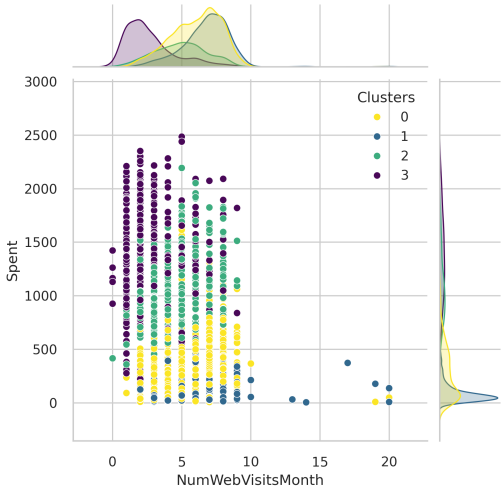


Figure 6.4: Joint plot btw Spent & Web Visits

- Cluster 3 (purple) consists of high-value customers with spending over 1500, while Cluster 0 (yellow) has lower spenders, mostly under 1000. Web purchases are concentrated between 5 to 10, with some high outliers (15-30) in Cluster 2, indicating occasional high-frequency buyers.(refer fig. 6.1).

- The plot shows four clusters (0–3) indicating spending behavior based on catalog purchases: clusters 0 and 1 (yellow, blue) have lower spending, while clusters 2 and 3 (green, purple) represent higher spenders, often exceeding 2000. Most customers make 0–10 catalog purchases, with a dense spending concentration of around 2000, though some outliers make up to 30 purchases without a proportional spending increase.(refer fig. 6.2).
- The plot shows four spending clusters (0–3) based on store purchases: clusters 0 and 1 (yellow, blue) have lower spending, typically under 1000, while clusters 2 and 3 (green, purple) indicate higher spenders, often exceeding 2000. Most customers make 0–10 purchases, with spending concentrated around 2000, but additional purchases beyond 10 do not lead to proportional spending increases.(refer fig. 6.3).

7 Similarity Measures

Similarity measures can include **Euclidean distance** and **cosine similarity**, among others, as these represent different ways to quantify similarity or dissimilarity between data points.

Euclidean Distance Matrix(2202 x 2202):

$$\begin{bmatrix} 0 & 10.1537 & 6.6814 & \cdots & 8.5038 & 7.8394 & 9.5600 \\ 10.1537 & 0 & 7.5841 & \cdots & 9.1069 & 6.1504 & 4.9113 \\ 6.6814 & 7.5841 & 0 & \cdots & 6.4695 & 4.3386 & 7.8867 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 8.5038 & 9.1069 & 6.4695 & \cdots & 0 & 7.8457 & 10.1244 \\ 7.8394 & 6.1504 & 4.3386 & \cdots & 7.8457 & 0 & 6.3704 \\ 9.5600 & 4.9113 & 7.8867 & \cdots & 10.1244 & 6.3704 & 0 \end{bmatrix}$$

Cosine Similarity Matrix(2202 x 2202):

$$\begin{bmatrix} 1 & -0.5103 & 0.4153 & \cdots & 0.2361 & 0.0760 & -0.2563 \\ -0.5103 & 1 & -0.4994 & \cdots & -0.3608 & -0.1424 & 0.4217 \\ 0.4153 & -0.4994 & 1 & \cdots & 0.3519 & 0.4347 & -0.5025 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0.2361 & -0.3608 & 0.3519 & \cdots & 1 & -0.0973 & -0.6012 \\ 0.0760 & -0.1424 & 0.4347 & \cdots & -0.0973 & 1 & -0.1230 \\ -0.2563 & 0.4217 & -0.5025 & \cdots & -0.6012 & -0.1230 & 1 \end{bmatrix}$$

8 Association Market Analysis

This report presents a market basket analysis conducted on various product categories: *Wines*, *Fruits*, *Meat*, *Fish*, *Sweets*, and *Gold*. Using the **Apriori algorithm**, we aimed to discover frequent item sets and association rules that highlight purchasing patterns. The analysis covers metrics such as *support*, *confidence*, and *lift*, providing insights for potential cross-selling opportunities.

8.1 Data Preparation

Each product category was converted into binary values, where a purchase (amount greater than zero) was marked as 1, and no purchase as 0. This binary transformation enabled a clear view of purchasing behavior across the transactions.

8.2 Frequent Itemsets and Association Rules

Frequent itemsets were identified using a minimum support threshold of 0.1 (10% of transactions). The following association rules were then generated with a minimum confidence of 0.5 (50%).

8.3 Top 10 Association Rules by Lift

Rule	Support (%)	Confidence (%)	Lift
{Fruits, Meat, Sweets} → {Gold, Wines, Fish}	64.62	89.22	1.109
{Gold, Wines, Fish} → {Fruits, Meat, Sweets}	64.62	80.35	1.109
{Fruits, Sweets} → {Gold, Wines, Fish}	64.62	89.22	1.109
{Gold, Wines, Fish} → {Fruits, Sweets}	64.62	80.35	1.109
{Gold, Wines, Meat, Fish} → {Fruits, Sweets}	64.62	80.35	1.109
{Sweets, Fish} → {Wines}	72.75	99.26	0.998
{Wines, Meat} → {Gold, Sweets, Fish}	71.39	71.81	0.998
{Wines} → {Gold, Sweets, Fish}	71.39	71.81	0.998
{Gold, Sweets, Fish} → {Wines}	71.39	99.24	0.998
{Gold, Sweets, Fish} → {Wines, Meat}	71.39	99.24	0.998

Table 8.1: Top 10 Association Rules by Lift

8.4 Interpretation of Rules

These rules indicate a strong relationship between product categories, with {Fruits, Meat, Sweets} and {Gold, Wines, Fish} often appearing together in transactions. High confidence values (e.g., 89.22%) and moderate lift values (1.109) suggest that these product bundles are common and can inform cross-selling strategies.

8.5 Product Purchase Frequencies

The purchase frequencies for each product category highlight their popularity:

Product	**Purchase Frequency (%)**
Wines	99.41
Fruits	82.15
Meat	99.95
Fish	82.79
Sweets	81.34
Gold	97.23

Table 8.2: Product Purchase Frequencies

8.6 Analysis and Insights

- **High Purchase Rates:** Most products demonstrate high purchase frequencies, particularly *Meat* (99.95%), *Wines* (99.41%), and *Gold* (97.23%).
- **Common Bundles:** Frequent itemsets suggest that certain products are often bought together, e.g., {Fruits, Meat, Sweets} and {Gold, Wines, Fish}.
- **Cross-Selling Potential:** The rules highlight opportunities to promote products as bundles. For instance, when a customer buys *Fruits, Meat, and Sweets*, promoting *Gold, Wines, and Fish* could be effective due to their common association.

8.7 Visualization

The following bar chart (Figure 8.1) provides a visual overview of the purchase frequencies across product categories.

The Apriori analysis reveals valuable patterns in customer purchasing behavior. Strong associations between products and high purchase frequencies suggest natural product groupings. Implementing cross-selling strategies based on these insights could drive sales by promoting complementary items.

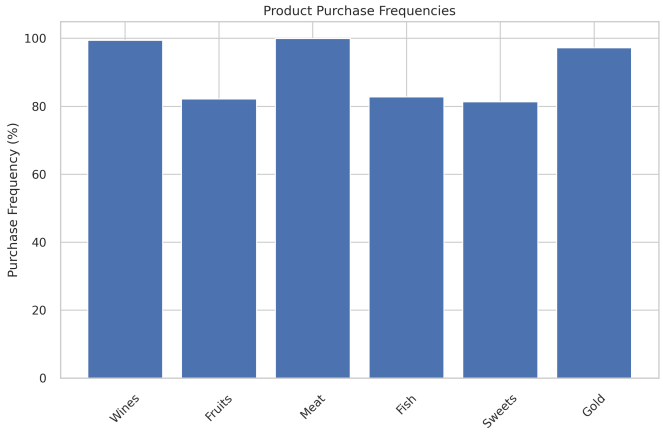


Figure 8.1: Product Purchase Frequencies

9 Work Distribution

Student	Work
Dasari Charithambika(210302)	Density estimation and Report making
Divya Gupta(210353)	Association market analysis & EDA & Report making
Soni Verma(211051)	Association market analysis & EDA & Report making
Chakravartula Vinay Kumar(231080029)	PCA & Clustering
Kajipally Sai Nihal(231080049)	Data cleaning preprocessing & Similarity measures