

## Task 2 : Customer Segmentation

## TABLE OF CONTENT

<b>1. Introduction.....</b>	<b>2</b>
<b>2. Dataset Description.....</b>	<b>2</b>
<b>3. Data Preprocessing.....</b>	<b>2</b>
3.1 Data Inspection.....	2
3.2 Handle Missing Data.....	2
3.3 Outlier Detection.....	3
<b>4. Exploratory Data Analysis (EDA).....</b>	<b>3</b>
4.1 Correlation Analysis.....	3
4.2 Distribution Analysis.....	4
4.3 Analysis based on insights given.....	4
4.4 Dimensionality Reduction .....	6
4.4.1 Data Scaling.....	6
4.4.2 UMAP Projection.....	7
4.4.3 PCA Projection.....	8
<b>5. Model Selection.....</b>	<b>8</b>
5.1 Determine optimal number of clusters.....	8
5.1.1 Elbow Method.....	8
5.1.2 Silhouette Score Analysis.....	9
5.1.3 Visual Inspection of UMAP and PCA projections.....	9
5.2 Determine best performing clustering model.....	9
<b>6. Model Evaluation.....</b>	<b>11</b>
<b>7. Identifying Clusters.....</b>	<b>11</b>
7.1 Key observations.....	12
<b>8. Conclusion.....</b>	<b>13</b>
<b>9. Challenges Faced.....</b>	<b>13</b>
<b>10. Suggestions for Improvement.....</b>	<b>14</b>
<b>11. References.....</b>	<b>14</b>

# 1. Introduction

The objective of this task was to identify distinct customer segments within an e-commerce platform based on their behavioral patterns. The provided dataset contained information about customer interactions, purchases, and browsing behavior. By applying clustering techniques, the objective was to uncover underlying customer groups, specifically targeting the identification of "Bargain Hunters," "High Spenders," and "Window Shoppers."

This analysis will provide valuable insights for targeted marketing strategies and improved customer engagement.

## 2. Dataset Description

The dataset contained customer behavior information across five key features:

- **total\_purchases:** Total number of purchases made by the customer.
- **avg\_cart\_value:** Average value of items in the customer's cart.
- **total\_time\_spent:** Total time (in minutes) spent by the customer on the platform.
- **product\_click:** Number of products viewed by the customer.
- **discount\_count:** Number of times the customer used a discount code.

The dataset is expected to have three hidden clusters, each representing a distinct customer segment: Bargain Hunters, High Spenders, and Window Shoppers.

## 3. Data Preprocessing

### 3.1 Data Inspection

This dataset consists of 999 unique customer details. All 5 key features include numerical values.

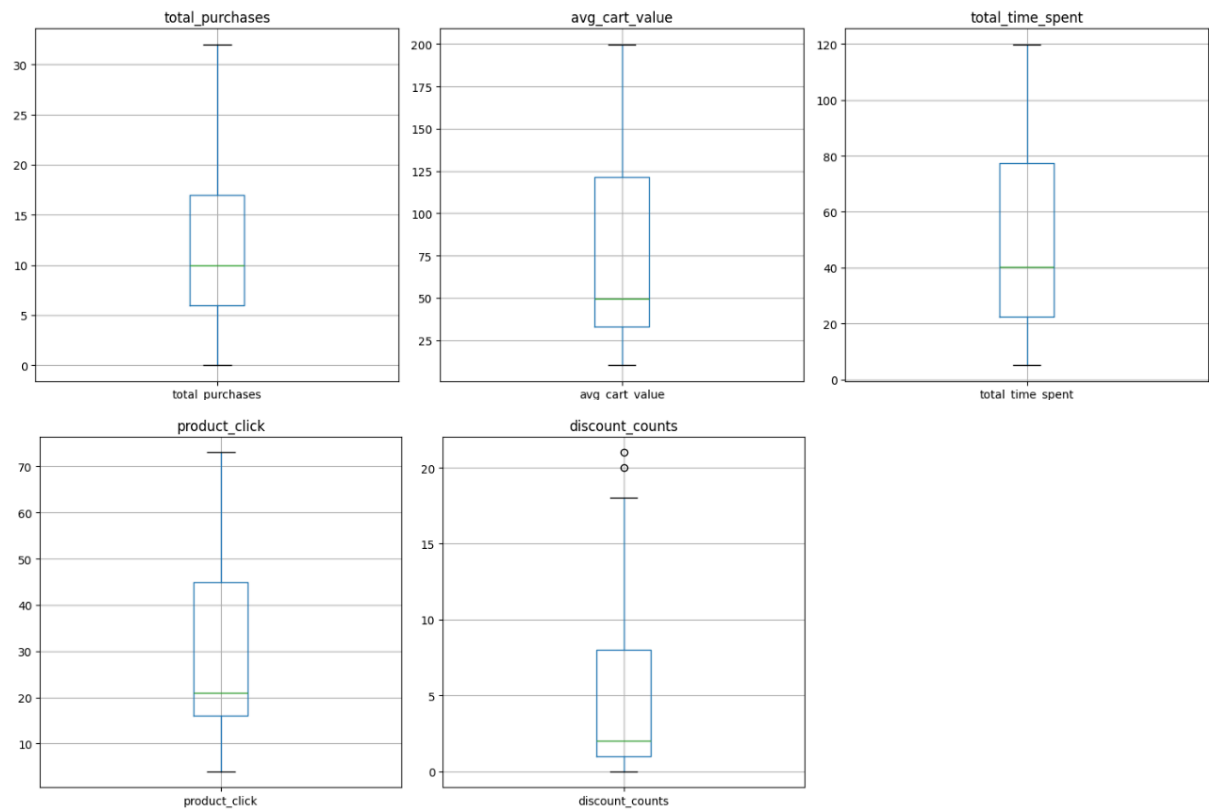
### 3.2 Handle Missing Data

By analysing the dataset, 20 null values were found. All those rows which contained null values were removed.

Then, no more missing values existed in the dataset.

### 3.3 Outlier Detection

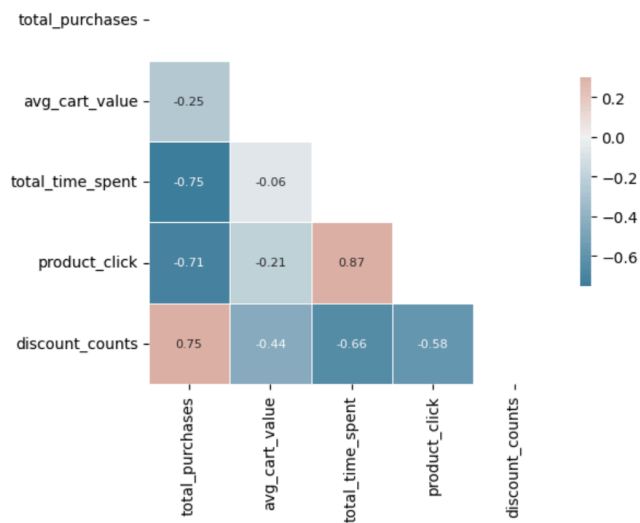
Boxplots were generated for each numerical column to visually identify potential outliers using IQR method.



No extreme outliers were observed.

## 4. Exploratory Data Analysis (EDA)

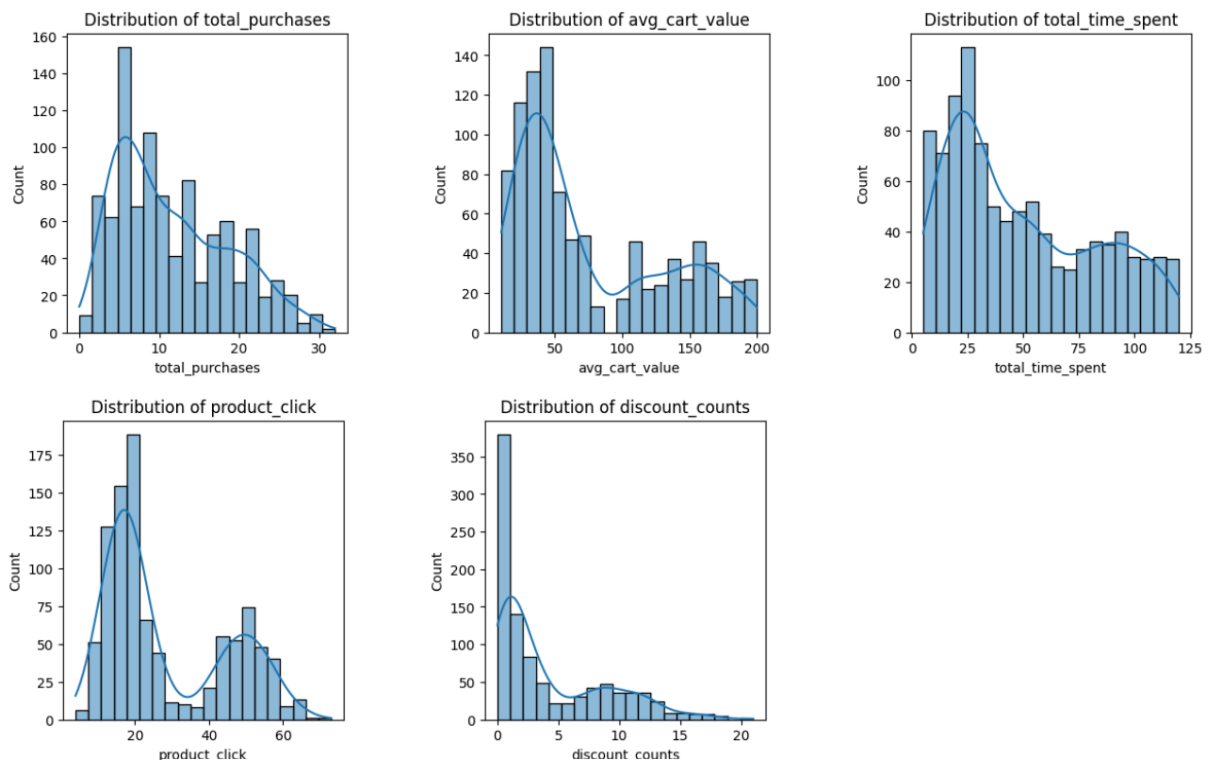
### 4.1 Correlation Analysis



A heatmap of the correlation matrix was generated to understand the relationships between features.

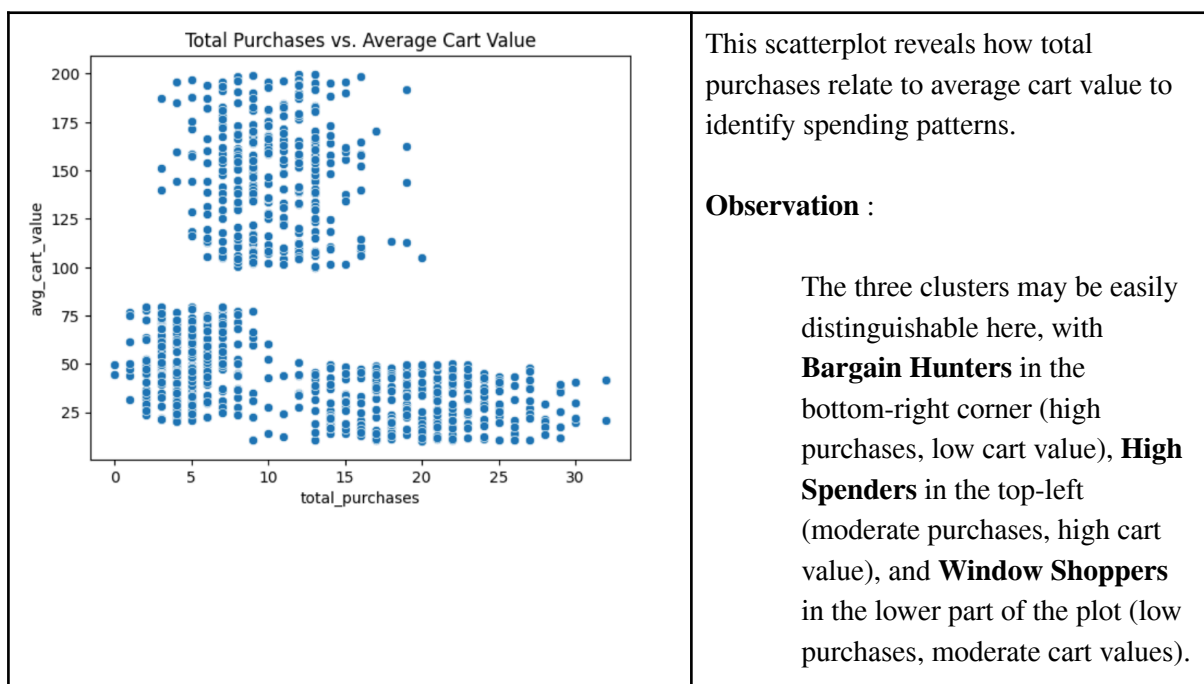
## 4.2 Distribution Analysis

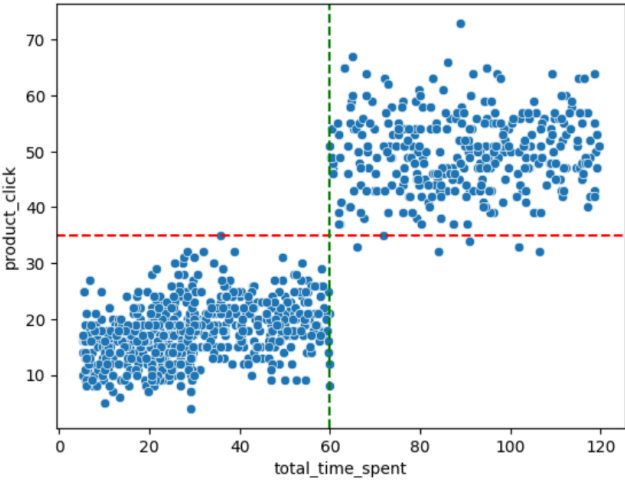
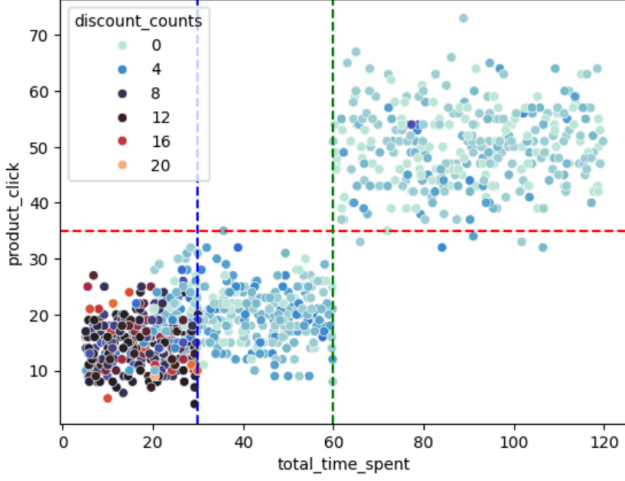
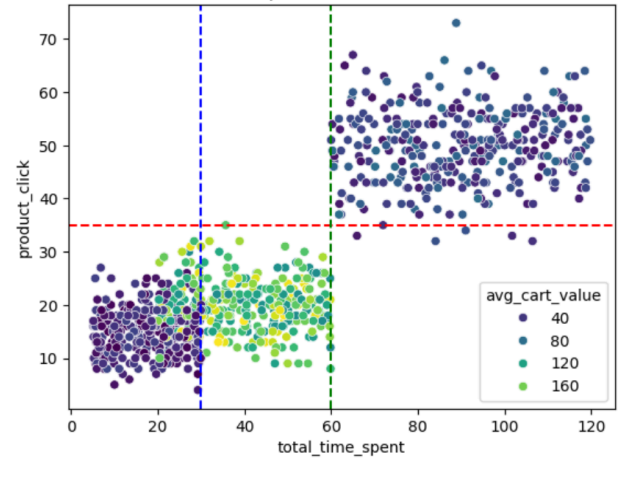
Histograms with kernel density estimation (KDE) were plotted for each numerical column to visualize their distributions.

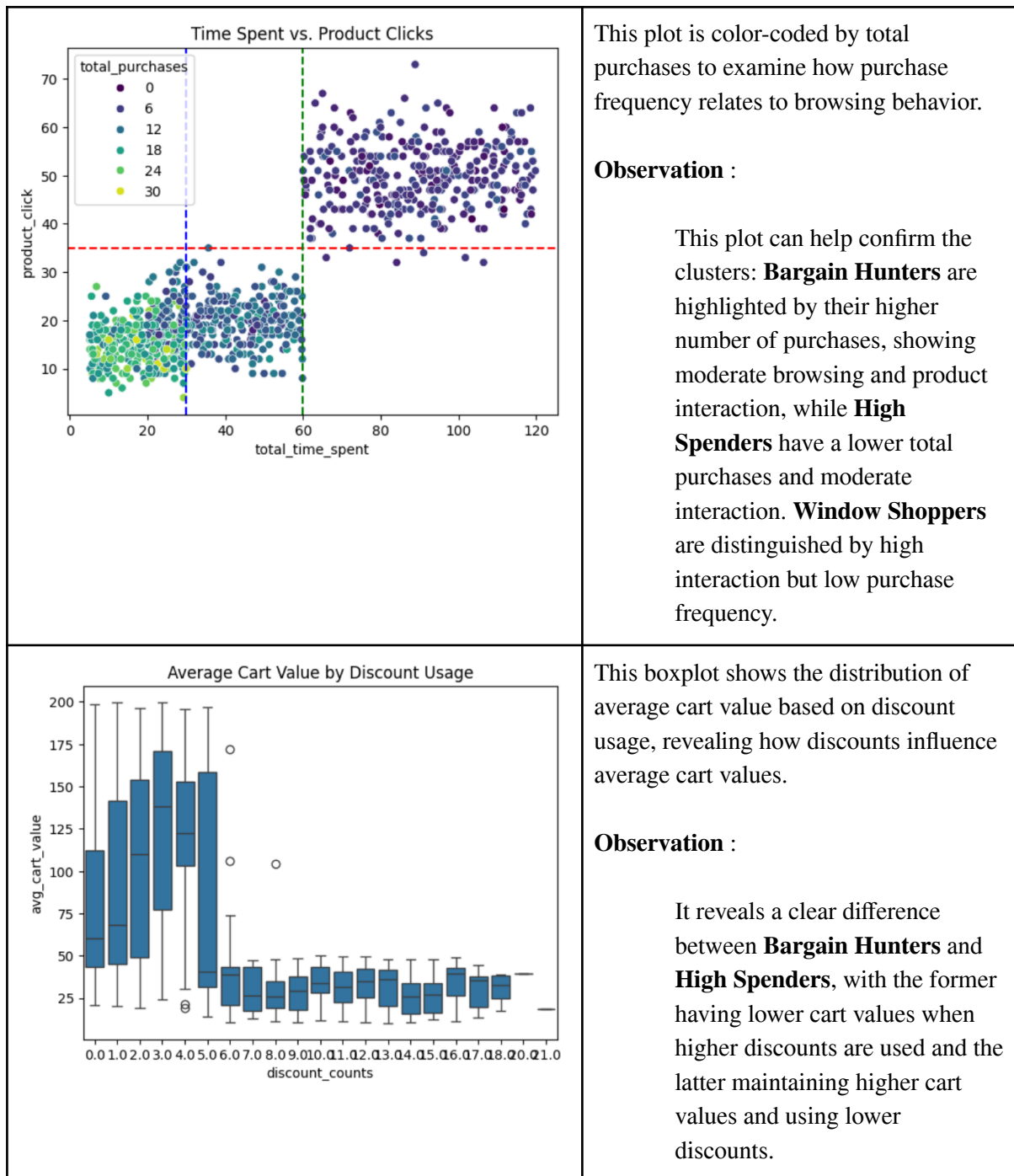


## 4.3 Analysis based on insights given

Scatterplots were generated to visualize the relationships between features.



	<p>This plot examines the relationship between time spent and the number of products clicked.</p> <p><b>Observation :</b></p> <p><b>Window Shoppers</b> have formed a distinct group with high values for both time spent and number of products clicked, while <b>Bargain Hunters</b> and <b>High Spenders</b> have clusters with moderate values for these features.</p>
	<p>This plot is color-coded by discount counts to highlight how frequently discounts are used while interacting with the platform.</p> <p><b>Observation :</b></p> <p>This plot can help confirm the three clusters: <b>Bargain Hunters</b> with moderate interaction and high discount usage, <b>High Spenders</b> with fewer interactions and lower discount usage, and <b>Window Shoppers</b> with extensive browsing but minimal discount usage.</p>
	<p>This plot is color-coded by average cart value .</p> <p><b>Observation :</b></p> <p>This visualization highlights the <b>High Spenders</b> as a distinct group with higher cart values but moderate interaction, while <b>Bargain Hunters</b> and <b>Window Shoppers</b> will have lower or similar cart values but differ in their browsing behavior.</p>



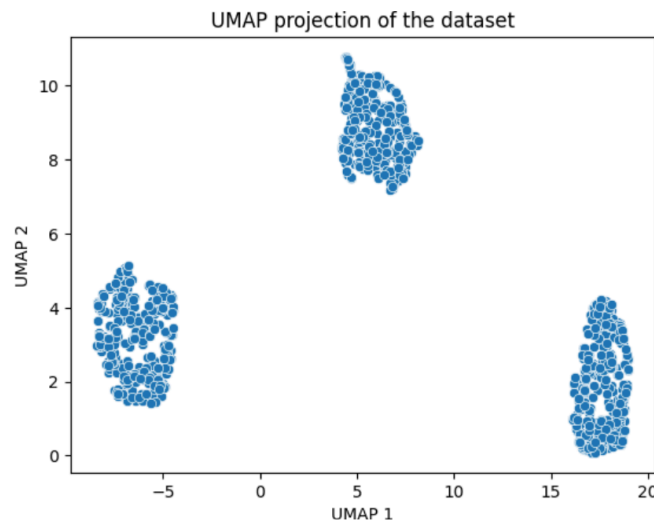
## 4.4 Dimensionality Reduction

In this analysis, dimensionality reduction techniques such as **UMAP** (Uniform Manifold Approximation and Projection) and **PCA** (Principal Component Analysis) were used to visualize the data distribution and identify potential clusters.

### 4.4.1 Data scaling

Before applying dimensionality reduction techniques, the dataset was standardized using **StandardScaler**. This step ensures that all features have a mean of 0 and a standard deviation of 1.

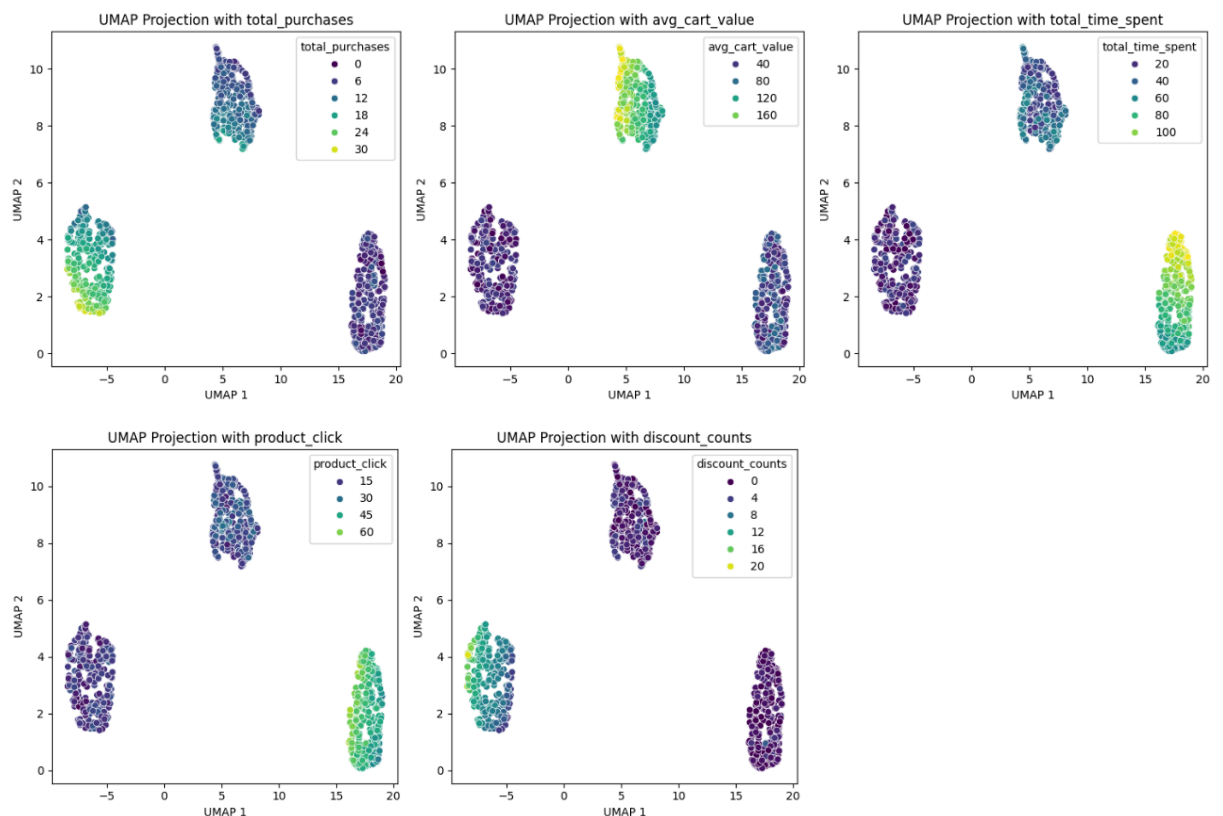
#### 4.4.2 UMAP projection



A 2D UMAP projection was generated for the dataset, revealing **a clear separation of the distinct clusters**.

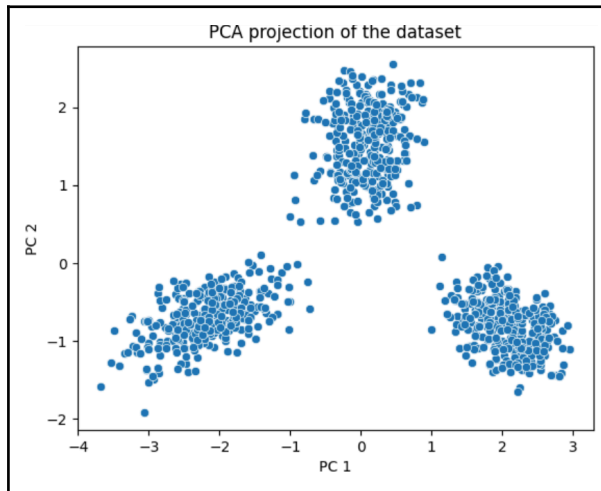
This UMAP plot provides an intuitive visualization of how customers are distributed into 3 clusters based on behavioral similarities. These clusters indicate that customers naturally form groups based on their purchasing habits, time spent on the platform, product engagement, and discount usage.

To better understand the contribution of each feature to these clusters, additional visualizations were created, where UMAP projections were colored based on individual features.





### 4.4.3 PCA projection



The 2D PCA projection of the dataset also showed a **clear separation of three clusters**, reinforcing the findings from UMAP.

Unlike UMAP, which captures non-linear relationships, PCA provides a global perspective of how variance is distributed in the dataset.

- By comparing UMAP and PCA projections, it's clear that there are **three well-separated customer groups** in the dataset.

## 5. Model Selection

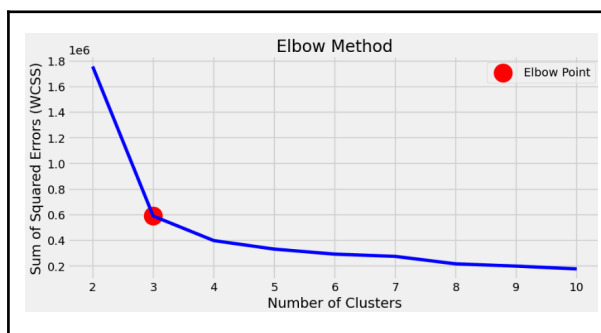
### 5.1 Determine optimal number of clusters

Before selecting a model, it was necessary to confirm that having **three clusters** was indeed the optimal choice for the number of clusters for this dataset.

To achieve this, multiple evaluation methods were used to determine the most optimal number of clusters.

#### 5.1.1 Elbow Method

The Elbow Method was applied, calculating *Within-Cluster Sum of Squares (WCSS)* to analyze the variance explained (inertia) for different values of  $k$ . The elbow point in the WCSS plot is used to determine the optimal  $k$ .

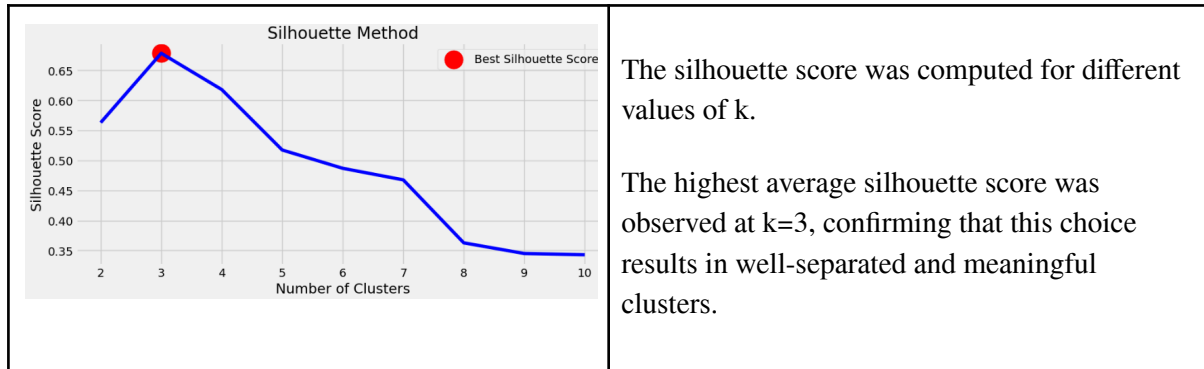


A noticeable "elbow" appeared at  **$k=3$** , indicating diminishing returns in variance reduction beyond this point.

This suggested that three clusters effectively balance compactness and separation in the data.

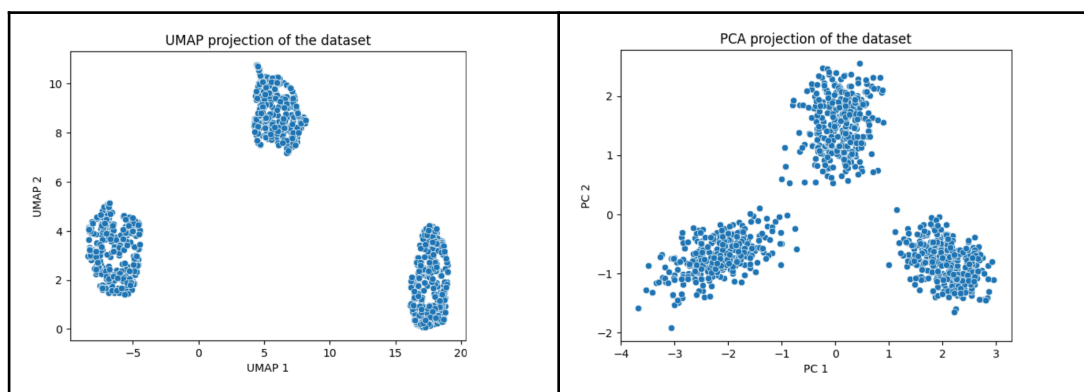
### 5.1.2 Silhouette Score Analysis

The Silhouette Score measures how well-defined clusters are by evaluating intra-cluster cohesion and inter-cluster separation.



### 5.1.3 Visual Inspection of UMAP and PCA Projections

Both projections consistently showed three distinct groups, further validating the choice of k=3.

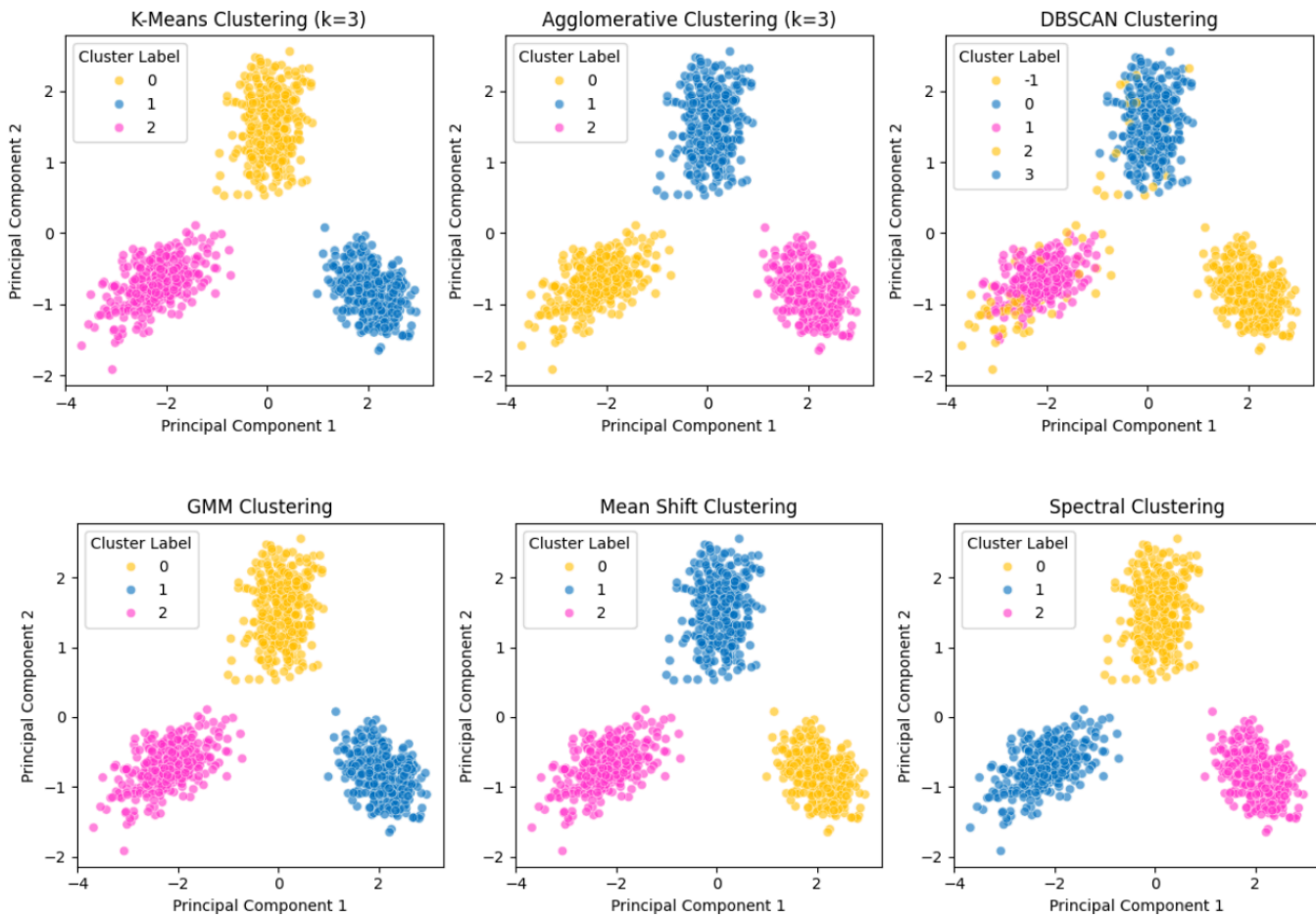


Based on these methods, **k=3** was confirmed as the **optimal number of clusters**, ensuring a robust segmentation of customers.

## 5.2 Determine best performing clustering model

To determine the most suitable clustering model for customer segmentation, the following clustering algorithms were tested and compared.

1. K-Means Clustering
2. Agglomerative Clustering
3. DBSCAN
4. Gaussian Mixture Model (GMM)
5. Mean Shift Clustering
6. Spectral Clustering



Visualizing Clusters	
K-Means Clustering (k=3)	The scatter plot for <b>K-Means clustering</b> showed three <b>well-separated clusters</b> , confirming its suitability.
Agglomerative Clustering (k=3)	Similar cluster patterns were observed, validating the <b>hierarchical structure</b> in the data.
DBSCAN	DBSCAN identified <b>noise points</b> , visible as unclustered data in the scatter plot. The method struggled with uniform customer distributions, making it less suitable.
Gaussian Mixture Model (GMM)	The model grouped customers into three well-separated clusters.
Mean Shift Clustering	Similar cluster patterns were observed
Spectral Clustering	Similar cluster patterns were observed

## 6. Model Evaluation

To assess the quality of clustering models, multiple evaluation metrics: **Silhouette Score**, **Calinski-Harabasz Index** and **Density-Based Clustering Validation (DBCV)** are utilized. These metrics provide insights into cluster cohesion, separation, and overall structure.

The computed evaluation scores for each clustering model are summarized below :

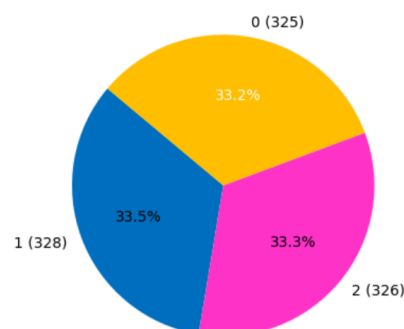
	Model	Silhouette Score	Calinski-Harabasz Index	DBCV Score
0	K-Means (k=3)	0.626018	2452.116045	0.494025
1	Agglomerative (k=3)	0.626018	2452.116045	0.494025
2	DBSCAN	0.365837	784.674716	0.394705
3	GMM	0.626018	2452.116045	0.494025
4	MeanShift	0.626018	2452.116045	0.494025
5	SpectralClustering	0.626018	2452.116045	0.494025

### Key Observations:

- **k-Means, Agglomerative Clustering, GMM, Mean Shift, and Spectral Clustering** all achieved the same scores across three evaluation metrics, indicating similar performance in separating the clusters.
- **DBSCAN had the lowest silhouette score (0.366)**, which is lower than that of K-Means, indicating **weaker separation** between clusters. Also it has the **lowest DBCV Score (0.395)**, which implies DBSCAN has **weaker clustering quality** than the others.
- **k-Means was chosen as the final model** because of its efficiency, ease of interpretation, and well-defined cluster boundaries, as seen in the PCA and UMAP visualizations.

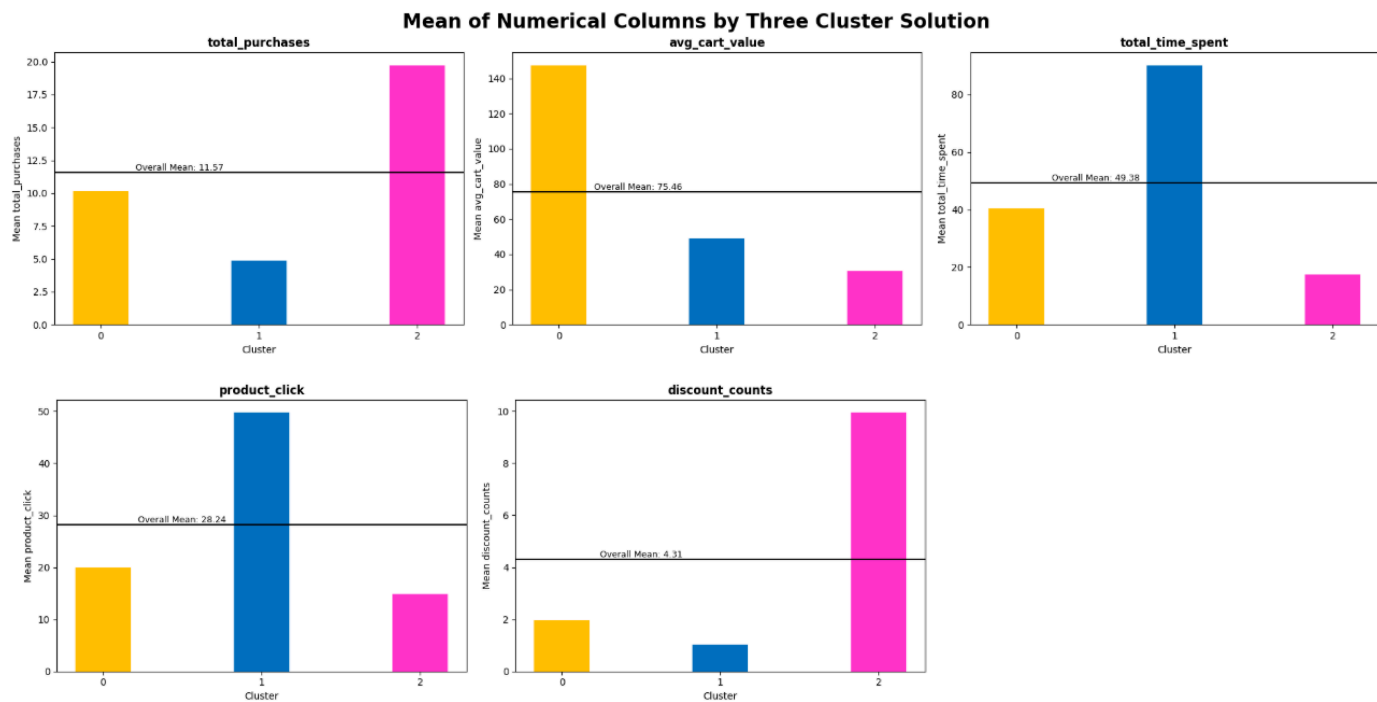
## 7. Identifying Clusters

To understand the distribution of customers across the identified clusters, **the proportion of customers in each group** was calculated..



This provides insights into which segment dominates the platform's user base and helps in targeting marketing strategies effectively.

To further analyze the differences between clusters, a bar plot analysis was conducted on numerical features. Below, bar plots were generated for each feature, comparing their average values across the identified clusters.



## 7.1 Key Observations

- **Total Purchases:** Cluster 2 had the highest total purchases, whereas Cluster 1 had the lowest.
- **Average Cart Value:** Cluster 0 showed significantly higher cart values, indicating a preference for premium products.
- **Total Time Spent:** Cluster 0 spent more time on the platform compared to Cluster 2, but Cluster 1 exhibited the highest engagement level.
- **Product Clicks:** Cluster 1 had a higher product exploration tendency, while Cluster 0 and Cluster 2 had a moderate engagement.
- **Discount Usage:** Discount redemption was most frequent among Cluster 2 customers, highlighting their responsiveness to promotional offers.

Based on the feature analysis, customer personas were assigned to each cluster:

Cluster	Observation	Identification
Cluster 0	<ul style="list-style-type: none"> <li>● Make moderate purchases with higher cart values.</li> <li>● Spend moderate time on the platform.</li> <li>● Rarely use discount codes.</li> </ul>	High Spenders

Cluster1	<ul style="list-style-type: none"> <li>• View more products before purchasing.</li> <li>• Spend a high amount per transaction.</li> <li>• Spend a significant amount of time on the platform.</li> </ul>	Window Shoppers
Cluster 2	<ul style="list-style-type: none"> <li>• Make frequent, low-value purchases.</li> <li>• Spend a moderate amount of time on the platform.</li> <li>• Engage deeply and often use discounts</li> </ul>	Bargain Hunters

## 8. Conclusion

The clustering analysis successfully segmented customers into three distinct groups based on their purchasing behavior, engagement, and discount usage. Through **UMAP and PCA projections**, the dataset showed a clear **three-cluster structure**, which was further validated using the **elbow method, silhouette scores, and model comparisons**.

Among the clustering models tested, **k-Means, Agglomerative Clustering, and GMM** performed best, each achieving a **silhouette score of 0.626, Calinski-Harabasz Index: 2452.12** and **DBCV Score: 0.494** indicating well-separated clusters. **DBSCAN performed poorly** due to its sensitivity to noise, while Mean Shift and Spectral Clustering showed similar patterns to K-Means.

These findings provide a data-driven foundation for targeted marketing strategies, allowing businesses to personalize promotions and optimize customer engagement.

## 9. Challenges Faced

### 1. Visualizing High-Dimensional Data:

While UMAP and PCA were effective in reducing dimensionality, interpreting the reduced components can be abstract. Understanding which original features contribute most to the principal components would have provided deeper insights.

### 2. Interpreting DBSCAN Results:

DBSCAN did not give a set number of clusters, which made it harder to compare to the other models.

## 10. Suggestions for Improvement

### 1. **Feature Engineering :**

Incorporating additional meaningful features, such as customer behavior trends, could improve clustering performance.

### 2. **Automated Hyperparameter Tuning :**

Using techniques like Grid Search or Bayesian Optimization to fine-tune clustering algorithms for improved results.

### 3. **Advanced Clustering Techniques :**

Exploring deep learning-based clustering (e.g., Autoencoders + K-Means) might provide better-defined clusters.

## 11. References

- [1] <https://scikit-learn.org/stable/modules/clustering.html>
- [2] <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>
- [3] [https://umap-learn.readthedocs.io/en/latest/basic\\_usage.html](https://umap-learn.readthedocs.io/en/latest/basic_usage.html)
- [4] <https://github.com/christopherjenness/DBCV>
- [5] <https://scikit-learn.org/stable/modules/clustering.html#calinski-harabasz-index>