Introduction

FreshMart is a fast-growing grocery retail chain based in the United States, serving thousands of customers across various cities and countries. Known for its wide product range and affordable pricing, FreshMart has built a strong presence in urban and suburban markets.

The company believes there are many untapped opportunities to grow sales. These may lie in:

- how product categories perform across months,
- Which types of customers spend more?
- how employees contribute to store-level sales,
- Which cities or countries have higher or lower sales, and
- How discounts are influencing buying behaviour.

Problem Statement

As the company prepares for its next growth phase, leadership wants to focus on adding new stores and increasing Total Sales Revenue from its existing network. This means a better understanding of what drives revenue, from which products perform well, to how different regions, customer segments, and sales staff contribute to the bottom line.

Analysis Approach

1. **Understand the Problem**: Define key questions and review the dataset to grasp the problem of viewer churn and engagement.
2. **Define Metrics and Hypotheses**: Identify revenue, quantity, and price metrics. Formulate hypotheses, e.g., "Higher price leads to more revenue"
3. **Data Cleaning**: Handle missing, duplicate, or inconsistent data. Standardise formats and remove outliers.
4. **Exploratory Data Analysis (EDA)**: Visualise trends in revenue, volume and price.
5. **Customer Segmentation**: Group customers by behaviour (e.g., regular shopper vs recent bulk buyer) and analyse differences.
6. **Summarise Insights: Present actionable findings to help creators optimise** strategies and retain viewers.

Dataset

This dataset contains six tables about FreshMart:

- Product Categories
- The cities they operate in
- Countries they operate in
- Customers
- Salespeople employed by them
- Products they sell
- Transactions from 1/1/2018 to 9/5/2018

Utilising tools such as Python and Tableau, this dataset can help uncover actionable insights to improve total sales revenue, ultimately giving sustainable growth for FreshMart.

The dataset includes the following tables:

1. Categories
2. Cities
3. Countries

4. Customers
5. Employees
6. Products
7. Sales

The column definitions for each table are:

1. Categories
    i. **CategoryID**
        1. Description: Unique identifier for each product category.
        2. Example: 1
    ii. **CategoryName**
        1. Description: Name of the product category.
        2. Example: Beverages
2. Cities
    i. **CityID**
        1. Description: Unique identifier for each city.
        2. Example: 101
    ii. **CityName**
        1. Description: Name of the city.
        2. Example: San Diego
    iii. **Zipcode**
        1. Description: Represents the population of the city.
        2. Example: 500000
    iv. **CountryID**
        1. Description: Reference to the corresponding country from countries.csv.
        2. Example: 1
3. Countries
    i. **CountryID**
        1. Description: Unique identifier for each country.
        2. Example: 1
    ii. **CountryName**
        1. Description: Name of the country.

        2. Example: United States

   iii. **CountryCode**

        1. Description: Two-letter country code.

        2. Example: US

4. Customers

   i. **CustomerID**

        1. Description: Unique identifier for each customer.

        2. Example: 1001

   ii. **FirstName**

        1. Description: First name of the customer.

        2. Example: Emma

   iii. **MiddleInitial**

        1. Description: Middle initial of the customer.

        2. Example: A

   iv. **LastName**

        1. Description: Last name of the customer.

        2. Example: Johnson

   v. **cityID**

        1. Description: City of the customer. Refers to cities.csv.

        2. Example: 101

   vi. **Address**

        1. Description: Residential address of the customer.

        2. Example: 123 Elm Street

5. Employees

   i. **EmployeeID**

        1. Description: Unique identifier for each employee.

        2. Example: 501

   ii. **FirstName**

        1. Description: First name of the employee.

2. Example: Michael

iii. **MiddleInitial**

1. Description: Middle initial of the employee.
2. Example: B

iv. **LastName**

1. Description: Last name of the employee.
2. Example: Davis

v. **BirthDate**

1. Description: Date of birth of the employee in YYYY-MM-DD format.
2. Example: 1985-07-14

vi. **Gender**

1. Description: Gender of the employee.
2. Example: Male

vii. **CityID**

1. Description: City where the employee is based. Refers to cities.csv.
2. Example: 103

viii. **HireDate**

1. Description: Date when the employee was hired.
2. Example: 2021-04-01

6. Products

i. **ProductID**

1. Description: Unique identifier for each product.
2. Example: 301

ii. **ProductName**

1. Description: Name of the product.
2. Example: Organic Apple

iii. **Price**

1. Description: The unit price of the product is in USD.
2. Example: 3.50

iv. **CategoryID**

   1. Description: Category reference for the product. Refers to categories.csv.
   2. Example: 2

v. **Class**

   1. Description: Classification type of the product (e.g., Standard, Premium).
   2. Example: Premium

vi. **ModifyDate**

   1. Description: Date when the product information was last updated.
   2. Example: 2023-06-01

vii. **Resistant**

   1. Description: Product resistance category.
   2. Example: Water-resistant

viii. **Is Allergic**

   1. Description: Indicates whether the item contains allergens.
   2. Example: No

ix. **VitalityDays**

   1. Description: Indicates the product's shelf life or freshness period.
   2. Example: 7

7. Sales

i. **SalesID**

   1. Description: Unique identifier for each sale.
   2. Example: 7001

ii. **SalesPersonID**

   1. Description: Employee responsible for the sale. Refers to employees.csv.
   2. Example: 501

iii. **CustomerID**

   1. Description: Customer making the purchase. Refers to customers.csv.
   2. Example: 1001

iv. **ProductID**

1. Description: Product being sold. Refers to products.csv.
2. Example: 301

v. **Quantity**

1. Description: Number of product units sold.
2. Example: 3

vi. **Discount**

1. Description: Discount applied to this sale, shown as a decimal.
2. Example: 0.10

vii. **TotalPrice**

1. Description: Final sale price after applying discount.
2. Example: 9.45

viii. **SalesDate**

1. Description: Date and time of the sale in YYYY-MM-DD HH:MM: SS format.
2. Example: 2024-05-15 14:32:00

ix. **TransactionNumber**

1. Description: Unique identifier for the transaction.
2. Example: TXN-20240515-0001

Metrics and Hypothesis

Key metrics are price, volume, revenue and discount.

Hypothesis:

1. Increasing the prices of products by 10% will increase the revenue by 10% as the Total volume per product is constant.
2. Adding extra salespeople to the stores will proportionally increase the revenue.

Data Cleaning & Preparation

Data cleaning and preparation are foundational steps in the data analysis process. The FreshMart analytics dataset, which includes various metrics such as price, volume, discount, salesdate, and categoryid, requires careful preprocessing to address missing values, inconsistent formats, and incorrect data types. This ensures the dataset is ready for exploratory data analysis and meaningful insights.

Step 1: Dataset Exploration & Overview

The first step involves gaining an understanding of the dataset.

- **Data type**: The data format (e.g., object, float, int).

- **Count of non-null values**: The number of entries not missing in each column.
- **Count of null values**: The number of missing entries in each column.
- **Number of unique values**: Distinct values in a column to identify categorical or numerical variables.
- **Percentage of null values**: The proportion of missing data in each column.

This exploration helps identify problematic columns. For example:

- Columns such as SalesDate may have significant null values.
- This step also includes generating basic statistical summaries to understand numerical column ranges, averages, and distributions.

Step 2: Handling Missing Values

Three columns in the dataset have null values.

- MiddleInitial will not have any imputed values as the column has no significance.
- SalesDate will be ignored for the analysis, which involves time; otherwise, it will be left as is, as only 0.97% of the values are missing.
- The missing value in CountryCode will be filled by assigning our code.

Step 3: Dropping Incorrect Columns

Specific columns, such as Class, do not have values matching their definition. These columns are dropped to simplify the dataset. The drop() function is used to remove these columns.

For example:

- Class: It denoted whether a product was a premium item, but the values indicated Medium, High, and Low, which do not match.
- Resistant: It denotes the resistance type of a product, eg water-resistant. But the values showed durable, weak, etc.
- VitalityDays: It showed the shelf life of a product, but it gave some perishable products a 0-day shelf life. I found this column unreliable, so I dropped it.

Step 4: Converting Data Types

Several columns, like SalesDate, were incorrectly formatted as objects when they were supposed to be a datetime. The table below will show which columns had their datatypes changed:

Table	**Column**	**Current Type**	**Required Type**
categories	CategoryName	object	category
cities	CityName	object	string
countries	CountryName	object	string
countries	CountryCode	object	string
customers	FirstName	object	string
customers	MiddleInitial	object	string
customers	LastName	object	string
customers	Address	object	string
employees	FirstName	object	string
employees	MiddleInitial	object	string
employees	LastName	object	string
employees	BirthDate	object	datetime
employees	HireDate	object	datetime

products	ProductName	object	string
products	ModifyDate	object	datetime
products	IsAllergic	object	bool
sales	SalesDate	object	datetime
sales	TransactionNumber	object	string

Step 5: Product–Category Mapping Correction

Some products, e.g., Onions - Cippolini, are categorised as poultry but should be classified as produce. For example, Cocktail Napkin Blue does not fit any category. To fix this issue, a proper rule-based classification was done to ensure better product-category mapping and the addition of another category called `Uncategorised` to handle products that don't belong in any of these.

Step 6: Customer Segmentation

Customers were segmented based on their shopping behaviour. The following factors were considered when segmenting a customer:

- Recency: When was the customer's last purchase
- Frequency: How often does a customer make a purchase
- Monetary: How much a customer spends in total
- Quantity: How many items does the customer purchase
- Amount: How much money do they spend on a purchase

These traits were examined, and the following segments were created:

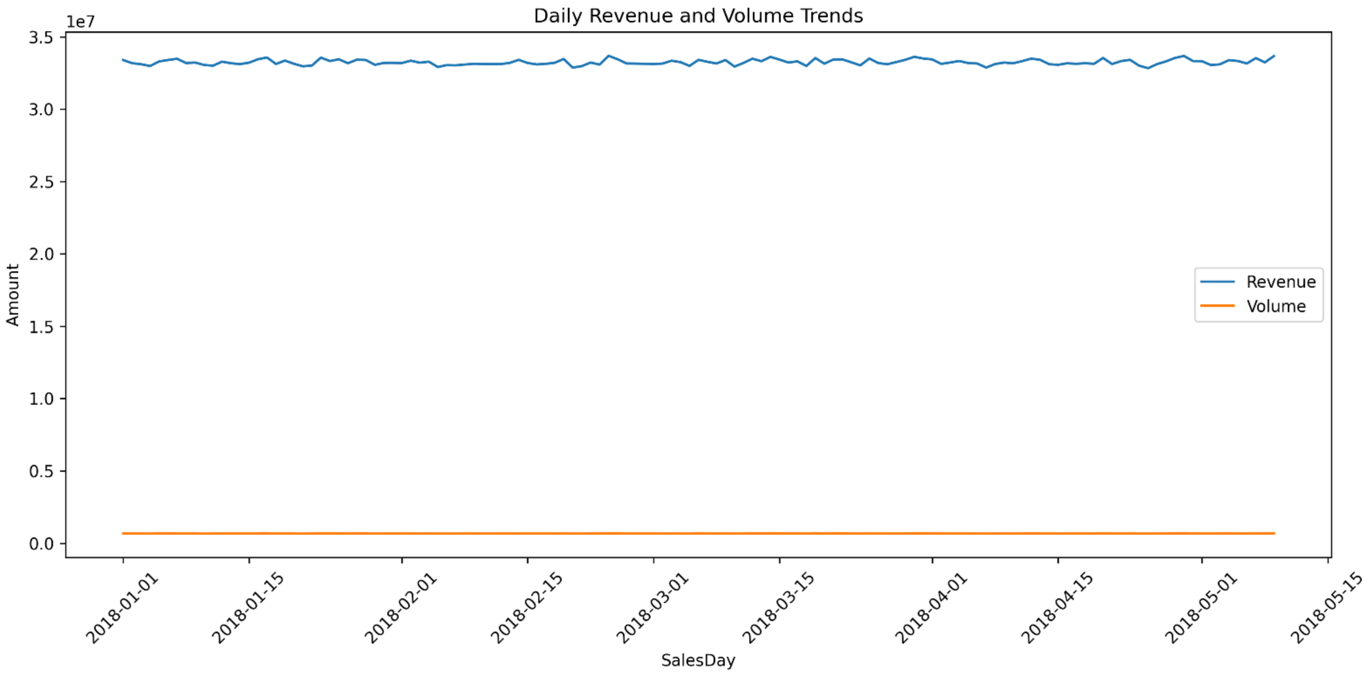
- Bulk Shopper (Low Frequency)
- Frequent Low-Spend Shopper
- Frequent Product Purchaser
- Infrequent Big Spender
- Lost or Inactive Customer
- Recent Bulk Buyer
- Recent Small Order Shopper
- Regular Shopper

EDA

Exploratory Data Analysis (EDA) is essential for understanding the dynamics of the FreshMart analytics dataset. This dataset captures crucial information like volume, price, discount, revenue, and customer dynamics

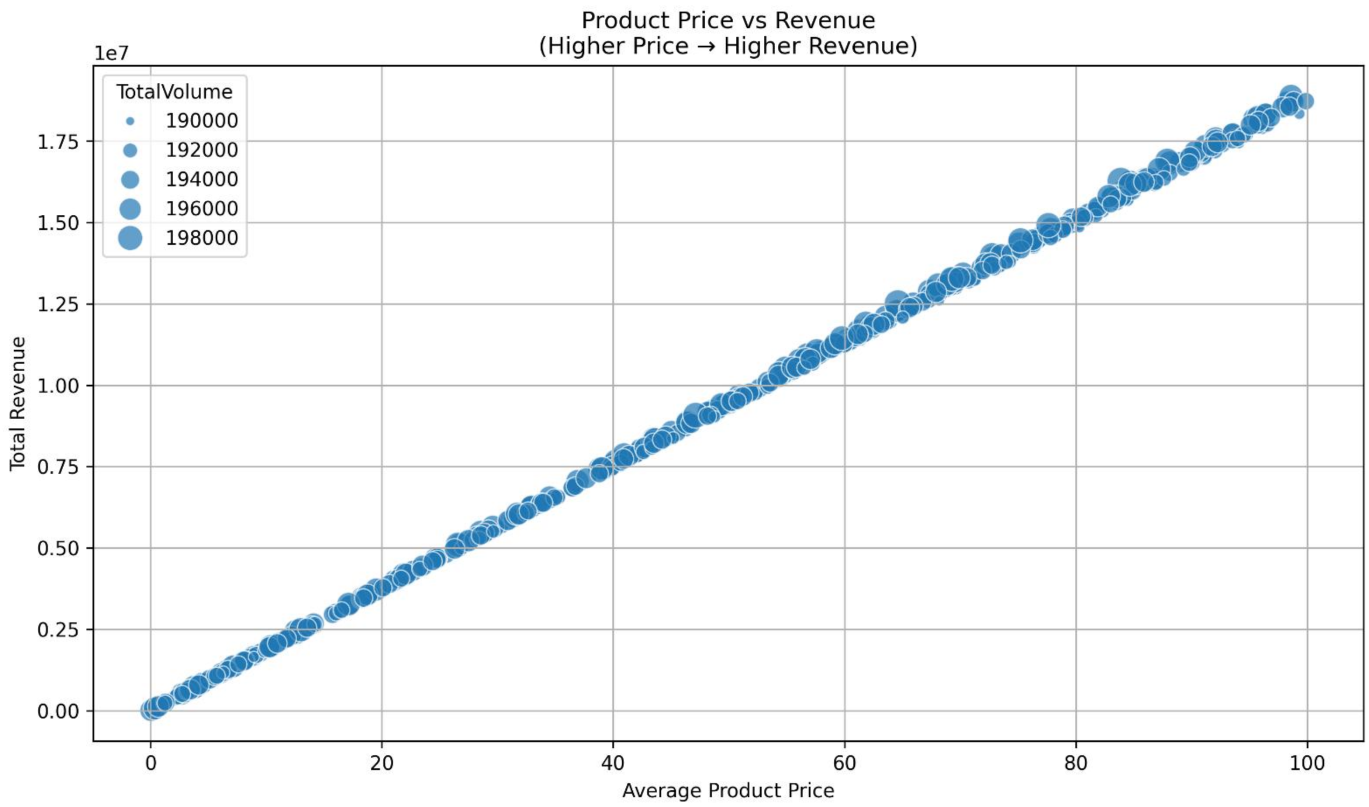
1. Daily Revenue and Volume Trends

Revenue and Volume are key metrics in maximising total sales revenue. The key findings were that revenue and volume were stable from 1/1/2018 to 9/5/2018.



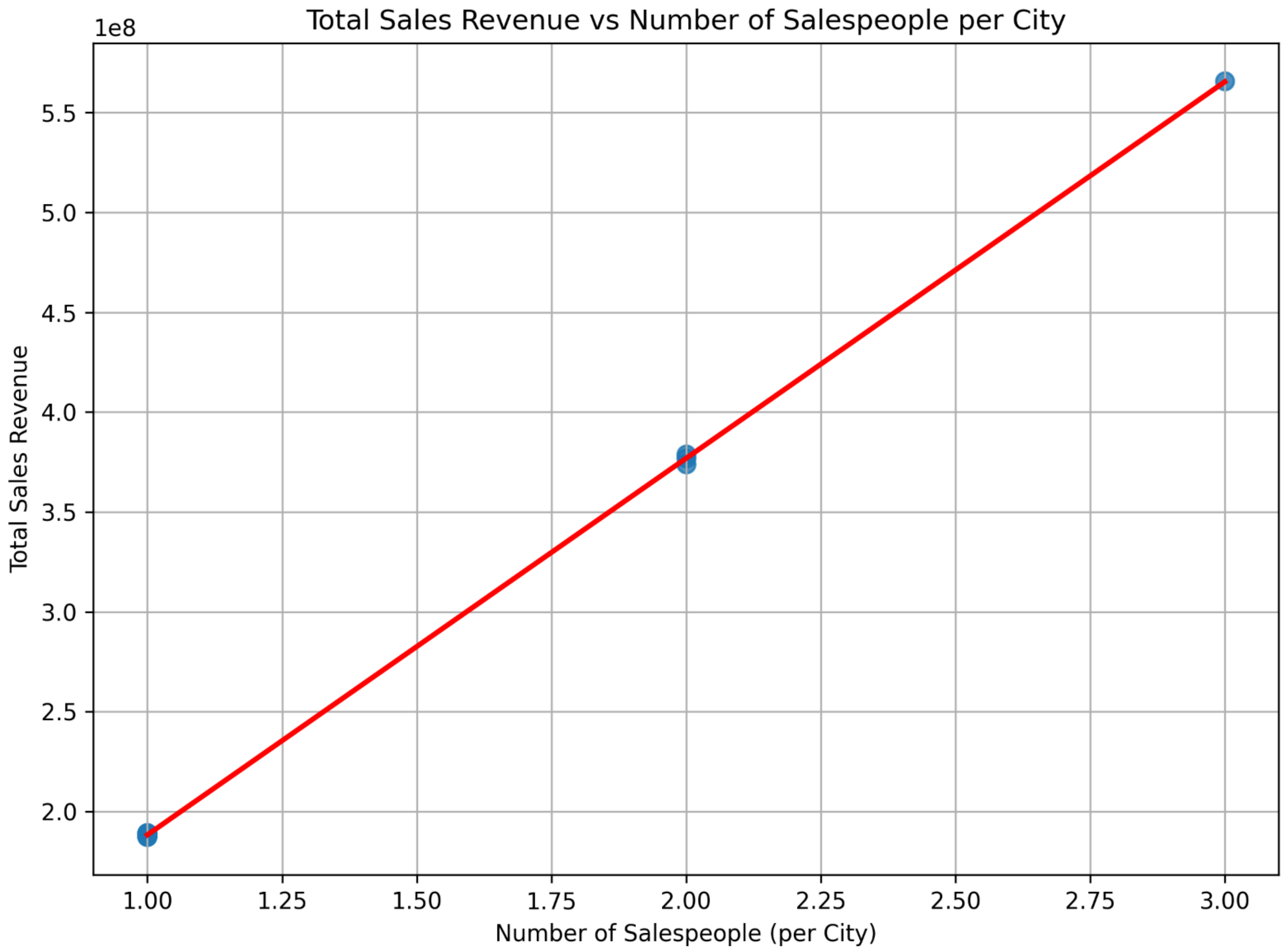
2. Product Price vs Revenue

The price of a product directly impacts FreshMart's total sales revenue. It was found that product price is the key driver of revenue, as volume and discount were constant across products. So, variation in revenue is driven by price.



3. Salespeople vs Revenue

Salespeople are the employees who work on the front lines and directly interact with customers. Their performance is critical to maximising total sales revenue. The key findings were that total sales revenue for a city is directly proportional to the number of salespeople a city has.



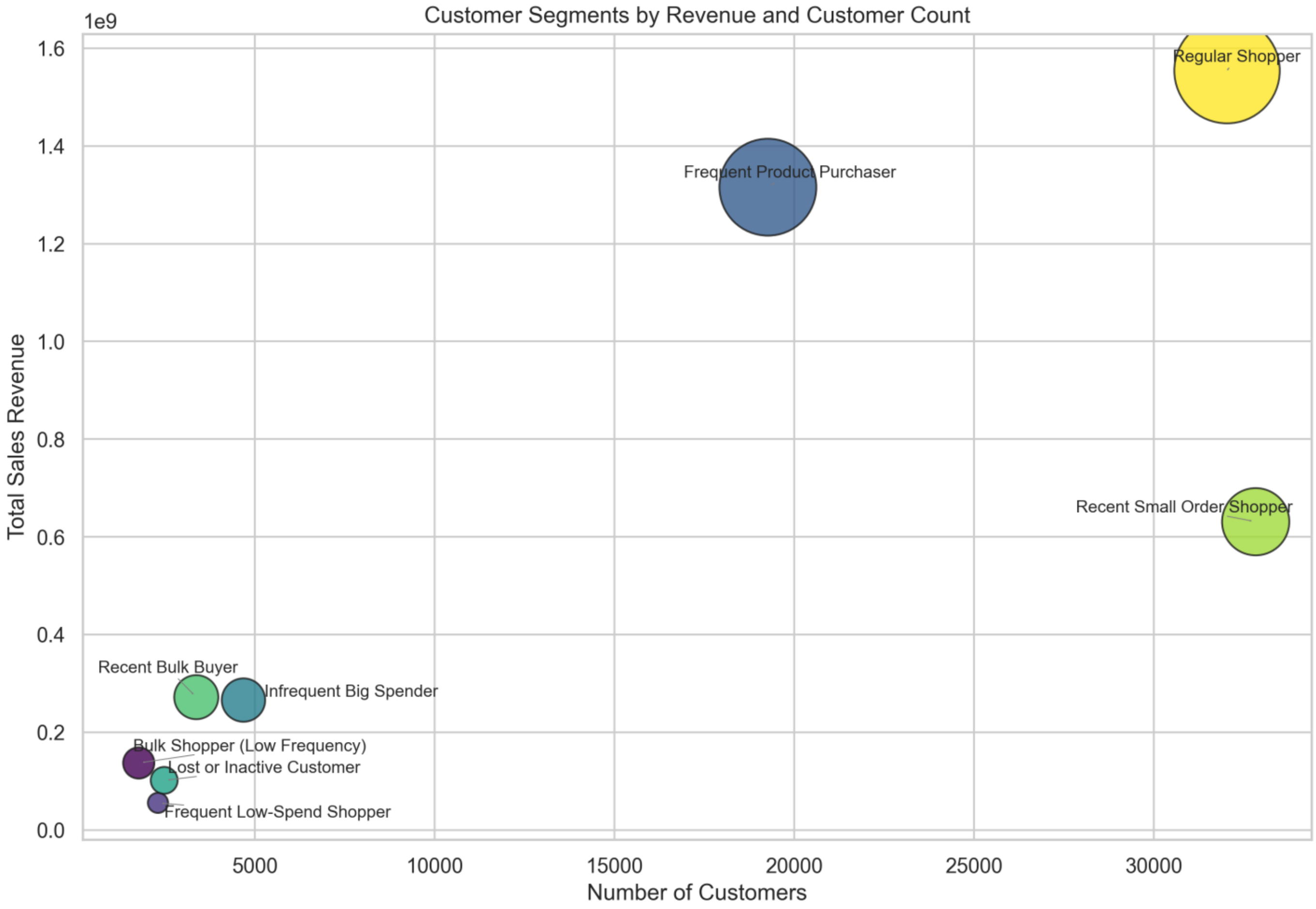
4. Customer Segmentation by Revenue and Customer Count

Analysing the revenue generated by each customer segment and how many customers are in each segment lets management decide how to use its resources to maximise total sales revenue.

Key Insights:

- Regular Shoppers are the Primary Revenue Drivers
  - Total Sales Revenue: USD 1.55B
  - Customer Count: 32,033
  - This segment is the largest in revenue and customer base, indicating that consistent, habitual buyers are the foundation of the business's revenue stream.
- Frequent Product Purchasers Also Generate Significant Revenue
  - Total Sales Revenue: USD 1.31B
  - Customer Count: 19,267
  - Although smaller in number than Regular Shoppers, this group generates a comparable amount of revenue, highlighting their high purchasing intensity or frequency.
- Recent Small Order Shoppers Present Growth Potential
  - Total Sales Revenue: USD 630.9M
  - Customer Count: 32,829 (the most significant segment)
  - Despite contributing significantly to revenue, the lower average transaction value suggests an opportunity to upsell or bundle products to increase value per transaction.
- Infrequent Big Spenders and Recent Bulk Buyers are High-Value Niche Segments
  - Infrequent Big Spenders TSR: USD 265.9M
  - Recent Bulk Buyer TSR: USD 271.8M

    - These groups, while smaller in customer count, contribute substantial revenue. They may be good candidates for targeted high-value campaigns or loyalty programs.
- Bulk Shoppers (Low Frequency) and Lost/Inactive Customers Contribute Lower Revenue
    - Bulk Shoppers TSR: USD 137.0M
    - Lost/Inactive TSR: USD 101.5M
    - These segments have low revenue and may either need re-engagement strategies or can be deprioritised depending on customer lifetime value.
- Frequent Low-Spend Shoppers Have Limited Revenue Contribution
    - TSR: USD 55.4M
    - Customer Count: 2,310
    - This group has a limited impact on total revenue and may incur higher service costs than revenue contribution.

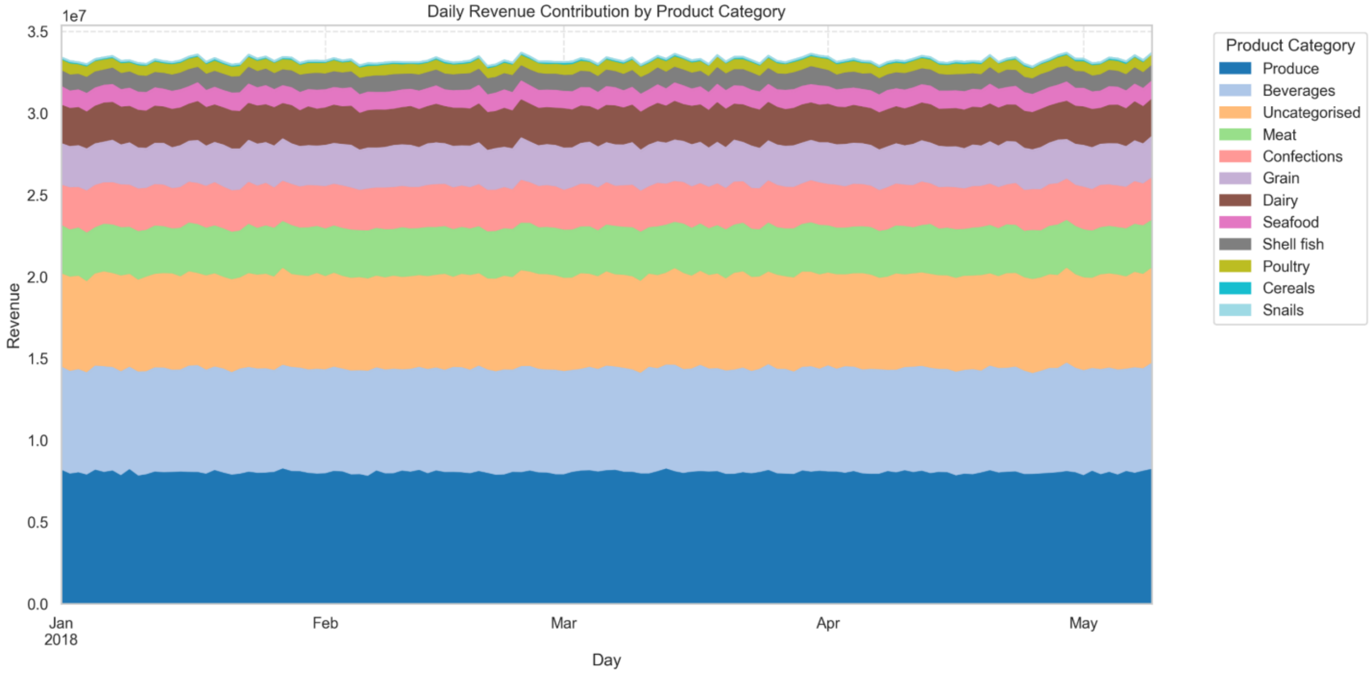


5. Daily Revenue Contribution by Category

Understanding which product categories generate revenue and when they generate revenue is key to understanding customer preferences.

Key Insights:

- Daily Revenue per Category is Consistent
    - Despite the differences in volume, classification, or features, each product category contributes daily, stable, predictable revenue.
    - There are no significant fluctuations or seasonal patterns within individual categories daily.
- Category Performance is Operationally Stable
    - This trend suggests that day-to-day demand shifts, promotions, or irregular events do not heavily impact most product categories.
    - Their performance appears to be decoupled from external volatility at the daily level.



Summary

The EDA of the FreshMart Dataset revealed unique insights into customer preferences, key revenue drivers and salesperson performance. Here are the key takeaways:

- **Revenue and Volume are Stable over Time**
  - Total Sales Revenue and Total Volume are consistent over time, as checked by calculating the Total Sales Revenue and Total Volume monthly, weekly, day of the week-wise, and hour of the day-wise.
- **Product price is the key Revenue driver**
  - Total Volume per product and the average discount offered were also constant. The differentiator in revenue was price. The more expensive the product is, the more revenue it generates.
- **Sales Revenue is Directly Proportional to the number of salespeople**
  - Revenue per salesperson is constant, so if there are more salespeople, there will be more revenue as long as customer demand supports it.
- **Regular Shoppers and Frequent Product Purchasers are FreshMart's key Revenue Sources**
  - Revenue from these segments is more than USD 1,000,000,000.

Strategies to Maximise Total Sales Revenue

- **Focus on High-Value Customer Segments**
  - **Regular Shoppers** (TSR: USD 1.55B) and **Frequent Product Purchasers** (TSR: USD 1.31B) are the primary revenue drivers.
  - Retention and personalised engagement strategies targeting these segments will directly support revenue sustainability.
  - Investment in loyalty, exclusive deals, and tailored communication can maintain and grow their contribution.
- **Optimise Engagement for High-Volume but Low-Value Segments**
  - **Recent Small Order Shoppers** have the most extensive customer base (32,829) but a lower average transaction value.
  - Upselling, bundling, and cross-selling strategies can increase revenue per customer.
  - Loyalty incentives or minimum-spend promotions could raise transaction size without relying on heavy discounts.
- **Target Niche High-Value Segments Strategically**

    - **Infrequent Big Spenders** and **Recent Bulk Buyers** contribute significant revenue despite smaller customer counts.
    - These customers are ideal for tailored high-value campaigns and premium loyalty programs.
    - Re-engagement efforts focused on frequency could unlock further growth.
- **Limit Spend on Low-Contribution Segments**
    - **Frequent Low-Spend Shoppers**, **Bulk Shoppers (Low Frequency)**, and **Lost/Inactive Customers** have low TSR.
    - Re-engagement campaigns should be low-cost and automated, based on potential customer lifetime value.
    - Marketing investment may be better reallocated to more profitable segments.
- **Prioritise High-Priced Products**
    - Product price is the primary revenue driver, with volume and discounts remaining stable across products.
    - Revenue growth should focus on shifting sales toward higher-priced or premium items.
    - Marketing efforts should prioritise visibility and search placement of high-revenue products.
- **Avoid Overreliance on Discounts**
    - Average discounting levels are stable and not strongly correlated with revenue variation.
    - This suggests discounting is not a central lever and should be used strategically, not broadly.
    - Maintaining pricing integrity helps protect long-term margins.
- **Expand the Salesforce Strategically**
    - Revenue is directly proportional to the number of salespeople.
    - Each additional salesperson contributes a predictable amount to revenue, enabling scalable, forecastable growth.
    - Justifies hiring and budget allocation for expanding sales operations where customer demand exists.

Operational Improvements

- **Daily Revenue Predictability Enables Accurate Forecasting**
    - Daily TSR is stable across product categories, enabling highly accurate short-term revenue forecasting.
    - Supports better cash flow, staffing, and inventory planning.
- **Streamlined Inventory and Fulfilment Planning**
    - With consistent volume and revenue patterns across time, operations can optimise procurement and logistics.
    - Reduces risk of stockouts or excess inventory.
    - Support labour or scheduling for fulfilment and customer service.
- **Marketing Optimisation**
    - Stable revenue patterns suggest low responsiveness to broad campaigns or promotions.
    - Marketing teams can focus resources on boosting underperforming segments or categories rather than maintaining already stable ones.
    - Product-level insights allow for targeted promotion of high-margin or high-revenue items.
- **Territory and Workforce Optimisation**
    - Salesperson distribution is uneven across cities.
    - Some territories may be underutilised, with only one salesperson despite strong revenue potential.
    - Reallocating or increasing staff in high-potential regions can unlock untapped revenue.
- **Segment-Driven Resource Allocation**
    - TSR and customer count by segment provide a data-backed foundation for investment decisions.

- Marketing, product, and service investments should be weighted toward segments with the highest return potential.
- Ensures efficient use of budget and maximises overall revenue contribution.