ibimbing

MAXIMIZING REVENUE WITH CUSTOMER SEGMENTATION

USING AMAZON SALES DATA FOR BETTER MARKETING

DS 27B Final Project

PROJECT OVERVIEW

ABOUT CUSTOMER SEGMENTATION

Customer segmentation is the process of dividing customers into different groups based on common characteristics or behaviors, such as how often they make purchases, how much they spend, and which products they prefer.

OBJECTIVE

The main objective of this project is to build a clustering model that will divide Amazon's customers into meaningful segments based on their purchasing behavior. Using this model, we aim to develop targeted marketing strategies to increase customer engagement and drive higher revenue.

GOAL

Automated Segmentation

Use clustering algorithms to automatically segment customers based on their buying behaviors, allowing for more effective targeting.

Increase Revenue

Develop marketing strategies tailored to each customer segment to increase purchases and average order value.

Customer Retention

Identify high-value customers and create retention strategies to maintain long-term relationships with them.

EVALUATION METRIC

- **Silhouette Score**
- **Davies-Bouldin Index**
- **Calinski-Harabasz Score**

DATASET OVERVIEW

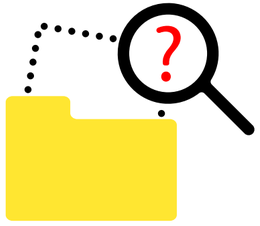
ABOUT

- This dataset contains sales transaction data from Amazon during 2020 and 2021. It includes various details related to the products purchased, order statuses, customer information, and shipping details.
- The dataset has 35 columns, covering information about products, transaction statuses, customers, and shipping locations.

KEY COLUMNS

- Transaction Details:
  - Order ID, Order Date, Item ID, Quantity Ordered, Price, Payment Method and Total Value.
- Customer Information:
  - Customer ID, Full Name, Gender, Age, Region, and Contact Information.
- Product Details:
  - SKU, Category, and Discount Information.
- Geographical and Temporal Data:
  - City, State, Region, Zip Code, Year, and Month of Purchase.

DATA CLEANING



HANDLE MISSING VALUE

Identify and address missing values using deletion or imputation methods.

Result: No missing values found.

HANDLE DUPLICATED

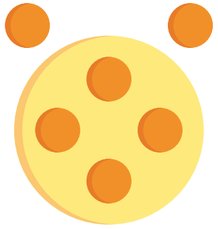
Detect and remove duplicate entries to ensure data integrity.

Result: No duplicates found.

HANDLE DATA TYPE

Ensure each column has the correct data type for accurate analysis.

Action: Fix date formats as needed.



HANDLE OUTLIER

Identify and manage outliers using techniques like Isolation Forest.

Action: Remove 6425 outliers
.

EXPLORATORY DATA ANALISYS

- DESCRIPTIVE ANALYSIS
- SALES PERFORMANCE ANALYSIS
- CUSTOMER INSIGHTS AND SEGMENTATION
- DISCOUNT AND PRICING STRATEGY

DESCRIPTIVE ANALYSIS

NUMERICAL COLUMN

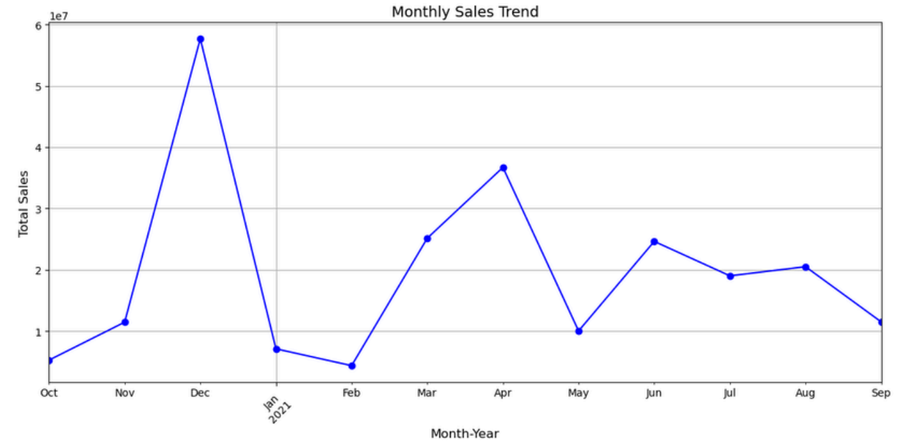
Column Name	Count	Mean	Std Dev	Min	25%	50%	75%	Max
qty_ordered	286392	3.01	4.57	1	2	2	3	100
price	286392	851.39	1741.75	0.00	49.90	99.90	199.80	9999.99
value	286392	885.88	2073.25	0.00	49.90	99.90	199.90	20000.00
discount_amount	286392	70.04	256.88	0.00	0.00	0.00	100.00	5000.00
total	286392	815.84	1983.58	0.00	49.90	99.90	199.90	10000.00
age	286392	46.49	16.67	18	32	46	60	120
Discount_Percent	286392	6.07	10.10	0	0	0	10	100

CATEGORICAL COLUMN

Column Name	order_id	status	sku	category	payment_method	bi_st	month	Gender	Sign in date	Phone No.	Place Name	County	City	State	Region
Count	286392	286392	286392	286392	286392	286392	286392	286392	286392	286392	286392	286392	286392	286392	286392
Unique	201716	13	47932	15	13	3	12	2	11647	64248	15892	2551	15892	51	4
Top	100476608	canceled	MATSAM59DB75ADB2F80	Mobiles & Tablets	cod	Gross	Dec-20	M	11/30/2005	217-861-7640	Dekalb	Jefferson	Dekalb	TX	South
Freq	43	112166	3775	61761	102916	112333	82528	146184	2536	2524	2525	3510	2525	17510	103482

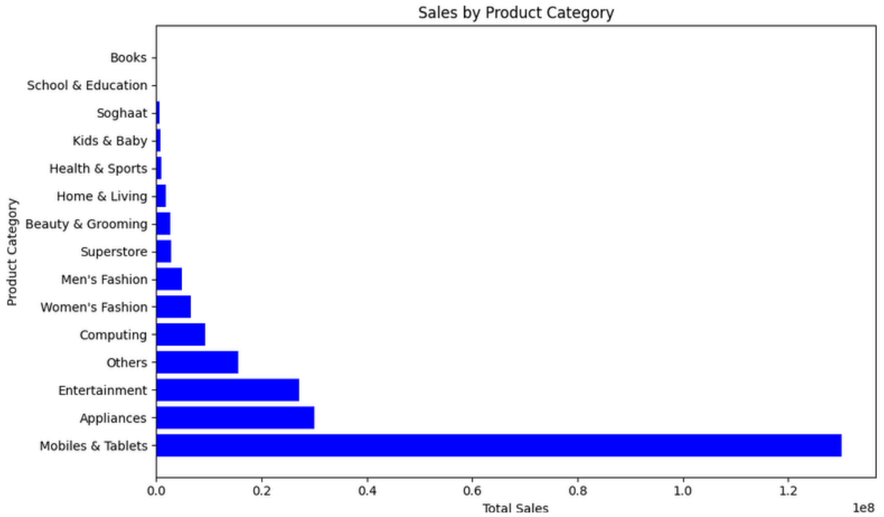
SALES PERFORMANCE ANALYSIS

TOTAL SALES PER PERIOD

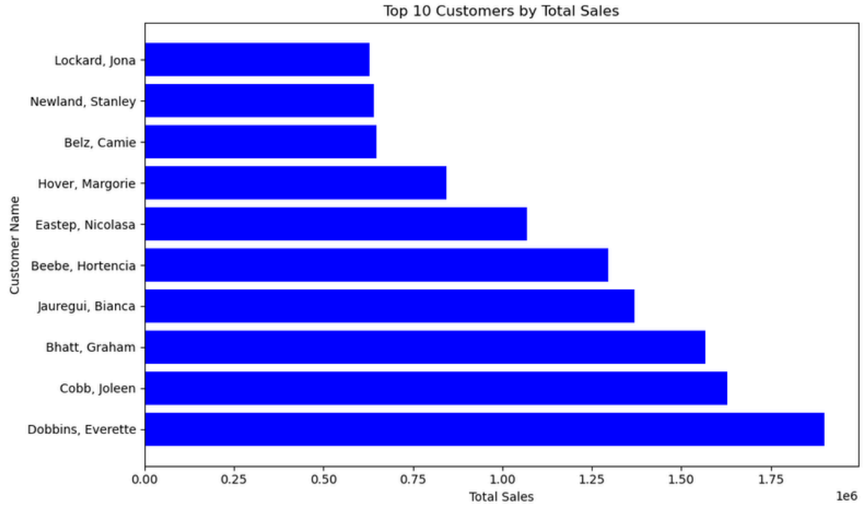


- **There is a significant peak in December, indicating a possible seasonal or promotional factor (e.g., holidays like Christmas or end-of-year sales).**
- **After December, sales drop sharply in January but recover steadily in subsequent months (February to April), with another smaller peak in April.**

SALES BY PRODUCT CATEGORY AND CUSTOMER



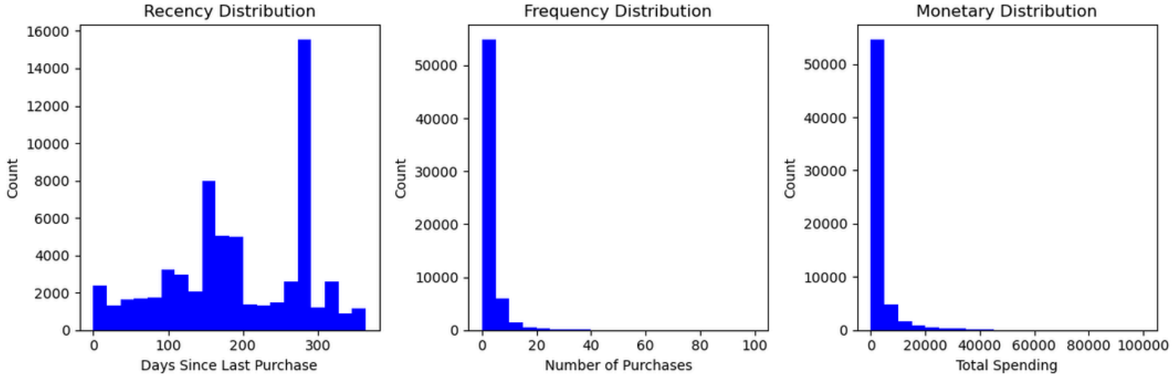
- **In Sales by Product Category, Mobiles & Tablets bring in the most sales, showing strong demand.**



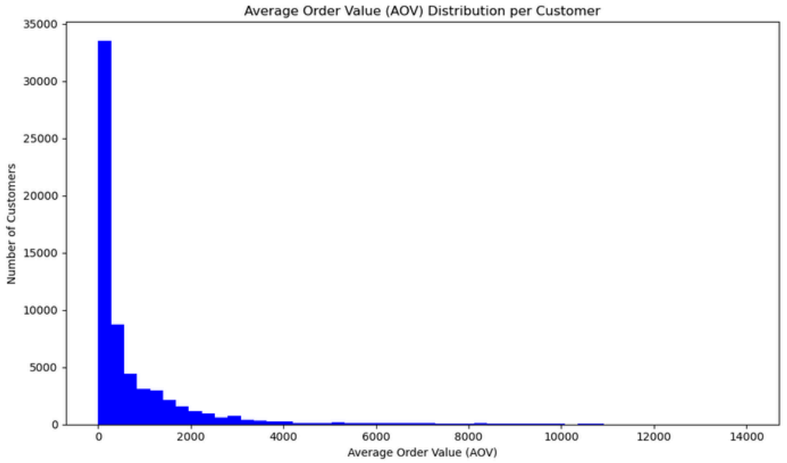
- **These are the customers who contribute most significantly to total sales with the highest spend.**

CUSTOMER INSIGHTS AND SEGMENTATION

RFM & AOV

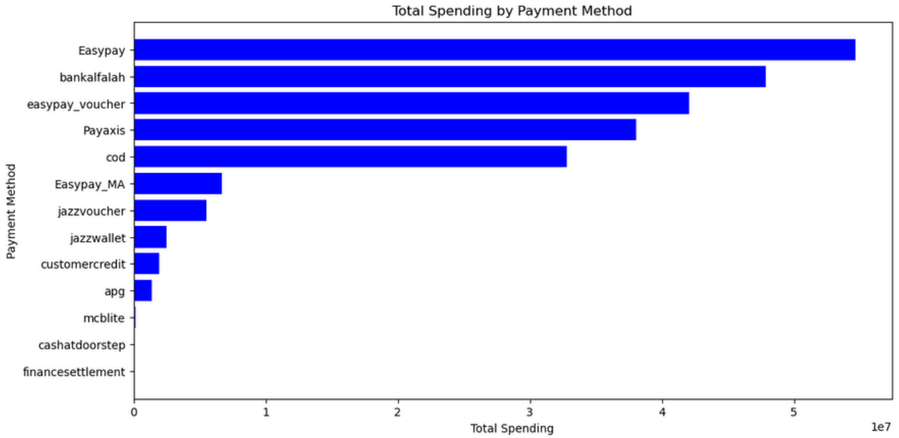


- **Recency Distribution:** Most customers made their last purchase between 200-300 days ago, with a smaller group purchasing recently.
- **Frequency Distribution:** A majority of customers made fewer than 10 purchases, indicating that most are occasional buyers.
- **Monetary Distribution:** The majority of customers have total spending below $20,000, with a small number contributing significantly higher amounts.



- The majority of customers have an **average order value below $2,000**, with a very small number of customers making significantly higher-value purchases. This indicates that most customers tend to make smaller transactions.

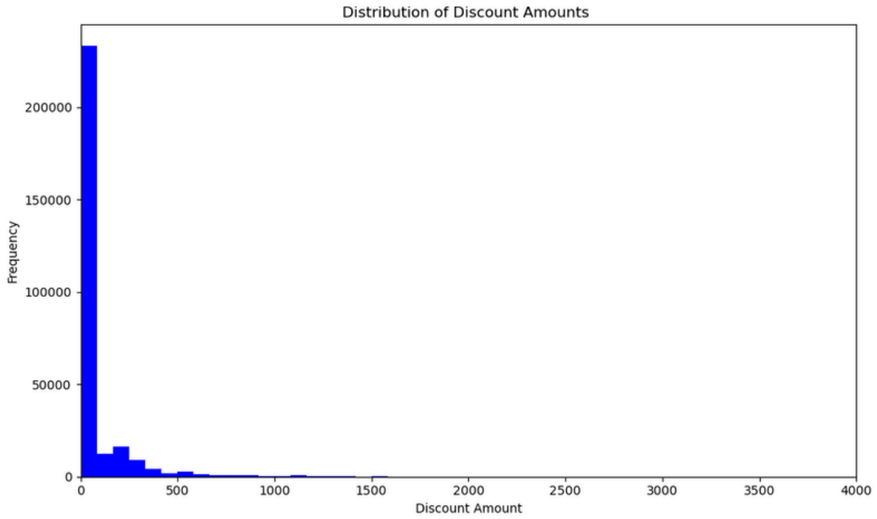
PAYMENT METHOD PREFERENCES



- **Easypay** is the most popular payment method, generating the highest total spending
- **Cash on Delivery (COD)** is still widely used but lags behind digital payment methods in terms of total spending.
- **Digital payment** options like Easypay and Payaxis dominate customer preferences, reflecting a shift towards online transactions.

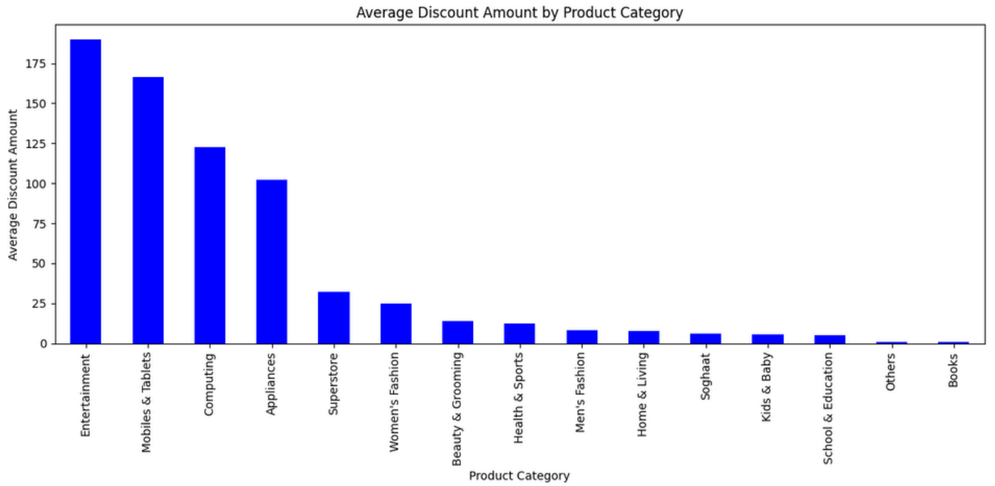
DISCOUNT AND PRICING STRATEGY

DISCOUNT AMOUNT DISTRIBUTION



- A large number of orders received low discounts, with most discounts concentrated below $500.
- There are very few instances of higher discounts, indicating that larger discounts are rare.

AVERAGE DICOUNT AMOUNT



- **Entertainment** has the highest average discount, significantly surpassing other categories.
- **Mobiles & Tablets** and **Computing** also show substantial average discounts, indicating competitive pricing in these sectors.
- Categories like **Books**, **School & Education**, and Others have much lower average discounts, suggesting less promotional activity or lower price reductions.

FEATURE ENGINEERING

- RFM (RECENCY, FREQUENCY, MONETARY)
- DISCOUNT FEATURES
- PURCHASE BEHAVIOR
- GEOGRAPHICAL FEATURES
- DEMOGRAPHIC FEATURES
- TEMPORAL FEATURES
- PROFITABILITY FEATURES
- PRODUCT INTERACTION
- LOYALTY AND ENGAGEMENT

RFM (RECENCY, FREQUENCY, MONETARY)

- **Recency Days**: Number of days since the customer's last transaction
  - Formula: Current Date - Last Purchase Date
- **Frequency**: Total number of purchases made by the customer.
  - Formula: Count of Unique Transactions
- **Monetary**: Total revenue generated by the customer.
  - Formula: Sum of Total Purchase Value

DISCOUNT FEATURES

- **Average Discount Percent**: Average discount percentage utilized by the customer.
  - Formula: Sum of Discount Percentages / Total Transactions
- **Total Discount Used**: Total discount value availed by the customer.
  - Formula: Sum of Discount Amounts

PURCHASE BEHAVIOR

- **Average Order Value**: Average value of a single purchase.
  - Formula: Total Monetary Value / Total Transactions
- **Average Quantity Ordered**: Average number of items purchased per transaction.
  - Formula: Sum of Quantities Ordered / Total Transactions
- **Diversity of Categories**: Total number of unique product categories bought by the customer.
  - Formula: Count of Unique Categories Purchased
- **Preferred Payment Method**: The most frequently used payment method by the customer (e.g., Credit Card, PayPal).

GEOGRAPHICAL FEATURES

- **Top Region**: The region where the customer made the most purchases.
- **Region Purchase Concentration**: Proportion of purchases from the customer's most preferred region.
  - Formula: Transactions in Top Region / Total Transactions

DEMOGRAPHIC FEATURES

- **Age Group**: Age of the customer grouped into ranges (e.g., 20-30, 30-40).

TEMPORAL FEATURES

- **Preferred Day of Week**: The day of the week when the customer most frequently makes purchases.

PROFITABILITY FEATURES

- **Customer Lifetime Value**: Total revenue generated during the customer's lifetime.
  - Formula: Sum of Total Purchases Over Lifetime

PRODUCT INTERACTION

- **Top Selling Category**: The product category purchased most frequently by the customer.
- **Product Diversity Index**: Ratio of unique categories to total purchases.
  - Formula: Diversity of Categories / Total Transactions

LOYALTY AND ENGAGEMENT

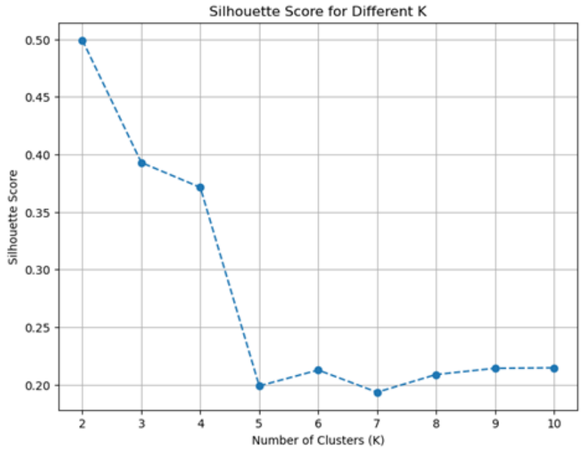
- **Customer Loyalty Index**: Measures customer loyalty based on frequency of transactions over time.
  - Formula: Frequency / Days Since Last Sign-In
- **Days Since Sign-In**: Number of days since the customer last logged in.
  - Formula: Current Date - Last Sign-In Date

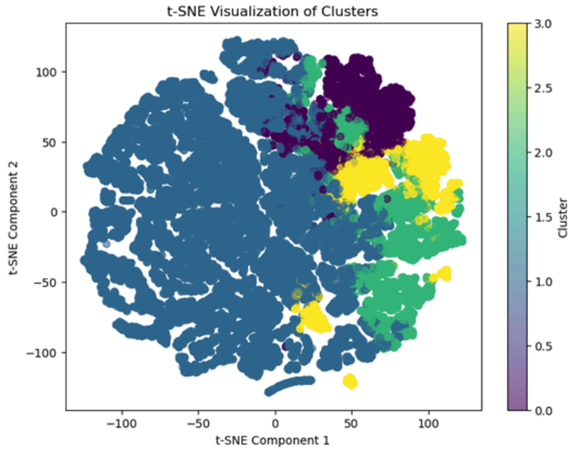
MODELLING

- CENTROID-BASED OR PARTITIONING CLUSTERING
- DENSITY-BASED CLUSTERING

CENTROID-BASED CLUSTERING

KMEANS CLUSTERING

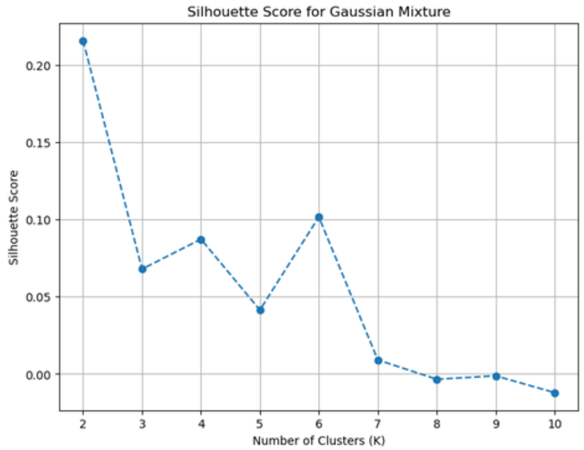


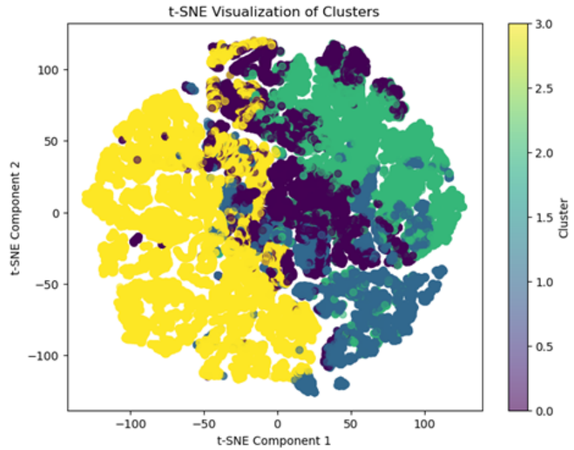


Amount of Data per Cluster:

Cluster	Amount
0	5346
1	41721
2	6558
3	4198

GAUSSIAN MIXTURE MODEL



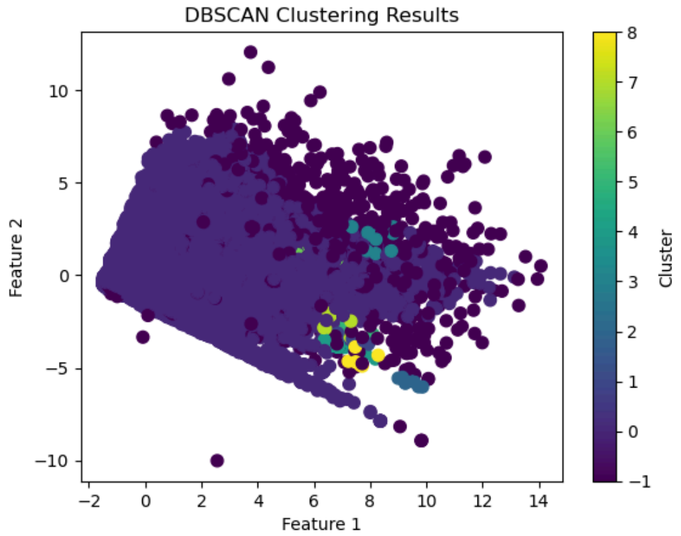


Amount of Data per Cluster:

Cluster	Amount
0	12152
1	9157
2	11091
3	25423

DENSITY-BASED CLUSTERING

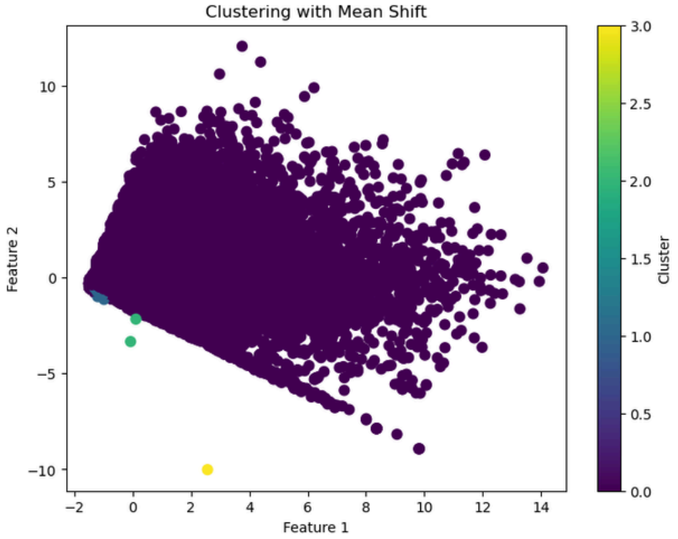
DBSCAN



Amount of Data per Cluster:

Cluster	Amount
-1	825
0	56910
1	16
2	8
3	11
4	23
5	6
6	8
7	9
8	7

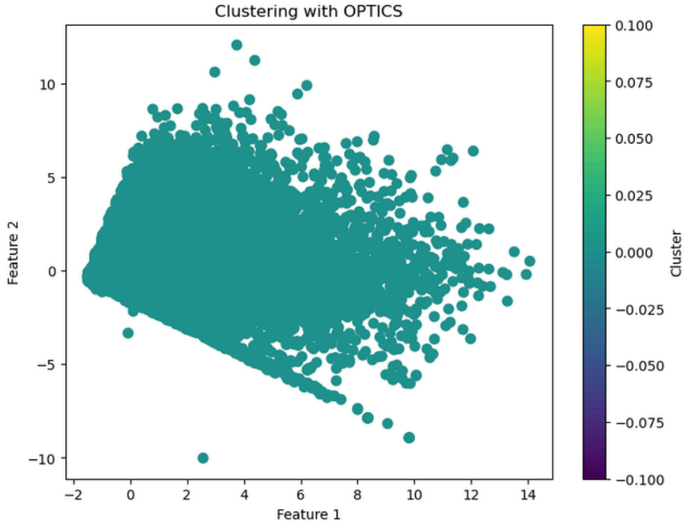
MEAN SHIFT



Amount of Data per Cluster:

Cluster	Amount
0	57803
1	17
2	2
3	1

OPTICS



Amount of Data per Cluster:

Cluster	Amount
0	57823

BEST MODEL : KMEANS

Why K-Means was Chosen:

- **Performance:**
  - K-Means achieved the highest silhouette score among all the tested clustering algorithms.
  - The silhouette score indicates how well-separated and cohesive the clusters are. A higher score means better-defined clusters.
- **Cluster Balance:**
  - K-Means provided a relatively balanced distribution of data points across the 4 clusters:
    - Cluster 0: 5,346 data points
    - Cluster 1: 41,721 data points
    - Cluster 2: 6,558 data points
    - Cluster 3: 4,198 data points
  - This balance is important for designing effective marketing and business strategies, as each cluster can be targeted with specific campaigns without overloading or ignoring any group.
- **Visualization:**
  - The t-SNE visualization shows distinct, well-separated clusters in the dataset. This makes it easier to interpret customer groups and their unique characteristics.

EVALUATION

SILHOUETTE SCORE : 0.34

This score indicates a moderate quality of clustering. Values closer to 1 indicate better clustering, while a score of 0.34 indicates that some points may be near cluster boundaries, indicating potential overlap.

DAVIES-BOULDIN INDEX : 1.47

This index measures the average similarity ratio of each cluster to its most similar cluster. Lower values are better, and 1.47 indicates that there is some overlap between clusters, but not too much.

CALINSKI-HARABASZ SCORE : 11969.79

This score reflects the ratio of the variance between clusters to the variance within clusters. A higher score indicates better-defined clusters. A score of 11969.79 indicates that the clusters are relatively well separated.

RESULT

- **CHARACTERISTICS PER CLUSTER**

CLUSTER OVERVIEW

Cluster	Total Revenue	Total Customers	AOV (Avg Order Value)	Key Characteristics
Cluster 0	7.42M	5,346	195.31	High loyalty, low purchase frequency.
Cluster 1	21.21M	41,720	334.81	High volume, largest customer group.
Cluster 2	15.06M	6,558	1,000	High discounts, focused on technology.
Cluster 3	32.43M	4,198	2,850	Premium customers with high value orders.

CLUSTER 0 : VALUE SEEKERS

CLUSTER

Select all

0

1

2

3

Total Revenue	Number of Purchases	Customer Lifetime Value	Average Quantity Ordered
7.42M	46K	12.55K	2.40

Customer Loyalty Index	Average Discount Percent	Total Discount Used	Average Order Value
12.55K	3.08	360.44K	195.31

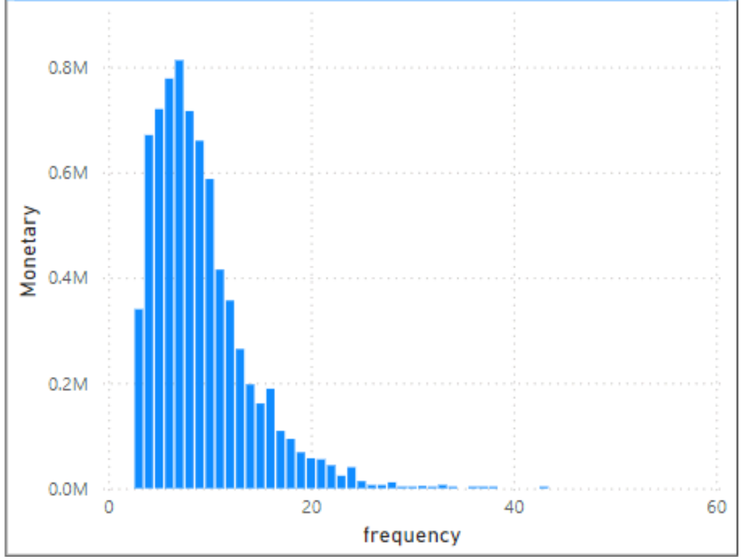
Total Customer

5346

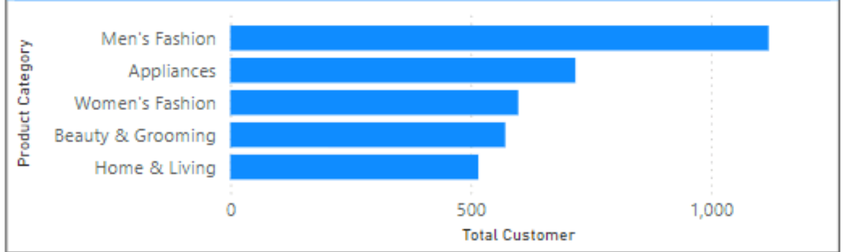
Days Since Last Purchase

166

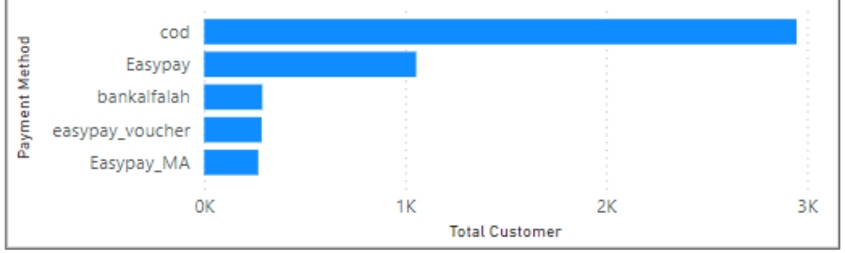
Monetary Distribution by Frequency



Top 5 Product Category



Top 5 Payment Method



CLUSTER 1 : MASS BUYERS

CLUSTER

Select all

0

1

2

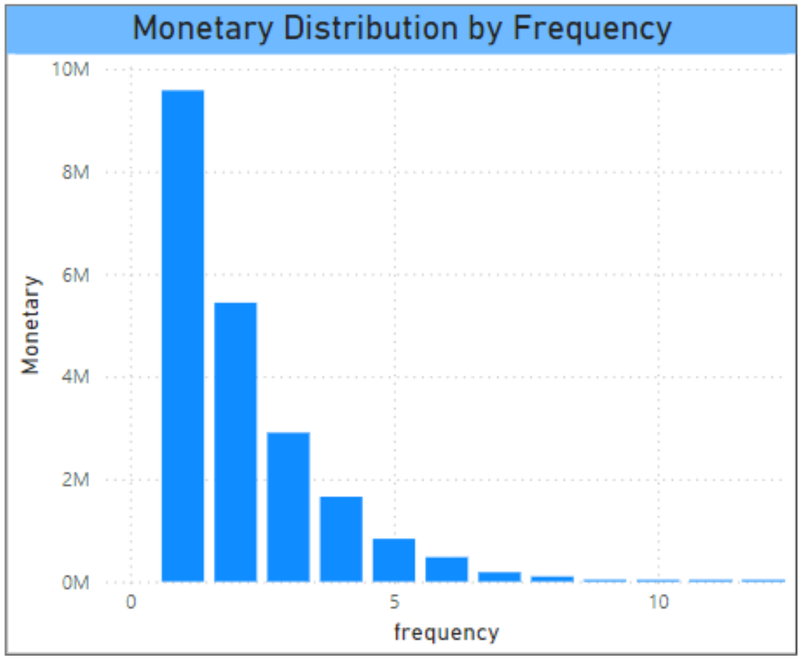
3

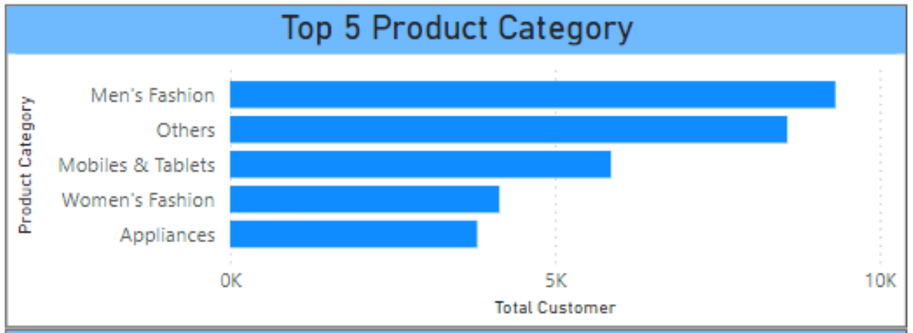
Total Revenue	Number of Purchases	Customer Lifetime Value	Average Quantity Ordered
21.21M	77K	1.08K	2.44

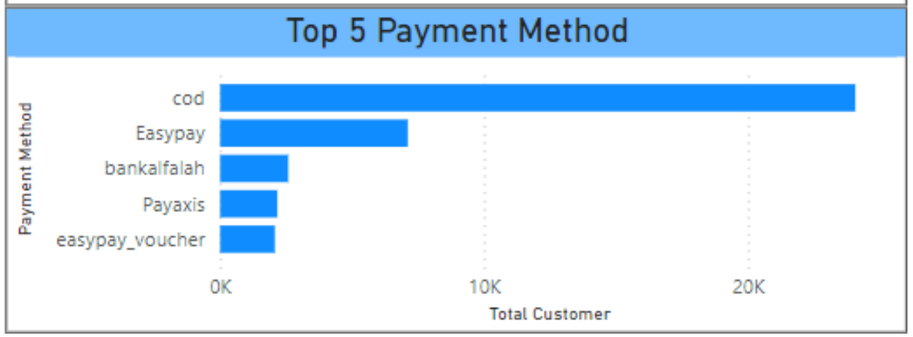
Customer Loyalty Index	Average Discount Percent	Total Discount Used	Average Order Value
1.08K	0.94	137.75K	334.81

Total Customer
41.72K

Days Since Last Purchase
187







CLUSTER 2 : DISCOUNT ENTHUSIASTS

CLUSTER

Select all

0

1

2

3

Total Revenue	Number of Purchases	Customer Lifetime Value	Average Quantity Ordered
15.06M	19K	7.33K	2.47

Customer Loyalty Index	Average Discount Percent	Total Discount Used	Average Order Value
7.33K	16.39	2.91M	1.00K

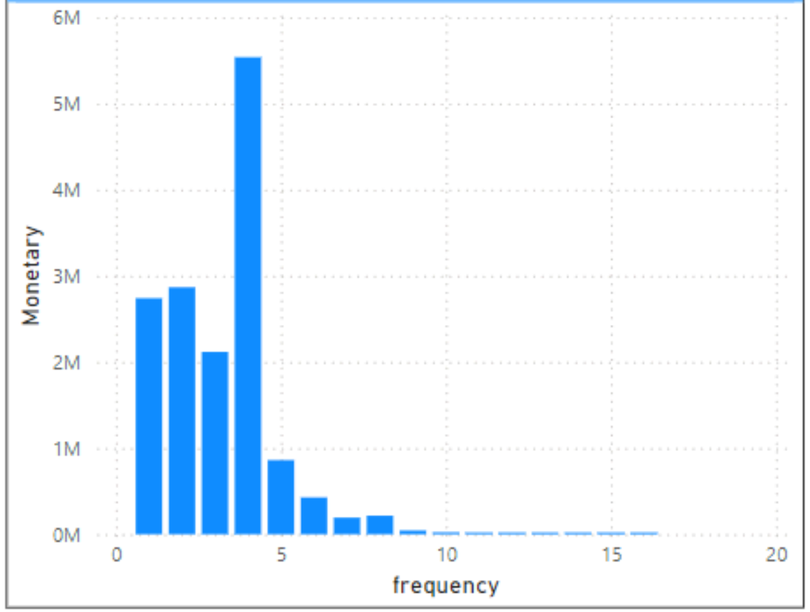
Total Customer

6558

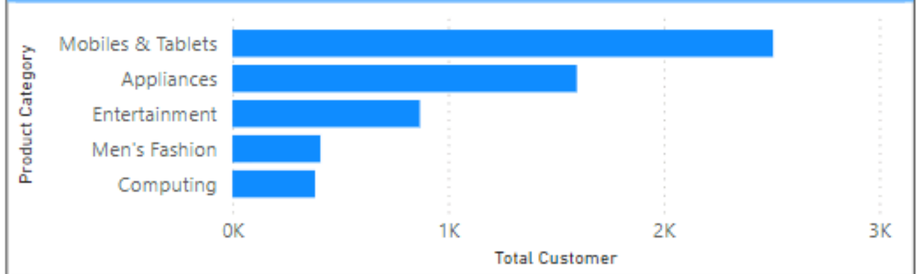
Days Since Last Purchase

277

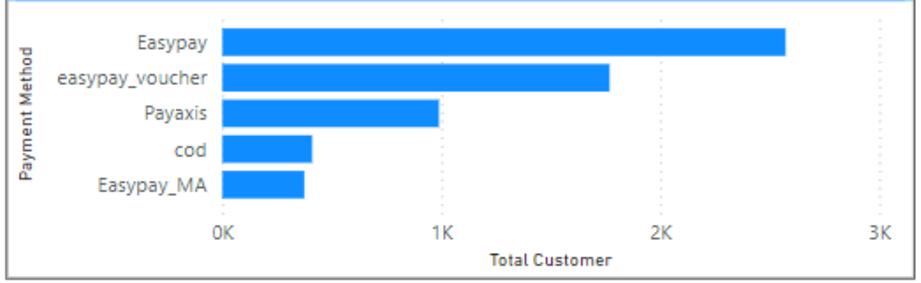
Monetary Distribution by Frequency



Top 5 Product Category



Top 5 Payment Method



CLUSTER 3 : PREMIUM LOYALISTS

CLUSTER

Select all

0

1

2

3

Total Revenue	Number of Purchases	Customer Lifetime Value	Average Quantity Ordered
32.43M	19K	34.65K	2.92

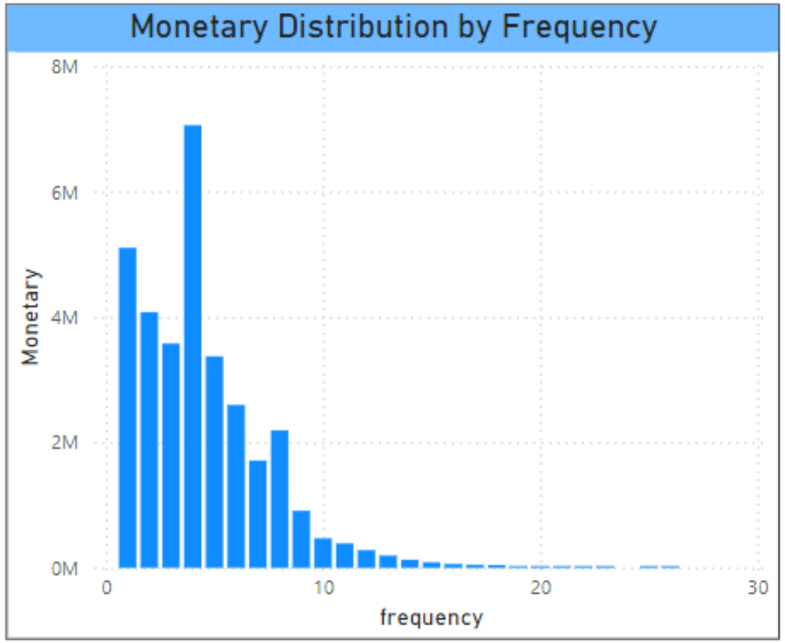
Customer Loyalty Index	Average Discount Percent	Total Discount Used	Average Order Value
34.65K	6.34	2.48M	2.85K

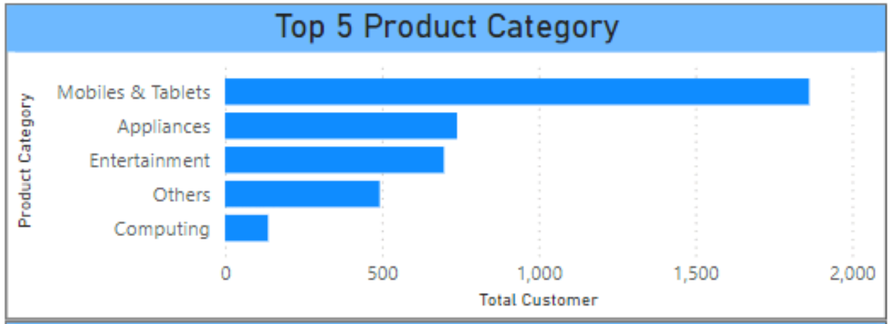
Total Customer

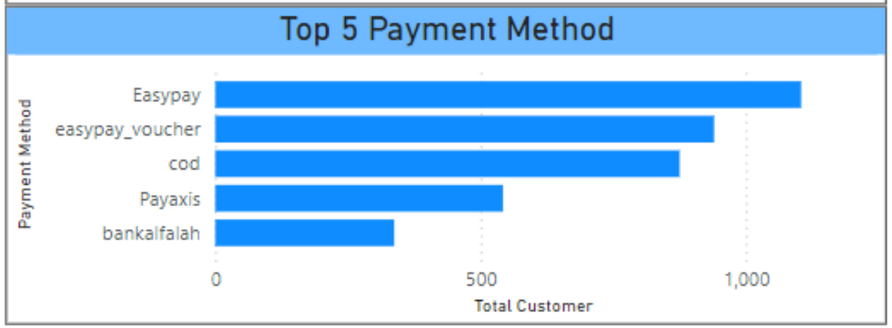
4198

Days Since Last Purchase

182







SIMULATION

- **STRATEGY FOR EACH CLUSTER**

CLUSTER 0 :

Goals:

1. Increase purchase frequency.
2. Increase Average Order Value (AOV).

Strategies:

- Offer small discounts (2%-5%) to boost order volume.
- Promote popular categories such as Men’s Fashion.

Why this work?

- Small discounts create incentives for customers to purchase more frequently without heavily impacting profit margins.
- This cluster has low AOV, so the strategy focuses on volume growth.

Simulation :

Assumptions:

- Current Data:
  - Total Customers: 5,346
  - Total Revenue: 7.42M
  - Number of Purchases: 46K
  - Average Order Value (AOV): 195.31
- Small discounts (e.g., 2%-5%) increase the frequency of purchases but have minimal impact on AOV.

Scenario	New Frequency	New AOV	New Revenue	Increase
Discount +2%	48.3K	195.31	9.43M	+26%
Discount +5%	48.3K	208.98	10.09M	+36%

CLUSTER 1 :

Goals:

1. Reduce customer churn.
2. Increase AOV through upselling.

Strategies:

- Introduce loyalty programs for dormant customers (>180 days).
- Upsell premium products such as Mobiles & Tablets.

Simulation :

Assumptions:

- Current Data:
  - Total Customers: 41,720
  - Total Revenue: 21.21M
  - AOV: 334.81
  - Number of Purchases: 77K.
- Loyalty programs help re-engage dormant customers (~5% of total customers).
- Upselling increases AOV by 10% through premium products.

Why this work?

- Loyalty programs re-engage customers who are unlikely to purchase without incentives.
- Upselling focuses on increasing basket size, leveraging existing customers without significantly increasing marketing costs.

Scenario	New Customers	New AOV	New Revenue	Increase
Loyalty Program	2.09K	368.29	29.89M	+41%
Premium Upselling	-	368.29	20.9M	+38.8%

CLUSTER 2 :

Goals:

1. Increase purchase frequency via bundling.
2. Increase AOV through premium product promotion.

Strategies:

- Create bundled offers for Mobiles & Tablets.
- Focus on customers using Easypay as the payment method.

Why this work?

- Bundling encourages customers to buy complementary products, increasing the total volume of purchases.
- Upselling focuses on increasing revenue per transaction, which is efficient for a small customer base with high AOV.

Simulation :

Assumptions:

- Current Data:
  - Total Customers: 6,558
  - Total Revenue: 15.06M
  - AOV: 1,000
  - Number of Purchases: 19K.
- Bundling increases purchase frequency by 20%.
- Upselling increases AOV by 10% through premium product add-ons.

Scenario	New Frequency	New AOV	New Revenue	Increase
Bundling Promotions	22.8K	1,000	22.8M	+51%
Premium Upselling	-	1,100	20.9M	+38.8%

CLUSTER 3 :

Goals:

1. Retain high-value customers.
2. Increase purchase frequency via VIP programs.

Strategies:

- Create exclusive VIP membership programs.
- Offer additional services like extended warranties and early product access.

Simulation :

Assumptions:

- Current Data:
  - Total Customers: 4,198
  - Total Revenue: 32.43M
  - AOV: 2,850
  - Number of Purchases: 19K.
- VIP programs increase frequency by 15%.
- Premium services increase AOV by 10%.

Why this work?

- VIP programs focus on rewarding high-value customers, increasing their engagement and frequency of purchases.
- Premium services are ideal for a high-value customer base, justifying the price increase.

Scenario	New Frequency	New AOV	New Revenue	Increase
VIP Program	21.85K	2,850	62.27M	+92%
AOV Increase	-	3,135	59.57M	+83.6%

CONCLUSION & RECOMMENDATIONS

CONCLUSION

This project successfully achieved its primary objective of segmenting Amazon's customers based on purchasing behaviors through a clustering model, with K-Means identified as the best algorithm for this purpose. The segmentation analysis uncovered four distinct customer clusters, each with unique characteristics and revenue contributions:

- Cluster 0: High loyalty with low purchase frequency, contributing moderate revenue.
- Cluster 1: The largest customer group with high volume and moderate average order value (AOV).
- Cluster 2: Technology-focused customers benefiting from high discounts, with a significant contribution to total revenue.
- Cluster 3: Premium customers with high-value orders, generating the highest revenue.

STRATEGIES AND RECOMMENDATIONS

Increase Purchase Frequency

Implement small discount strategies (2%-5%) to incentivize more frequent purchases, projected to increase revenue by up to 36%.

Increase Average Order Value (AOV)

Introduce premium product upselling and bundling promotions for categories like Mobiles & Tablets, which could boost revenue by 38%-51%.

Customer Retention

Deploy loyalty programs for dormant customers and VIP programs for high-value customers, potentially increasing revenue by up to 92%.

Targeted Marketing

Focus on personalized marketing for specific clusters, leveraging their unique characteristics (e.g., promoting premium services for Cluster 3)

ibimbing

THANK YOU