**Tathya AI**

# DATA-DRIVEN CUSTOMER SEGMENTATION FOR STRATEGIC MARKETING DECISIONS

**Prepared by Pratik Sharma**

Tathya AI
www.tathyaai.in

# Executive Summary

This report presents a data-driven approach to customer segmentation designed to strengthen strategic marketing decisions. Using Python for data cleaning, transformation, and RFM (Recency, Frequency, Monetary) modeling, customers were segmented based on their purchase behavior and value contribution. The processed dataset was stored and managed in **SQL Server on Microsoft Azure**, enabling secure, scalable access for analytics and dashboard integration.

Through **K-Means clustering**, distinct customer groups were identified, revealing key behavioral patterns such as frequency of purchases, spending capacity, and engagement trends. These insights were visualized using an **interactive Power BI dashboard**, allowing marketing and business teams to dynamically explore customer segments and derive actionable strategies.

By integrating cloud-based data management with advanced analytics and visualization, this project demonstrates how organizations can shift from intuition-driven decisions to **evidence-based marketing**, leading to improved targeting, customer retention, and overall business growth.

# Introduction

In dynamic industries such as **FMCG, retail, and logistics**, understanding customer behavior is crucial for maintaining competitiveness and driving profitability. These sectors operate in high-volume, low-margin environments where success depends on efficient resource allocation, optimized supply chains, and precisely targeted marketing strategies.

**Customer segmentation** enables businesses to categorize their customers into distinct groups based on their buying frequency, spending patterns, and engagement levels. In the **FMCG and retail sectors**, segmentation helps identify loyal customers, frequent buyers, and price-sensitive shoppers—allowing marketers to design personalized campaigns, optimize inventory, and introduce relevant product bundles. For the **logistics sector**, segmentation provides valuable insights into client types, shipment frequency, and service-level expectations, helping firms enhance delivery efficiency and customer satisfaction.

By leveraging data-driven segmentation, companies can transition from intuition-based decisions to **evidence-based marketing and operational planning**. This approach empowers organizations to prioritize high-value customers, tailor retention strategies, and improve profitability while maintaining agility in a rapidly changing marketplace.

This project applies analytical techniques such as **RFM modeling, clustering, and Power BI visualization**, integrated through **SQL Server on Azure**, to demonstrate how customer segmentation can guide strategic decision-making across FMCG, retail, and logistics domains.

# Project Goals and Expected Outcomes

The primary goal of this project is to leverage **data analytics and cloud-based tools** to develop a **data-driven customer segmentation model** that supports strategic marketing, sales, and operational decisions in FMCG, Retail, and Logistics sectors.

**Project Goals**

1. **Analyze Customer Behavior:**
   Understand customer purchasing patterns using RFM (Recency, Frequency, Monetary) analysis to quantify engagement and loyalty levels.
2. **Segment Customers Effectively:**
   Apply clustering algorithms (such as K-Means) to group customers into distinct segments based on their transactional and behavioral data.
3. **Integrate Cloud Data Infrastructure:**
   Utilize **SQL Server on Microsoft Azure** to securely store, manage, and retrieve processed data for scalable analytics and visualization.
4. **Develop Interactive Dashboards:**
   Build a **Power BI dashboard** that presents customer segments, KPIs, and trends in an intuitive, decision-friendly format for business stakeholders.
5. **Enable Data-Driven Decision-Making:**
   Provide actionable insights that empower marketing, sales, and logistics teams to optimize campaigns, improve retention, forecast demand, and allocate resources efficiently.

**Expected Outcomes**

- Identification of high-value, at-risk, and low-engagement customer groups.
- Improved marketing ROI through targeted and personalized strategies.
- Enhanced supply chain planning in FMCG and logistics via demand insights.
- A scalable, cloud-enabled analytics framework for future predictive modeling.
- Greater alignment between customer insights and business strategy.

# About the Data

The dataset used in this project is sourced from transactional records that capture customer purchase activity. It is representative of real-world scenarios in **FMCG, retail, and logistics** operations, where large volumes of transactions occur daily.

**Key Attributes of the Dataset**

The dataset contains the following core fields:

- **Customer ID:** Unique identifier for each customer.
- **InvoiceDate:** Date of purchase or shipment.
- **Description:** Description of purchased goods or services.
- **Quantity:** Number of items purchased in a transaction.
- **Price:** Price per item.
- **Invoice:** Unique identifier for each transaction.
- **StockCode:** Product (item) code. Nominal. A 5-digit integral number uniquely assigned to each distinct product.
- **Country:** The name of the country where a customer resides.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1067371 entries, 0 to 1067370
Data columns (total 8 columns):
 #   Column       Non-Null Count    Dtype  
---  ------       --------------    -----  
 0   Invoice      1067371 non-null  object 
 1   StockCode    1067371 non-null  object 
 2   Description  1062989 non-null  object 
 3   Quantity     1067371 non-null  int64  
 4   InvoiceDate  1067371 non-null  object 
 5   Price        1067371 non-null  float64
 6   Customer ID  824364 non-null   float64
 7   Country      1067371 non-null  object 
dtypes: float64(2), int64(1), object(5)
memory usage: 65.1+ MB
```

# Methodology

This project follows a structured analytical workflow that integrates **Python-based data processing**, **cloud data management via Azure SQL**, and **interactive visualization through Power BI**. The end-to-end methodology ensures scalability, accuracy, and practical business relevance across the **FMCG, retail, and logistics** sectors.

**1. Data Preparation (Python)**

The initial dataset was imported into Python for preprocessing and cleaning. This step involved:

- **Handling missing and duplicate values** to maintain data quality.
- **Filtering invalid transactions**, such as cancelled or zero-quantity orders.
- **Standardizing date and currency formats** for consistency.
- **Calculating total transaction value** (Revenue = Quantity × UnitPrice) for monetary analysis.

The cleaned data was then exported for RFM scoring.

### 2. RFM Modeling

To quantify customer engagement and value, **RFM (Recency, Frequency, Monetary)** analysis was performed:

- **Recency:** Number of days since the last purchase.
- **Frequency:** Number of purchases made by each customer.

```
# Convert 'InvoiceDate' to datetime objects
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])

# Set a reference date (usually the max InvoiceDate + 1 day)
reference_date = df['InvoiceDate'].max() + pd.Timedelta(days=1)

rfm = df.groupby('Customer ID').agg({
    'InvoiceDate': lambda x: (reference_date - x.max()).days,  # Recency
    'Invoice': 'nunique',                                      # Frequency
    'Price': 'sum'                                             # Monetary
}).reset_index()

rfm.rename(columns={
    'InvoiceDate': 'Recency',
    'Invoice': 'Frequency',
    'Price': 'Monetary'
}, inplace=True)
```

- **Monetary:** Total spending by each customer.

Each metric was calculated using Python's pandas library, and customers were scored on each dimension. This provided a clear numerical representation of their loyalty and activity.

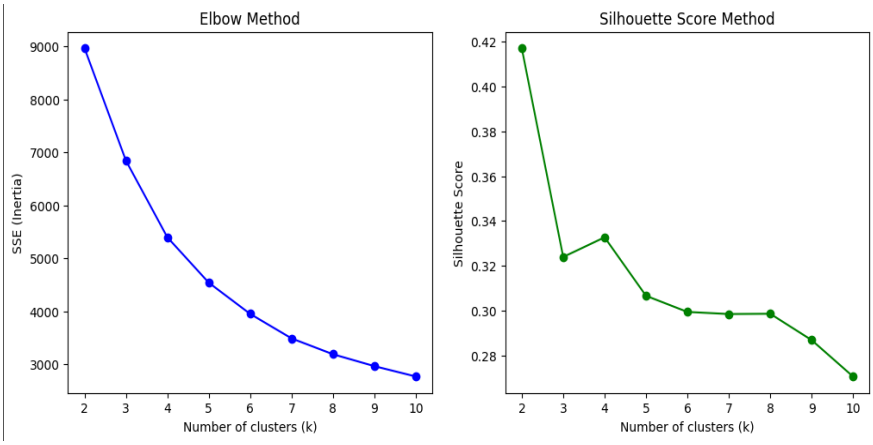
### 3. Clustering and Segmentation

The RFM data was normalized using **Min-Max scaling** to ensure equal weight across metrics. Next, **K-Means clustering** was applied to segment customers into distinct groups based on similar purchasing behavior.

The **Elbow Method** and **Silhouette Score Method** was used to determine the optimal number of clusters, balancing model accuracy and interpretability.
Each cluster was labeled based on its RFM profile—for example:

- *Valuable but slipping customers.*
- *Lost- low value Customer*
- *Recent and valuable customers*
- *Top champions*

These clusters provided the foundation for actionable business insights.

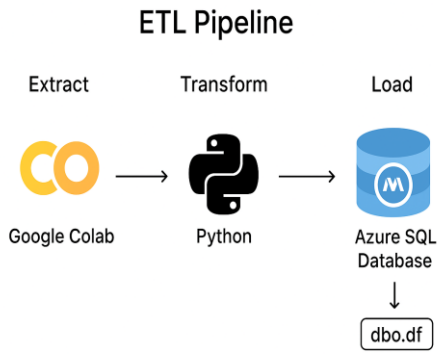



### 4. Cloud Integration with Azure SQL

After segmentation, the final dataset was uploaded to **Microsoft Azure SQL Database**.
This cloud integration enabled:

- Centralized and secure data storage.
- Real-time access to analytical outputs.
- Seamless connection with **Power BI** for dashboard creation.

Azure's scalability and performance made it suitable for large transaction datasets typical of FMCG and retail operations.




### 5. Power BI Dashboard Development

The segmented dataset stored in Azure SQL was connected directly to **Power BI Service**.
An **interactive dashboard** was designed to visualize:

- Segment distribution and size.
- Total revenue contribution by segment.
- Recency and frequency trends.
- Regional and product-level insights (for FMCG/Retail).
- Customer demand and shipment patterns (for Logistics).

Dynamic filters and drill-through features were incorporated, enabling business teams to explore customer patterns intuitively and identify opportunities for marketing optimization, resource allocation, and operational efficiency.

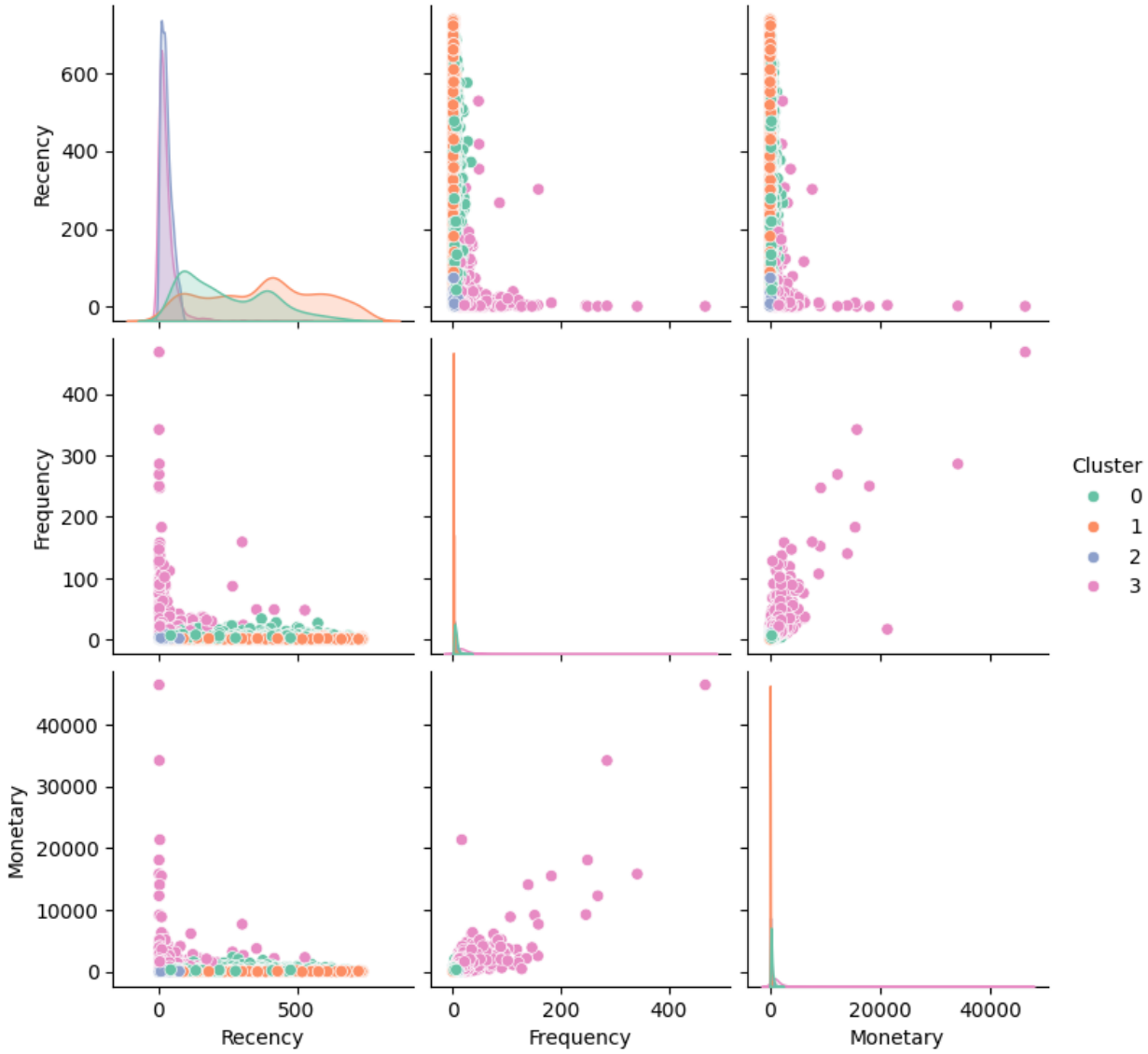

# Analysis and Findings

After applying RFM modeling and K-Means clustering, customers were segmented into distinct groups based on their purchasing frequency, spending behavior, and recency of transactions. The segmentation revealed clear behavioral patterns that provide actionable insights for marketing, sales, and operational strategy across FMCG, retail, and logistics sectors

## 1. Cluster Overview

| Cluster | Segment Name | Key Traits | Strategic Focus |
|---|---|---|---|
| 0 | Valuable but Slipping Customers | Customers who have contributed significant revenue historically but show declining engagement or reduced purchase frequency. | Re-engagement campaigns, personalized reminders, or targeted discounts to win them back. |
| 1 | Lost Low-Value Customers | Low-spending customers who have not purchased for a long time; minimal contribution to revenue. | Cost-effective outreach or minimal targeting; focus resources elsewhere. |
| 2 | Recent and Valuable Customers | New or recently active customers showing high spending potential and growing engagement. | Strengthen relationship with personalized offers and onboarding campaigns to build loyalty. |
| 3 | Top Champions | Highly loyal customers with frequent purchases, strong monetary value, and recent engagement. | Maintain loyalty through exclusive offers, early access programs, or premium benefits. |

**This segmentation highlights the diversity of the customer base and allows for precise, data-driven marketing strategies.**




**General Observations:**

- Cluster 0 (Green): Very low frequency & monetary, with high recency → Dormant/One-time buyers.
- Cluster 1 (Orange): Moderate recency, low frequency & low spending → At-risk customers.
- Cluster 2 (Blue/Purple): Very recent purchases but still low frequency & monetary → New customers / Potential Loyalists.
- Cluster 3 (Pink): Higher frequency & monetary spread, more recent → Champions & Loyal customers.

**Breakdown by RFM:**

**Recency:**

- Cluster 0 (Green): Long time since last purchase (high recency) → Churned/dormant.
- Cluster 1 (Orange): Not very recent either → At-risk.
- Cluster 2 (Blue): Very recent → New customers.
- Cluster 3 (Pink): Mostly recent → Active champions/loyal**.**

**Frequency:**

- Cluster 0 & 1: Low frequency.
- Cluster 2: Slightly better but still low → They've just started buying.
- Cluster 3: Clearly the highest → Frequent buyers.

**Monetary:**

- Cluster 0 & 1: Low spenders.
- Cluster 2: Low spenders but potential to grow.
- Cluster 3: High spenders (biggest revenue source).

**Interpretation:**

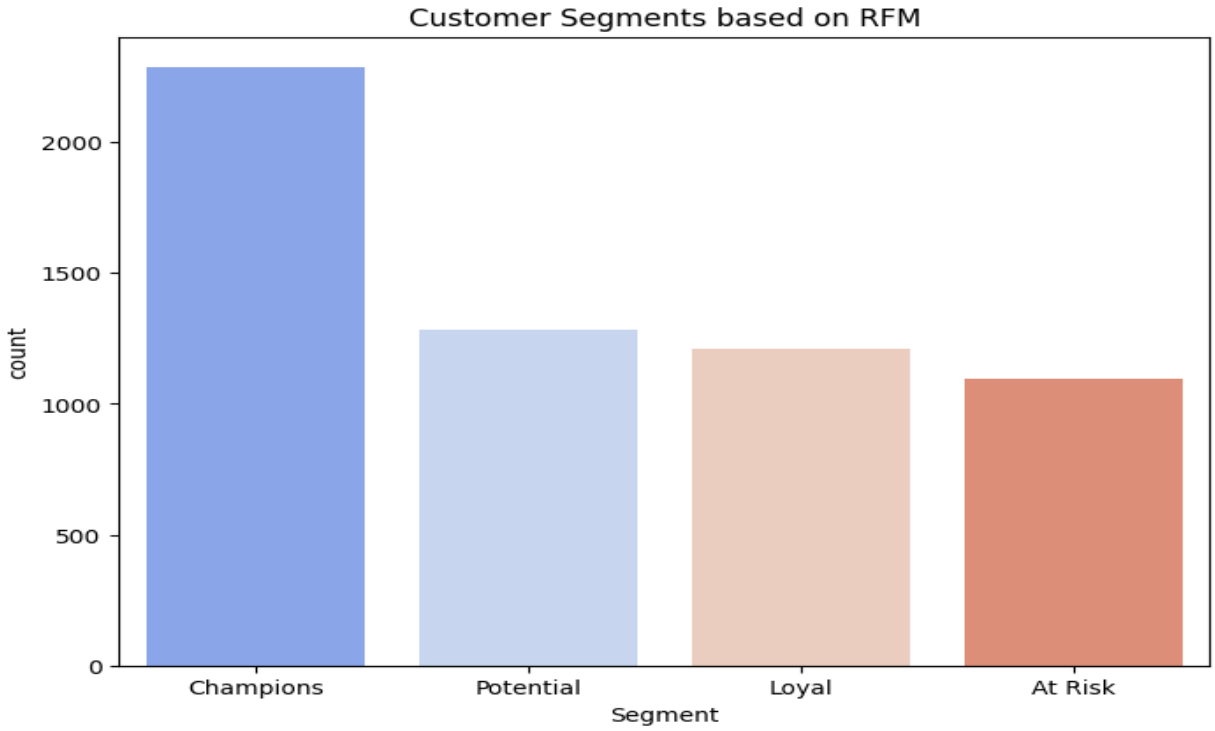
- Cluster 0 (Green) → Dormant/One-time buyers Haven't bought in a long time, low value.
- Cluster 1 (Orange) → At-risk customers Used to buy, but less recent now.
- Cluster 2 (Blue) → New customers / Potential loyalists Just bought recently, haven't built frequency yet.
- Cluster 3 (Pink) → Champions & Loyal customers Recent, frequent, high spenders (most valuable**)**

# 2. Behavioral Patterns and Insights

**Recency Trends:**
Loyal customers consistently engage with recent transactions, while at-risk groups show extended inactivity periods—highlighting the need for retention efforts.

**Frequency Distribution:**
In FMCG and retail, frequent purchasers often align with small but regular orders, reflecting routine consumption behavior. In logistics, repeat clients tend to represent steady B2B shipments, indicating service reliability.

**Monetary Value Insights:**
A small percentage of customers contribute a large share of total revenue (typical Pareto pattern). Prioritizing this group can enhance marketing ROI and profitability.



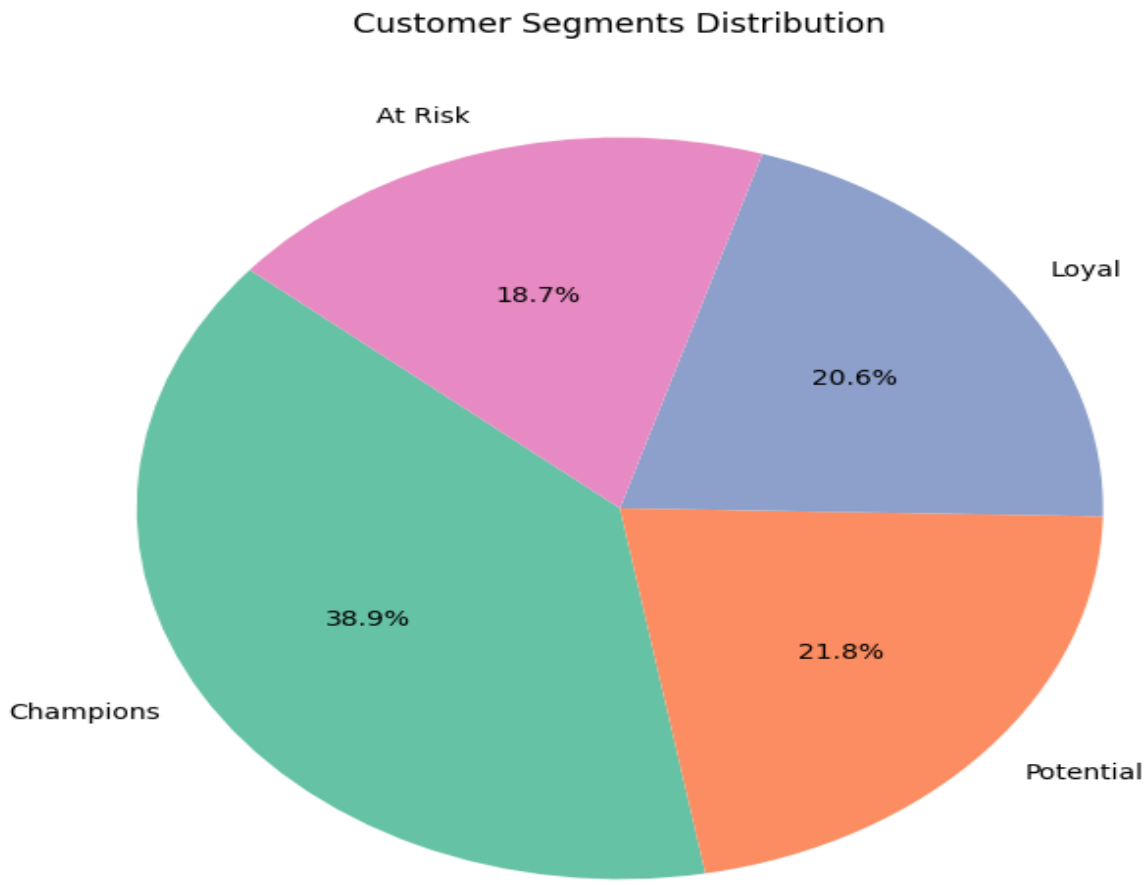

The above chart shows the segmentation of customers based on Behavioral analysis (RFM).

**Champions (2000+ customers) --**These are your most valuable customers → bought recently, buy frequently, and spend a lot.

**Loyal Customers (~same as Potential) --**Loyal: buy often, but maybe not the highest spenders.

**Potential Loyalists--** newer customers showing promise (good recency & frequency, not yet big spenders).

**At Risk (~1000 customers)** --Once valuable but haven't bought recently.




**This Pie Chart clearly shows that-**

Nearly **40% of revenue** is concentrated among **Champions**, emphasizing the importance of customer retention.

**At-Risk and Potential** customers together form ~40% of the base — the **biggest growth opportunity**.

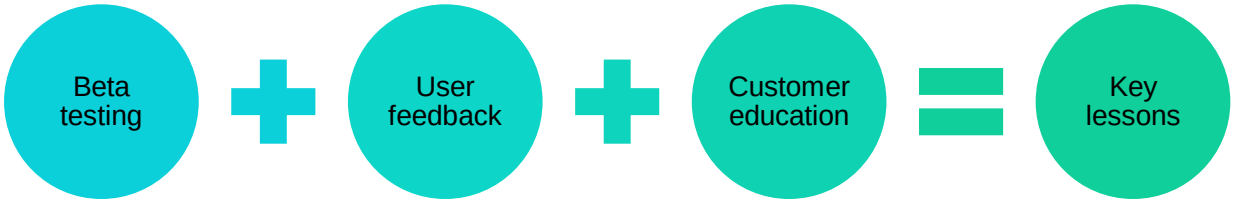
The segmentation highlights clear **action paths** for marketing teams to design targeted retention and upselling strategies.

## 4. Challenges and adjustments

The primary challenge encountered was a higher-than-expected customer support volume due to initial bugs. The team quickly addressed these issues by deploying a patch update within two weeks, improving app stability. Customer education through webinars helped mitigate confusion regarding app features.

## 5. Lessons learned

Key lessons include the importance of rigorous beta testing prior to launch and the value of proactive customer education. Incorporating user feedback early in the development phase ensured a more refined final product.




## 6. Next steps

Future plans include introducing AI-powered financial insights and expanding Taskerr to new international markets. A loyalty rewards program is also being developed to incentivize continued app usage.