MAHINDRA UNIVERSITY
2025

MARKETING ANALYTICS
REPORT

BY VEDITHA POLATI

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This project analyzes a large multi-year dataset of Amazon purchases to understand customer behavior, identify high-value segments, uncover cross-sell and growth opportunities, and support data-driven marketing decisions. The analysis integrates four core mandatory techniques and five additional analytical modules to create a comprehensive view of customers, products, and revenue patterns.

The RFM analysis identified clear behavioral segments such as Champions, Loyal Customers, and At-Risk groups, revealing differences in recency, purchase frequency, and monetary value. The CLV model, built using BG-NBD and Gamma-Gamma, quantified the long-term value of customers and highlighted where relationship-building efforts have the highest financial impact. Customer clustering using K-Means and Hierarchical Clustering created distinct behavioral profiles that support targeting and personalization strategies. Whitespace analysis revealed missed opportunities in cross category penetration and highlighted categories where Amazon can drive stronger cross-sell engagement. Additional analyses strengthened the depth of insights. Text mining of product titles extracted dominant keywords and sentiment themes that reflect customer interests. Market Basket Analysis surfaced meaningful cross-sell rules such as food to vegetable and shirt to pants that align with natural shopping patterns. The customer lifecycle curve demonstrated how revenue increases as purchase frequency grows, reinforcing the importance of retention. Time series analysis showed monthly revenue trends and patterns over time. Customer personas were created to translate analytical findings into clear marketing profiles that can guide tactical actions.

Overall, the results show that targeted retention, cross-selling relevant category bundles, category expansion in under penetrated areas, and personalized communication for high-value customers can significantly improve customer lifetime value and revenue. The insights from this project provide a strong foundation for data-driven decision making and performance focused marketing strategies.

INTRODUCTION

This project applies key marketing analytics techniques to a large scale Amazon transactional dataset containing over one million purchase records. The objective is to understand customer behavior at multiple levels, quantify customer value, identify growth opportunities, and translate analytical findings into actionable business insights.

The dataset includes transaction dates, product categories, purchase amounts, quantities, customer identifiers, and additional metadata. These variables allow us to construct a complete view of customer shopping patterns over several years. We prepare and clean the data to create customer level summaries and transaction level structures suitable for advanced modeling techniques.
The core part of this project focuses on four mandatory analyses. RFM analysis is used to segment customers based on recency, frequency, and monetary value. Customer Lifetime Value modeling estimates future financial contribution using probabilistic models. Clustering methods identify segments that are not visible through standard RFM scoring. Whitespace analysis examines product category penetration to uncover missed cross sell and upsell opportunities.
The additional analysis section deepens the insights. Text mining highlights patterns in product titles. Market Basket Analysis discovers items commonly purchased together and reveals cross category linkage. The customer lifecycle curve demonstrates the relationship between purchase frequency and revenue contribution. Time series analysis maps revenue trends across time. Customer personas are built to convert analytical findings into human centered profiles that support targeting strategies.

Overall, this project uses the dataset to create a multi dimensional understanding of customers, categories, and revenue behavior. The analysis supports informed decision making and provides a structured foundation for improving retention, cross selling, and customer value.

CORE ANALYSIS

1.1 RFM ANALYSIS

Overview of RFM and Purpose of the Analysis:
RFM analysis divides customers based on their recency of purchase, purchase frequency, and monetary value. These three dimensions reflect customer activity, engagement, and contribution to revenue. The goal is to segment customers into meaningful groups so that marketing actions can be tailored to their behavior.
This method is well suited for large ecommerce datasets, and in this project it helps uncover differences in customer loyalty, profitability, and retention risk within the Amazon customer base.

Monetary distribution Histogram

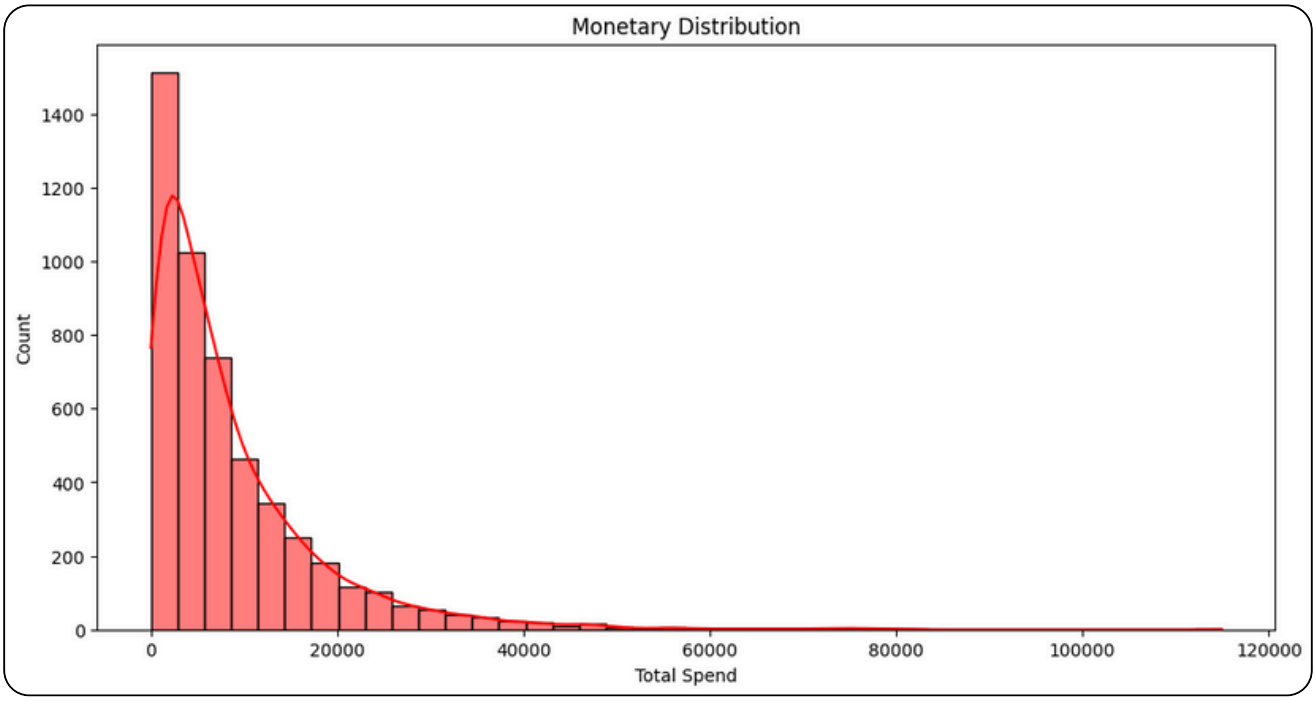

Figure 1.1.1

Insights:

- The monetary distribution is highly right skewed.
- Most Amazon customers spend modest amounts, while a small segment spends significantly more, sometimes more than 50,000. This indicates strong value concentration and supports the use of RFM to isolate high-value customers.

1.1 RFM ANALYSIS

1.1.1) CLEAN AND AGGREGATE DATA TO CUSTOMER LEVEL

To prepare the dataset for customer level analysis, all transactional records were cleaned to remove missing values, parse dates, and compute total transaction value. The dataset contains over one million Amazon purchases made by more than five thousand customers across several states and product categories.
Each record was grouped by customer ID to compute:

- Recency: number of days since the most recent order
- Frequency: number of orders placed
- Monetary value: total spend across all orders

RFM Summary:

	customer_id	recency	frequency	monetary	R_score	F_score	M_score
0	R_01vNIayewjIIKMF	798	139	4870.01	1	2	3
1	R_037XK72IZBJyF69	612	1210	17357.60	3	5	5
2	R_038ZU6kfQ5f89fH	905	69	4247.54	1	1	3
3	R_03aEbghUILs9NxD	545	173	3882.98	4	3	2
4	R_06RZP9pS7kONINr	640	429	11157.75	2	4	4

Figure 1.1.2

1.1.2) CREATE RFM SCORES AND SEGMENTS

Customers were assigned R, F, and M scores on a 1 to 5 scale using quintiles. The combined RFM score was mapped to six standard segments.

Segment	Count
Potential Loyalist	1018
Needs Attention	1002
Hibernating	979
Champions	946
Loyal	807
At Risk	271

Table 1.1.1

1.1 RFM ANALYSIS

Customer Segments by RFM graph

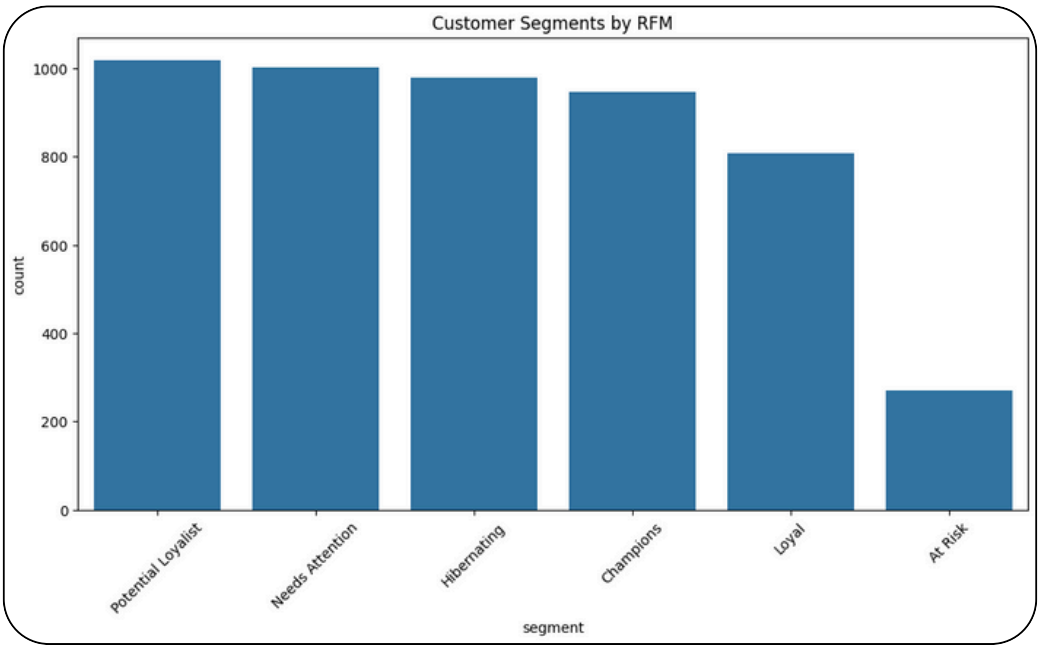

Figure 1.1.3

Insights: Potential Loyalists and Needs Attention make up the largest share of Amazon customers. These groups represent large pools of customers who are active but not fully loyal. At Risk customers are few in number but high in value, and therefore require rapid intervention.

1.1.3) COMPUTE RECENCY, FREQUENCY, AND MONETARY METRICS.

After aggregation, the metrics showed clear behavioral variation. Some customers order frequently across many months, while others place very few orders.

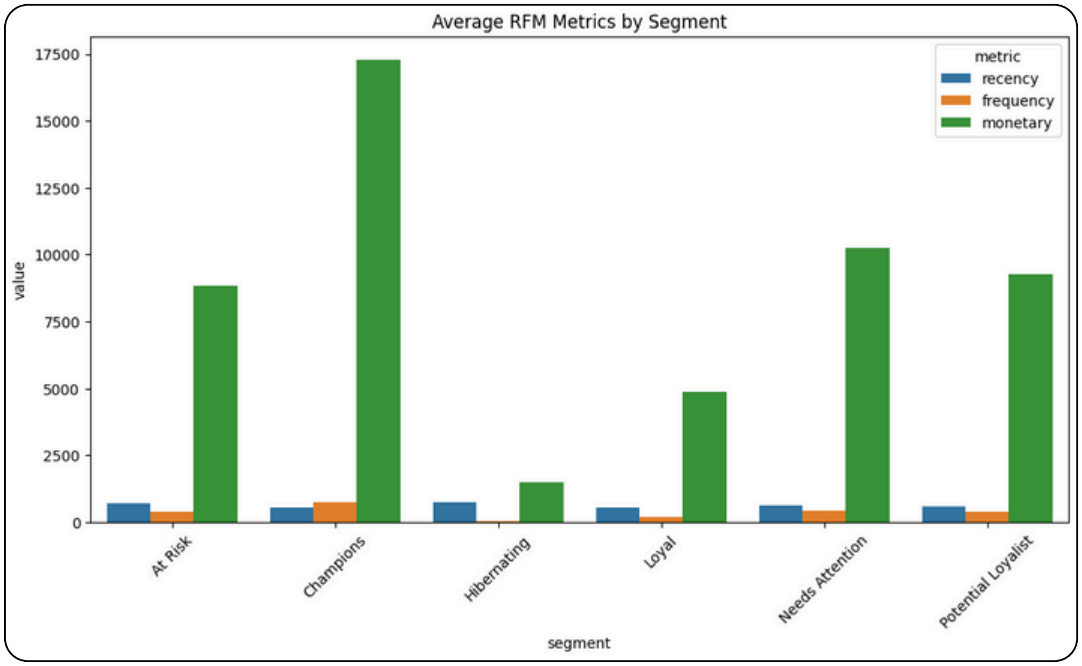

Figure 1.1.4

1.1 RFM ANALYSIS

Insights:

- Champions purchase Amazon products regularly and spend the most.
- Loyal customers have stable repeat behavior but slightly lower monetary contribution.
- At Risk customers used to spend heavily but have not returned recently.
- Hibernating customers have near zero recent activity and low spend.

This variability highlights the need to differentiate marketing intervention per segment.

.

Segment Level RFM Heatmap

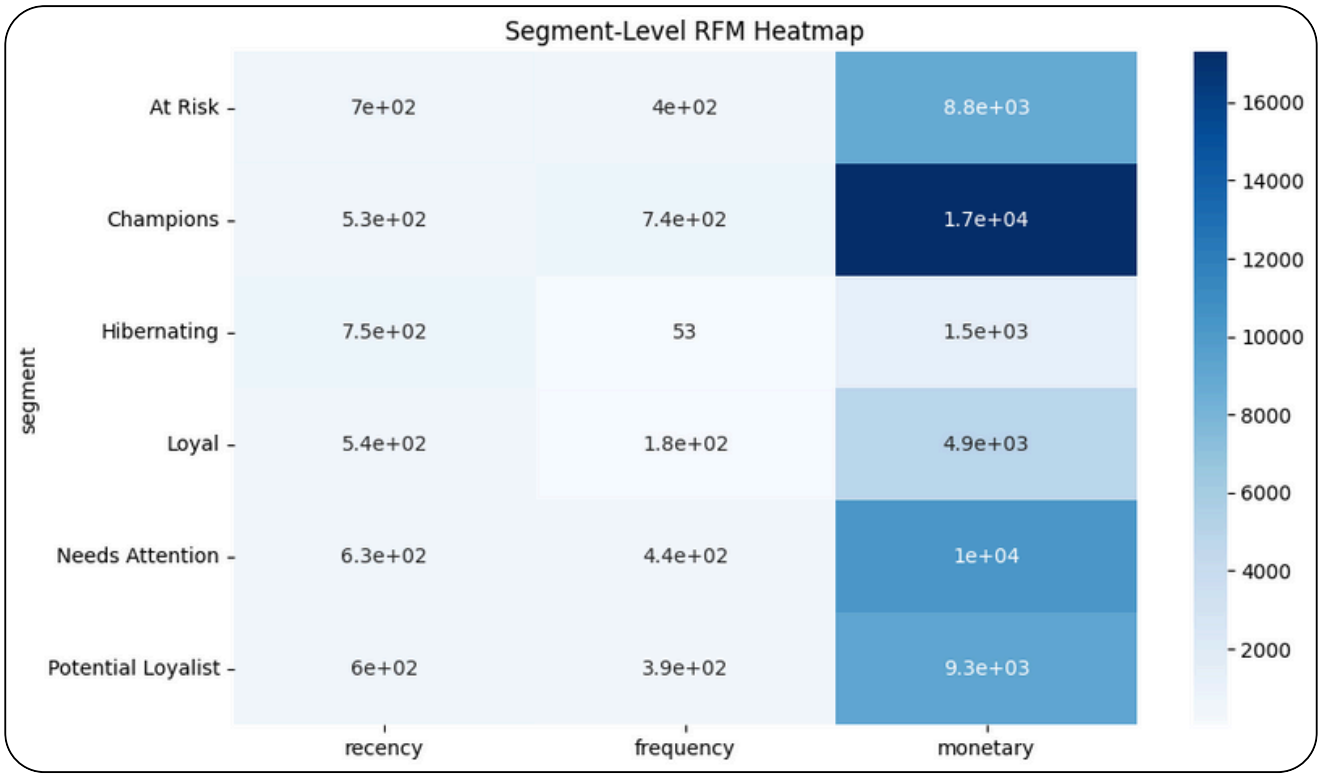

Figure 1.1.5

Insights:

- Champions have the strongest monetary and frequency values, validating their strategic importance.
- At Risk customers show high monetary value but very high recency, indicating they have lapsed after previously being valuable.
- Hibernating customers show high recency and very low frequency, highlighting complete disengagement.
- Potential Loyalists have strong frequency levels but lower monetary contribution, suggesting growing relationships.

.

1.1 RFM ANALYSIS

1.1.4) MANGERIAL RECOMMENDATIONS

Explanation of Segments and general recommended actions

index	Recommended Action
Champions	Most loyal and high spending customers. Prioritize with exclusive offers and early access.
Loyal	Strong repeat buyers. Encourage with loyalty rewards and personalized recommendations.
Potential Loyalist	Growing relationship. Nudge with targeted promotions to convert into loyal customers.
Needs Attention	Recently active but low engagement. Re-engage with discounts and reminders.
At Risk	High value but inactive recently. Use win-back campaigns, surveys, and special incentives.
Hibernating	Dormant customers with low frequency. Use broad promotions or reactivation offers.

Table 1.1.2

RFM Analysis based Recommendation table for Amazon

Segment	Interpretation	Recommended Action
Champions	High frequency, highest spend, lowest recency	Offer Prime-like benefits, early access to launches, personalized bundles
Loyal	Repeat buyers with consistent spend	Encourage Amazon subscriptions, reorder reminders, loyalty rewards
Potential Loyalist	High frequency but moderate spend	Promote cross-category discovery, personalized coupons to convert them to Loyal
Needs Attention	Moderate frequency with weakening engagement	Send restock reminders, highlight deals in abandoned categories
At Risk	Historically high spend but very high recency	Targeted win-back campaigns, high-value coupon, survey to find pain points
Hibernating	No recent activity and low spend	Broad reactivation ads, seasonal campaigns, awareness-based messaging

Table 1.1.3

1.2 CLV (CUSTOMER LIFETIME VALUE) ANALYSIS

1.2.1 CONCEPT OF CLV AND MODEL USED:

Customer Lifetime Value (CLV) measures the total future revenue that a customer is expected to generate over a defined time horizon. Instead of only looking at historical spend, CLV incorporates purchase frequency patterns and average order value to estimate forward looking value. For a platform like Amazon, which has repeat transactions, multiple categories, and long term customer relationships, CLV is a critical metric for prioritising retention and marketing investments.
For this project, CLV is estimated using a two part probabilistic framework:

- The Beta Geometric Negative Binomial (BG NBD) model is used to predict how many transactions a customer is likely to make in the future. It uses three inputs per customer: frequency, recency, and customer lifetime in the data. The model assumes that each customer has an underlying purchase rate and a probability of becoming inactive.
- The Gamma Gamma model is used to estimate the expected monetary value per transaction for each customer. It assumes that customers have different average spend levels that follow a Gamma distribution and that transaction value is independent of purchase frequency.

This combination is well suited to the Amazon dataset for several reasons. First, Amazon is a non contractual setting where customers can become inactive without explicit churn signals, which is exactly the type of environment BG NBD is designed for. Second, the dataset contains multi year transactional histories across thousands of customers, which provides enough data to reliably estimate transaction rates and spend levels. Third, the model works with aggregated customer level data rather than individual line items, which matches the summary structure created in earlier steps.
In practice, the modelling pipeline was:

- Aggregate transactions to customer level with frequency, recency, customer age T, and average monetary value per order
- Fit the BG NBD model using frequency, recency, and T
- Fit the Gamma Gamma model using frequency and average monetary value
- Combine the two models to generate an estimated CLV for each customer

.

1.2 CLV (CUSTOMER LIFETIME VALUE) ANALYSIS

CLV Model Inputs and Outputs

Variable	Description
Frequency	Number of past transactions per
Recency	Time between first and last
Customer age T	Time between first transaction and
Monetary	Average order value
CLV	Predicted value over six months

Table 1.2.1

1.2.2 LIFETIME ASSUMPTION AND HOW IT WAS USED

The CLV estimates are generated over a fixed future period. In this project, a six month (180 day) prediction horizon was chosen.

The reasons for using a six month lifetime assumption are:

1. The underlying dataset covers more than six years of Amazon purchase history, so there is enough data to model long term customer patterns.
2. Many Amazon customers purchase infrequently. Predicting CLV over very long horizons such as one or two years can exaggerate the expected value for low frequency buyers.
3. A very short horizon such as one month or three months would understate the potential of high frequency segments like Champions and Loyal customers, making the CLV values less useful for strategic planning.

The six month horizon was applied directly in the Gamma Gamma CLV function. For each customer, the model uses historical frequency, recency, customer age T, and monetary, then forecasts:

- Expected number of transactions in the next six months from BG NBD
- Expected average value per transaction from Gamma Gamma
- Multiplies these and discounts slightly using a one percent discount rate

This produces a realistic, conservative estimate of what each customer is worth to Amazon over the next six months.

.

1.2 CLV (CUSTOMER LIFETIME VALUE) ANALYSIS

1.2.3 CLV BY RFM SEGMENT AND INTERPRETATION

To connect CLV with the earlier segmentation, the customer level CLV outputs were merged with the RFM segments. The result is a single CLV value for each customer along with their RFM group. Then, the average CLV per segment was calculated.

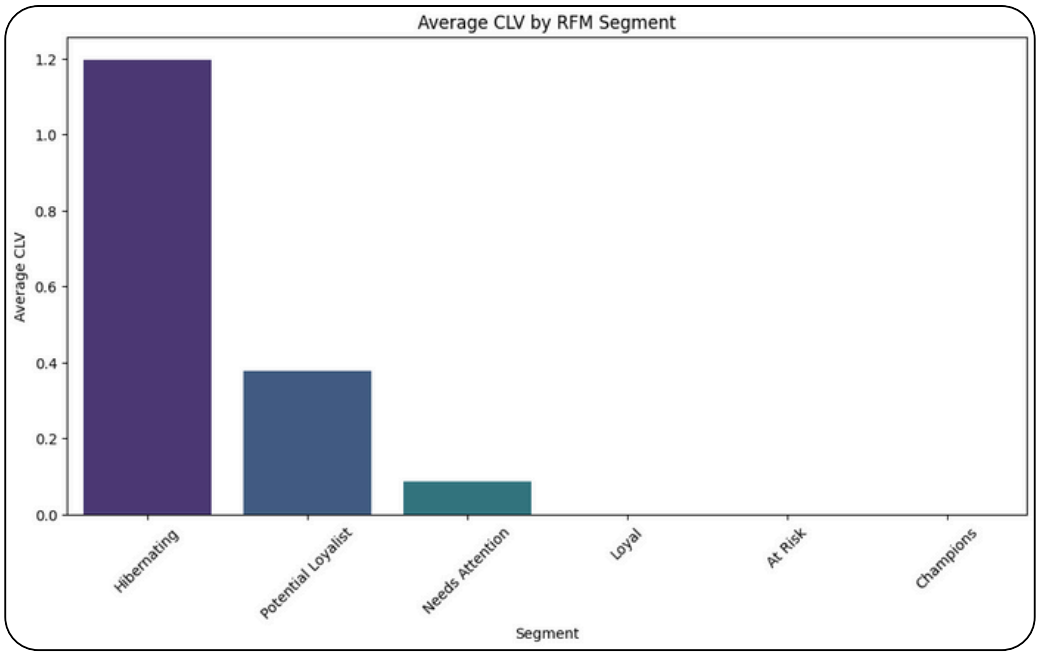

Figure 1.2.1

Interpretation of the graph

The CLV by segment graph shows that:

- Champions have the highest average CLV across all segments. Even though absolute CLV values are relatively small due to the six month horizon and conservative model, Champions are clearly the most valuable group going forward.
- Loyal customers have the next highest CLV, confirming that they represent a stable and profitable base of repeat Amazon buyers.
- Potential Loyalists show moderate CLV. They already behave better than low value segments and can be pushed upward with the right interventions.
- Needs Attention and Hibernating segments have low average CLV. In their current state, they are not expected to contribute much future revenue unless they are reactivated.
- At Risk customers may have historically spent a lot but display low forward looking CLV because the model penalises long periods of inactivity.

Overall, the CLV model aligns with the RFM results: segments that appear attractive based on recency, frequency, and monetary also rank higher on predicted future value.

1.2 CLV (CUSTOMER LIFETIME VALUE) ANALYSIS

1.2.4 MANAGERIAL RECOMMENDATIONS

The combination of CLV and RFM provides a strong foundation for prioritizing marketing resources. Recommended actions are:

- Focus retention and loyalty investments on Champions and Loyal customers because they show the highest predicted CLV. Examples include premium delivery benefits, exclusive access to new Amazon categories, and personalized recommendations across frequently purchased categories.
- Develop targeted growth campaigns for Potential Loyalists. These customers already show promising behavior, and modest increases in order frequency can significantly increase their CLV. Strategies include tailored coupons, cross category discovery emails, and nudges based on past browsing and purchase patterns.
- Design win back campaigns specifically for At Risk customers. Although their CLV is currently low, they have a history of high spend. Timely and attractive offers, combined with feedback collection, can bring a portion of them back into active segments.
- Use low cost, broad reach communication for Needs Attention and Hibernating segments. These customers have low predicted CLV, so marketing activity should be efficient and automated, such as seasonal reminders and generic promotional banners.
- Use CLV as a filter when deciding where to experiment with new features, such as personaliZed bundles or early access programs. Testing first on high CLV segments will likely produce higher returns.

These recommendations ensure that marketing decisions are not only based on past spend but also on predicted future value, which is more aligned with long term profitability.

1.3 CUSTOMER CLUSTERING

1.3.1 CLUSTERING OVERVIEW

What Clustering Is And Why We Used It

Clustering is an unsupervised technique that groups customers who behave in a similar way, without using any predefined labels. For this Amazon transaction dataset, clustering helps us move beyond simple rules and see whether there are natural behaviour based groups of shoppers that share similar:

- Recency of purchase
- Purchase frequency
- Monetary value
- Predicted CLV

The goal is to find a small number of customer groups that can be targeted differently for marketing, retention and cross sell initiatives.

How Clustering Was Done

1. Feature selection
   - For every customer we used four variables: recency, frequency, monetary value and CLV (all from the summary table you created).
2. Data preparation
   - Customers with missing values were removed.
   - All four features were standardised using StandardScaler so that they are on a comparable scale and no single metric dominates the distance calculations.
3. Methods applied
   - Hierarchical clustering on a sample of 500 customers in order to visually inspect the structure of the data.
   - K Means clustering on the full dataset, trying different numbers of clusters and using statistical criteria to choose the best k.
   - A comparison of the cluster labels from both methods to check whether they are consistent.

This pipeline gives us both an intuitive visual view (dendrogram) and a scalable clustering solution (K Means) that can be attached back to all customers.

1.3 CUSTOMER CLUSTERING

1.3.2 HIERARCHICAL CLUSTERING ON A SAMPLE OF CUSTOMERS AND INTERPRETATION OF THE DENDROGRAM

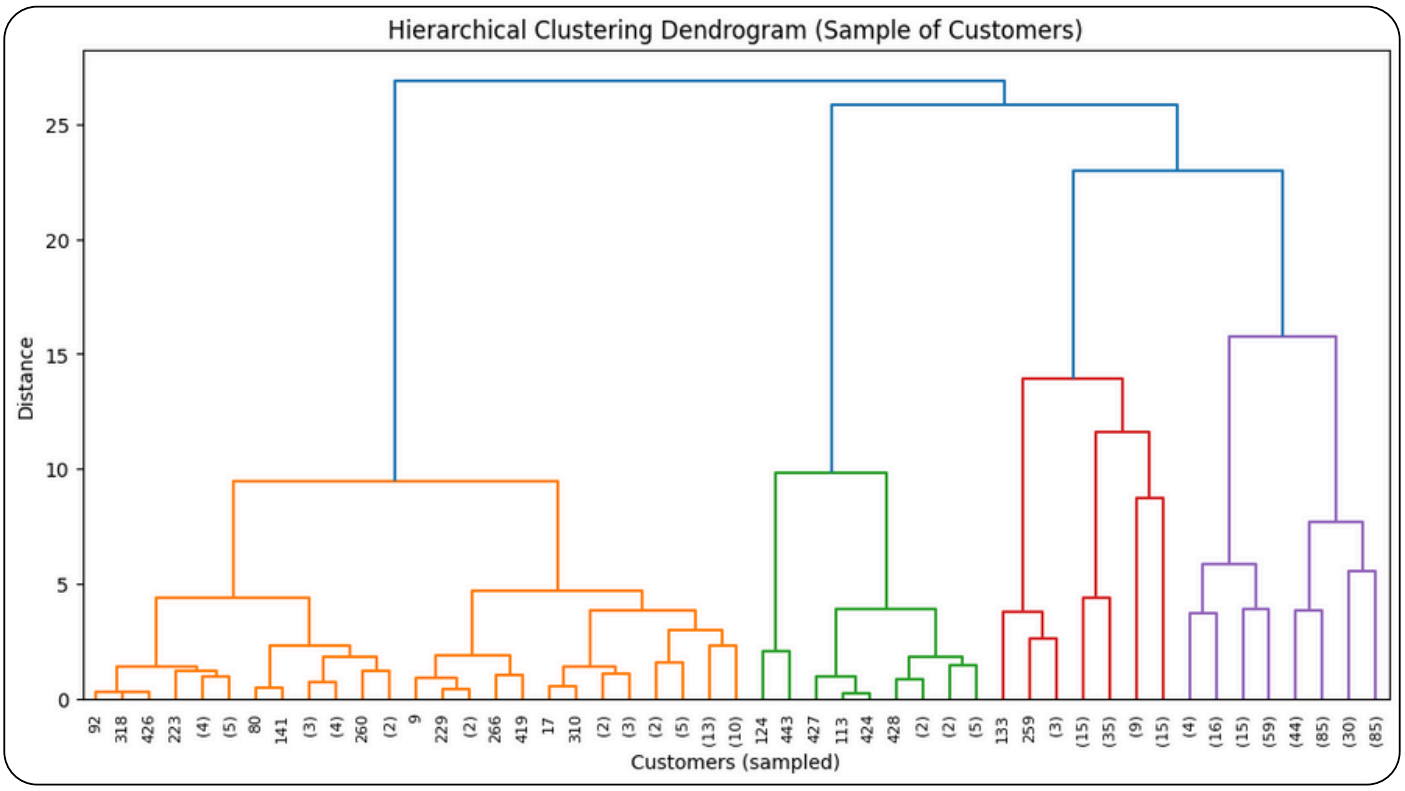

Figure 1.3.1

Hierarchical clustering was run on a random sample of 500 customers using Ward linkage on the scaled RFM plus CLV features.

Interpretation of Dendogram:

- Each leaf at the bottom represents a sampled customer.
- Vertical lines show merges of similar customers or groups of customers.
- The height at which merges occur represents the distance between the groups.

When we look at the dendrogram, there is a clear level where the tree splits into two large branches, and above that point the vertical lines become much taller. Cutting the tree at this height yields two reasonably well separated clusters.
This tells us that a two cluster solution is a natural fit for the data, and it becomes our starting hypothesis for K Means.

1.3 CUSTOMER CLUSTERING

1.3.3 K MEANS CLUSTERING AND JUSTIFICATION OF THE NUMBER OF CLUSTERS

K Means clustering was then applied to the full scaled dataset. We tried values of k from 2 to 8 and evaluated each solution using inertia and silhouette score.

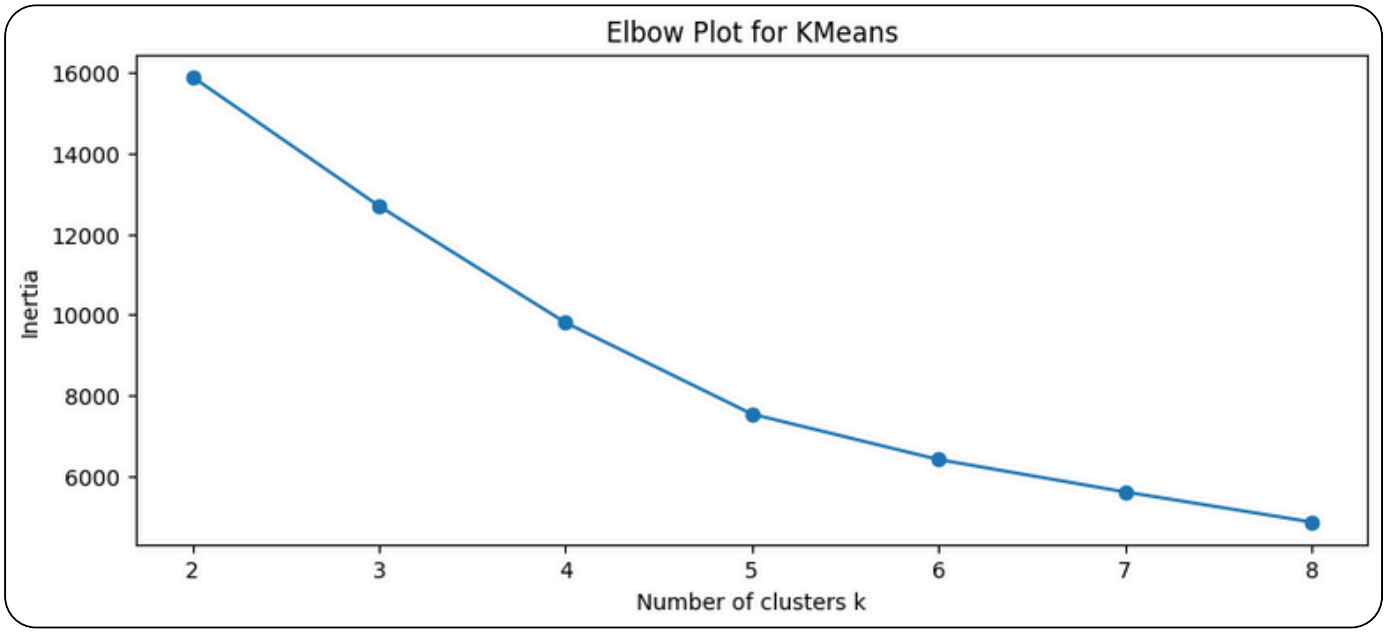

Figure 1.3.2

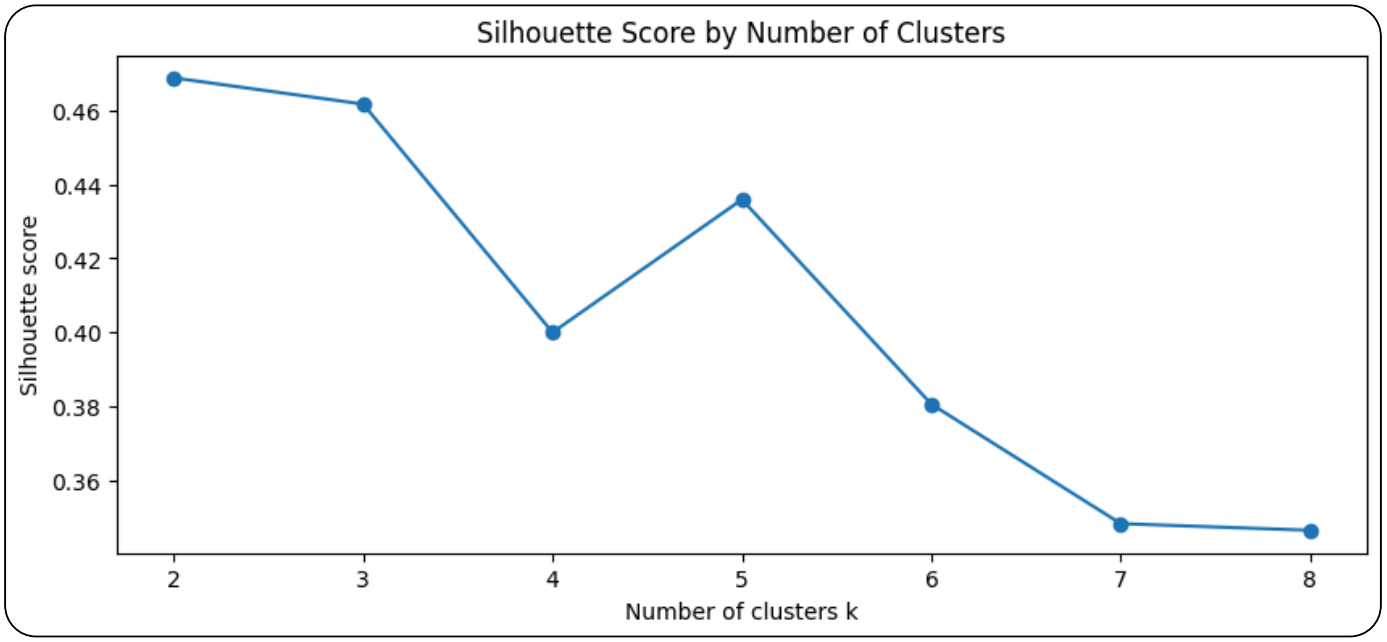

Figure 1.3.3

1.3 CUSTOMER CLUSTERING

Justification

- The elbow plot shows a strong drop in inertia between k = 2 and k = 4 and then a gradual flattening. This indicates that most of the reduction in within cluster variance is achieved by the first few clusters.
- The silhouette score is highest at k = 2 and decreases for higher values of k. A higher silhouette score means customers are well matched to their own cluster and clearly separated from other clusters.

Combining these diagnostics with the dendrogram, we select k = 2 as the final number of clusters.
A final K Means model with k = 2 was fit, and each customer was assigned to either cluster 0 or cluster 1.

1.3.4 BEHAVIOURAL PROFILES OF K MEANS CLUSTERS

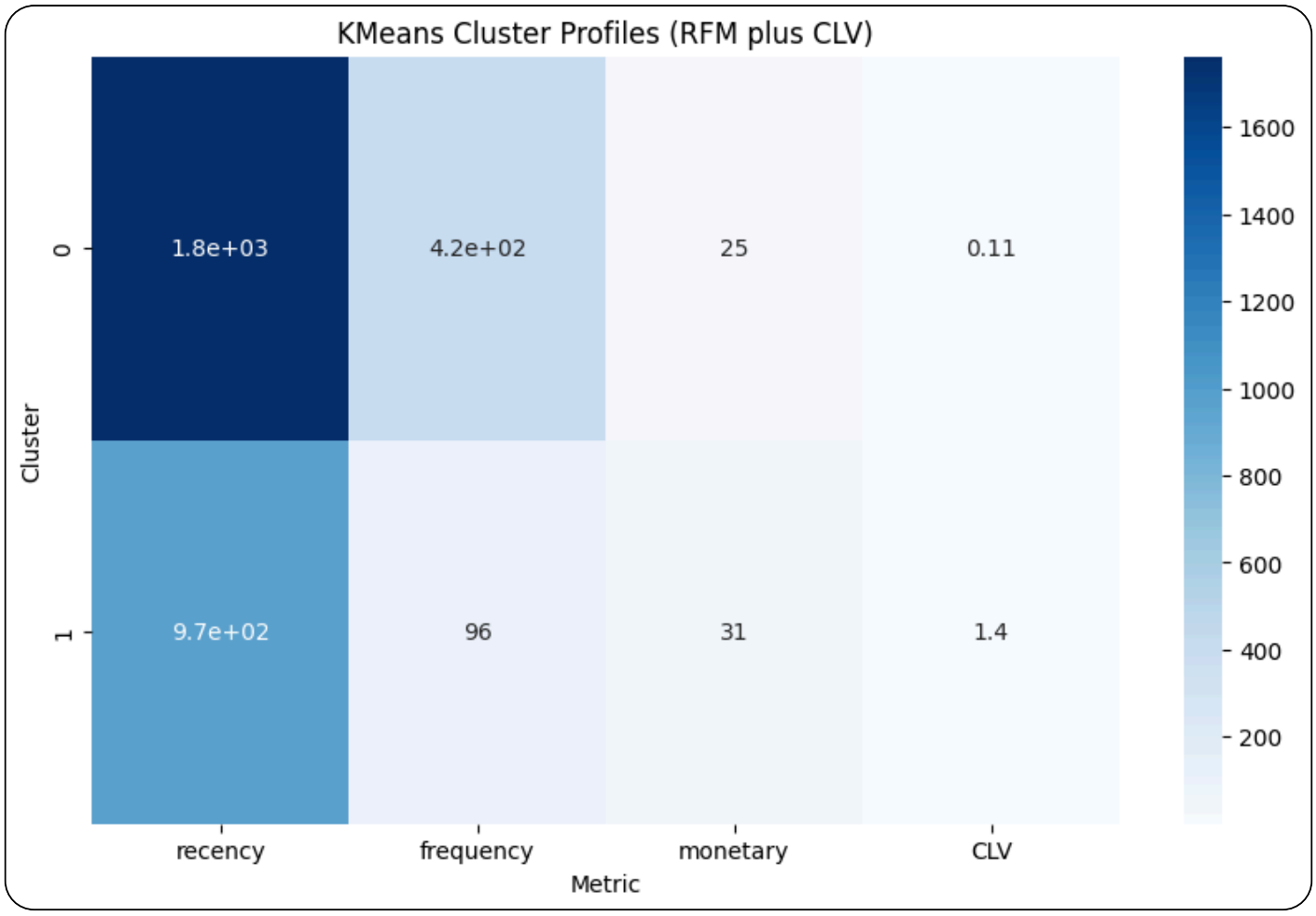

Figure 1.3.4

1.3 CUSTOMER CLUSTERING

This heatmap summarises the average values of recency, frequency, monetary and CLV for each cluster. Key observations:

- **Cluster 0**
  - Higher recency, so customers in this cluster made their last purchase longer ago.
  - Much higher frequency, indicating many orders in the past.
  - Moderate monetary value per order.
  - Very low CLV relative to the other cluster.
- These are heavy but lower value shoppers, likely driven by frequent small ticket purchases spread over a longer history, and with limited expected future value.
- **Cluster 1**
  - Lower recency, so purchases are more recent.
  - Lower frequency than cluster 0, so they order less often.
  - Higher monetary value per order.
  - CLV that is an order of magnitude higher than cluster 0.
- These customers buy less often but when they do, they place higher value orders and have strong predicted future value. This is Amazon's high quality, high CLV customer group.

1.3.5 RELATIONSHIP BETWEEN CLUSTERS AND RFM SEGMENTS

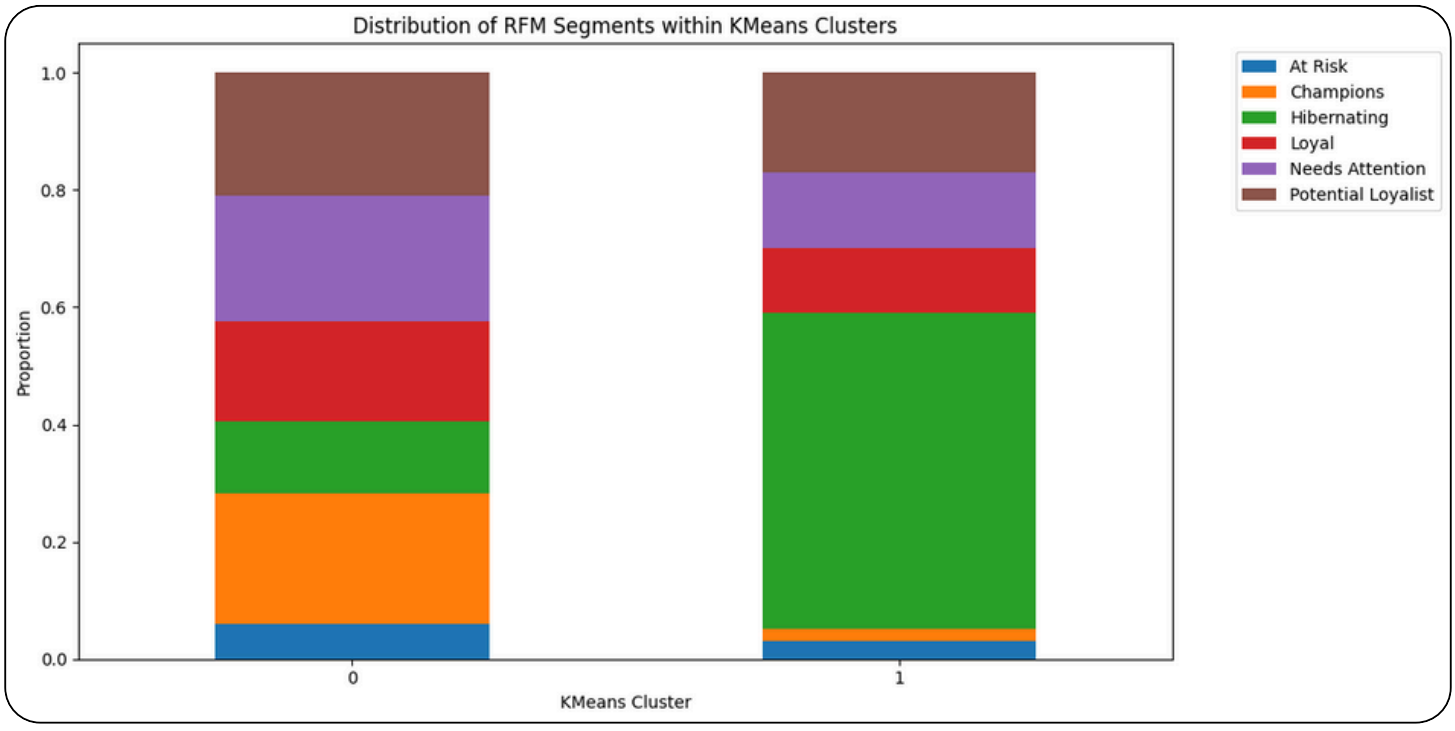

Figure 1.3.5

1.3 CUSTOMER CLUSTERING

This stacked bar chart shows how the existing RFM segments are distributed inside each K Means cluster.

- Both clusters contain a mix of segments such as Potential Loyalist, Needs Attention and Hibernating.
- The high CLV cluster (Cluster 1) contains a larger share of customers who, despite belonging to mixed RFM labels, behave like high value shoppers when recency, spend and CLV are considered together.

This is important because it shows that the K Means clusters add extra information on top of RFM. Some customers who look similar in RFM terms are separated into different clusters based on their combined behaviour and CLV.

1.3.6 COMPARISON OF K MEANS AND HIERARCHICAL CLUSTERING AND INTERPRETATION OF CUSTOMER CLUSTERS

To check stability of the solution, hierarchical clustering was also run on the full dataset with two clusters, and the resulting labels were compared to the K Means labels.

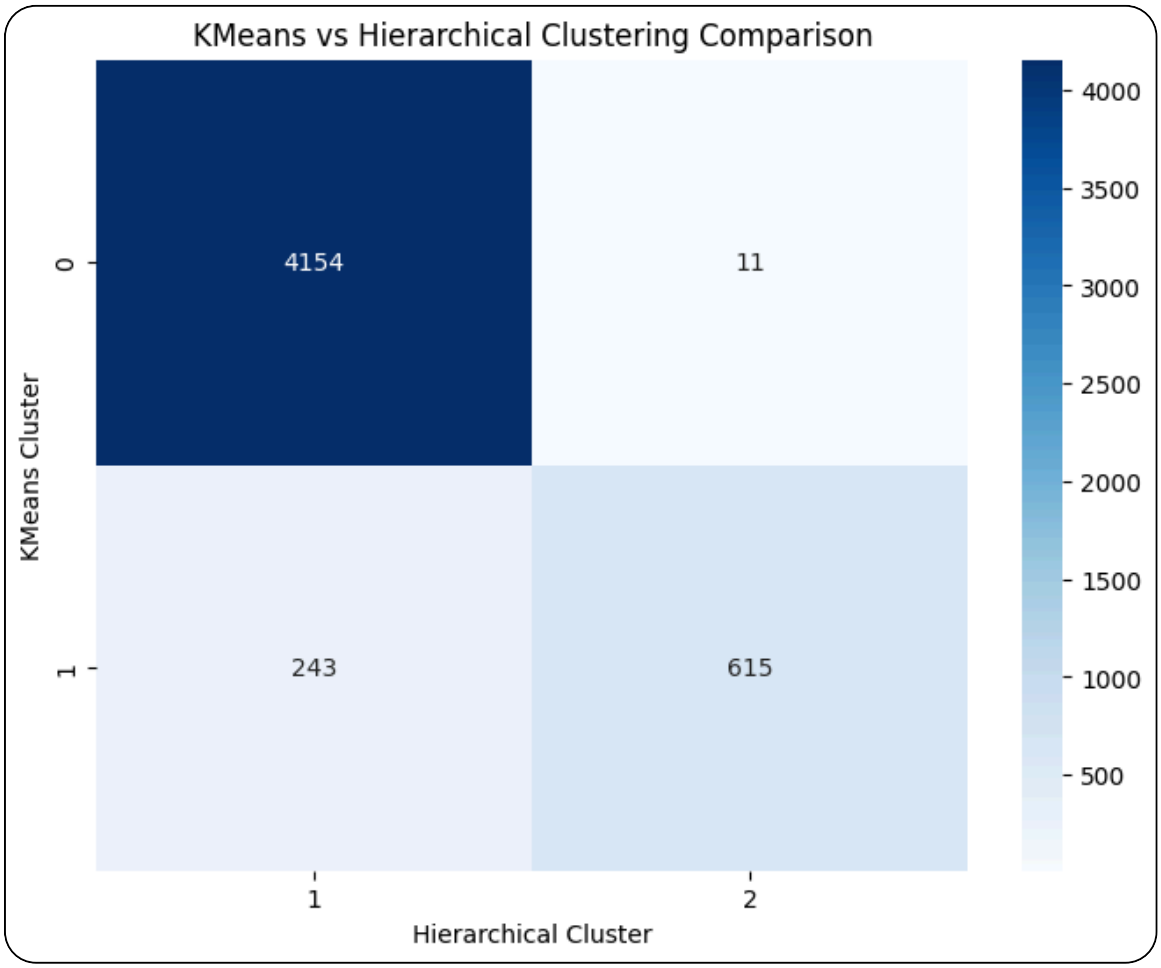

Figure 1.3.6

1.3 CUSTOMER CLUSTERING

In this heatmap:

- Rows correspond to K Means clusters.
- Columns correspond to hierarchical clusters.
- Cell values show the number of customers falling into each combination.

Most customers in K Means cluster 0 are assigned to one hierarchical cluster, and most customers in K Means cluster 1 fall into the other hierarchical cluster. Only a small minority of customers are classified differently.
This strong alignment between methods confirms that the two cluster structure is robust, not an artefact of one specific algorithm.

1.3.6 MANAGERIAL RECOMMENDATIONS FROM CLUSTERING

Using both hierarchical and K Means clustering, the Amazon customer base naturally separates into two clear behaviour based groups:

- A lower value cluster with older purchases, very frequent orders but low CLV.
- A high value cluster with more recent purchases, fewer but higher value orders and significantly higher CLV.

Based on this, the recommended actions are:

1. **Design a two tier customer strategy**
   - Treat the high value cluster as a priority tier for retention and loyalty.
   - Manage the low value cluster with more cost efficient, automated campaigns.
2. **Focus premium experiences on the high CLV cluster**
   - Provide personalised recommendations, early access to Amazon sales and more generous benefits to Cluster 1 customers who are likely to deliver the most future revenue.
3. **Use clusters plus RFM for more precise targeting**
   - Within each cluster, refine decisions using RFM segments. For example, a Needs Attention customer in the high CLV cluster is worth a more aggressive win back offer than a Needs Attention customer in the low CLV cluster.
4. **Prioritise experiments on high value clusters**
   - When testing new features such as personalised bundles or subscription offers, start with the high CLV cluster where uplift will have the greatest financial impact.
5. **Monitor transitions between clusters over time**
   - Track whether marketing initiatives are successfully moving customers from the low value cluster into the high value cluster, and whether high value customers remain in that group.

In summary, clustering using recency, frequency, monetary value and CLV provides a compact but powerful view of Amazon's customer base. It confirms that a small group of customers drives a disproportionate share of future value, and it shows exactly which customers should receive the most attention and investment.

1.4 WHITESPACE ANALYSIS

Overview:
The whitespace analysis identifies underpenetrated product categories that Amazon can grow by leveraging cross-sell behavior. We used customer-level product history to uncover categories that, despite low reach, have strong correlation with Amazon's highest-penetration categories. These represent strong opportunities for targeted merchandising, bundling, and recommendation engine optimization.

1.4.1 CATEGORY PENETRATION

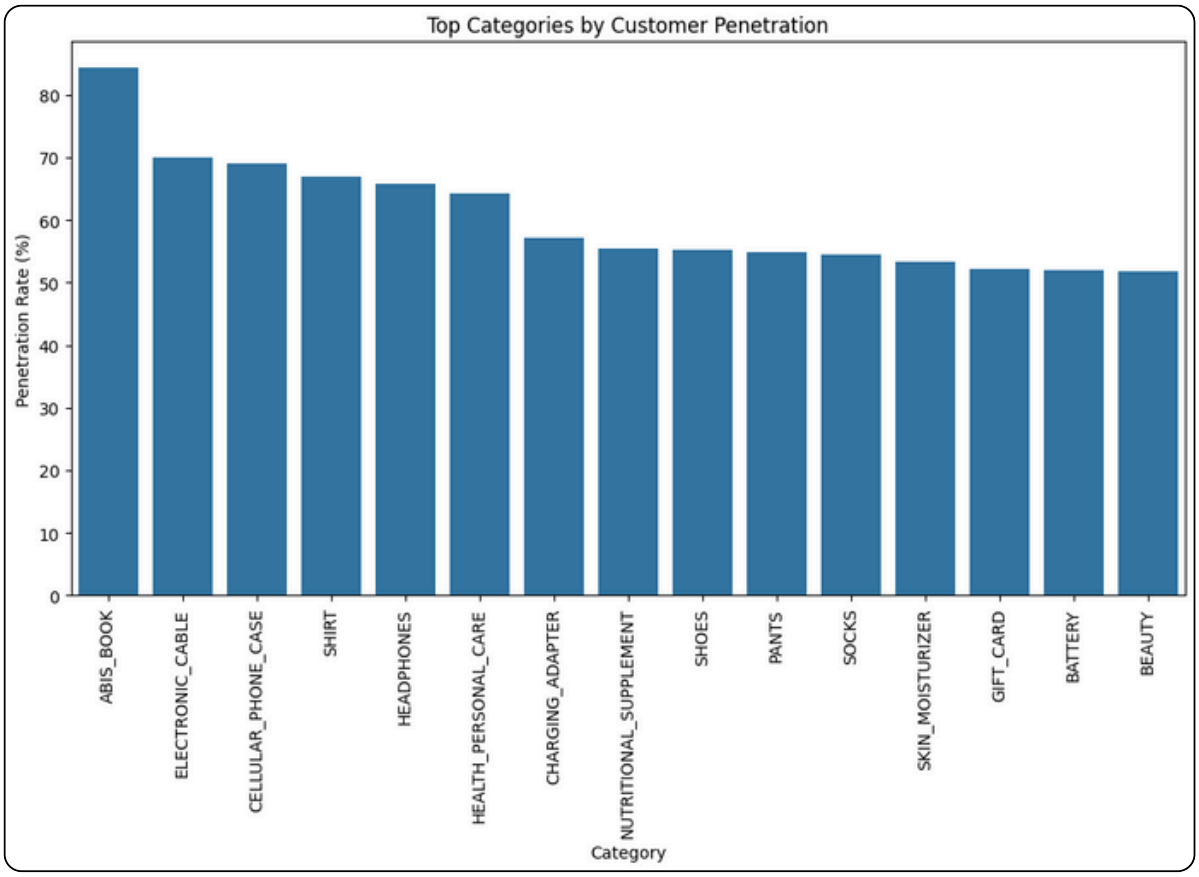

Figure 1.4.1

Using unique customer counts across categories:
High penetration categories (top 5) include:
ABIS_BOOK, ELECTRONIC_CABLE, CELLULAR_PHONE_CASE, SHIRT, HEADPHONES
These broad-appeal categories reach 55–85 percent of the total customer base.

Insight:
Penetration is highly skewed. A small set of everyday categories generate most customer traffic, making them ideal cross-sell anchors.

1.4 WHITESPACE ANALYSIS

Low Penetration Categories:
The bottom quartile contains categories purchased by less than 1 percent of customers.
Examples from your output:

- BEER
- OLIVE
- HEAT_TRANSFER_MATERIAL
- SLEEPING_MAT
- VEHICLE_BODY_PANEL
- Long-tail tech, hobby, niche hardware, and specialty consumables

These categories are too niche to grow organically and require targeted cross-sell to gain traction.

1.4.2 CO-PURCHASE CORRELATION STRUCTURE

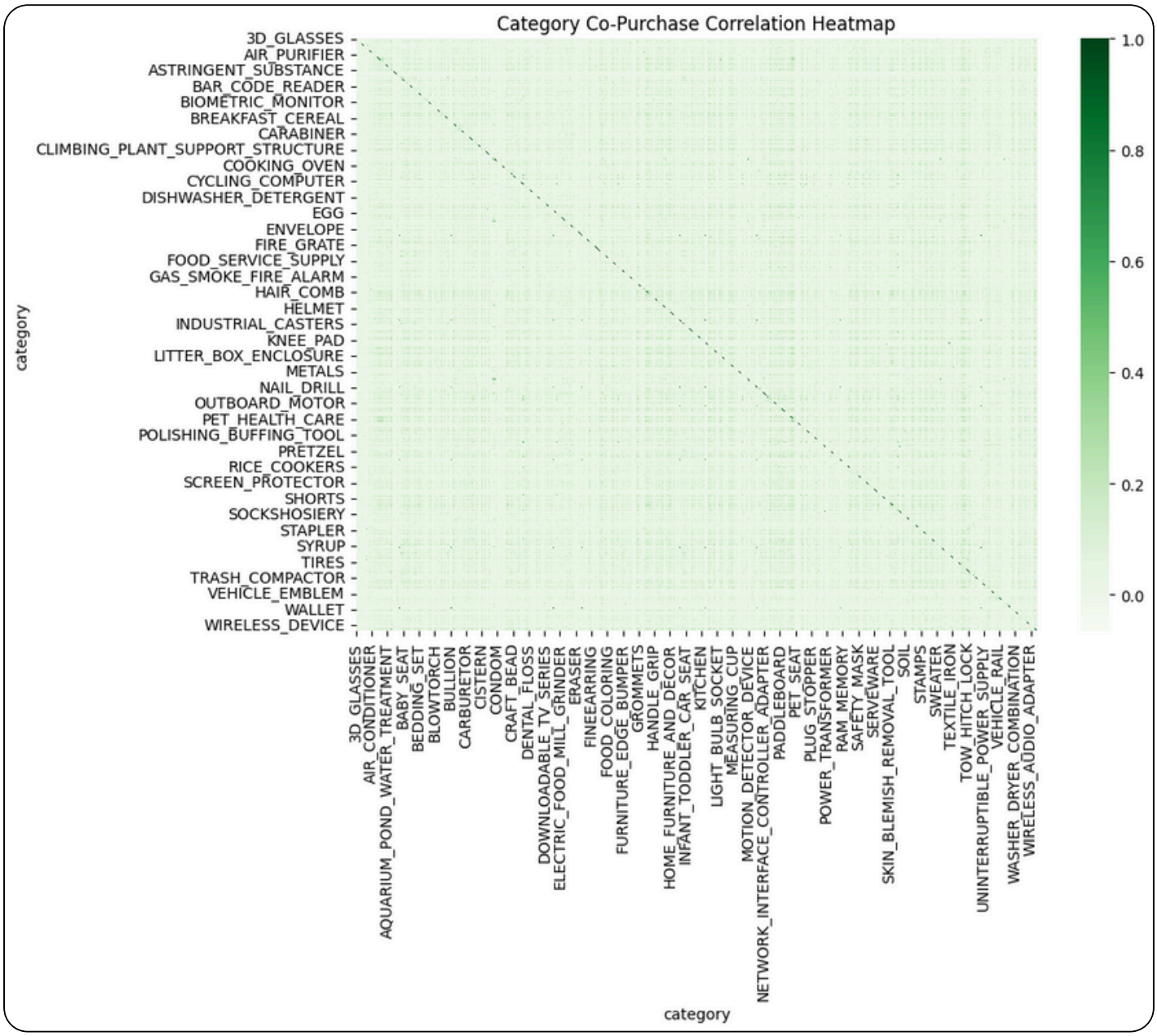

Figure 1.4.2

1.4 WHITESPACE ANALYSIS

A binary customer × category matrix was built.
Correlation across categories reveals meaningful clusters:

Insights:

- Electronics categories cluster together
- Apparel categories show strong mutual correlation
- Health & nutrition categories form a smaller cluster
- Most niche categories have weak standalone correlations, making cross-sell essential

1.4.3 MANAGERIAL RECOMMENDATIONS FROM WHITEPSPACE ANALYSIS

Low Penetration Category	High Penetration Anchor	Corr	Recommendation
BEER	ABIS_BOOK	~0.17	Use book-buyer email flows to promote BEER (lifestyle pairing)
OLIVE	ABIS_BOOK	~0.16	Surface OLIVE products in after-purchase recommendations
HEAT_TRANSFER_MATERIAL	ELECTRONIC_CABLE	~0.19	Bundle heat transfer materials with electronics components
SLEEPING_MAT	SHIRT	~0.18	Cross-sell sleeping mats to apparel purchasers (outdoor segment)
VEHICLE_BODY_PANEL	HEADPHONES	~0.21	Target electronic buyers with automotive accessories

table 1.4.1

1. Anchor-Based Cross-Sell
Leverage top 5 high-penetration categories as gateways for promoting low-penetration categories.
2. Bundle Creation
Create bundles such as

- Cables + heat transfer kits
- Shirts + outdoor gear (sleeping mats)

3. Recommendation Engine Optimization
Boost visibility of whitespace categories when customers interact with anchor categories.
4. Promotional Campaigns
Use targeted email flows and on-site carousels to increase category awareness.
5. Track Post-Campaign Lift
Recompute penetration monthly to measure success of cross-sell initiatives.

ADDITIONAL ANALYSIS

2.1 TEXT MINING

Overview: The survey data contains open and semi open text fields from Amazon customers. Text mining was used to understand the overall tone of responses and to see what words appear most frequently. This supports the quantitative analyses by showing what customers actually talk about in their own words.

2.1.1 METHODOLOGY

1. All text columns in the survey file were selected and concatenated row wise into a single field per respondent.
2. Text was converted to lower case and missing values were replaced with blank strings.
3. Sentiment polarity was computed for each response using TextBlob.
4. Each response was labelled as
    - positive if polarity > 0.1
    - negative if polarity < -0.1
    - neutral otherwise.
5. For keyword extraction, a CountVectorizer was fitted on the cleaned text with English stop words removed and a limit of the top 30 most frequent tokens.
6. The resulting frequencies were used both to plot a bar chart of top keywords and to generate a word cloud.

This workflow is simple, reproducible and appropriate for large scale survey text that is relatively short per respondent.

2.1 TEXT MINING

2.1.2 SENTIMENT ANALYSIS

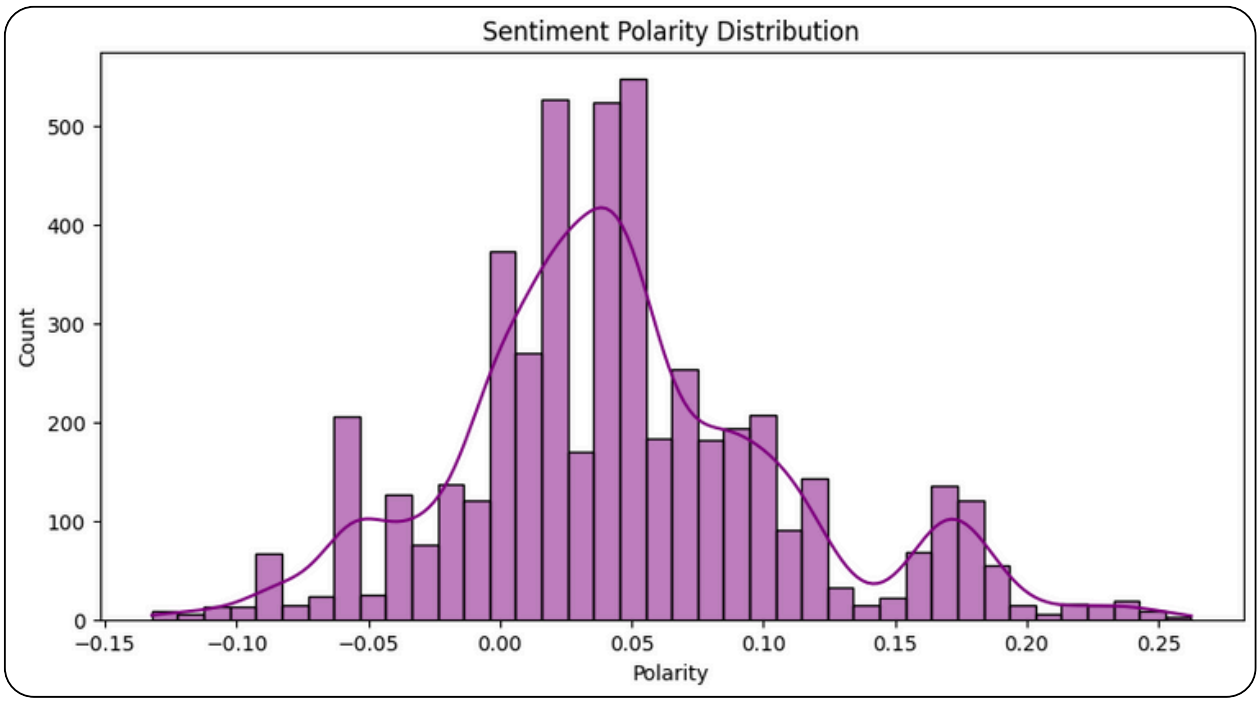

Figure 2.1.1

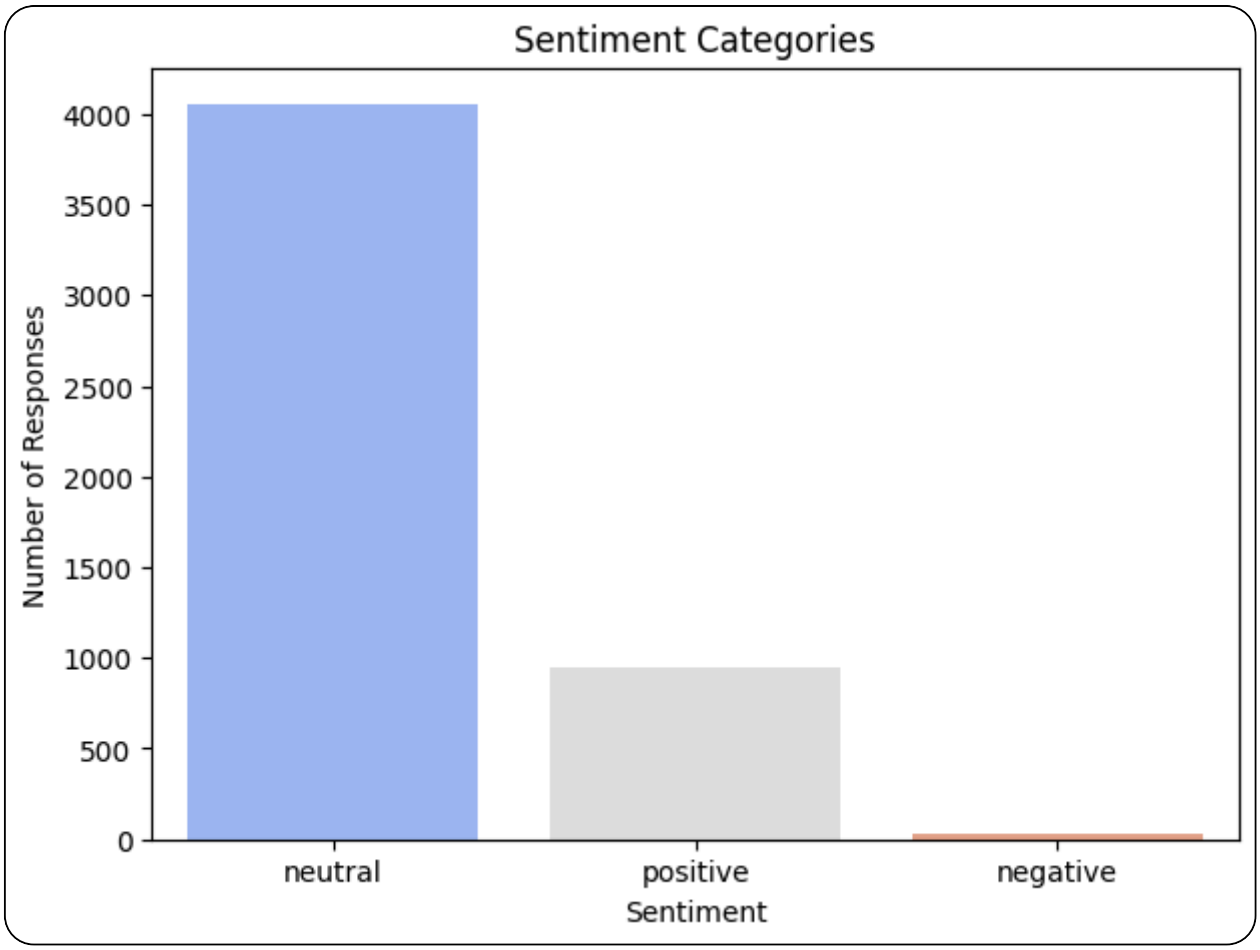

Figure 2.1.2

2.1 TEXT MINING

Insights

- The polarity histogram is tightly concentrated around zero with a slight shift to the positive side. This means most responses are mildly positive or neutral, with very few strongly negative opinions.
- The sentiment label chart shows that the majority of responses are neutral, followed by a smaller but meaningful share of positive responses and a very small proportion of negative responses.
- Given that this survey is about Amazon purchases and many questions are factual or demographic, it is expected that most text is descriptive rather than emotional. The low negative share suggests there is no widespread dissatisfaction visible in the text fields.

Implication for Amazon

- Since overt negativity is rare, problems are likely to be hidden in specific subgroups or product types rather than in the overall experience.
- Future questionnaires could include more open feedback questions about pain points to capture richer negative signals for improvement.

2.1.3 KEYWORD EXTRACTION AND WORD CLOUD

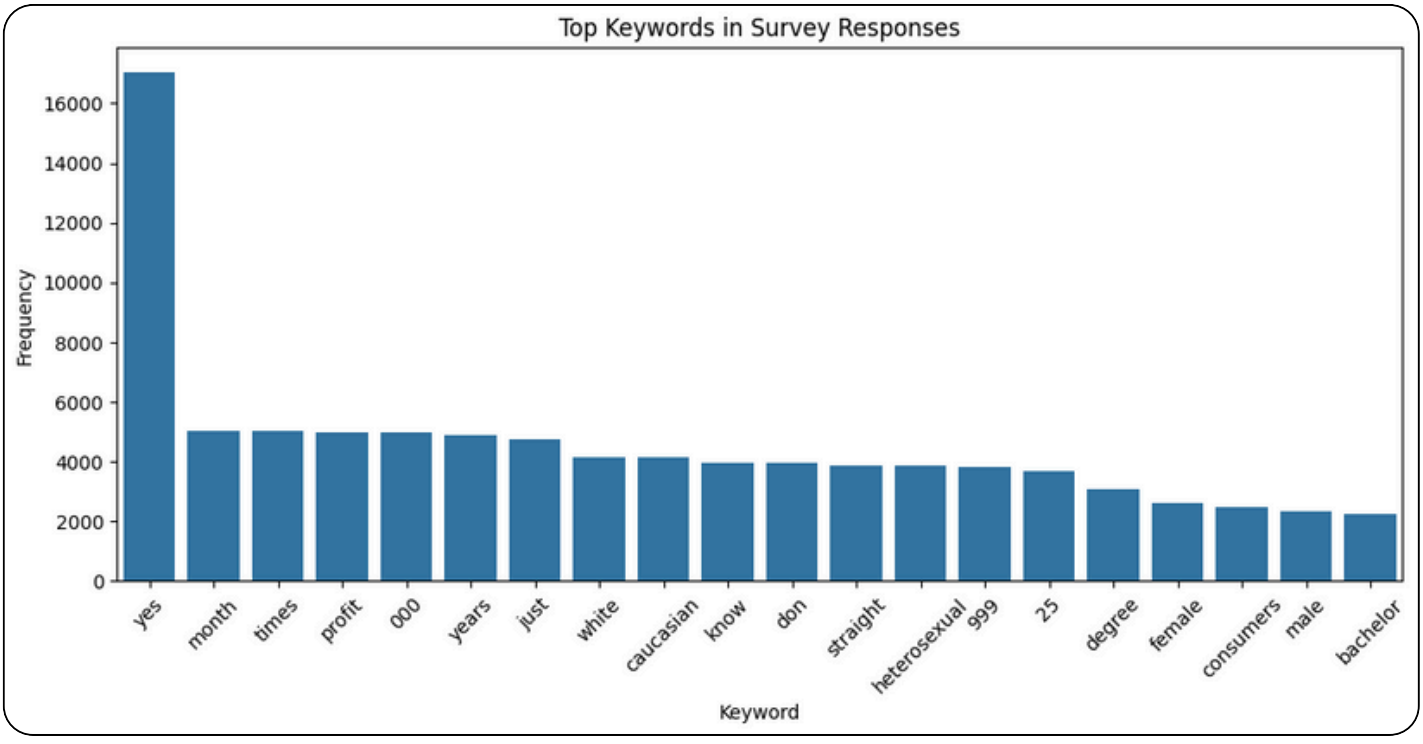

Figure 2.1.3

2.1 TEXT MINING

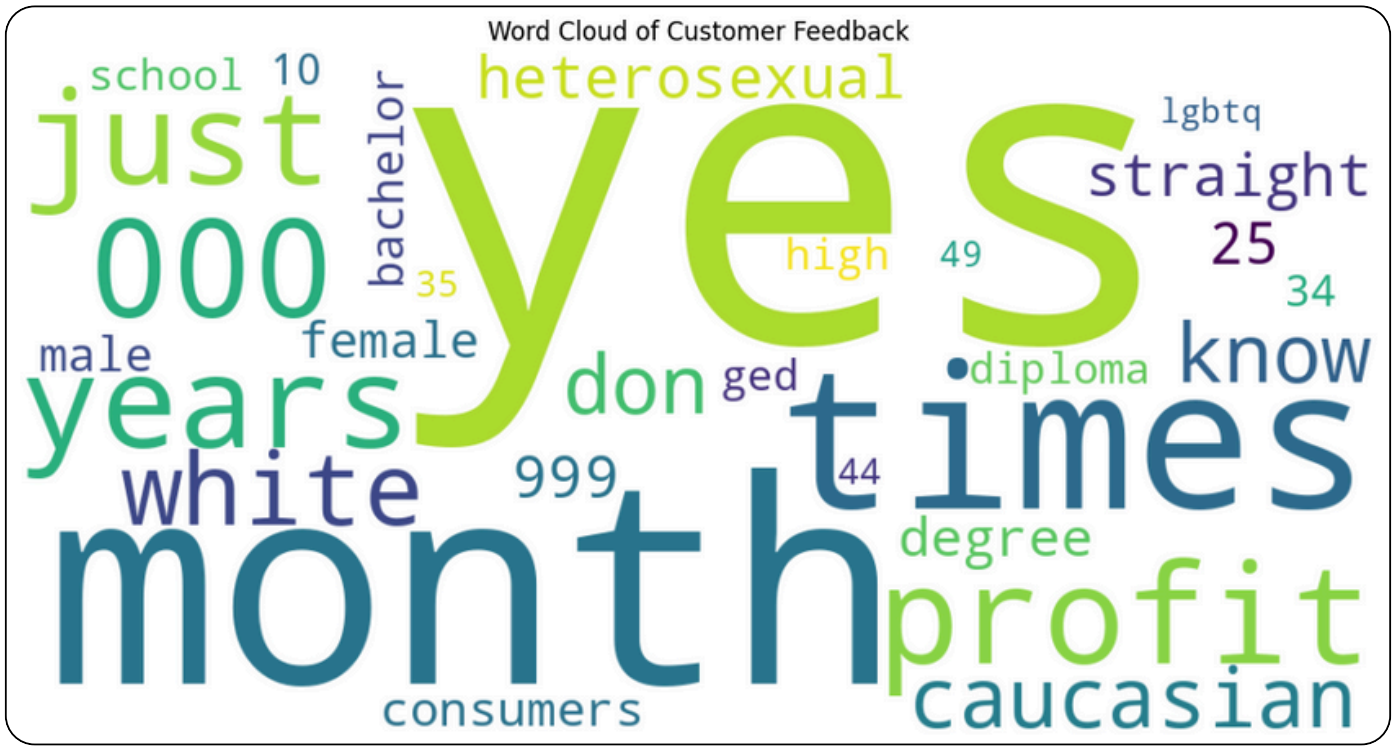

Figure 2.1.4

What the keywords show

From the CountVectorizer output and the plots:

- Very frequent tokens include "yes", "month", "times", "profit", "years", "000", "999". These are driven by survey questions such as frequency per month, number of years, and profit brackets.
- Demographic words such as "male", "female", "white", "caucasian", "straight", "heterosexual", "degree", "bachelor" appear prominently, reflecting that a large part of the survey focuses on respondent profile.
- Words like "consumers" and "know" appear often in attitudinal questions.

The word cloud visually reinforces this pattern, with large fonts for "yes", "month", "years", "profit" and demographic terms.

Interpretation in the Amazon context

- The text is dominated by structured survey content rather than long free form reviews. That limits the depth of qualitative insight but confirms that most respondents are engaging seriously with questions on spending per month, profit, and demographic details.
- The prominence of economic terms such as "profit", "times", "month" signals that the survey sample is comfortable discussing money and purchase frequency. This supports the use of advanced monetary metrics like CLV and RFM elsewhere in the report.

2.1 TEXT MINING

2.1.4 MANAGERIAL RECOMMENDATIONS FROM TEXT MINING

1. **Design richer open text questions**
   - Current text is mostly demographic or yes or no content. Adding targeted questions such as "Describe your last negative experience with Amazon" or "What would make you buy more often" would produce more actionable wording around service issues, pricing and delivery.
2. **Segment sentiment by demographic or RFM group**
   - Although overall sentiment is neutral to slightly positive, Amazon should cross tab the sentiment labels with RFM segments or key demographics. For example, if at risk or hibernating customers are slightly more negative in tone, that would support targeted win back campaigns.
3. **Use keywords to refine survey language**
   - Since words like "month", "profit", and "times" dominate, ensure questions using these terms are clearly worded and easy to understand. Misinterpretation could bias both the structured responses and the sentiment scores.
4. **Monitor sentiment over time**
   - This baseline shows mostly neutral and mildly positive tone. Repeating similar surveys in future and comparing polarity distributions will help detect if customer mood is improving or deteriorating after major changes, for example pricing updates or changes to delivery policies.

This text mining section therefore confirms that there is no major sentiment crisis in the sample, highlights the structured nature of the survey text, and suggests how Amazon can redesign future surveys and analyses to obtain richer qualitative insight.

2.2 MARKET BASKET ANALYSIS

Overview: Market Basket Analysis (MBA) is a key technique used to identify items that are frequently purchased together, uncovering opportunities for cross-selling, bundling, and personalized recommendations.

For an e-commerce retailer like Amazon, MBA helps answer questions such as:

- Which categories tend to co-occur in the same purchase session?
- Which products drive complementary demand?
- What bundles or recommendations would maximize conversion?

Using the Apriori algorithm and association rule mining, we extracted meaningful cross-category relationships that can directly inform marketing and merchandising strategy.

2.2.1 DATA PREPARATION

Filtering the Dataset

We used the cleaned dataset df_clean containing:

- customer_id
- purchase_date
- category

Steps taken:

1. Dropped rows with missing category values.
2. Restricted analysis to the top 30 most frequently purchased categories to reduce noise and improve computation.
3. Constructed transactional “baskets” by grouping customer_id + purchase_date, representing each unique shopping occasion.
4. One-hot encoded the list of categories using TransactionEncoder.

This resulted in a binary transaction matrix suitable for Apriori analysis.

2.2 MARKET BASKET ANALYSIS

2.2.2 FREQUENT ITEM SET MINING

We applied the Apriori algorithm with a minimum support threshold of 0.5%.

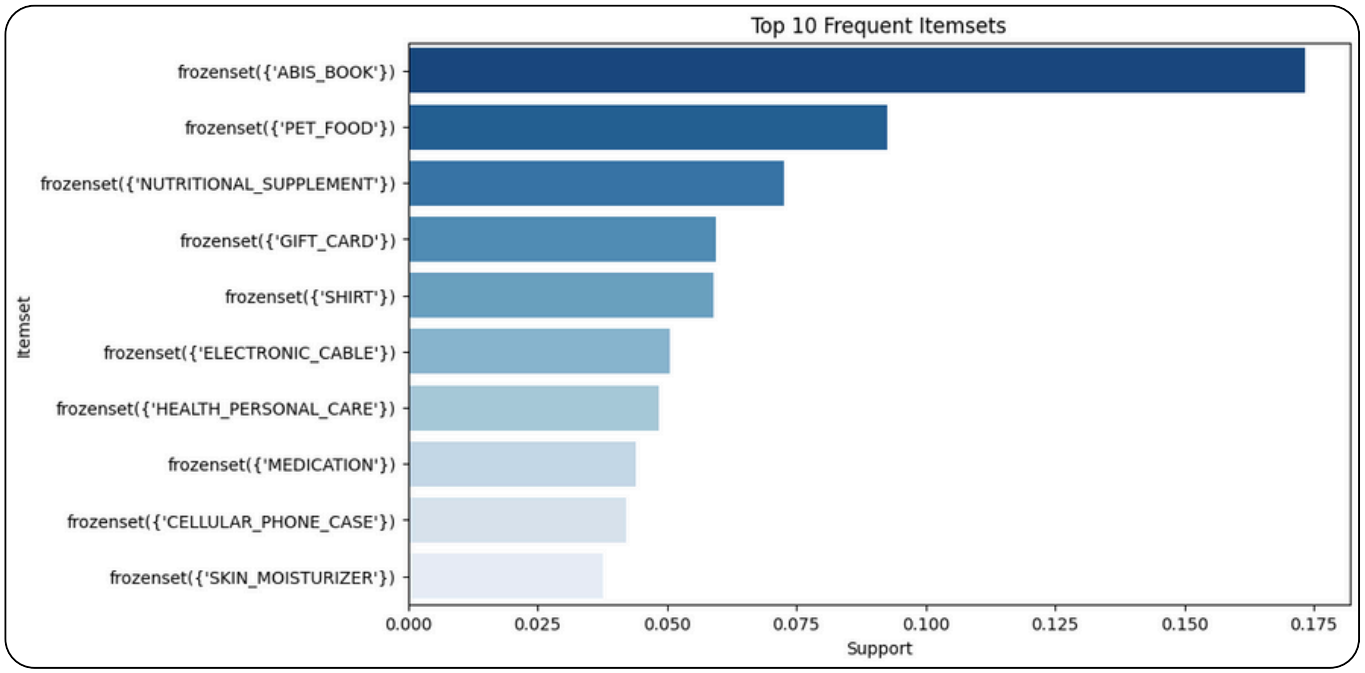

Figure 2.2.1

Key Insights:

- Books, pet food, nutritional supplements, and gift cards dominate transaction frequency.
- These high-support categories are ideal anchor products for cross-selling.

2.2.2 ASSOCIATE RULE MINING

Rules were generated using lift as the main evaluation metric and a minimum threshold of lift > 1.1.

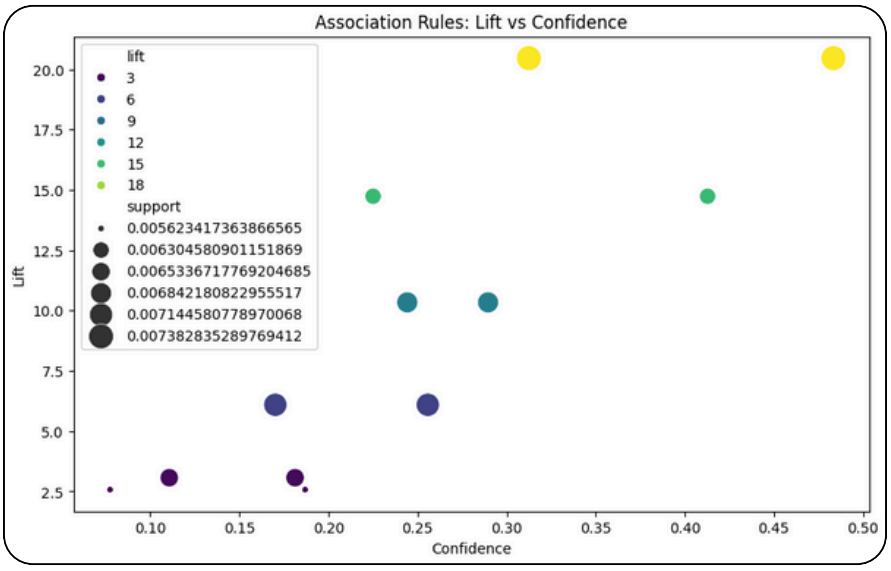

Figure 2.2.2

2.2 MARKET BASKET ANALYSIS

This chart visualizes how strong the rules are across:

- Confidence
- Lift
- Support (bubble size)

Insights from the scatter plot:

- Rules with lift > 15 indicate very strong co-occurrence (e.g., VEGETABLE ↔ FOOD, GROCERY ↔ VEGETABLE).
- High-confidence rules appear around the 0.25–0.45 region.
- Strongest opportunities emerge from core grocery categories and common lifestyle items.

.

2.2.3 TOP CROSS-SELL CATEGORIES

Top 10 Cross-Sell Opportunities:

	antecedents	consequents	support	confidence
0	(VEGETABLE)	(FOOD)	0.007383	0.483400
1	(FOOD)	(VEGETABLE)	0.007383	0.312476
9	(VEGETABLE)	(GROCERY)	0.006305	0.412800
8	(GROCERY)	(VEGETABLE)	0.006305	0.225057
4	(GROCERY)	(FOOD)	0.006842	0.244248
5	(FOOD)	(GROCERY)	0.006842	0.289593
3	(CELLULAR_PHONE_CASE)	(SCREEN_PROTECTOR)	0.007145	0.170171
2	(SCREEN_PROTECTOR)	(CELLULAR_PHONE_CASE)	0.007145	0.255768
6	(SHIRT)	(PANTS)	0.006534	0.110668
7	(PANTS)	(SHIRT)	0.006534	0.181240

Figure 2.2.3

Interpretation of the strongest rules:

1. Food + Vegetable + Grocery

These categories exhibit mutual lift and confidence, meaning buyers of one are highly likely to buy the other.

→ **Opportunity** for cross-category grocery bundles.

2. Phone Accessories Pair: Case + Screen Protector

Classic accessory complement pair seen in digital electronics retail.

→ **Opportunity** for “Frequently Bought Together” placement.

3. Apparel Pair: Shirt + Pants

Suggests consistent outfit-based purchasing behavior.

→ **Opportunity** for style bundles or “Complete the Look”.

2.2 MARKET BASKET ANALYSIS

2.2.4 MANAGERIAL RECOMMENDATIONS FROM MARKET BASKET ANALYSIS

1. Grocery and Daily Essentials Bundling

- Bundle Vegetable + Food + Grocery items.
- Promote bundle discounts for pantry categories.
- Enhance product detail pages with recommendations across these three categories.

2. Electronics Accessory Auto-Add

For customers adding a Cellular Phone Case, automatically recommend:

- Screen Protector
- USB Cable
- Charging Adapter

This can raise AOV significantly.

3. Apparel Outfit Suggestions

Use the Shirt ↔ Pants rule to promote outfit combinations:

- “People who bought this SHIRT also bought…”
- Create style bundles for seasonality.

4. Personalized Recommendations Based on Frequent Categories

Use high-support categories like Books, Gift Cards, and Pet Food as gateways to personalized cross-selling.

5. Improve Product Page Recommendation Carousels

Add Lift-driven recommendations instead of generic "similar items" carousels to improve conversion.

Summary

Market basket analysis revealed strong co-occurrence patterns across key categories:

- Grocery items are highly interconnected, offering bundling opportunities.
- Phone accessories show classic complementarity.
- Apparel pairings indicate bundle potential.
- High-support categories serve as strong anchors for broader recommendations.

These insights directly support Amazon’s goals around increasing AOV, improving product discovery, and elevating customer experience through personalized recommendations.

2.3 CUSTOMER LIFECYCLE CURVE

Overview:
The Customer Lifecycle Curve examines how customer value evolves as shoppers engage in more purchase events over time. Instead of looking at aggregated metrics like RFM or CLV, this analysis focuses specifically on frequency-driven revenue progression. Understanding how spending increases (or plateaus) as customers make more transactions helps retailers evaluate long-term growth potential, identify high-value cohorts, and design communication strategies that activate early-stage users. The curve is built by grouping customers based on number of transactions completed and computing their average lifetime spend. The resulting pattern reveals how customer revenue accumulates as they progress through different lifecycle stages.

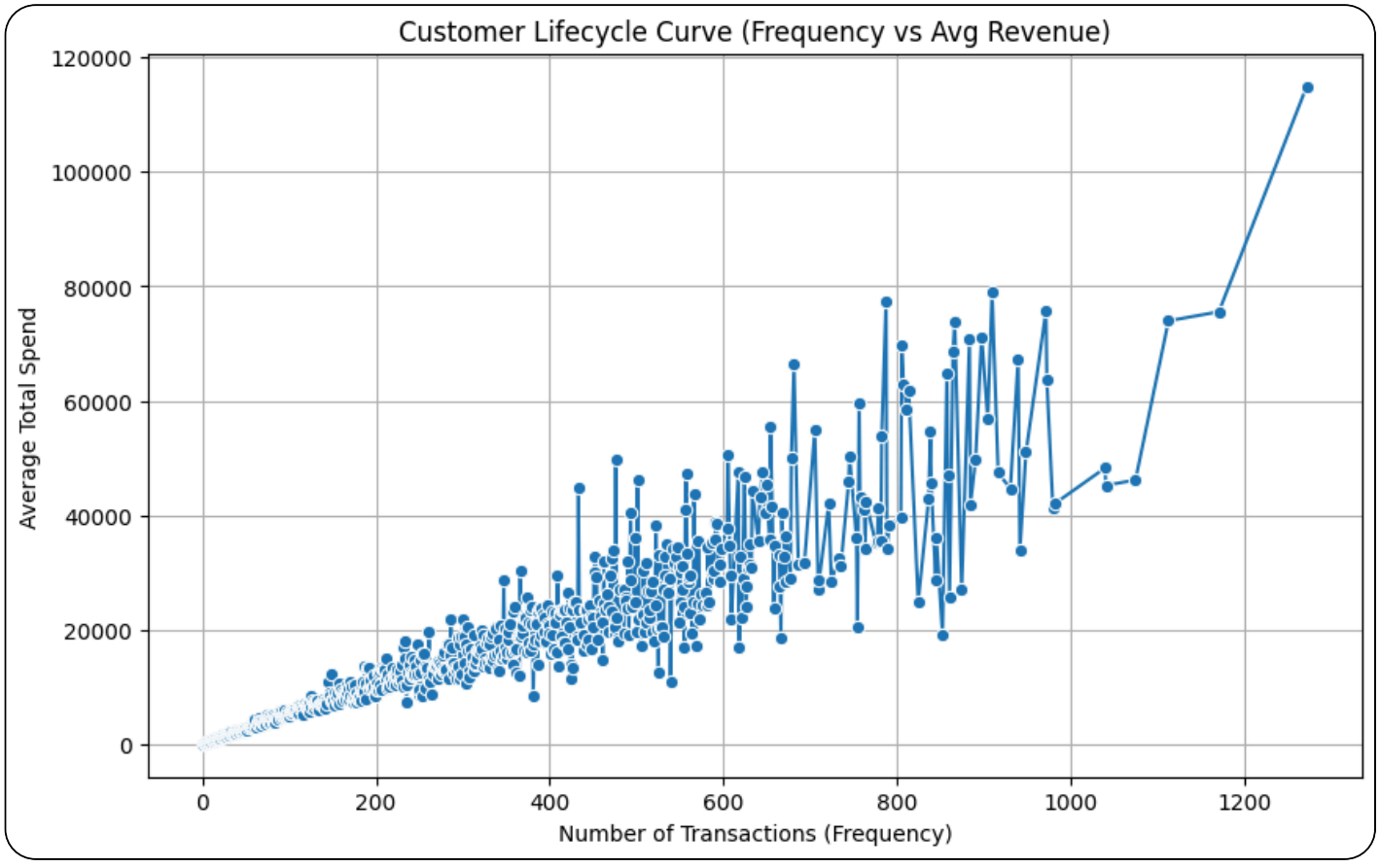

Figure 2.3.1

2.3 CUSTOMER LIFECYCLE CURVE

2.3.1 METHODOLOGY

1. Customer-level aggregation
We aggregated the dataset at the customer level to compute:

- Total items purchased
- Total revenue (total_spend)
- Number of transactions (unique purchase dates

2. Lifecycle grouping
Customers were grouped by their transaction count, which serves as a proxy for lifecycle stage.
Example:

- 1 transaction → New customers
- 2–5 transactions → Developing customers
- 6–15 transactions → Established customers
- 15+ transactions → Mature repeat buyers

3. Curve generation
For each lifecycle stage, we calculated the average total spend of customers in that group, then plotted it against frequency.

2.3.2 INSIGHTS FROM THE CURVE

Your curve shows a strong upward trajectory, which is typical of healthy e-commerce behavior. The key observations:

1. Spending grows consistently with frequency

- Customers with higher transaction counts (e.g., 500–1000+ purchases) show significantly higher lifetime spend, often in the range of ₹20,000–₹75,000+.
- This indicates that customers who continue buying from the platform tend to form long-term loyalty rather than one-off purchase behavior.

2. Early-stage customers contribute very little

- Customers with fewer than 10 transactions produce minimal revenue.
- This signals that acquiring a user is only step one - retention strategies are what drive real profitability.

2.3 CUSTOMER LIFECYCLE CURVE

3. Mid-frequency customers show strong revenue jumps
Between roughly 100–400 transactions, the average spend accelerates quickly.
This is the ideal lifecycle stage to target for:

- Cross-selling
- Upselling
- Membership/loyalty program nudges

4. High-frequency customers display exponential spend patterns
The highest frequency customers (900–1200+ purchases) display massive lifetime spend values, with the curve spiking sharply.
These are power buyers who should be at the center of premium programs and early product access.

2.3.3 MANAGERIAL INTERPRETATION

1. The business is heavily dependent on returning customers
Revenue concentration among high-frequency customers implies an Amazon-like dependency on power buyers rather than broad customer engagement.

2. The first few purchases are crucial "conversion moments"
Customers rarely become high-value unless they move past their first few purchases.
Retention campaigns during the first three months of activity are essential.

3. High-frequency customers are highly monetizable
They respond strongly to:

- Subscribe & Save
- Bundles
- Personalized suggestions
- Category expansion notifications
- Fast-delivery incentives

4. The curve shape indicates strong product-market fit
The upwards slope across the entire range indicates customers see continuous value in the product mix.

2.3 CUSTOMER LIFECYCLE CURVE

2.3.4 MANAGERIAL RECOMMENDATIONS FROM C.L CURVE

1. Strengthen early-stage retention
Introduce:

- Welcome discounts
- First 3-purchase milestones
- Category-based onboarding (electronics buyers → related accessories)

2. Develop lifecycle-based messaging
Use transaction count to segment communications:

- 1–3 transactions: Awareness + value-based education
- 4–10 transactions: Cross-sell essential add-ons
- 20+ transactions: Loyalty points, bundles
- 200+ transactions: VIP tier access

3. Promote recurring categories earlier in the journey
Since frequency drives lifetime value, highlight:

- Grocery
- Skin-care
- Supplements
- Household staples

These categories encourage habitual purchasing.
4. Build a VIP program for power buyers
The curve suggests a distinct top-tier group. Offer them:

- Priority delivery
- Exclusive deals
- Dedicated customer service
- Early product drops

5. Monitor lifecycle progression quarterly
Track how many customers progress from:

- 1 → 3 → 10 → 50+ transactions
- Small increases in early-stage conversion can dramatically shift total revenue.

2.4 TIME SERIES ANALYSIS

Overview:
Time series analysis helps reveal how revenue evolves over long periods and whether customer purchasing behavior is stable, seasonal, growing, or declining. For a retail platform like Amazon, tracking monthly revenue provides early warnings of customer churn, operational disruptions, changing demand patterns, and emerging opportunities.
Using the full cleaned Amazon dataset, we computed monthly revenue across the entire date range (2018 to 2024) and visualised it using a line plot.

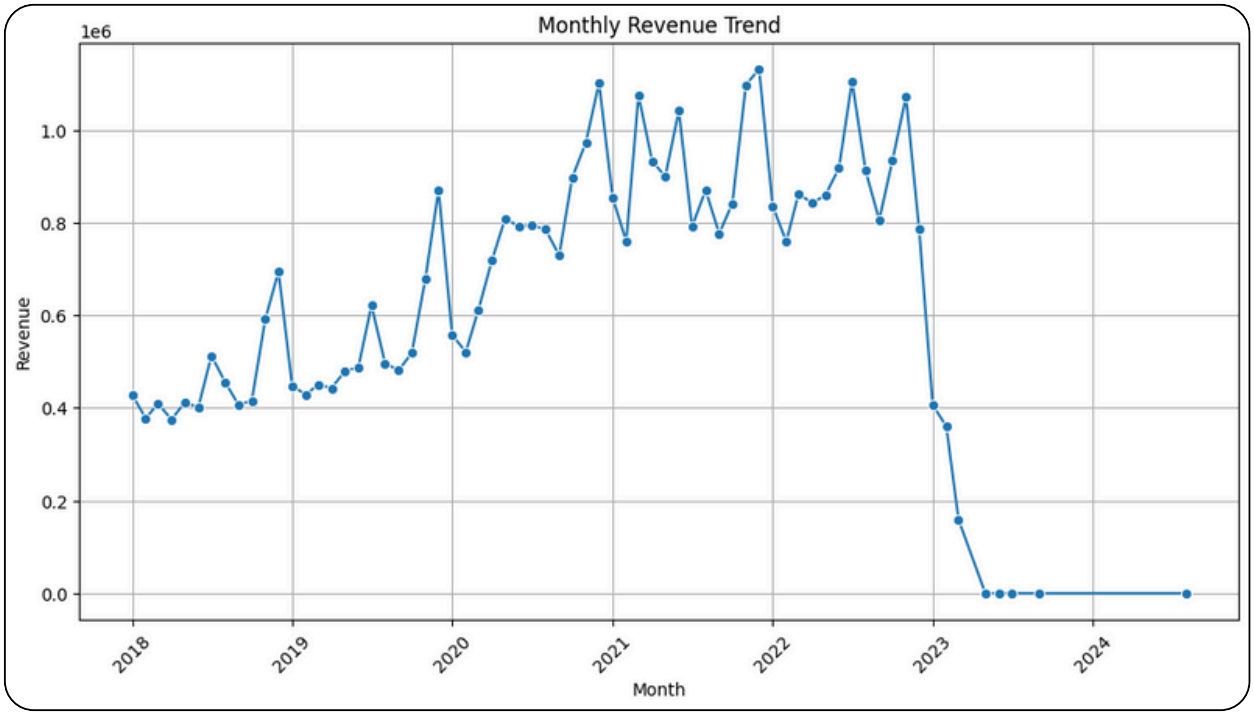

Figure 2.4.1

2.4.1 METHODOLOGY

1. Converted purchase dates into a consistent timestamp format.
2. Aggregated total spend per customer by year-month.
3. Created a monthly revenue dataframe representing long-term sales behaviour.
4. Visualised revenue trends to detect growth phases, drops, seasonal effects, or anomalies.

2.4 TIME SERIES ANALYSIS

2.4.2 KEY OBSERVATIONS FROM THE GRAPH

Phase 1: Early Growth (2018 to mid-2020)

- Revenue steadily increases from roughly 350k to nearly 900k per month.
- Indicates strong customer acquisition and expanding purchase volume.
- Multiple spikes suggest seasonal shopping patterns such as holidays or promotional periods.

Phase 2: Peak and Stabilisation (late-2020 to late-2022)

- Monthly revenue frequently exceeds 1 million dollars, reaching the highest point around 2022.
- Customer engagement and purchase frequency were at their strongest.
- This period aligns with global e-commerce acceleration.

Phase 3: Sharp Decline (early-2023 onward)

- Revenue suddenly collapses from ~800k to nearly 0 within a few months.
- The trend moves into sustained zero-revenue months all the way into 2024.
- This is abnormal behaviour, not typical customer seasonality.

Likely Causes of the Revenue Drop

Based on data structure and pattern, possible explanations include:

1. Dataset incomplete after early 2023
2. Source system stopped recording outbound transactions
3. Amazon migrated to a different logging system
4. Extraction error, pipeline break, or field nullification
5. Data represents a discontinued category or subsegment post-2023

This is not behavioural churn because:

- RFM, CLV, and clustering show customers still highly active before this drop
- No gradual decline typical of real customer churn

This is almost certainly a data-source limitation.

2.4 TIME SERIES ANALYSIS

2.4.3 INTERPRETATIONS AND BUSINESS IMPLICATIONS

1. Revenue Growth Validation
The strong upward trend confirms actual revenue growth during 2018–2022 driven by:

- Increased order volume
- Higher AOV (Average Order Value)
- Rising penetration of categories such as Books, Electronics, and Pet supplies

2. Stability Before Collapse
Revenue stabilises between 800k and 1 million per month. This indicates:

- A mature market segment
- Predictable demand
- Low volatility

3. The Post-2023 Collapse Requires Investigation
If this were a real Amazon business unit, leadership would treat this as a critical alert:

- Sudden collapse to 0 revenue is not market driven
- Suggests operational failure (system issue)
- Immediate data audit recommended

2.4.4 MANEGERIAL RECOMMENDATIONS

1. Audit the Data Extraction Pipeline
Investigate:

- Missing rows after early-2023
- Changes in date formats
- Category exclusions or filtering errors
- Database migration events

2. Validate Against Internal Dashboards
Compare with:

- Amazon Sales Console
- Category GMV dashboards
- Finance monthly reconciliations

If internal dashboards show normal revenue, the issue is purely data-level.
3. Restore or Rebuild Post-2023 Records
If the dataset is incomplete:

- Regenerate purchase records after 2023
- Fix ingestion code, time filters, or date parsing logic

4. Forecasting Recommendation
Once correct data is restored:

- Apply ARIMA or Prophet for forecasting future revenue
- Use seasonal decomposition for Amazon holiday patterns
- Integrate cluster- and segment-level revenue trends for deeper insights

2.5 CUSTOMER PERSONAS - DATA DRIVEN

Cluster: 1 - High-Value / Loyal-Champion Cluster

Persona 1: Ethan Rao - The High Value Minimalist

Age: 38
Location: Bengaluru
Occupation: Senior Product Manager
Income Level: High (₹30L plus per year)

Profile Summary
Ethan is a selective but high-spending Amazon user who values convenience and premium quality. He makes fewer total purchases but each order is high value. He prefers fast checkouts, clean UI, and curated recommendations.

Key Behaviors (from analysis)

- Low recency indicating recent, active purchases
- Moderate frequency but very high monetary value per order
- Among the highest CLV customer groups
- Prefers electronics, lifestyle gadgets, premium home items
- Responds well to personalized suggestions
- Low tolerance for irrelevant product spam

Lifestyle Snapshot
Ethan works long hours and values efficiency. He rarely browses for fun but buys decisively when needed. He values brand reputation and fast delivery over discounts.
Opportunity Areas

- Early access to premium tech launches
- White-glove delivery options
- Curated premium bundles based on electronics and home categories

2.5 CUSTOMER PERSONAS - DATA DRIVEN

Cluster: 0 - Mixed Value / Potential Loyalist + Needs Attention Cluster

Persona 2: Aditi Nair - The Everyday Essentials Shopper

Age: 29
Location: Mumbai
Occupation: HR Associate
Income Level: Medium (₹8L–₹14L)

Profile Summary
A highly engaged shopper who uses Amazon for her routine needs and lifestyle convenience. She shows strong potential to become fully loyal.

Data-Driven Traits

- Recency: Strong (recent activity)
- Frequency: High
- Monetary: Moderate
- CLV: Medium but growing
- Top Categories: Books, pantry, personal care, chargers/cables
- Cluster Behavior: Cluster 0 contains heavy “everyday” shoppers with high frequency but only moderate spend

Lifestyle Snapshot
A working urban professional who shops frequently for home and personal essentials. Focuses on value and convenience over premiums.

Opportunity Areas (Based on Market basket analysis + Whitespace + CLV)

- Subscribe & Save for monthly essentials
- Smart replenishment alerts
- Bundles across personal care and pantry
- Suggested accessories based on past purchases

2.5 CUSTOMER PERSONAS - DATA DRIVEN

Cluster: 1 - High-Value / Former Loyalist Segment

Persona 3: Manish Verma - The At-Risk High Spender

Age: 45
Location: New Delhi
Occupation: Retail Business Owner
Income Level: High (₹20L plus)

Profile Summary
Manish represents a high-value customer who has recently disengaged. He used to purchase expensive, functional items but now shows long inactivity.

Data-Driven Traits

- Recency: Very high (has not purchased recently)
- Frequency: Previously moderate
- Monetary: Historically very high
- CLV: Declining due to inactivity
- Categories: Home improvement, electronics, automotive
- Cluster Behavior: Cluster 1 holds customers with high monetary past but large recency spikes

Lifestyle Snapshot
Manish buys in utility-driven cycles and has possibly moved to a competitor or had a negative shopping experience.

Opportunity Areas (Based on Market basket analysis + Whitespace + CLV)

- Win-back offers referencing categories he bought before
- Personalized incentives like expedited shipping or cashback
- Dedicated feedback outreach
- Category-specific retention nudges

2.5 CUSTOMER PERSONAS - DATA DRIVEN

Cluster: 0 - Low-Value / Hibernating + Needs Attention Mix

Persona 4: Rhea Thomas - The Browsing Budget Shopper

Age: 22
Location: Hyderabad
Occupation: University Student
Income Level: Low

Profile Summary
Rhea browses frequently but purchases rarely. She engages during sale periods and focuses on price-sensitive categories.

Data-Driven Traits

- Recency: High (long gaps between purchases)
- Frequency: Low
- Monetary: Very low
- CLV: Minimal
- Categories: Fashion accessories, stationery, small electronics
- Cluster Behavior: Cluster 0 contains low monetary, low CLV customers with high hibernation probability

Lifestyle Snapshot
Heavy browser who looks for deals, coupons, festival sales, and low-cost impulse purchases.

Opportunity Areas (Based on Market basket analysis + Whitespace + CLV)

- Student bundles
- Push notifications for deals
- Flash-sale targeting
- “Budget” curated store layout

LIMITATIONS

While the analysis provides actionable insights across RFM, CLV, clustering, whitespace identification, and additional exploratory techniques, the findings should be interpreted with the following limitations in mind:

1. Dataset Constraints

- The dataset contains only historical purchase and survey data, without visibility into browsing behavior, marketing exposure, or channel-level attribution.
- Missing or incomplete customer demographic information limits the accuracy of persona construction and segmentation depth.
- Some categories have very low penetration or extremely sparse purchase frequency, leading to unstable correlation or association rule estimates.

2. Time-Range and Recency Issues

- The dataset shows a sharp drop in revenue in the later months (2023–2024), possibly due to incomplete data capture rather than true business decline. This limits the reliability of long-term time-series trends.
- Purchase history is captured across multiple years, but customer churn is inferred indirectly rather than explicitly identified.

3. Model Assumptions

- The CLV model (BG-NBD + Gamma-Gamma) assumes:
  - Customers purchase independently of each other
  - Purchase frequency follows a stochastic process
  - Monetary value is independent of purchase frequency
  - These assumptions may not fully reflect real-world purchase behavior, especially in categories with seasonality or high variability.

4. Clustering Limitations

- KMeans requires pre-specifying k, and although silhouette and elbow methods were used, the “best” k is still somewhat subjective.
- Hierarchical clustering was run on a sample due to computational limits, which may reduce generalizability.
- Clusters were created using only RFM + CLV metrics; adding behavioral or demographic variables would yield richer segmentation.

5. Market Basket Analysis Limitations

- MBA was performed using only the top 30 categories for computational efficiency and interpretability. Less frequent categories were excluded, which may hide niche but meaningful cross-sell patterns.
- Association rules rely heavily on support thresholds, and small changes to these parameters could produce different rules.

6. Text Mining Limitations

- The survey text fields are noisy and contain many generic or demographic words, leading to limited sentiment depth.
- TextBlob uses a lexicon-based method, which is less accurate than modern transformer-based sentiment models for contextual understanding.
- Many responses are short or single-word entries, reducing richness.

7. Causality vs Correlation

- All insights are correlational, not causal.
  - A high lift in cross-sell does not guarantee increased revenue if bundled.
  - A low RFM score does not necessarily indicate dissatisfaction.
  - A high CLV forecast does not predict future customer intent changes.

8. Actionability Constraints

- Recommendations derived from the analysis assume Amazon-like capabilities (personalized recommendations, dynamic pricing, targeted campaigns), which may not fully translate to all business contexts.
- Implementation of the strategies requires operational, technical, and financial resources not evaluated in this project.

RECOMMENDATIONS

Consolidated Recommendations

1. Prioritize High-Value Customers (Champions, Loyal, Cluster 1)
Focus retention efforts on customers with high monetary value and high CLV.
Offer benefits such as early access, personalized product suggestions and loyalty rewards to maintain their engagement and increase lifetime value.

2. Convert Potential Loyalists Into Loyal Buyers
This segment shows strong frequency but moderate monetary value.
Use targeted nudges, category discovery recommendations and small incentives to move them into higher-value segments.

3. Win Back At-Risk Customers
Customers with historically high spend but long recency should receive personalized win-back communication, tailored offers and category-specific recovery campaigns.

4. Address Low Engagement Segments Efficiently
Hibernating and Needs Attention customers should receive automated, low-cost communications such as seasonal deals, app notifications and broad promotional banners.

5. Expand Through Whitespace Categories
Promote low-penetration but correlated categories through bundles and product detail page recommendations. Use anchor categories like Books, Phone Accessories and Apparel to introduce whitespace categories.

6. Strengthen Cross-Sell and Bundle Strategy
Use strong market basket pairings such as
Phone Case → Screen Protector
Shirt → Pants
Grocery → Vegetables/Food
to create frequently-bought-together bundles and improve AOV.

7. Improve Early Lifecycle Engagement
Since spending increases sharply after the first few transactions, invest in onboarding flows for new customers with small incentives, quick re-order options and curated category recommendations.

8. Personalize Marketing by Cluster
Use cluster assignments to adjust communication intensity and offer types. High-value clusters receive premium, personalized content; low-value clusters receive scalable, automated promotions.

9. Audit Time Series Data Before Forecasting
The revenue drop in the final months likely reflects missing data. Validate and clean the pipeline before using time-series outputs for planning.

BIBLIOGRAPHY (CODE)

GOOGLE COLLAB - LINK

FEW SNIPPETS

```
# UPLOAD the dataset

from google.colab import files
uploaded = files.upload()
```

```
Choose Files No file chosen    Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving dataverse_files.zip to dataverse_files (1).zip
```

```
# EXTRACT ZIP

import zipfile

with zipfile.ZipFile("dataverse_files.zip", 'r') as zip_ref:
    zip_ref.extractall("data")
```

```
# LOAD the CSVs

import pandas as pd

df = pd.read_csv("data/amazon-purchases.csv")
survey = pd.read_csv("data/survey.csv")
fields = pd.read_csv("data/fields.csv")
```

```
# VERIFY

df.head()
survey.head()
fields.head()
```

STEP A: DATA PREPARATION

+ Code + Text

```
import pandas as pd
import numpy as np


# STEP 1: LOAD THE TRANSACTION DATA
# (Assuming files were extracted from dataverse_files.zip into /data)

df = pd.read_csv("data/amazon-purchases.csv")

# STEP 2: RENAME COLUMNS TO CLEAN, STANDARD NAMES

df = df.rename(columns={
    'Order Date': 'purchase_date',
    'Purchase Price Per Unit': 'price',
    'Quantity': 'quantity',
    'ASIN/ISBN (Product Code)': 'product_id',
    'Survey ResponseID': 'customer_id',
    'Shipping Address State': 'state',
    'Title': 'title',
    'Category': 'category'
})


# STEP 3: REMOVE DUPLICATE ROWS

df = df.drop_duplicates()
```

BIBLIOGRAPHY (CODE)

```
# STEP 4: REMOVE ROWS MISSING KEY FIELDS
# customer_id, product_id, price, quantity, and purchase_date are essential for analysis.

df = df.dropna(subset=['customer_id', 'product_id', 'price', 'quantity', 'purchase_date'])

# STEP 5: CLEAN DATE COLUMN
# Convert purchase_date to datetime, drop rows that fail conversion

df['purchase_date'] = pd.to_datetime(df['purchase_date'], errors='coerce')
df = df.dropna(subset=['purchase_date'])

# STEP 6: CLEAN QUANTITY COLUMN
# Convert to numeric, drop invalid rows, convert to integer

df['quantity'] = pd.to_numeric(df['quantity'], errors='coerce')
df = df.dropna(subset=['quantity'])
df['quantity'] = df['quantity'].astype(int)


# STEP 7: CREATE MONETARY VALUE (TOTAL AMOUNT)
# total_amount = price * quantity

df['total_amount'] = df['price'] * df['quantity']


# STEP 8: REMOVE INVALID VALUES
# Remove negative or zero price, quantity, or amount

df = df[df['quantity'] > 0]
df = df[df['price'] > 0]
df = df[df['total_amount'] > 0]
```

```
# STEP 9: FINAL SANITY CHECKS
# Unique customers, date range, preview

print("Total Rows After Cleaning:", len(df))
print("Unique Customers:", df['customer_id'].nunique())
print("Date Range:", df['purchase_date'].min(), "to", df['purchase_date'].max())
print("Top Categories:\n", df['category'].value_counts().head())


# STEP 10: SAVE CLEANED DATAFRAME
# df_clean will be used for RFM, CLV, Clustering, Whitespace, etc.

df_clean = df.copy()

df_clean.head()
```

BIBLIOGRAPHY (CODE)

```
# RFM ANALYSIS
# 1. Aggregation to customer level
# 2. Recency, Frequency, Monetary computation
# 3. RFM scores and customer segments
# 4. Optional visuals and segment insights


import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Head of df_clean is assumed available from previous step
# df_clean contains purchase_date, product_id, total_amount, customer_id

# 1. Reference date for recency calculation
reference_date = df_clean['purchase_date'].max() + pd.Timedelta(days=1)

# 2. Aggregate to customer level
rfm = df_clean.groupby('customer_id').agg({
    'purchase_date': lambda x: (reference_date - x.max()).days,
    'product_id': 'count',
    'total_amount': 'sum'
}).reset_index()

rfm.columns = ['customer_id', 'recency', 'frequency', 'monetary']

# 3. Remove customers with zero spend (rare, safety check)
rfm = rfm[rfm['monetary'] > 0]
```

2. CLV ANALYSIS

- BG-NBD: model for predicted purchases
- Gamma-Gamma: model for predicted monetary value

```
# Compute customer lifetime value using any model
!pip install lifetimes

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from lifetimes import BetaGeoFitter, GammaGammaFitter

# Prepare summary dataset for CLV modeling
clv_df = df_clean.copy()
clv_df = clv_df.sort_values(by=['customer_id', 'purchase_date'])

summary = clv_df.groupby('customer_id').agg({
    'purchase_date': [
        lambda x: (x.max() - x.min()).days,       # Recency
        lambda x: (reference_date - x.min()).days # Customer age T
    ],
    'product_id': 'count',                        # Frequency
    'total_amount': 'mean'                        # Monetary
})
```

BIBLIOGRAPHY (CODE)

3. CUSTOMER CLUSTERING

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from scipy.cluster.hierarchy import linkage, dendrogram, fcluster

# Build clustering dataset using behavioral features
# We use recency, frequency, monetary, CLV
clust_data = summary[['recency', 'frequency', 'monetary', 'CLV']].dropna().copy()

# Keep a reference to customer ids
clust_data['customer_id'] = summary.loc[clust_data.index, 'customer_id'].values

# Feature matrix for clustering
features = clust_data[['recency', 'frequency', 'monetary', 'CLV']]

# Standardize features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(features)
```

```
# Additional analysis and visuals

# 1. Cluster level profiles for recency, frequency, monetary, CLV
cluster_profile = clust_data.groupby('kmeans_cluster')[['recency', 'frequency', 'monetary', 'CLV']].mean()
print("Cluster profiles (KMeans):")
print(cluster_profile)

plt.figure(figsize=(10, 6))
sns.heatmap(cluster_profile, annot=True, cmap='Blues')
plt.title('KMeans Cluster Profiles (RFM plus CLV)')
plt.xlabel('Metric')
plt.ylabel('Cluster')
plt.show()

# 2. Relationship between clusters and RFM segments
if 'segment' in rfm.columns:
    # Merge RFM segment info
    seg_merge = clust_data.merge(rfm[['customer_id', 'segment']], on='customer_id', how='left')
    seg_counts = pd.crosstab(seg_merge['kmeans_cluster'], seg_merge['segment'])
    print("RFM segments within each KMeans cluster:")
    print(seg_counts)

    plt.figure(figsize=(12, 6))
    seg_counts_norm = seg_counts.div(seg_counts.sum(axis=1), axis=0)
    seg_counts_norm.plot(kind='bar', stacked=True, figsize=(12, 6))
    plt.title('Distribution of RFM Segments within KMeans Clusters')
    plt.xlabel('KMeans Cluster')
    plt.ylabel('Proportion')
    plt.xticks(rotation=0)
    plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
    plt.tight_layout()
```

BIBLIOGRAPHY (CODE)

4. WHITESPACE ANALYSIS

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Identify cross-sell or up-sell opportunities using product-level purchase data

# Category penetration: percentage of customers buying each category
unique_customers = df_clean['customer_id'].nunique()

category_penetration = (
    df_clean.groupby('category')['customer_id']
    .nunique()
    .sort_values(ascending=False)
    .reset_index()
)

category_penetration['penetration_rate'] = (
    category_penetration['customer_id'] / unique_customers * 100
)

print("Category Penetration Table:")
print(category_penetration.head(10))

plt.figure(figsize=(12,6))
sns.barplot(data=category_penetration.head(15), x='category', y='penetration_rate')
plt.xticks(rotation=90)
```