Grocery Sales Analysis

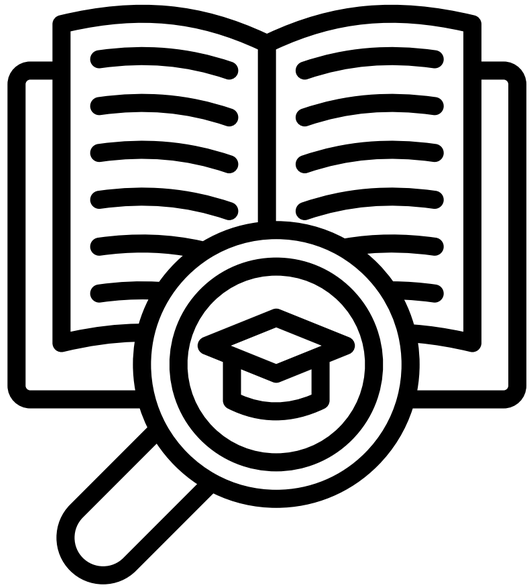
EXPLORATORY DATA ANALYSIS
& CUSTOMER BEHAVIOUR

Iftanul Ibnu Rochman

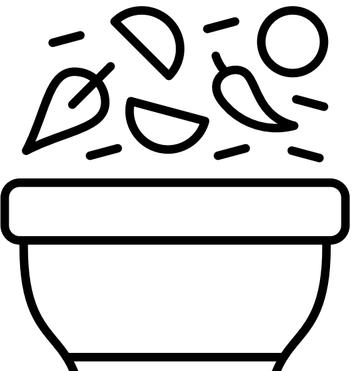
INTRODUCTION

In today’s fast-moving consumer goods (FMCG) landscape, understanding sales performance and customer behaviour is essential for driving profitability, optimizing inventory, and designing effective marketing strategies. This project begins with an Exploratory Data Analysis to build a clear understanding of the underlying sales patterns, product dynamics, and customer interactions before moving into deeper analytical methods such as segmentation and behavioural modeling.

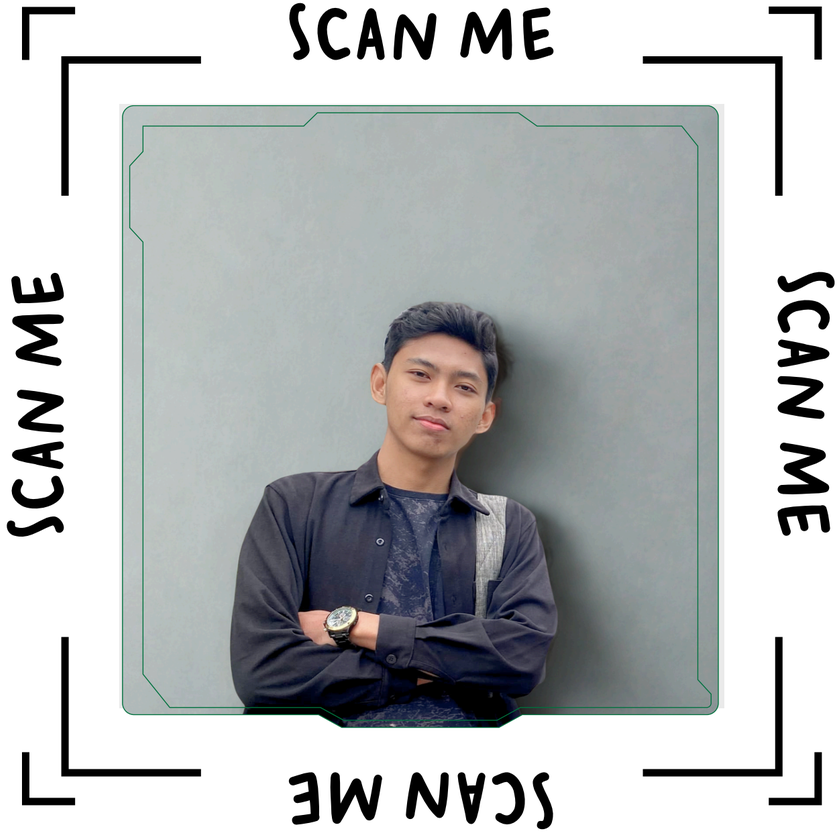
OUTLINE



- Business Understanding
- Dataset Overview
- Data Preprocessing
- Exploratory Data Analysis
- RFM Analysis
- Key Insight & Recomendation



ABOUT ME



"Data is the new oil; valuable but useless if unrefined, it only becomes a great power when processed into insight."

EXPERIENCE

Jan - July 2023

Administration & Analysis Data Intern
Dinas Bapenda, Pemkab Malang

EDUCATION

Sept 2025 - Present

Data Science Bootcamp
dibimbing.id

2020 - 2025

Bachelor of Mathematics
UIN Malang

Iftanul Ibnu Rochman

BUSSINES PROBLEM

No Insight Into Product and Category Performance

There is no evaluation of which products are revenue drivers, which ones
underp

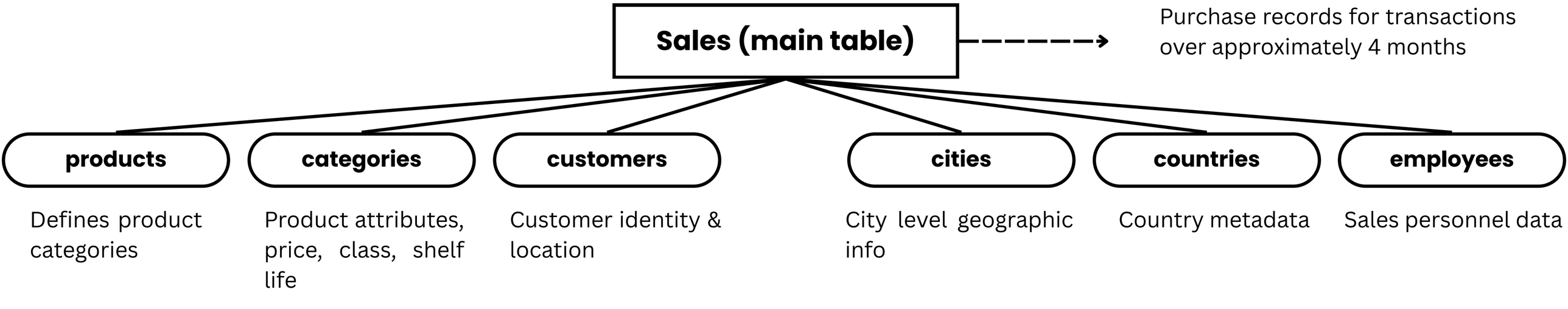
Lack of Trend Awareness for Operational Planning

The company does not have a clear view of how sales fluctuate over time. Monthly, weekly, and seasonal trends are not analyzed, causing difficulties in forecasting demand, planning inventory, and timing promotions effectively.

Customer Value & Behaviour Are Unknown

The company lacks visibility into who its most valuable customers are. There is no understanding of purchase frequency, spending patterns, or which customers are likely to churn.

DATASET UNDERSTANDING



6.7M Sales

98.7K Cust

452 Product

11 Category

96 Cities

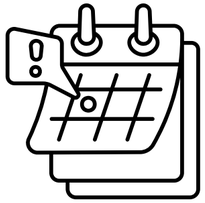
206 Country

Jan–May 2018

23 employees

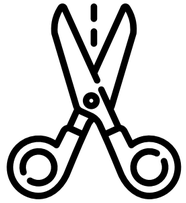
DATA PREPROCESSING

Table Integration
Merged 7 tables into one analytical dataset.



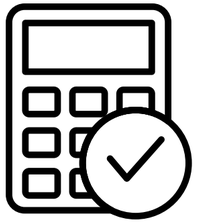
Data Type Correction
Converted SalesDate to datetime and Standardized numeric fields.

Missing Value Treatment
~1% missing SalesDate fixed using Mode Imputation

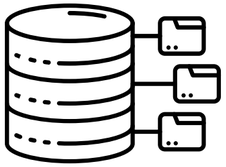


Column Refinement
Removed redundant IDs, unused employee attributes, and retained 14 essential analytical columns.

Outlier Assessment
Price, Quantity, Revenue, VitalityDays reviewed.
Skewed values confirmed as valid business behaviour



Revenue Correction
Original TotalPrice field invalid (all zeros).
Recalculated accurate revenue: Price × Quantity × (1 − Discount).



Final Analytical Dataset
clean and ready include 6.7 M rows and 14 columns

EKSPLORATORY DATA ANALYSIS

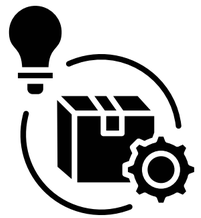
This EDA explores transaction patterns, customer behaviour, product performance, and temporal trends across 6.75 million sales records. The Goals are to:

Sales Patterns

Analyze transaction scale, monthly trends, and category contributions.

Customer Behaviour

Explore purchasing frequency, spending levels, and early churn indicators.



Product & Category Performance

Evaluate top-selling products, pricing patterns, and shelf-life impact (VitalityDays).

SALES PATTERNS

Total Transaction
6.67 M

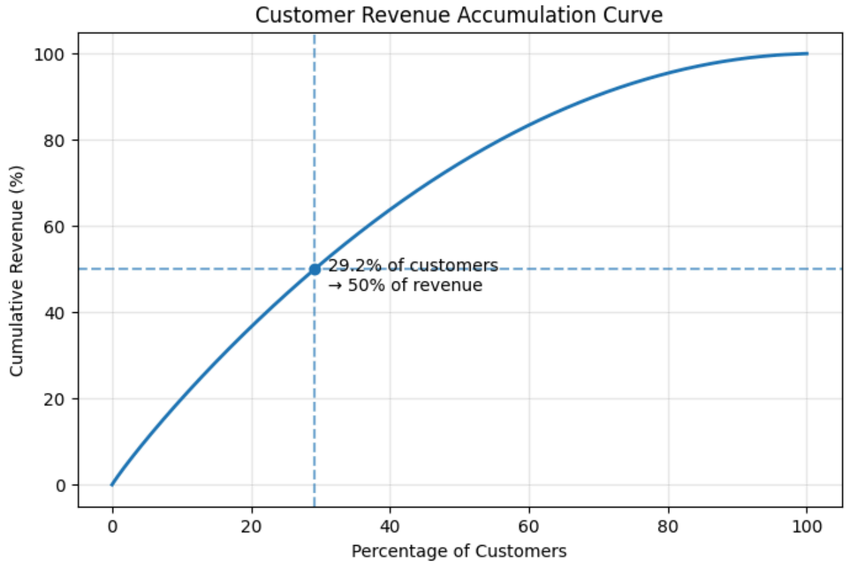
Total Revenue
$4.33B

Total Customers
98.7K

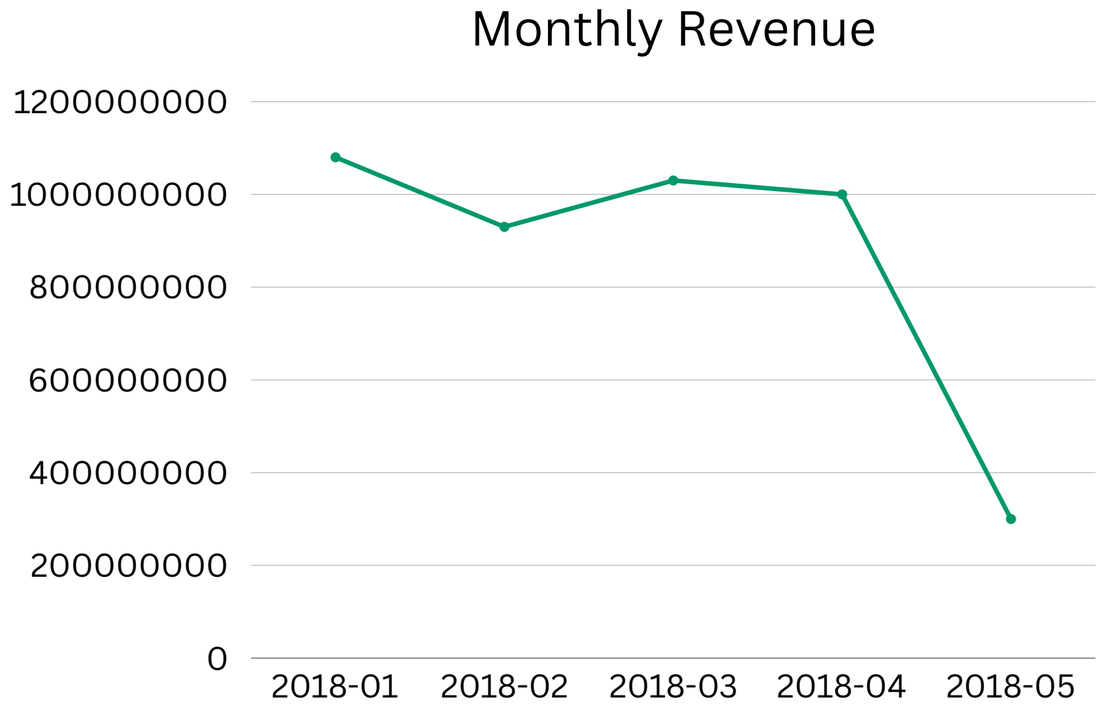
Average of Value
$641.07

- The company fundamentally relies on a high frequency of transactions and repeat business, rather than on single, very large transactions. Customer loyalty and daily shopping habits are key drivers of growth.

- This business demonstrates outstanding financial performance with total revenue of $4.33 billion driven by massive transaction volume (6.76 million), yet served by a relatively small customer base of only about 98,000. This data indicates a business model highly reliant on high transaction frequency and strong customer loyalty, where each customer makes dozens of transactions and has a very high customer lifetime value (CLV), supported by a stable average order value (AOV) of around $641. This suggests strong profitability and predictable purchasing behaviour, with a primary focus on customer retention and repeat transactions rather than mass customer acquisition.



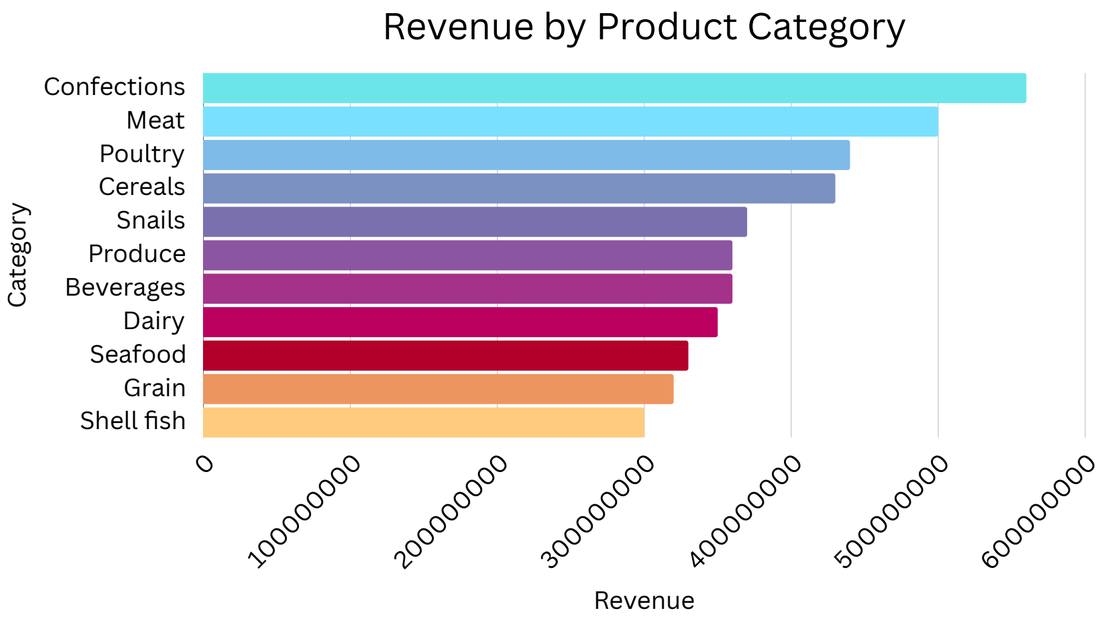
The current revenue distribution displays a moderate Pareto effect, where the top 20% of customers generate only around 37% of total revenue, which is significantly lower than the 80% typically observed in high-loyalty retail models. This suggests a lack of strong, high-value customer segments and limited repeat purchasing, making the business overly reliant on constant new customer acquisition to maintain its revenue stream.



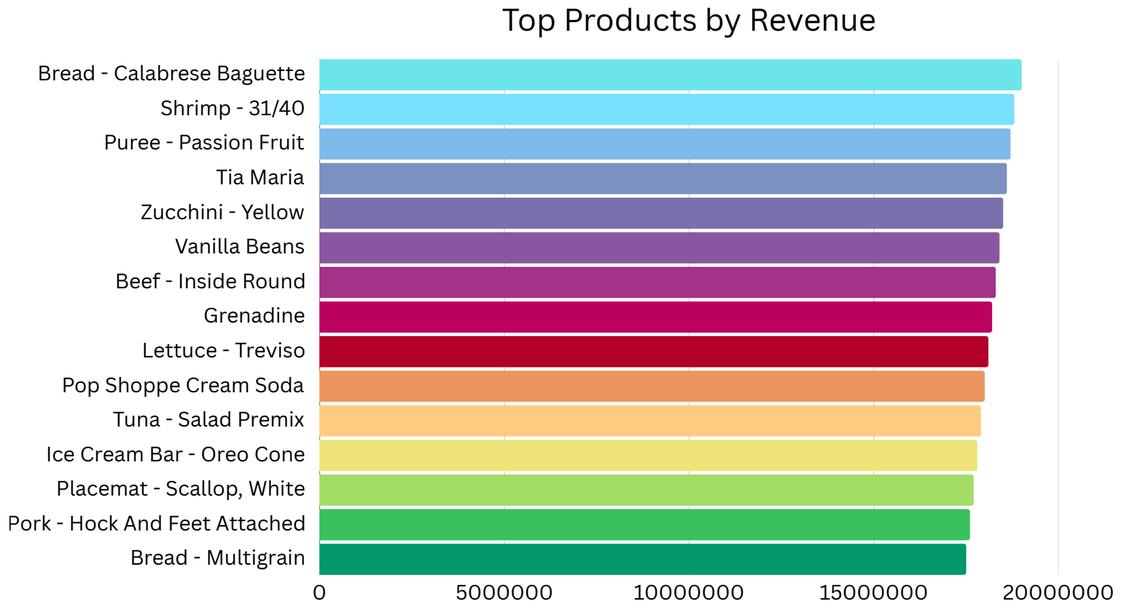
Understanding monthly trends allows the business to plan inventory, staffing, and promotions more effectively. The stability across January–April suggests reliable demand, while incomplete May data should not be mistaken for operational decline.

EXPLORATORY DATA ANALYSIS

PRODUCT & CATEGORY PERFORMANCE

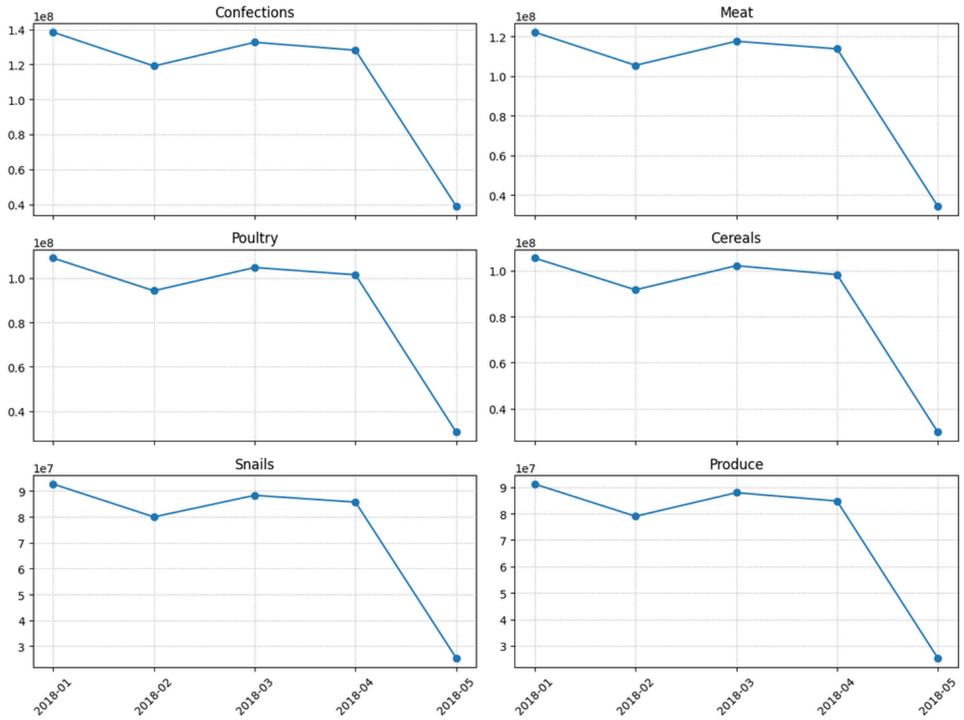


- Revenue distribution is uneven across categories.
- Confections, Meat, and Poultry are the top-performing categories, contributing the largest share of total revenue.
- Categories such as Shell Fish and Grain contribute significantly less and show underperformance compared to the top categories.
- This imbalance indicates dependency on a few key categories while other categories remain under-leveraged.



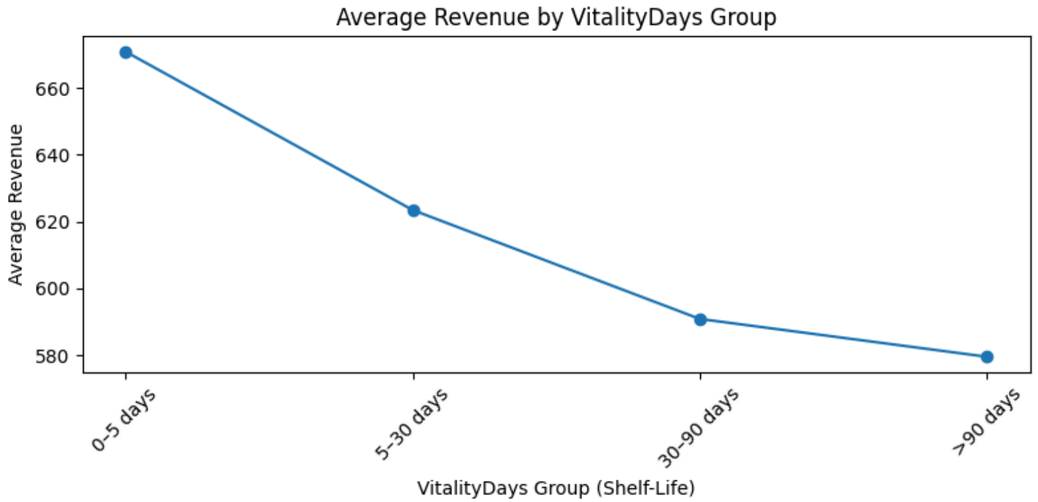
- The chart highlights products with the highest revenue contribution.
- These products consistently generate higher revenue compared to others.
- They represent key revenue drivers within the product portfolio.

EXPLORATORY DATA ANALYSIS



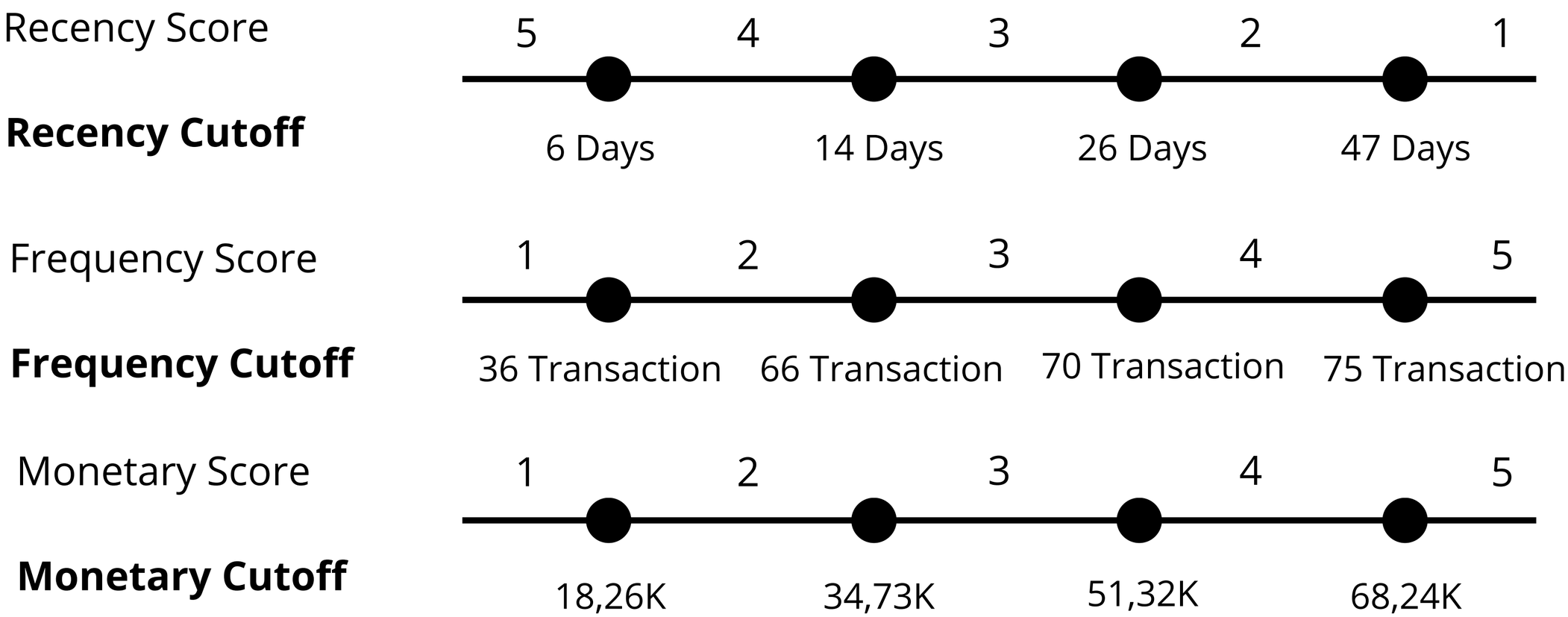
- All major categories exhibit similar monthly patterns: strong performance in January, dip in February, recovery in March, and stability in April.
- The drop observed in May is due to incomplete data (ending May 9), not an operational decline.
- The high alignment across categories indicates consistent demand behaviour across the product portfolio.
- No category shows abnormal deviation, suggesting stable category-level performance.

Products with very short shelf lives (0–5 days) generate the highest average revenue per transaction, and this revenue steadily decreases as shelf life (VitalityDays) increases, with long shelf-life products (>90 days) yielding the lowest average revenue. This pattern indicates a customer preference for fresh or highly perishable items and highlights that, despite operational challenges like storage, spoilage, and stock rotation, these perishable goods contribute significantly to the business's overall transaction value.



EXPLORATORY DATA ANALYSIS

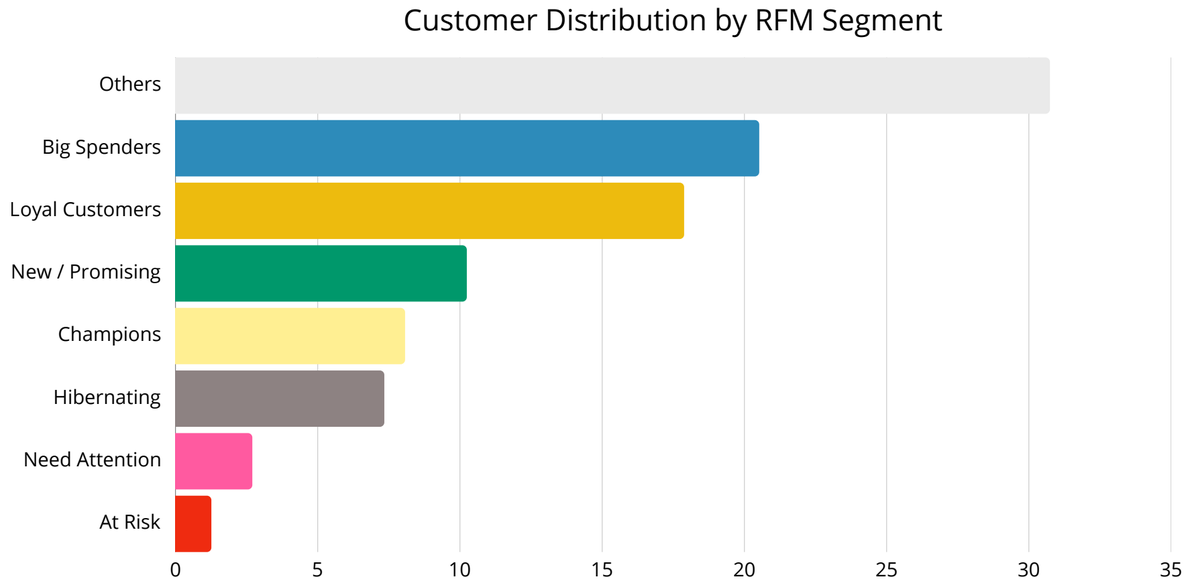
CUSTOMER BEHAVIOUR

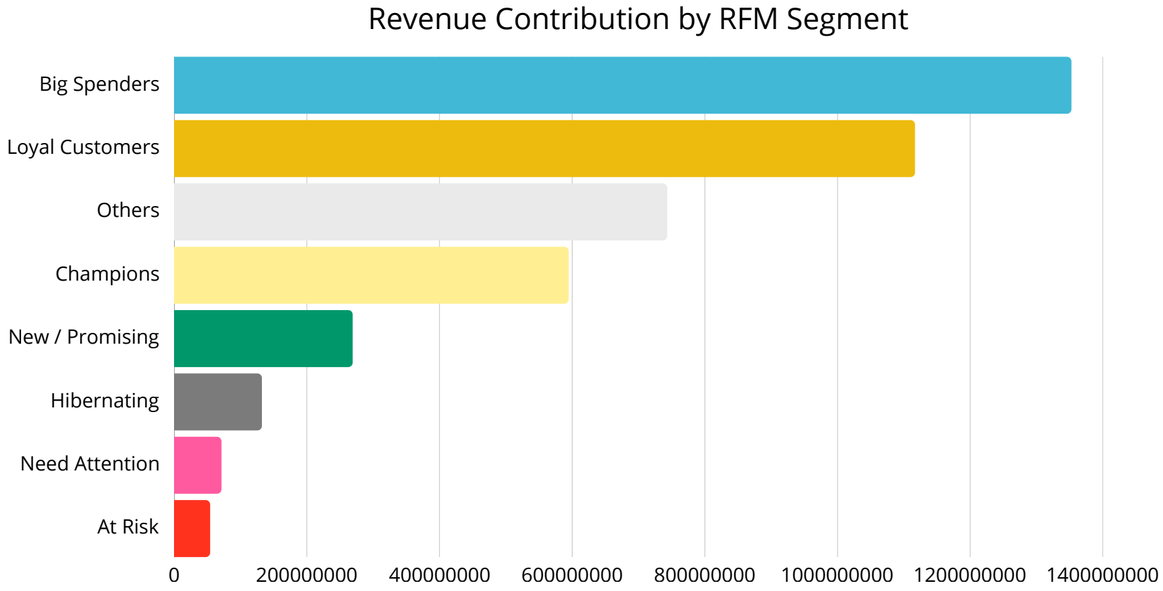


EXPLORATORY DATA ANALYSIS

Segment	Description	Recommendation
Champions	Customers with high recency, frequency, and monetary value. Highly active, loyal, and the strongest contributors to revenue.	Focus on retention strategies, exclusive loyalty programs, early access, and personalized rewards.
Loyal Customers	Customers with high purchase frequency and solid spending, though not always the most recent buyers. A stable and valuable customer base.	Strengthen engagement through subscriptions, bundles, and personalized communication.
Big Spenders	Customers with high monetary value but lower purchase frequency. Significant revenue potential per transaction.	Prioritize upselling and cross-selling, especially premium or high-value products.
New / Promising	Recently acquired customers with growth potential, but still limited frequency and spending.	Optimize onboarding experience, product education, and introductory offers to encourage repeat purchases.
At Risk	Previously valuable customers whose activity has declined due to low recency. High risk of churn.	Execute win-back campaigns, personalized incentives, and timely reminders.
Hibernating	Long-inactive customers with low frequency and low spending. Minimal engagement with the brand.	Apply light reactivation campaigns (e.g., deep discounts) or deprioritize to optimize marketing spend.
Need Attention	Customers with declining activity but still showing potential to re-engage.	Use targeted reminders, limited-time offers, and behavioral nudges to increase activity.
Others	Customers who do not strongly fit into key RFM segments; activity and value are generally low or inconsistent.	Focus on activation campaigns or monitor for potential segment upgrade.

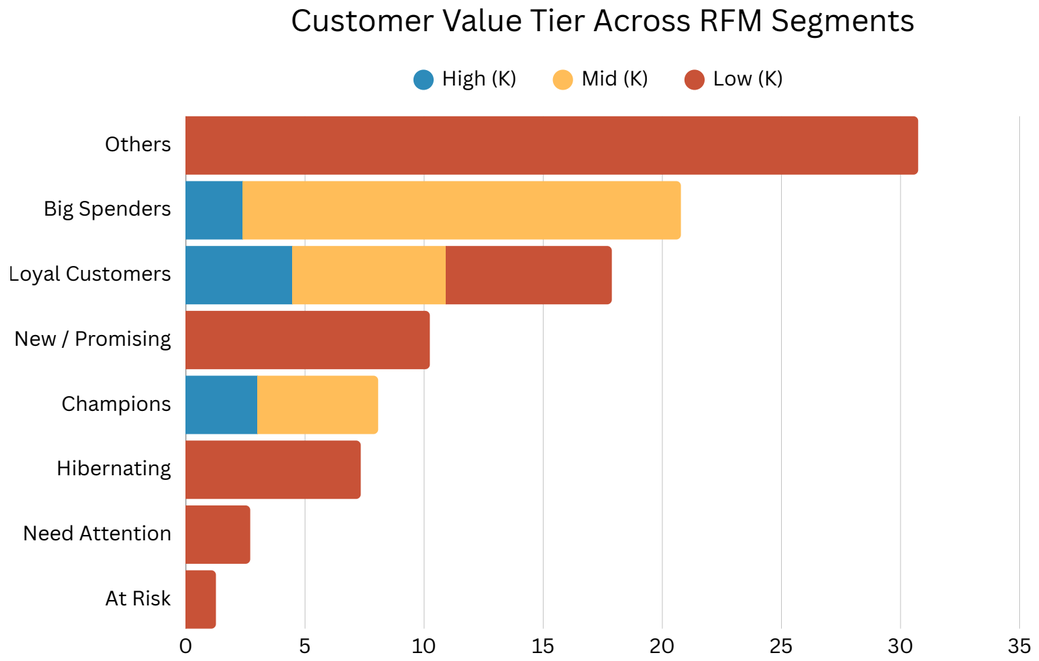
The "Others" segment has the largest number of customers, but its revenue contribution is relatively lower compared to the high-value segments. This insight confirms that a large customer base does not always directly correlate with the business value generated.



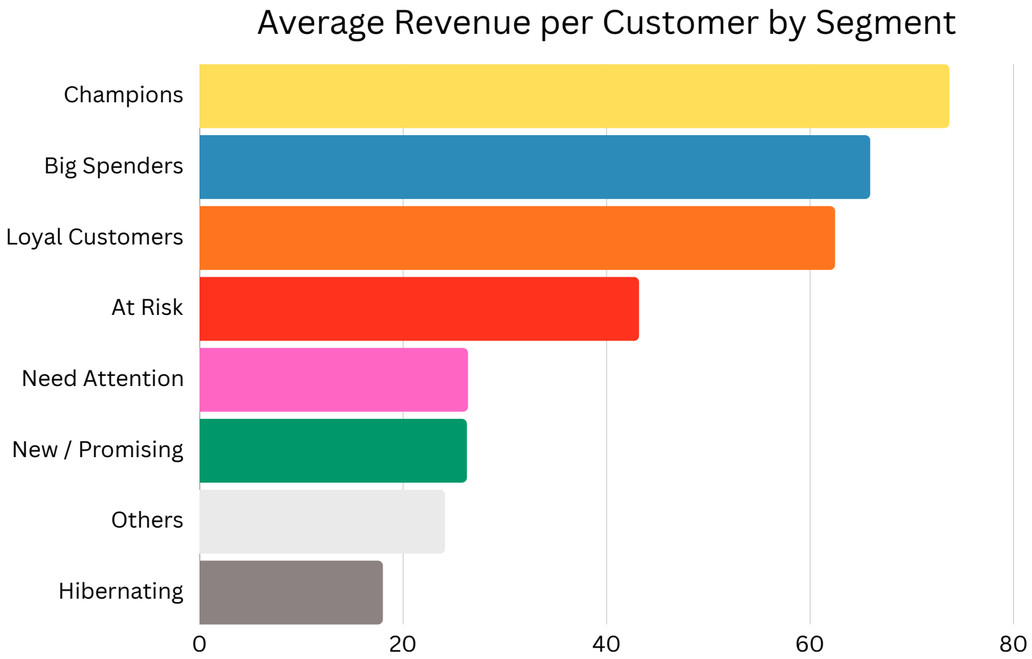


Revenue is highly concentrated in the Big Spenders and Loyal Customers segments, which together account for the majority of total revenue. This indicates that a small percentage of customers have a significant financial impact on the business.

EXPLORATORY DATA ANALYSIS



Customer distribution is still dominated by the Mid and Low Value segments, with the largest concentration in the Others segment. This indicates a significant opportunity to increase customer value through activation and upgrading strategies.



The Champions and Big Spenders segments generate the highest average revenue per customer, even though they have a relatively smaller number of customers. This insight confirms that customer quality has a greater impact on revenue than quantity.

EXPLORATORY DATA ANALYSIS

KEY INSIGHT

Sales Patterns

- The company relies on transaction frequency, not large transactions

✗ High-Value ✓ High-Volume

- For over 4 months, it was relatively stable, with a slight fluctuation in February (Sessional Effect).
- Revenue depends on minority customers; 29% of customers generate 50% of revenue, meaning losing a significant portion of customers would have a major impact on revenue (Sales pattern is not secure without retention).

Categoty & Product Performance

- Revenue and majority are dominated by sweet/processed foods.
- The value of the product/category is high in terms of short vitality days, fresh products = high value, low frequency (more expensive, bought less often, high value per purchase).
- There is no category where sales are declining on their own, they are stable every month, meaning processed foods = habitual demand (core consumption behaviour).

Customer Behaviour

- The majority of revenue is driven by high-value segments (Champions, Big Spenders, Loyal Customers) despite their relatively small number of customers.
- The segment with the largest number of customers (Others) has low revenue contribution and value per customer, so customer volume is not directly proportional to business performance.
- High-value customers are already concentrated in the right segment, not hidden in other segments, indicating a healthy and controlled customer value structure.

RECOMENDATIONS

1 Prioritise Retention as the Main Strategy for Revenue Protection

Focus business strategy on retaining high-value customers (Champions, Big Spenders, Loyal Customers) through loyalty programs, personalised offers, and ongoing engagement.

Reason

Because 29% of customers generate 50% of revenue and sales patterns depend on frequency, losing even a small percentage of high-value customers will have a significant impact on revenue.

2 Optimise Purchase Frequency for Core Products (Processed/Sweet Food)

Encourage repeat purchases in the processed food category through bundling, light subscriptions, and promotions based on consumption habits.

Reason

The processed category indicates stable and recurring demand, so increasing frequency will directly boost revenue.

3 Position Fresh Products as a High-Value Complement, Not a Volume Driver

Use fresh products as an upsell or premium add-on, not as a target for increasing frequency.

Reason

Fresh produce has a high transaction value but low frequency, making it more effective to maximise from a value perspective rather than volume.

4 Shift the Growth Strategy from Mass Acquisition to Customer Upgrading

Reduce the focus on volume-based new customer acquisition and shift towards activating and upgrading customers in the "Others" segment to a higher-value segment.

Reason

The "Others" segment dominates customer numbers but contributes little revenue, while high-value customers are already clearly concentrated.

5 Conduct Revenue Planning with Anticipation of Seasonal Effects

Incorporate seasonality factors (February) into promotion planning, inventory management, and sales targets.

Reason

Sales patterns are stable with seasonal fluctuations in February, which can be anticipated in business planning.

TERIMA KASIH

Dashboard Interaktif : Tableu Public