The Manhattan Project

presents

Challenge-01

The Team

Sukrit Mukherjee

Final-year Computer Science and Engineering student with a strong interest in Data Science and collaborative projects.

The Team

Sukrit Mukherjee

Final-year Computer Science and Engineering student with a strong interest in Data Science and collaborative projects.

Anindita Saha

Final-year Computer Science and Engineering student with a strong interest in Data Science and collaborative projects.

The Team

Sukrit Mukherjee

Final-year Computer Science and Engineering student with a strong interest in Data Science and collaborative projects.

Anindita Saha

Final-year Computer Science and Engineering student with a strong interest in Data Science and collaborative projects.

Soumyajit Paul

Final-year Computer Science and Engineering student with a strong interest in Data Science and collaborative projects.

The Tea

Sukrit Mukherjee

Final-year Computer Science and Engineering student with a strong interest in Data Science and collaborative projects.

Anindita Saha

Final-year Computer Science and Engineering student passionate about Data Science, Machine Learning, and turning data into actionable insights.

Soumyajit Paul

Final-year Computer Science and Engineering student and technology enthusiast, driven by curiosity across AI, data-driven systems, and innovative software applications.

Debayudh Khag

Final-year Computer Science and Engineering student with a keen interest in teamwork, collaborative projects, and building impactful digital solutions.

Problem Statement

The Retail Challenge: Optimizing Customer Experience and Sales at OmniMart Retailers

Objectives

To make OmniMart the most customer-centric retailer by delivering personalized experiences that drive growth, strengthen loyalty, and build lasting relationships with every generation and income group of customers

Data Description

Customer Information

- Customer ID – Unique identifier
- Demographics – Name, Age, Gender, Income, Customer Segment
- Contact & Location – Email, Phone, Address, City, State, Country, Zipcode

Transaction Details

- Last Purchase Date
- Total Purchases
- Amount Spent

Product Information

- Product Category (Electronics, Clothing, etc.)
- Product Brand
- Product Type

Transaction Logistics

- Shipping Method (Standard, Express, Same-Day)
- Payment Method (Credit Card, Wallets, etc.)
- Order Status (Shipped, Delivered, Canceled)

Feedback

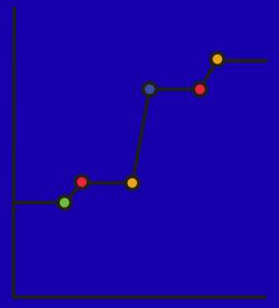
Customer Ratings & Feedback

Analysis

01

Univariate Analysis

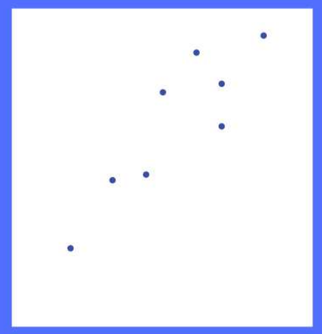
Analysis of a single variable to understand its distribution, central tendency, and spread.



02

Bivariate Analysis

Analysis of the relationship between two variables



03

Multi-variate Analysis

Analysis of more than two variables simultaneously to study complex relationships

04

Outliers Analysis

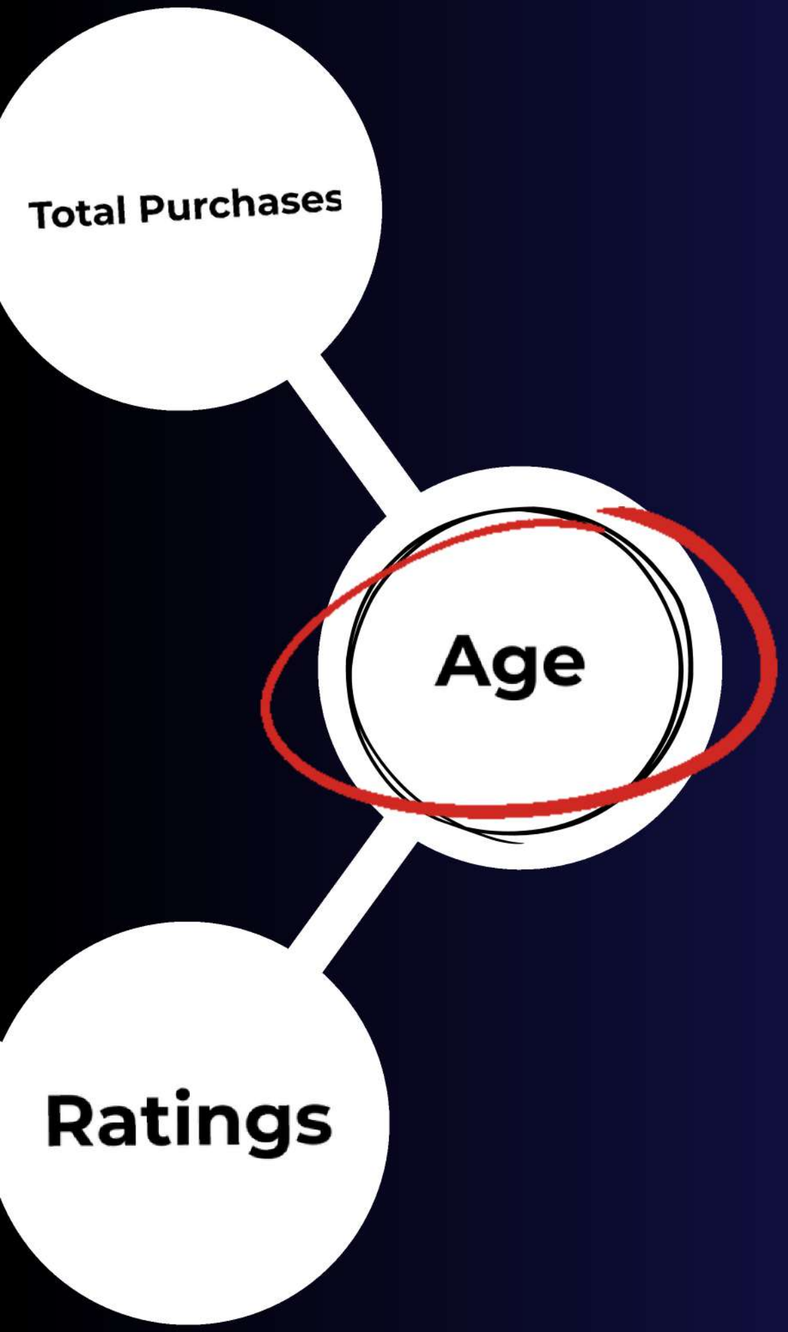
Detecting and studying data points that deviate significantly from the rest of the dataset

05

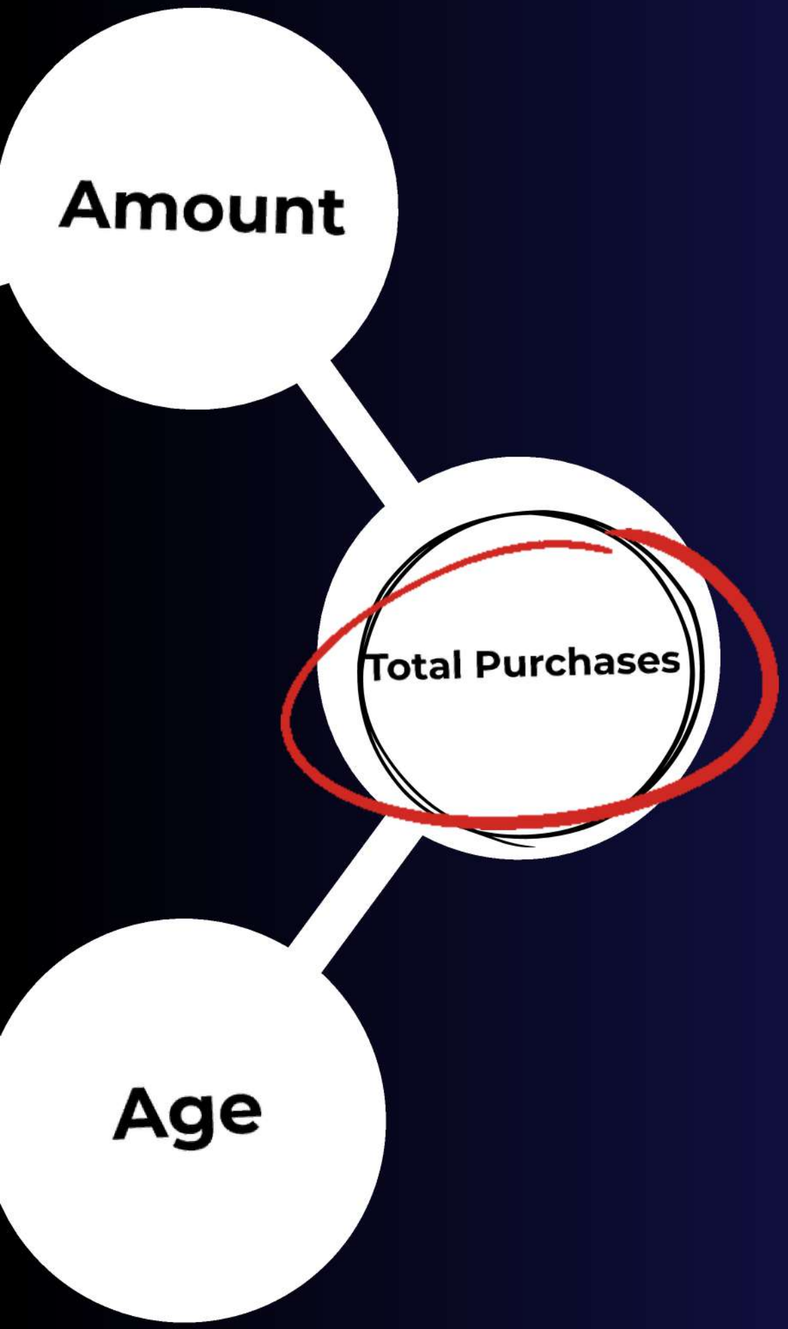
Time-series Analysis

Analysis of data over time to identify trends, patterns, and seasonality

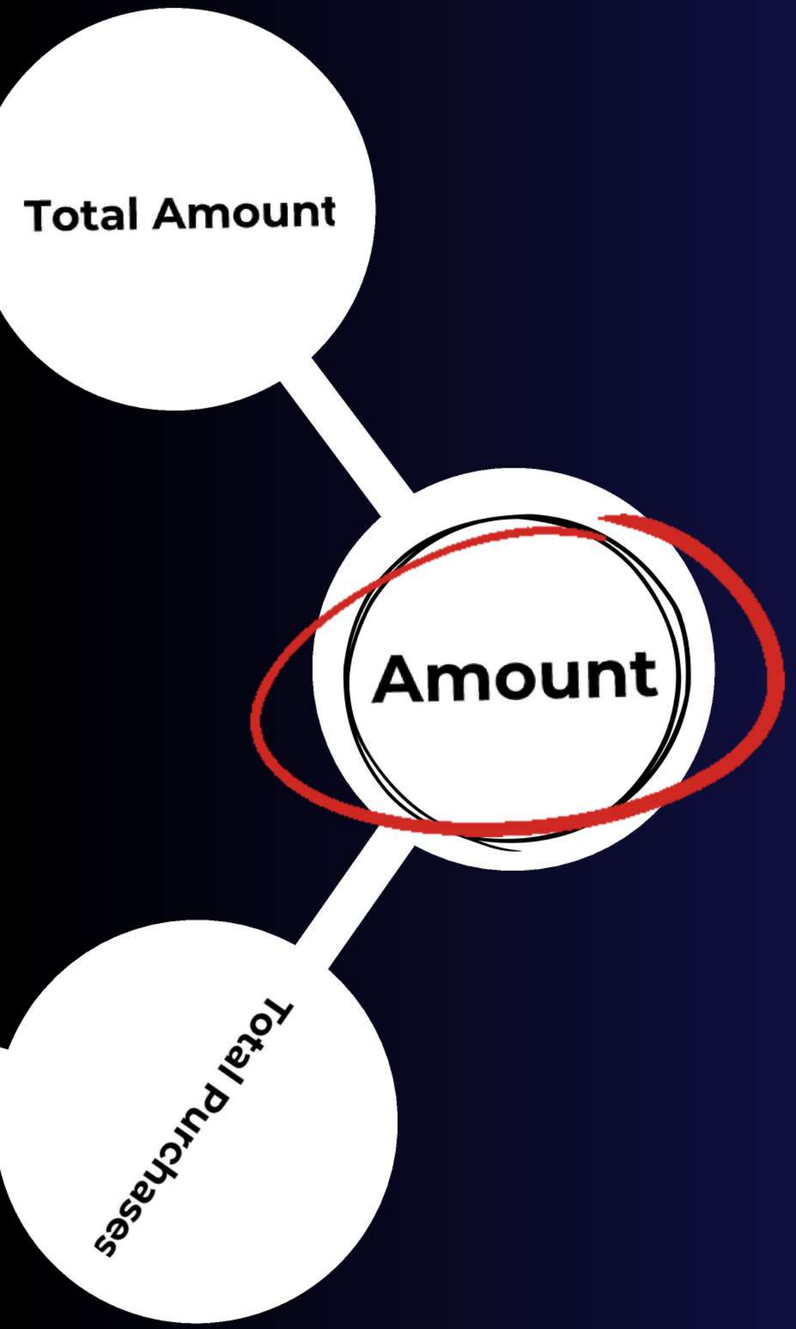
Statistical Analysis of Variables



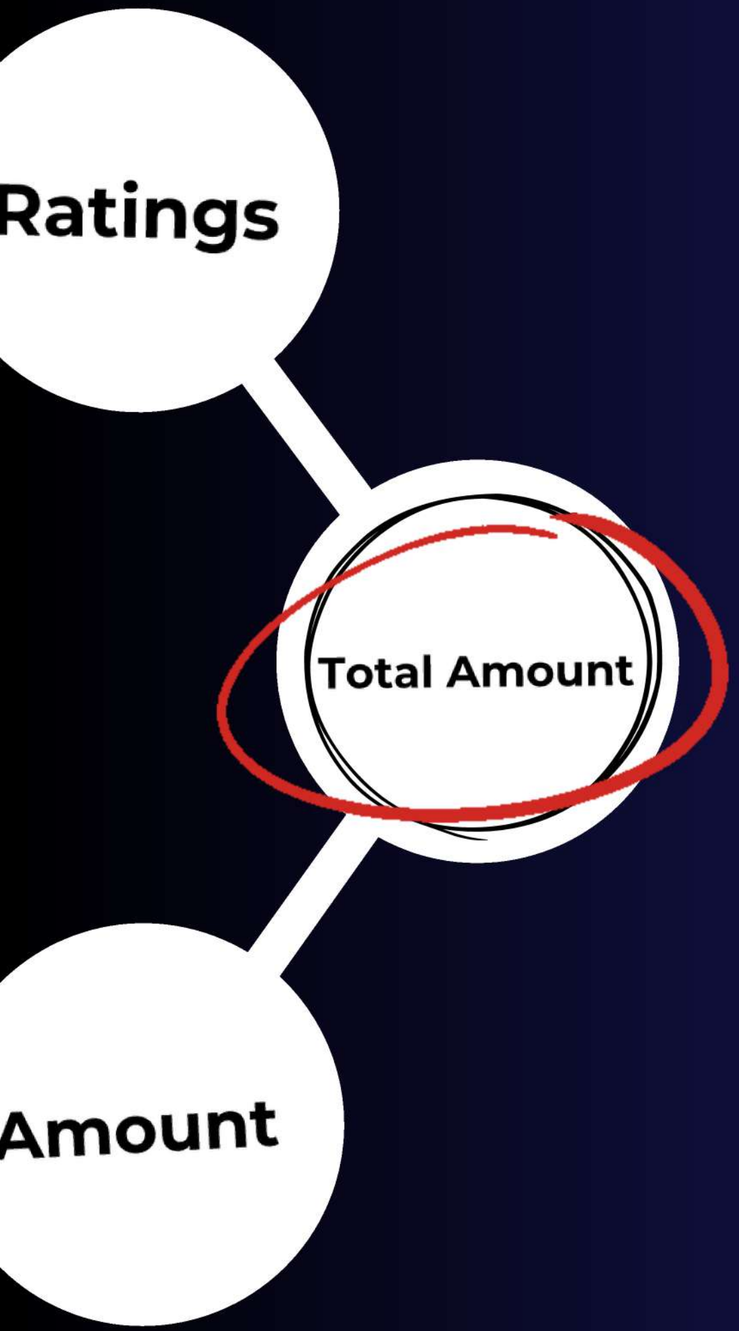
Mean	**31.67**
Median	26.00
Mode	20.0
Standard Deviation	13.22
25%ile	21.00
50%ile	26.00
75%ile	46.00
minimum	18.0
maximum	70.0
Skewness	0.99
Kurtosis	-0.09



Mean	5.37
Median	5.00
Mode	5.0
Standard Deviation	2.87
25%ile	3.00
50%ile	5.00
75%ile	8.00
minimum	1.0
maximum	10.0
Skewness	0.07
Kurtosis	-1.22



Mean	255.18
Median	255.32
Standard Deviation	141.42
25%ile	132.98
50%ile	255.32
75%ile	377.68
minimum	10.00
maximum	499.997911
Skewness	-0.00
Kurtosis	-1.20



Mean	1369.73
Median	1042.04
Standard Deviation	1130.02
25%ile	439.87
50%ile	1042.04
75%ile	2030.90
minimum	10.003
maximum	4999.625796
Skewness	0.97
Kurtosis	0.17

Mean	3.02
Median	3.00
Mode	4.0
Standard Deviation	1.34
25%ile	2.00
50%ile	3.00
75%ile	4.00
minimum	1.0
maximum	5.0
Skewness	-0.09
Kurtosis	-1.26

Age

- Most customers are young adults in their 20s(18-25)
- The spread is quite high (Std Dev = 13.22, Range = 18–70)
- There is a long right tail (customers up to 70), showing presence of older but fewer buyers
- Your core customer segment = 20–30 years
- Marketing campaigns should target younger audiences (students, working professionals) but also retain older loyal customers with age-relevant offers

Total Purchases

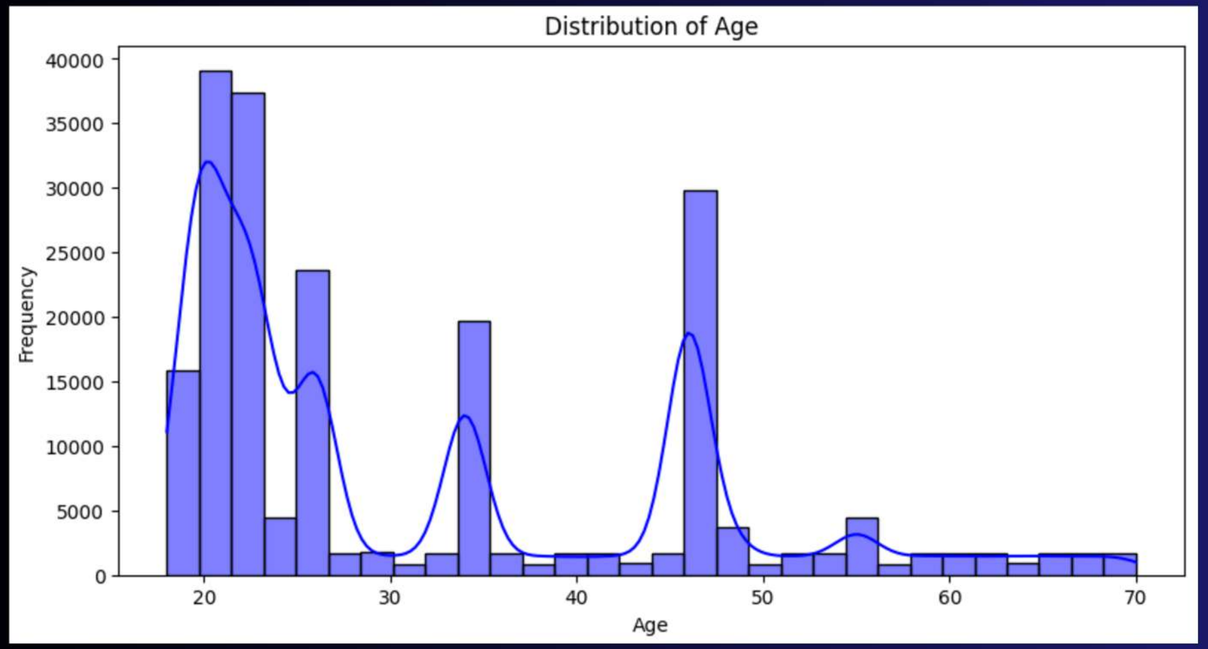
- Distribution is nearly symmetric (Skewness = 0.07)
- Most customers make around 5 purchases
- Encourage mid-frequency buyers (3–5 purchases) to move into the high-frequency group with rewards, discounts, or VIP benefits
- promote bundles or add-ons to increase basket size.

Total Amount

- Many customers spend ~1000 total, but a few heavy spenders push the mean up.
- Your top 25% of customers drive majority of revenue
- For buyers below 1000, provide them with rewards that could encourage them to increase their purchases
- Focus retention/relationship building on the top 25% spenders.

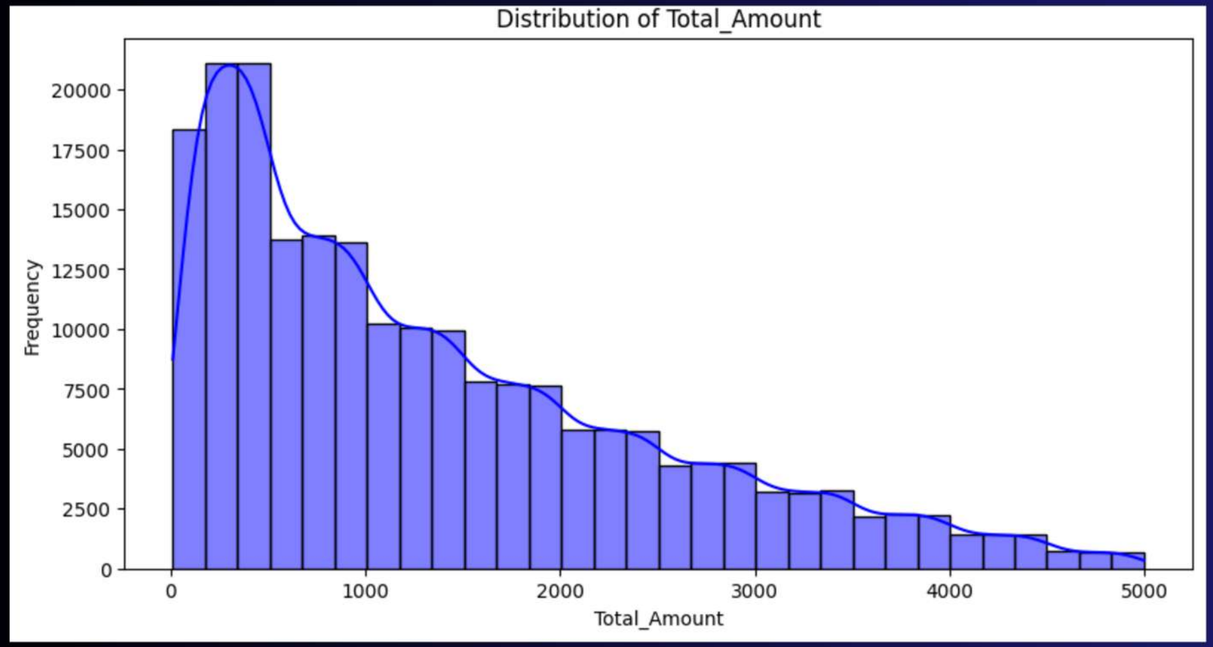
Ratings

- Customers are neutral-satisfied, but not extremely happy (few 5s).
- Distribution is flat (Kurtosis = -1.26) → ratings are spread out.
- Identify the pain points of customers with ratings 1-2



Age Distribution

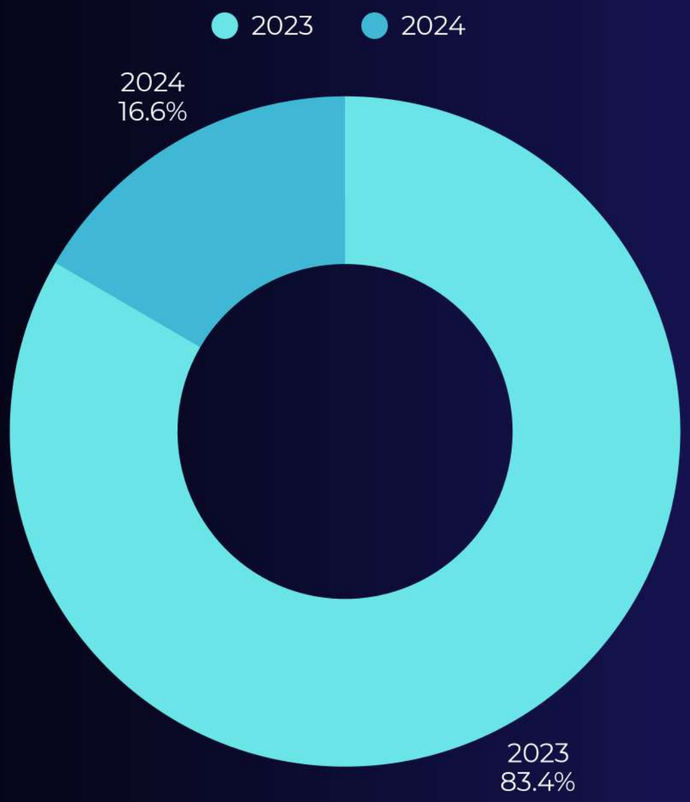
Your core customer segment = 20–30 years



Total amount Distribution

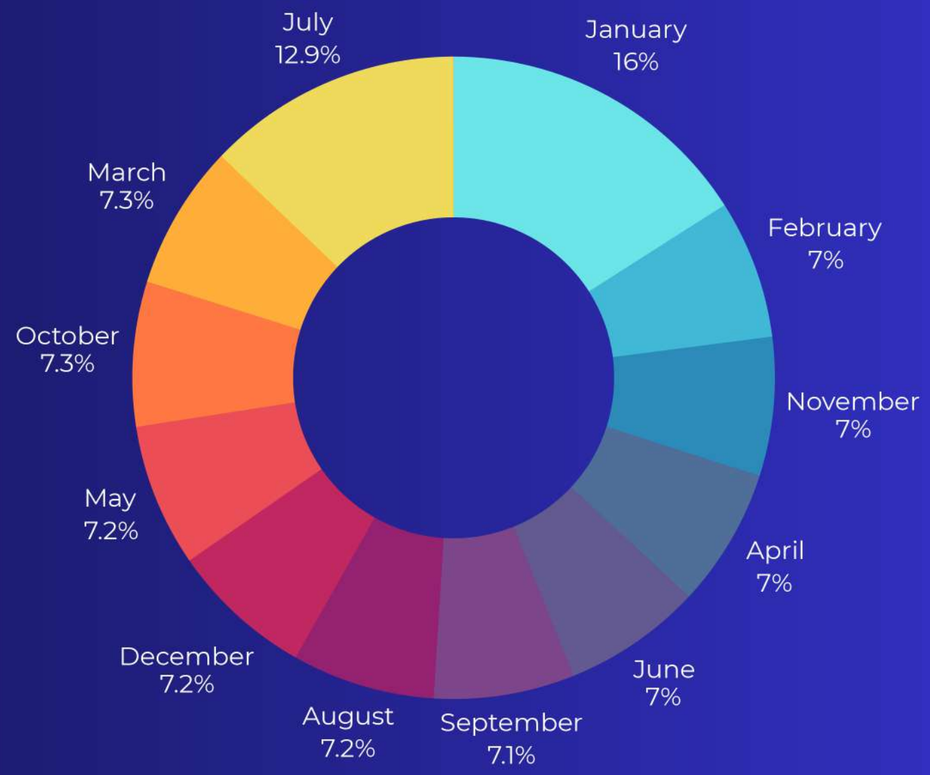
Your top 25% of customers drive majority of revenue

Year-wise Data distribution



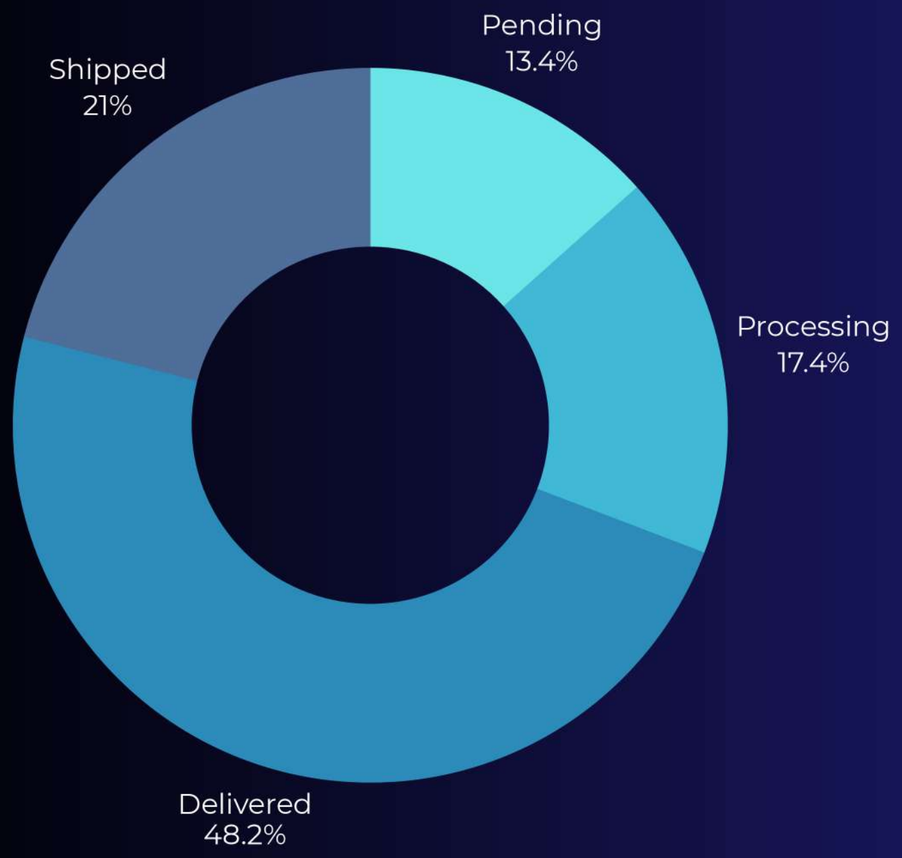
Data for the whole year of 2024 is unavailable in this data set which is the reason for such distribution

Month-wise Data distribution



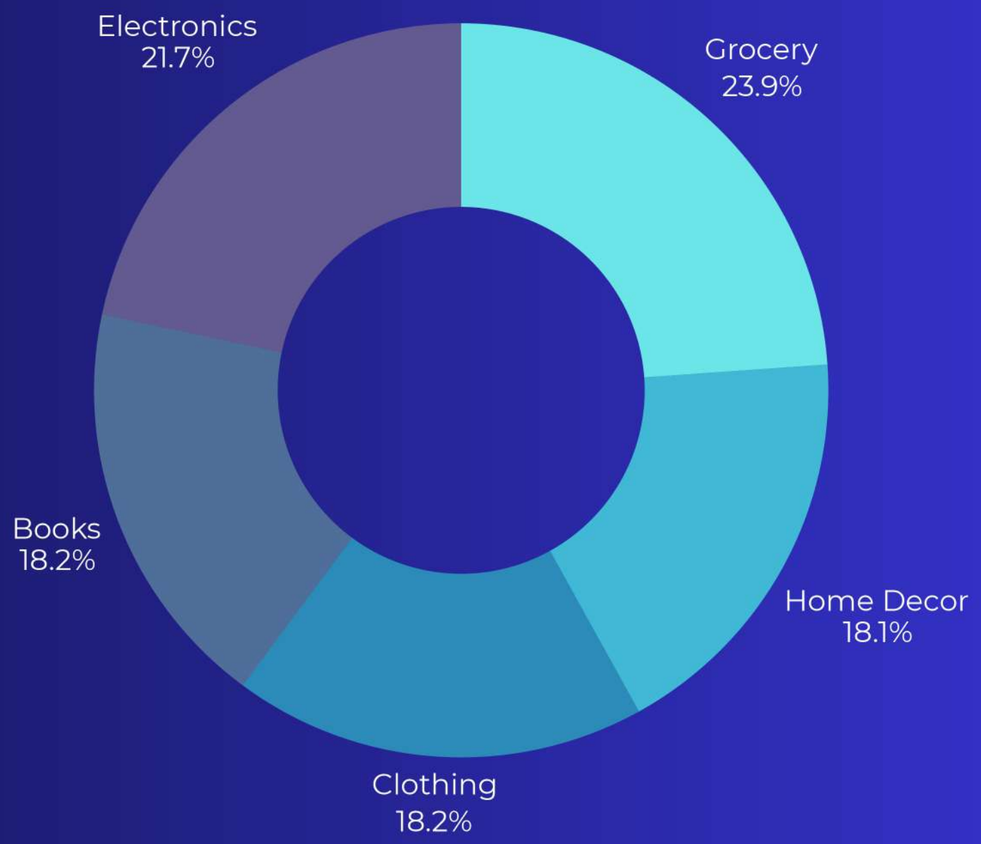
Shows the percentage of monthly entries in the data set

Data distribution based on Order-status



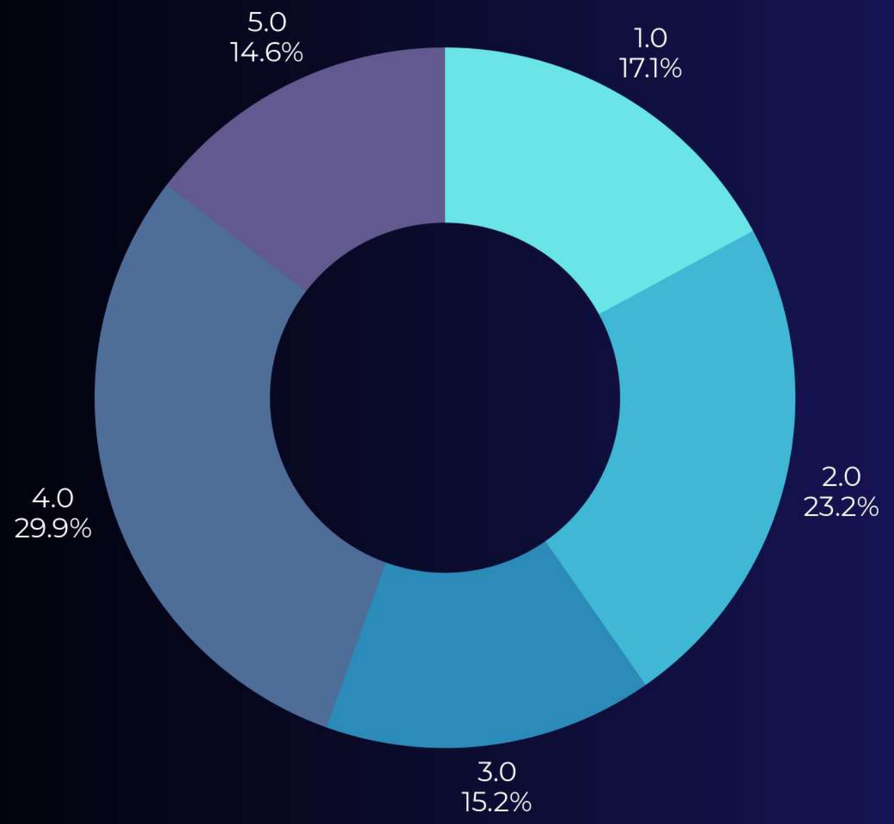
Business has a strong delivery rate, but almost half of transactions are stuck in pre-delivery stages

Product-category-wise data distribution



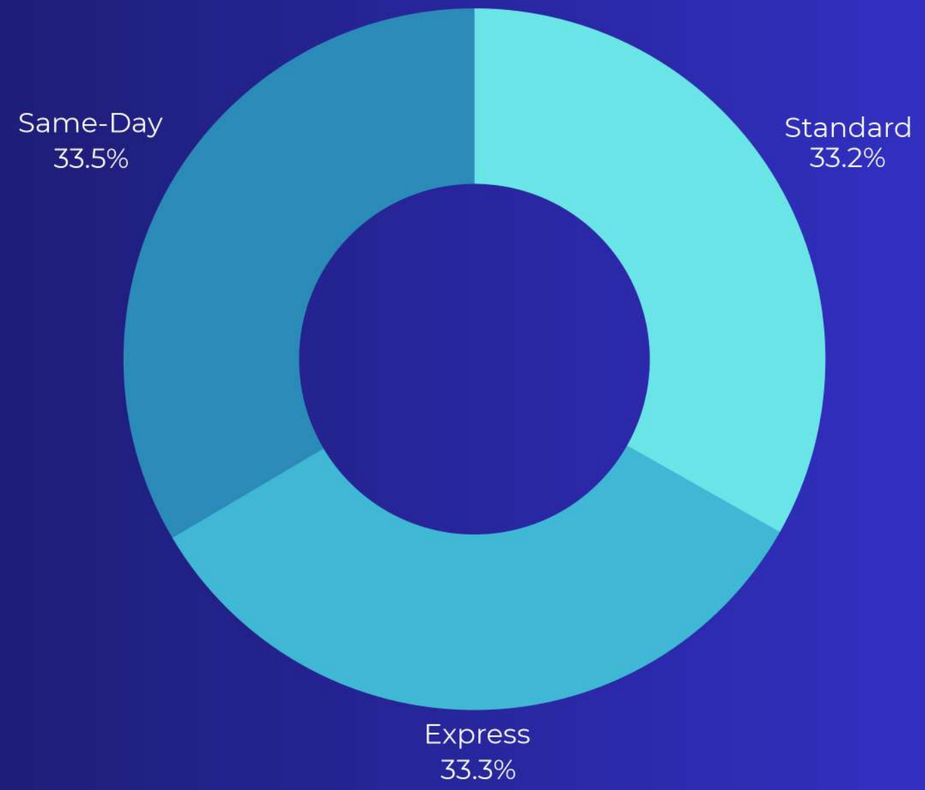
Grocery dominates, followed by Electronics

Ratings-wise data distribution



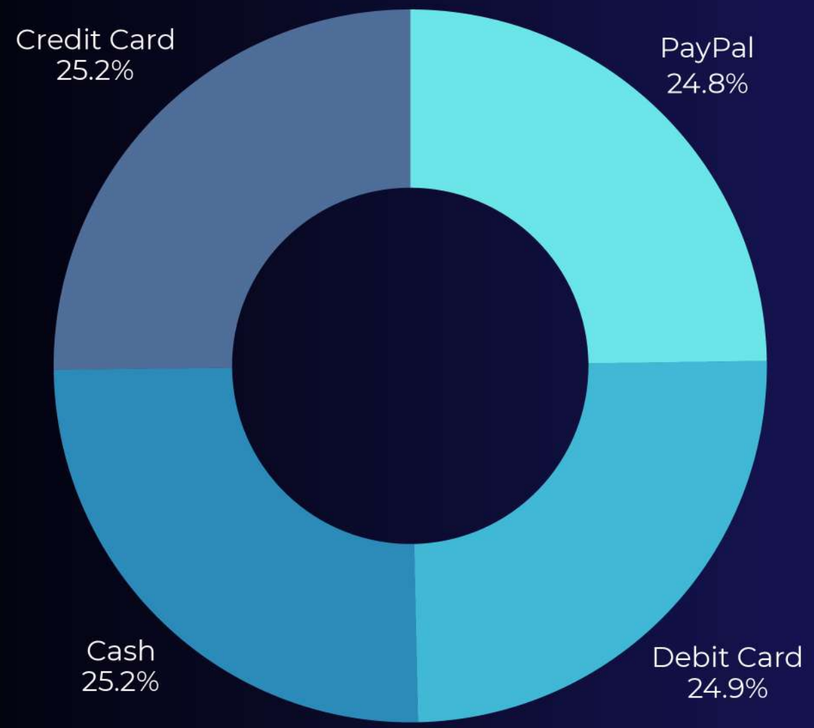
The significant finding is that the vast majority of customers are happy. However a major chunk is unhappy about 40%(1-star 2-star ratings)

Shipping Method-wise data distribution



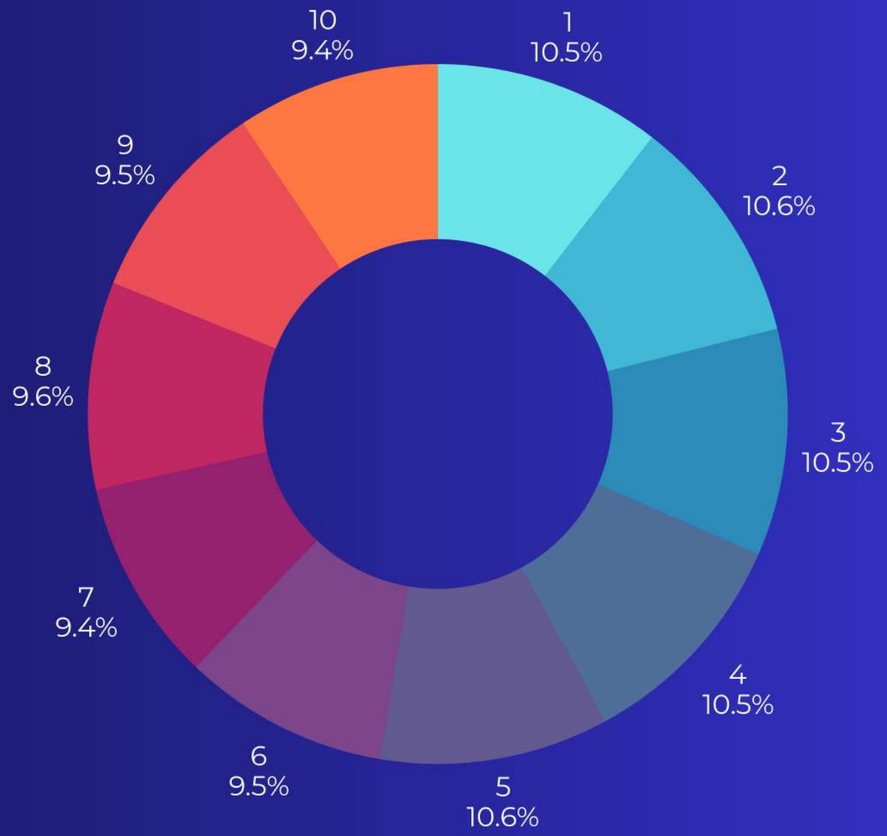
Almost evenly distributed

Payment-Method-wise Data distribution



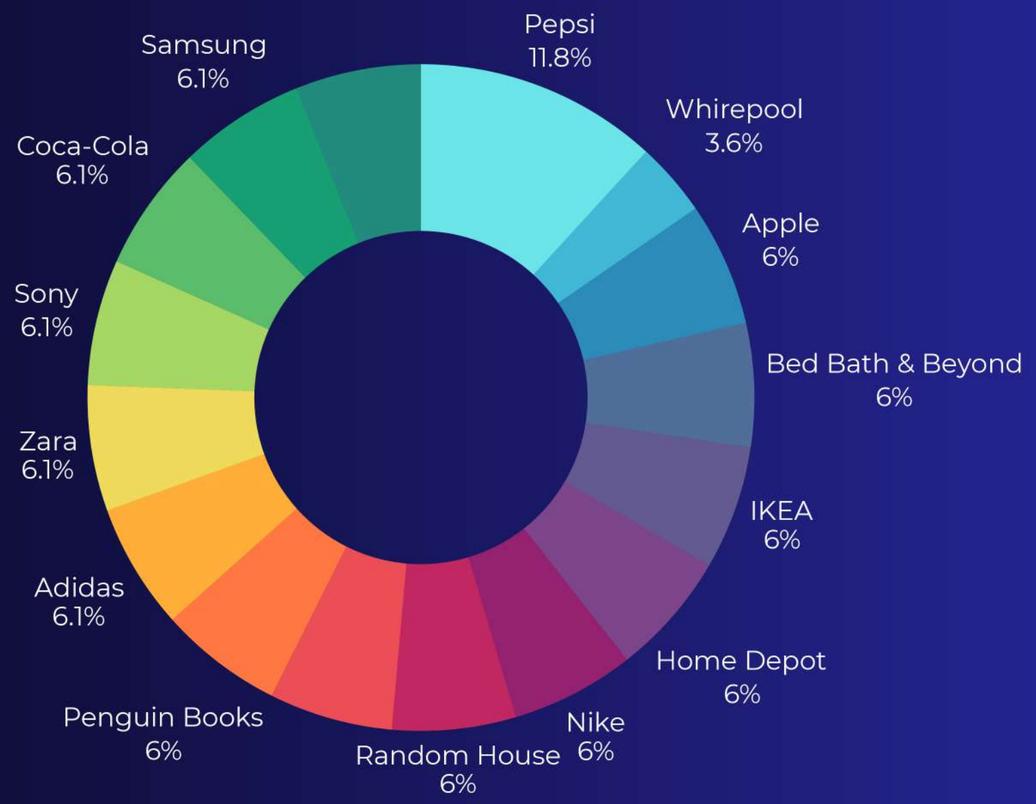
Evenly distributed

Purchase-amount-wise Data distribution



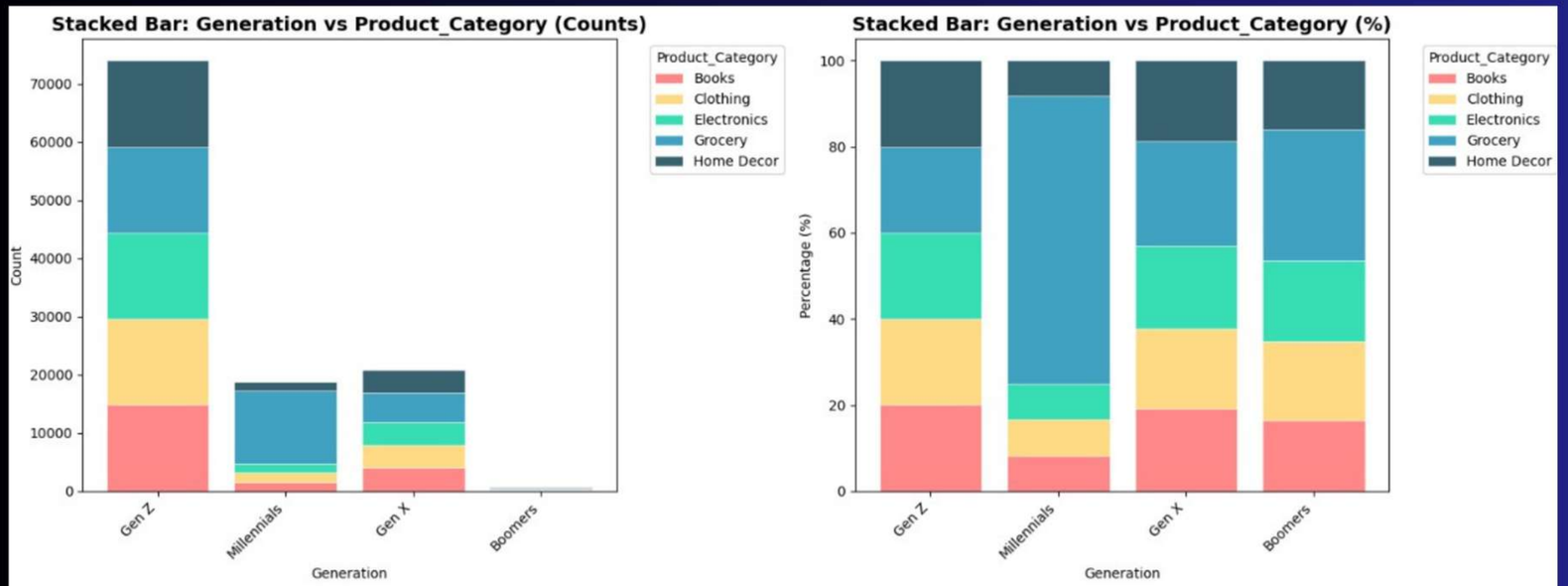
Most people purchases in the range of 1-5

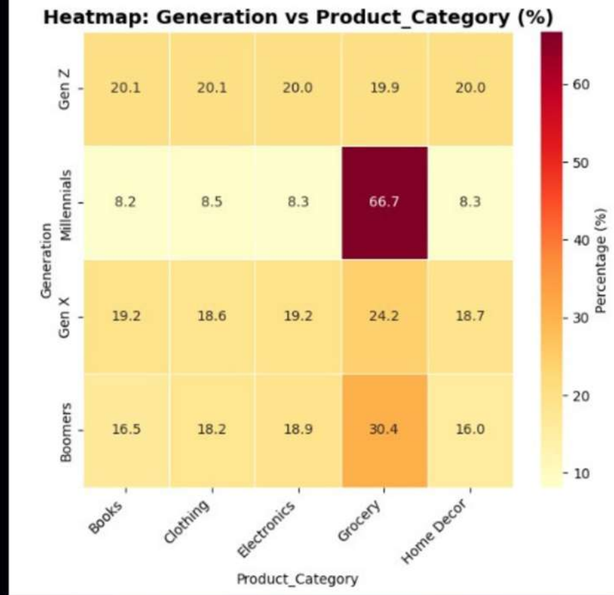
Brand-wise Data Distribution

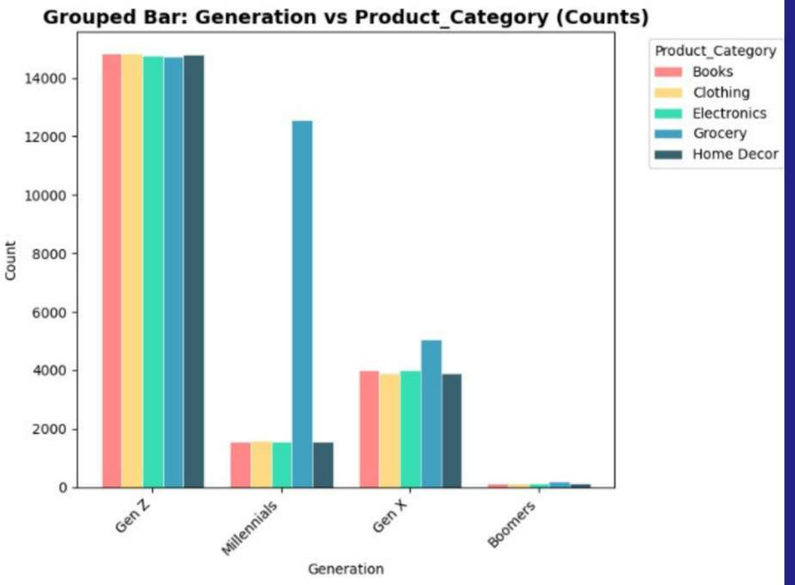


Pepsi has the largest share.Whirlpool has the lowest share.
Customer attention is evenly distributed across multiple brands

Generation vs Product category Analysis





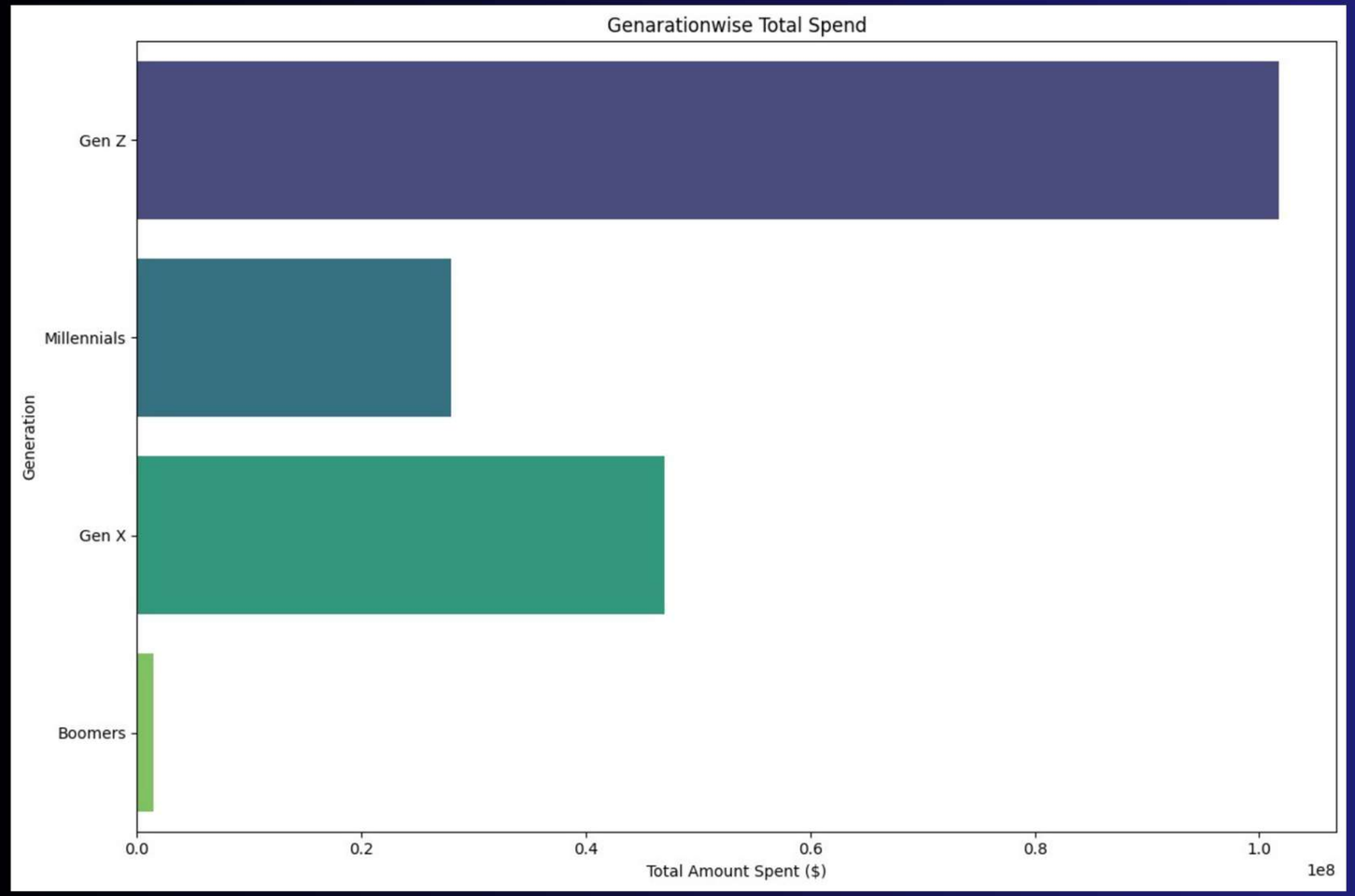


Gen Z dominates → very high counts across almost all categories. Gen Z distribution is balanced for every category ~20% each.

Millennials spike → huge count for Groceries compared to other categories. Groceries ~67% of all their purchases.

Gen X shows mid-level engagement with all categories, slightly higher for Groceries.
Boomers negligible across all categories, no clear leader

Generation-wise Total Spend(Revenue)



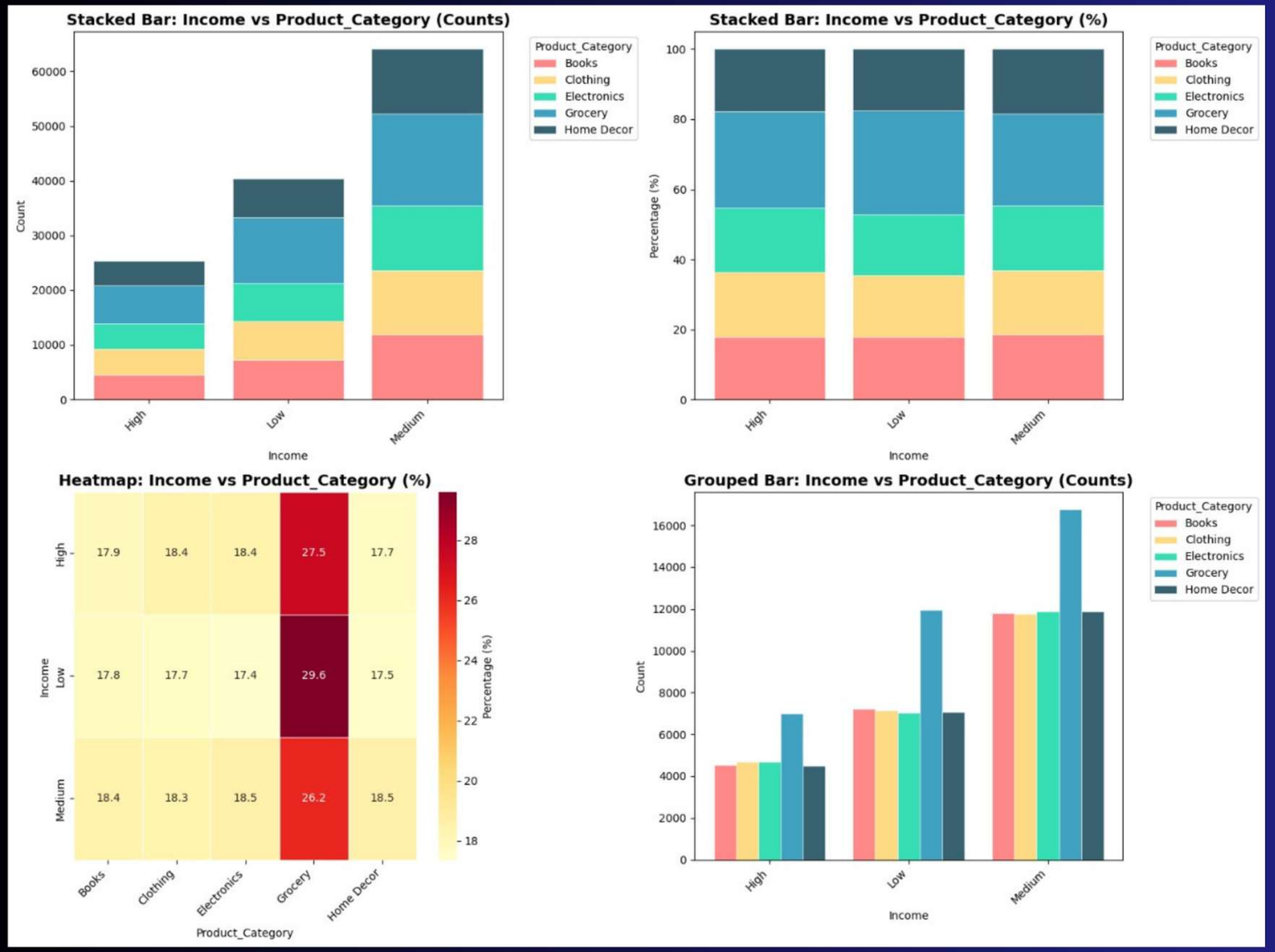
Gen Z is the largest, most balanced contributor across categories → main growth driver.

Millennials are highly grocery-focused and Pepsi-loyal.

Gen X is steady, mid-level, and highly satisfied → strong promoters.

Boomers are the smallest group but spend more per purchase → premium segment opportunity.

Product Category vs Income Analysis

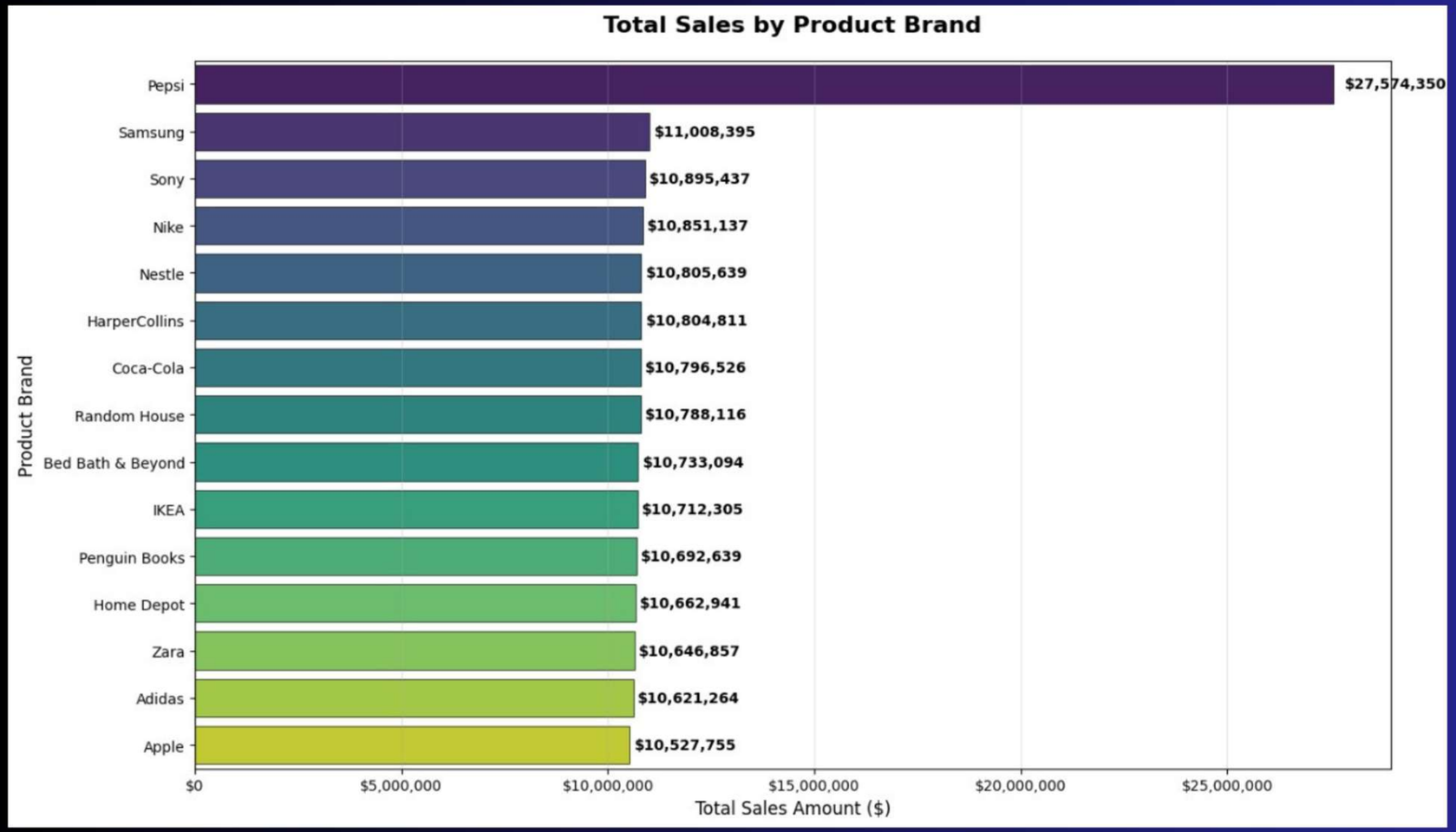


Medium-income customers dominate total purchases and are balanced across essentials + discretionary.

Low-income customers are necessity-driven (groceries ~30% of spend).

High-income customers are small in number but have premium grocery & lifestyle potential.

Product Brand vs Total Spend Analysis



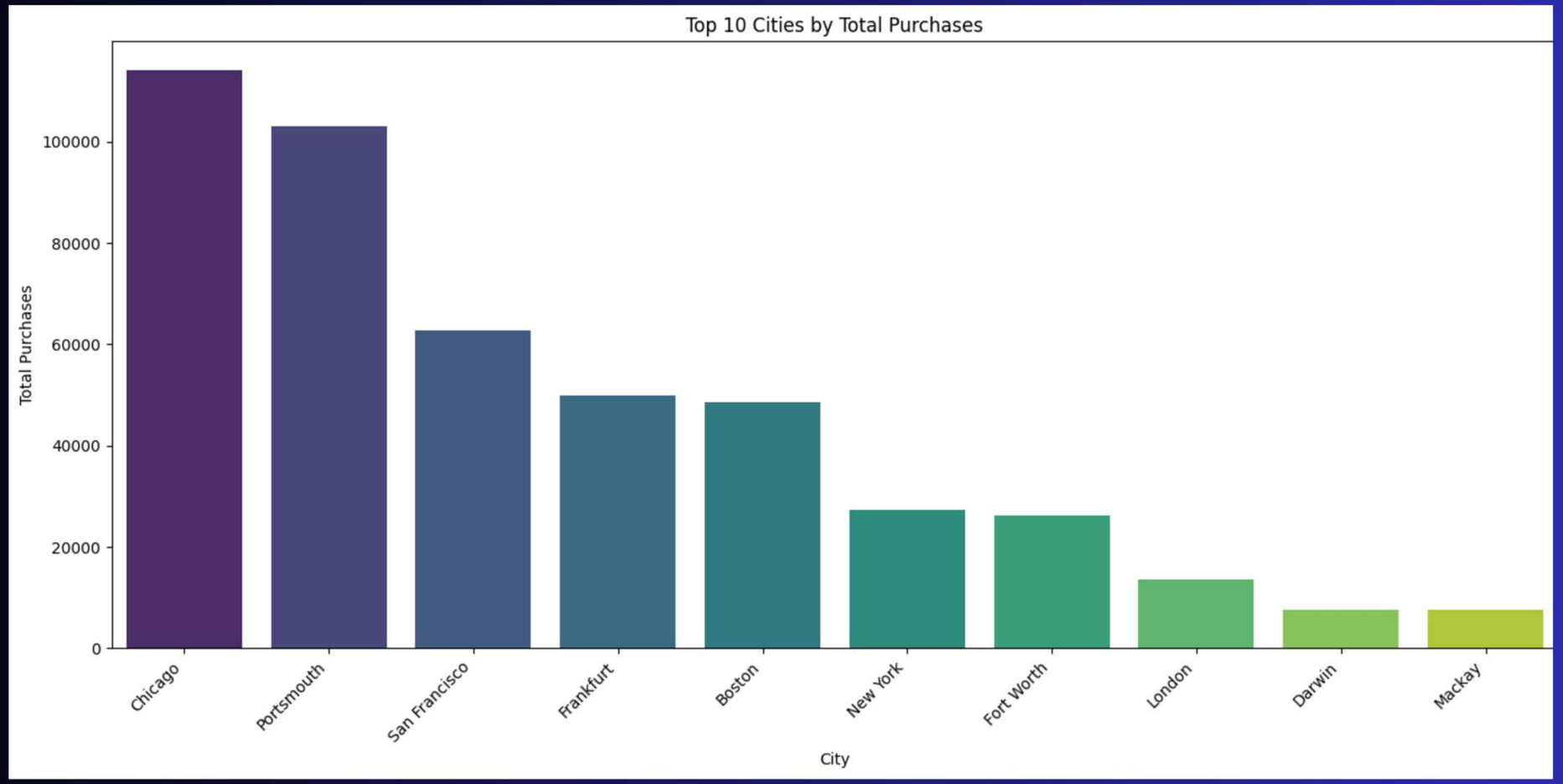
Pepsi dominates with ~$27.6M sales, 2.5× higher than any other brand.

Other brands (Samsung, Sony, Nike, Coca-Cola, Nestle, etc.) cluster around ~$10.7M → flat competition.

Apple underperforms despite premium status.

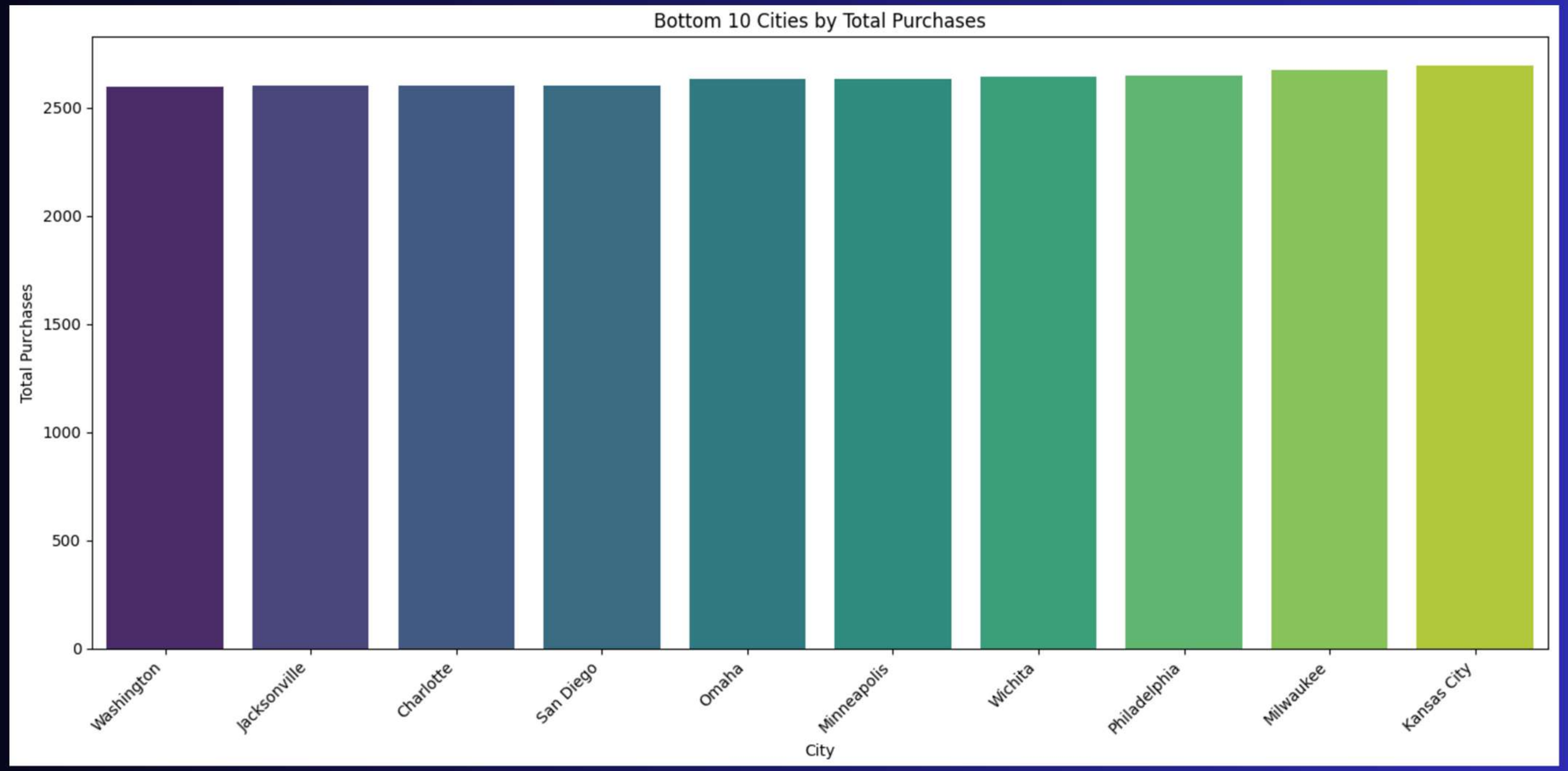
Book publishers (HarperCollins, Penguin, Random House) surprisingly match global brands in revenue.

Top 10 Cities by Total Purchases



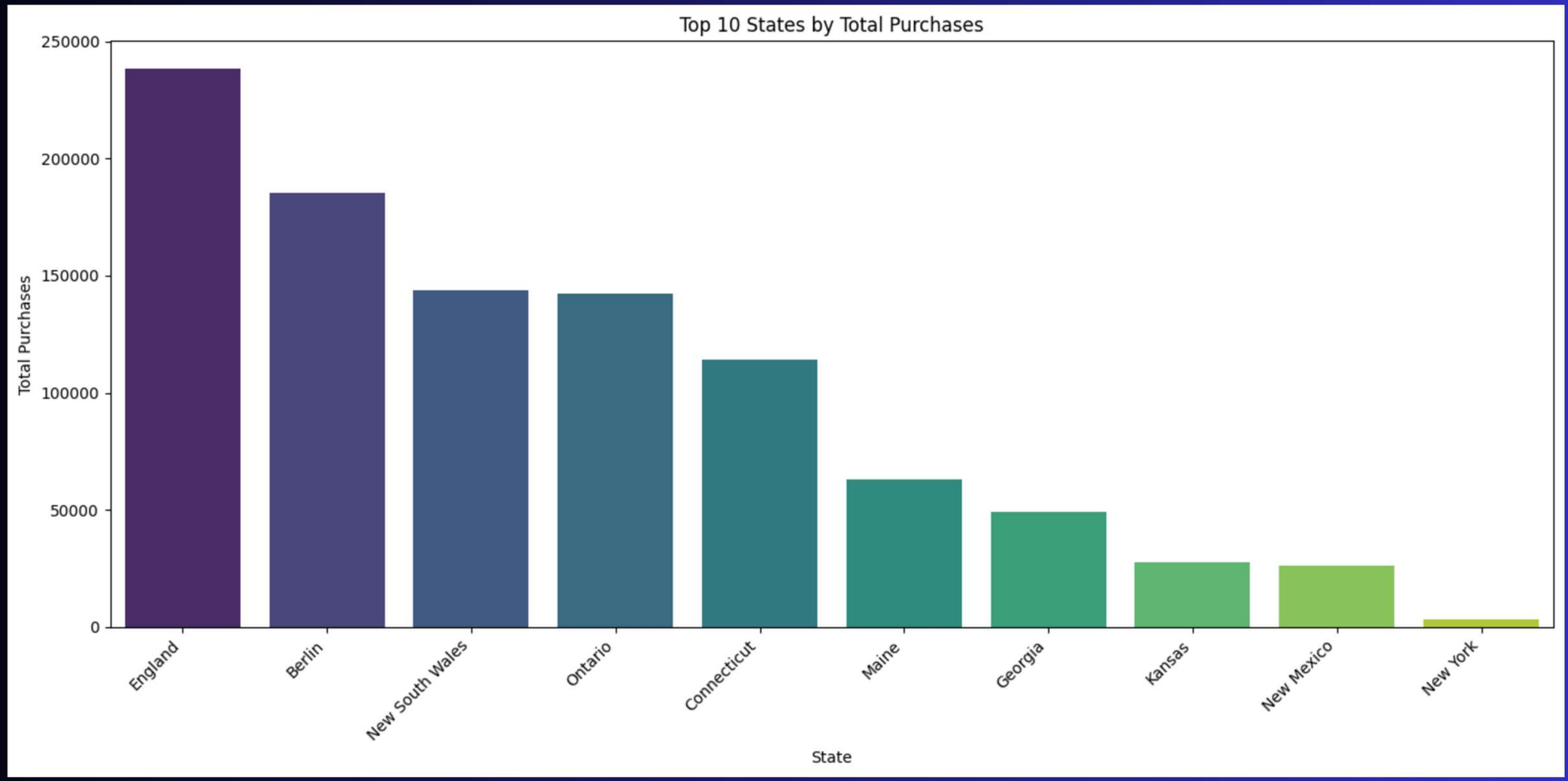
The top 3 cities (Chicago, Bergenouth, San Francisco) account for a massive portion of the total revenue shown here.

Bottom 10 Cities by Total Purchases

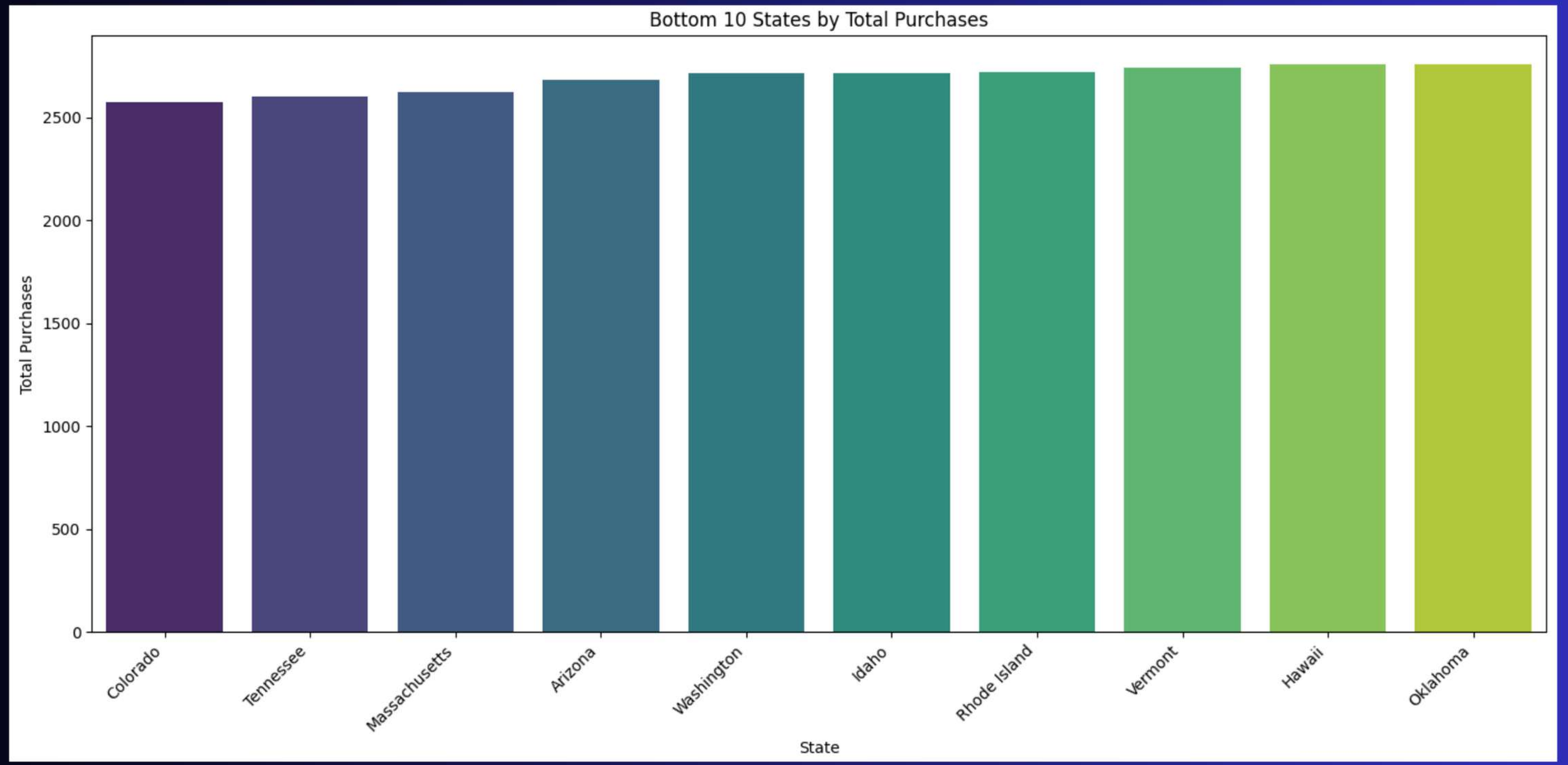


The bottom 10 cities indicate the cities that are untapped somehow. The presence of Washington, Philadelphia, San Diego, Charlotte, and Kansas City on this list is a five-alarm fire.

Top 10 States by Total Purchases



Bottom 10 states by Total Purchases

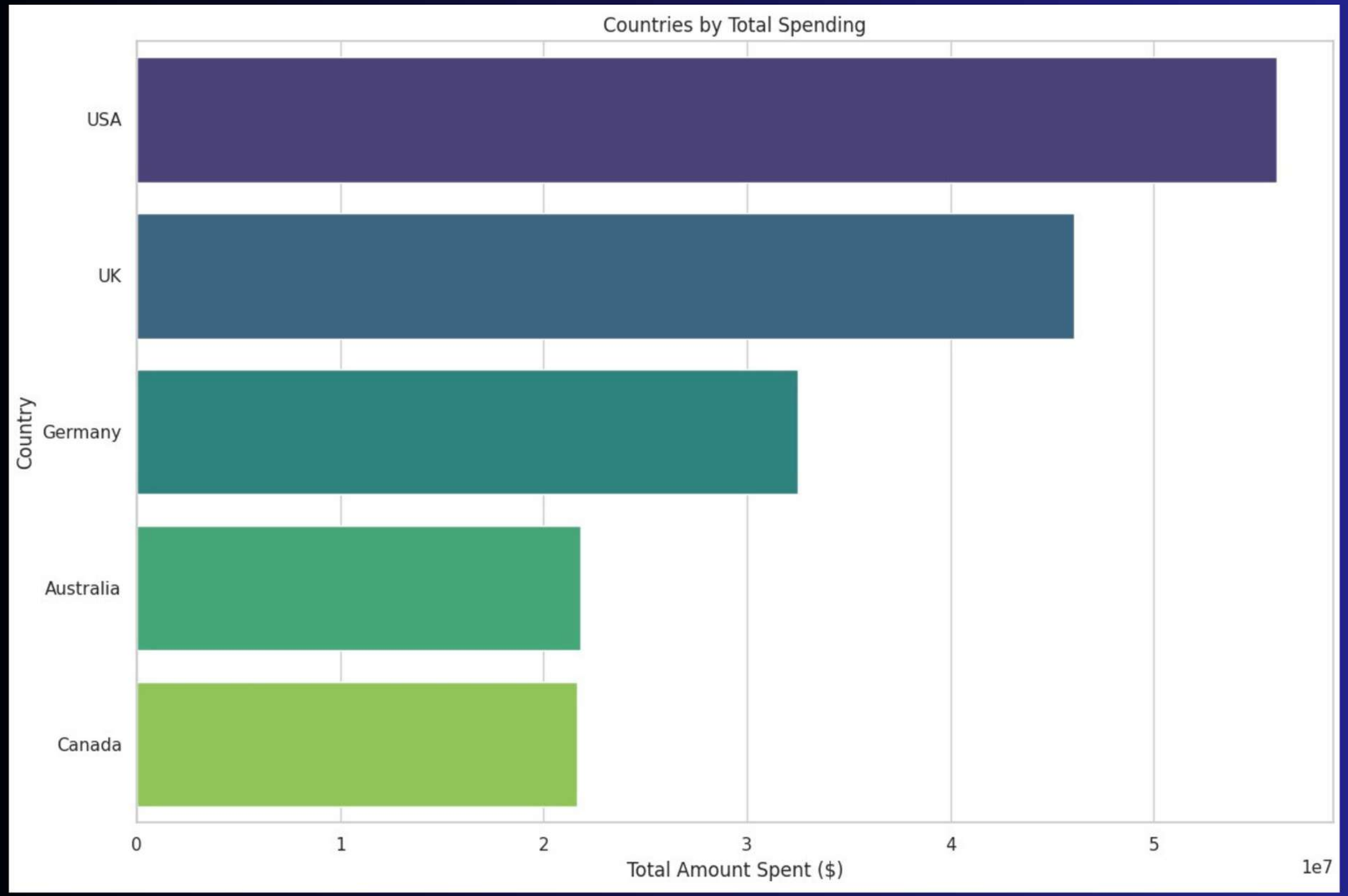


Many of these states (Idaho, Hawaii, Rhode Island, Oklahoma, Vermont) have smaller or more dispersed populations compared to your top states

This explain the root causes : Lower Market Size; Logistical and Economic Barriers

Red Flag inludes the presence of Washington state

Countries vs Total Spending

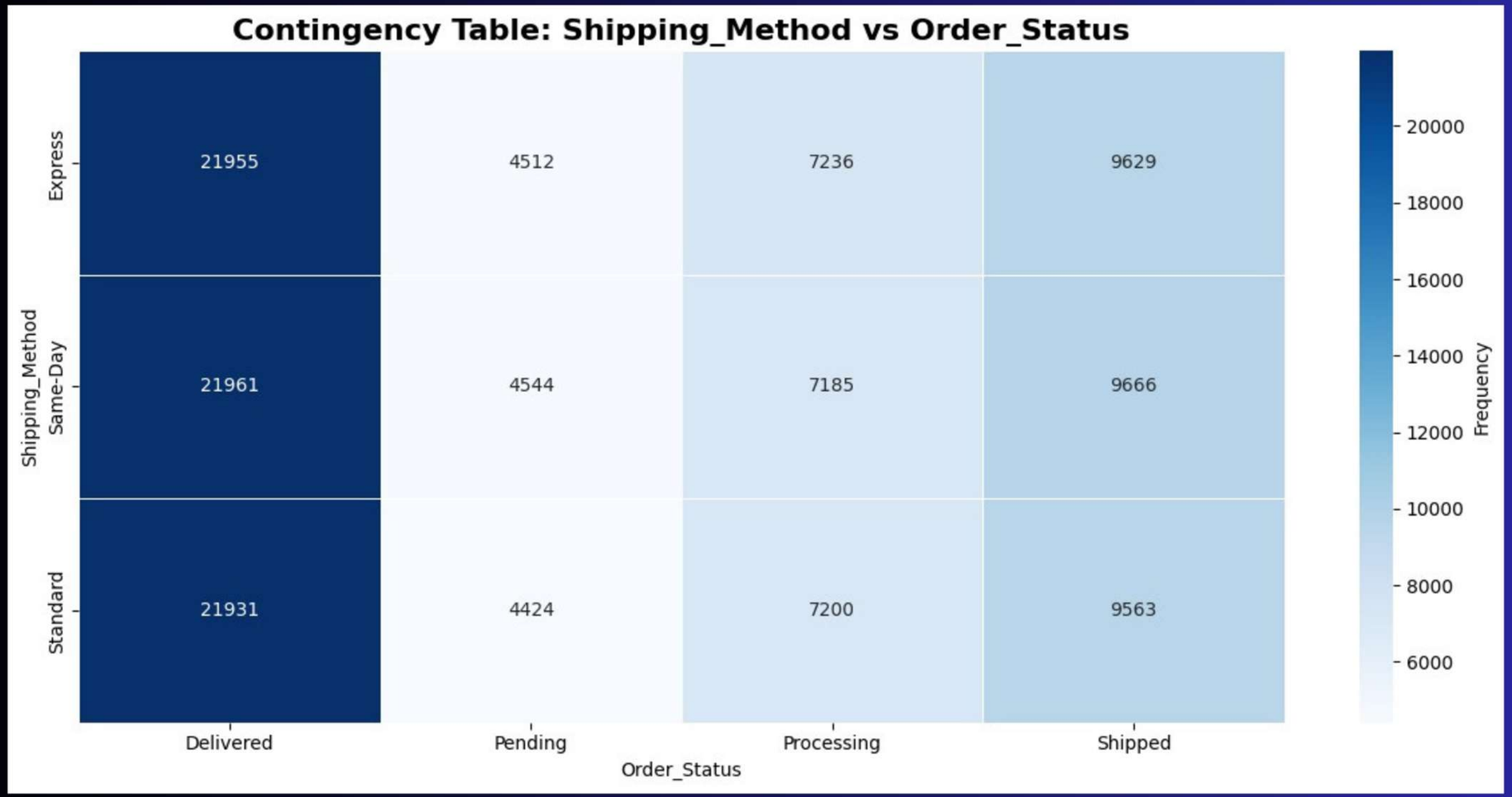


The very low revenue in Australia (an English-speaking country with a culture and consumer habits similar to the US and UK) is a surprising and important red flag.

The dominance of USA is expected, as it is often the home market for businesses, benefiting from established brand awareness, optimized logistics, and cultural familiarity.

Germany represents a successful initial international expansion

Shipping vs Order Status



Delivered orders dominate (~21k per shipping method).

Shipped but not delivered (~9.6k per method) → key bottleneck in last-mile delivery.

Consistent performance across Express, Same-Day, and Standard.

AOV-Average Order Value

```
AOV by Customer Segment:
 Customer_Segment
New         1369.299925
Premium     1364.774504
Regular     1371.364497
Name: Total_Amount, dtype: float64
```

Remarkable Consistency Across Segments
Strong First Impression
The most critical and surprising insight is that Premium Customers do not have the highest Average Order Value (AOV). In fact, they have the lowest AOV

USA is a red-flag here
Consider launching high-valued yet useful products for the premium customers for USA
Lack of Marketing campaign or population might be the reason for Canada's low AOV

```
AOV by Country:
 Country
Australia    1374.634199
Canada       1364.505787
Germany      1372.176861
UK           1381.415428
USA          1362.323932
Name: Total_Amount, dtype: float64
```

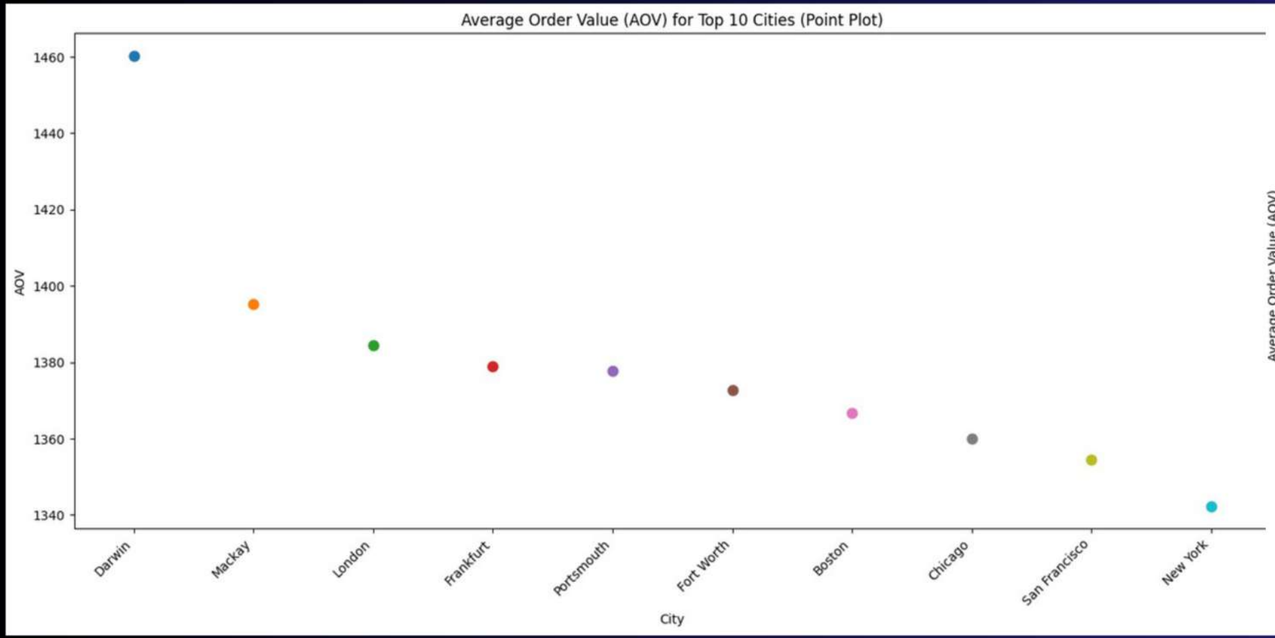
```
AOV by Shipping Method:
 Shipping_Method
Express      1368.556558
Same-Day     1371.523542
Standard     1369.047533
Name: Total_Amount, dtype: float64
```

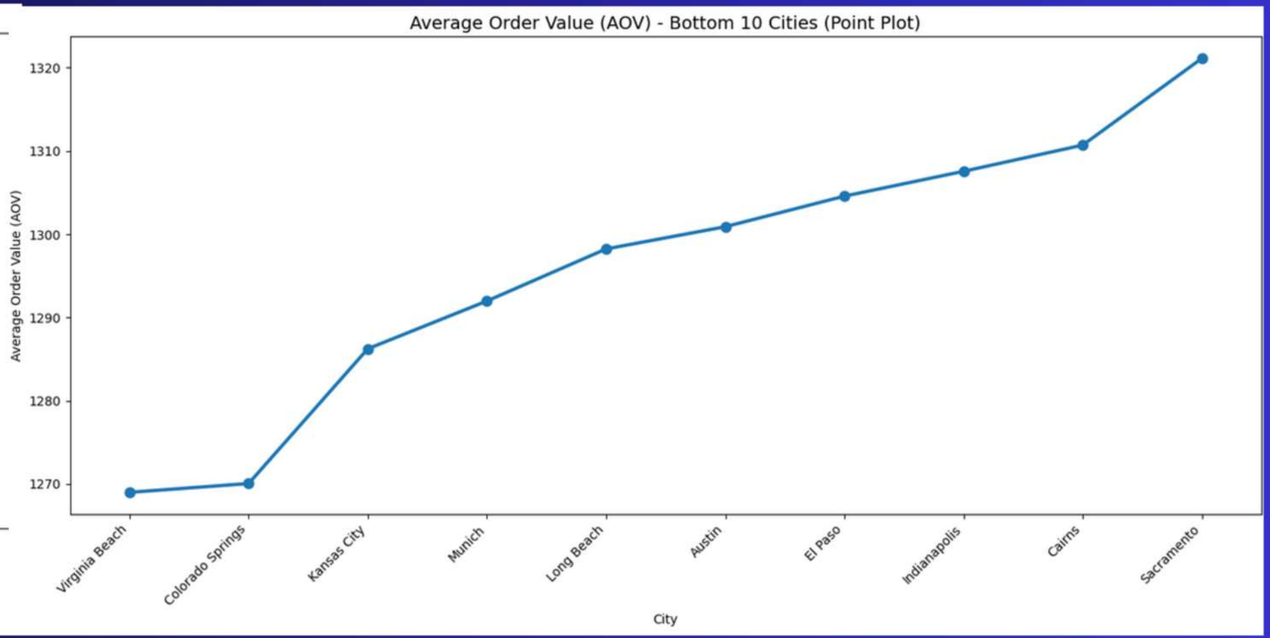
Since shipping choice doesn't affect product revenue, you can now evaluate your shipping strategies purely on a cost-to-serve basis
Understand that offering free "Same-Day" shipping as a promotion is a very expensive customer acquisition tool that does not lead to a higher return per order. Use it sparingly and strategically for high-value customer segments only.

AUG,MAR,APRIL sees the highest AOV suggesting higher trends during the Summer Holidays. December's AOV rebounds significantly to become the third-highest month(Christmas)

```
AOV by Month:
 Month
April        1380.494460
August       1386.763844
December     1377.801124
February     1372.683216
January      1358.543157
July         1370.914202
June         1356.291170
March        1385.842182
May          1374.944233
November     1358.066377
October      1364.662022
September    1362.830635
Name: Total_Amount, dtype: float64
```

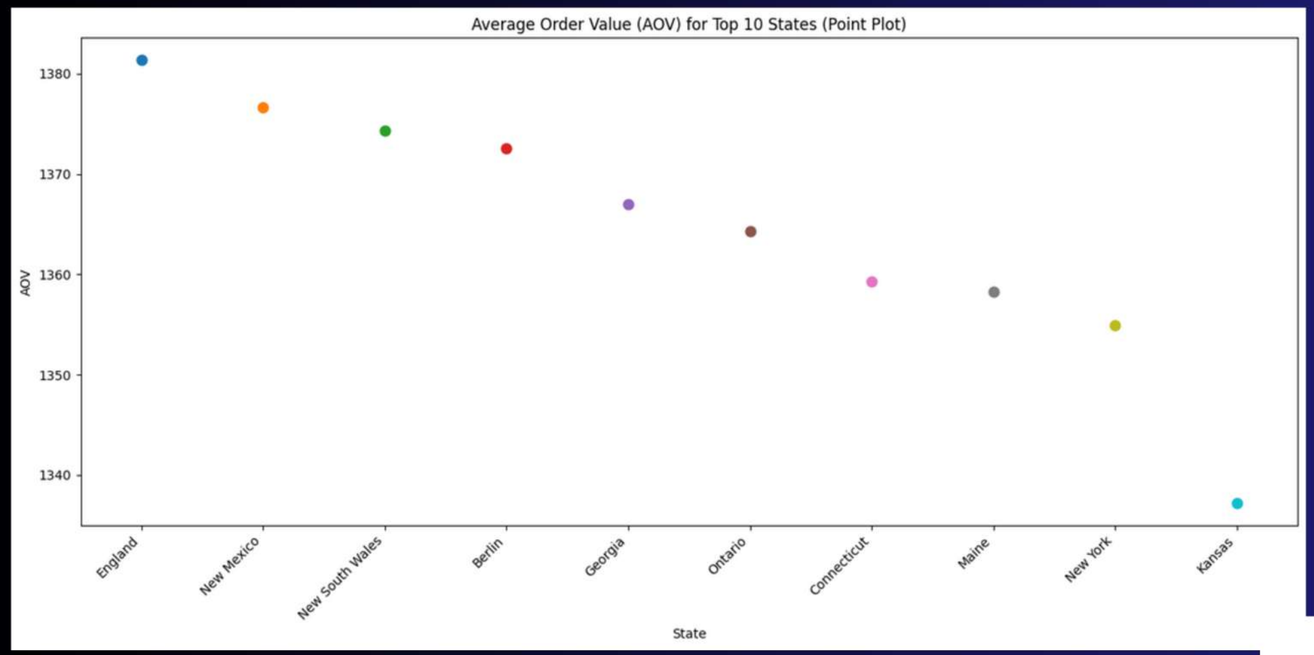
AOV for Top 10 & bottom 10 cities



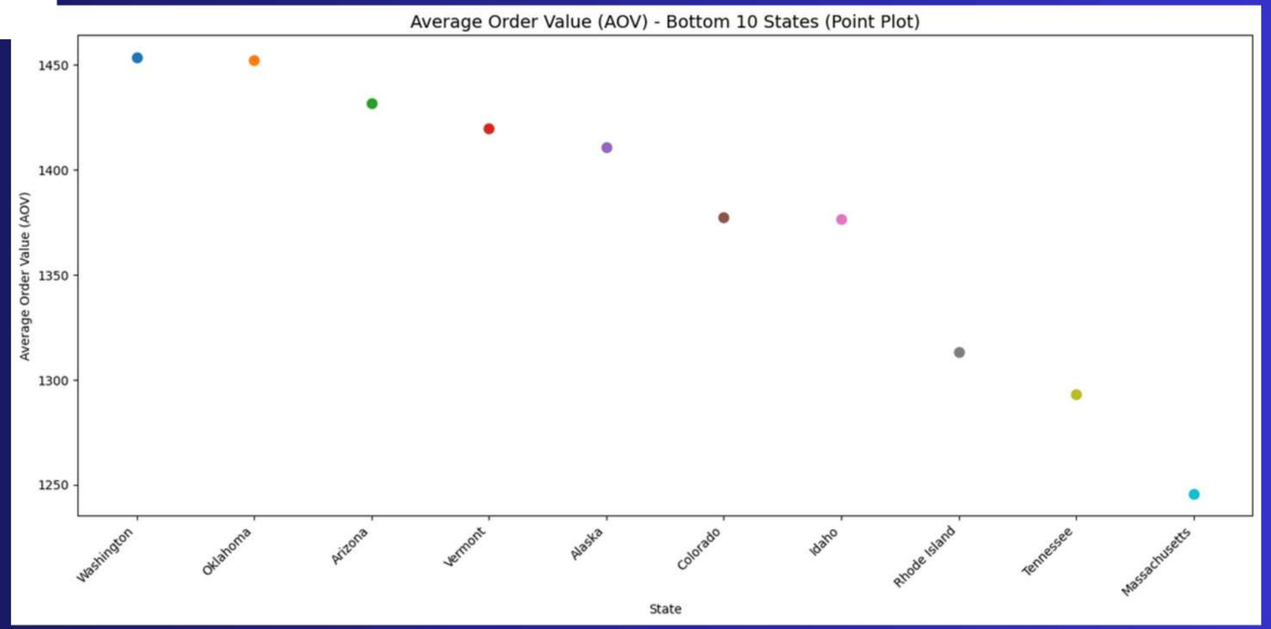


The "Philadelphia" and "San Diego" Paradox: These are major U.S. metropolitan areas appearing on the bottom list for volume but with a mid-tier AOV. This is a critical signal. It means that when you do attract customers in these cities, they spend a reasonable amount

AOV for top 10 and bottom 10 states



Similar Observations



General Recommendations

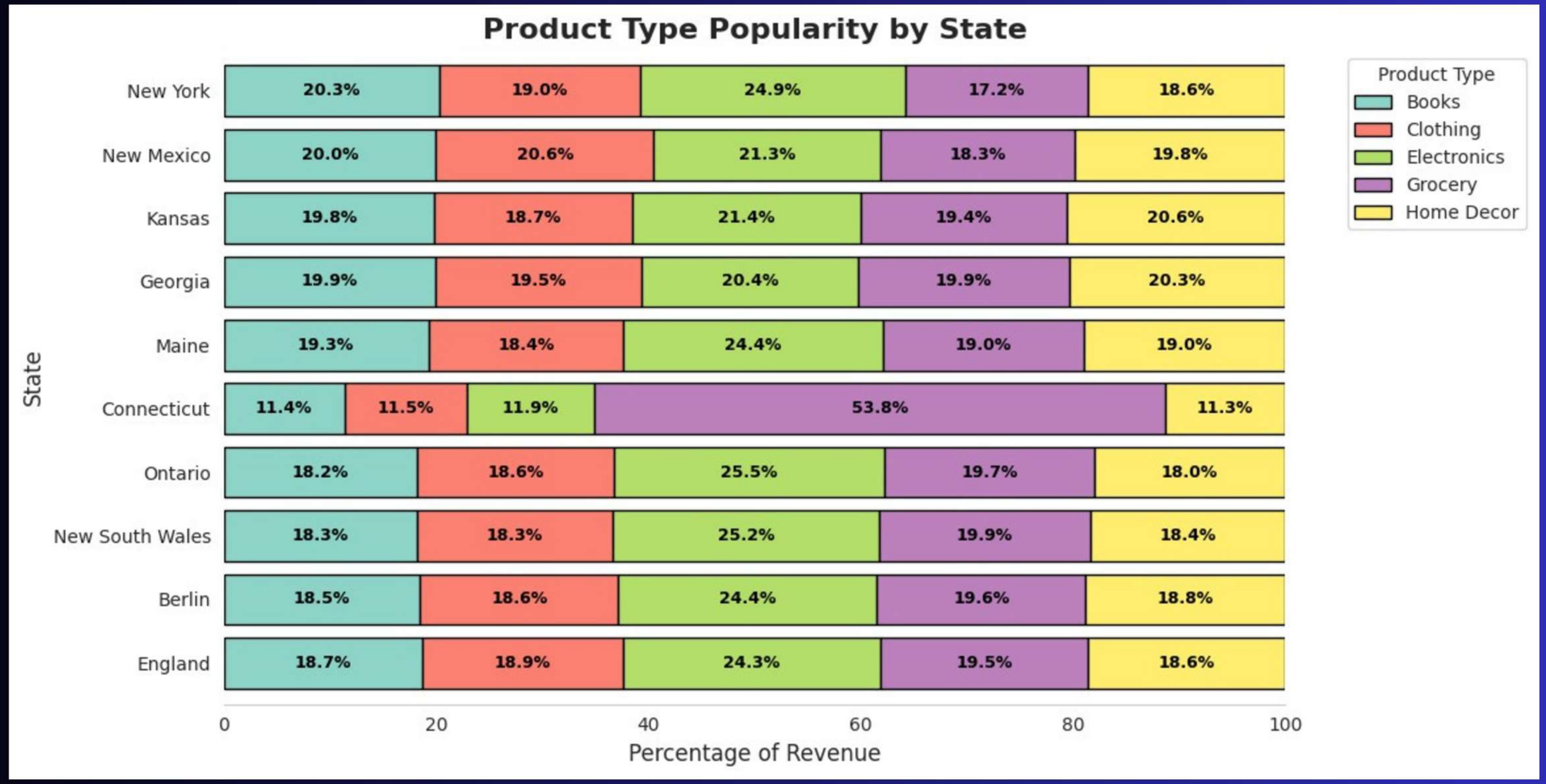
Leverage Shipping as a Strategic Tool

Start exclusive promotions from the start of the year that might led to higher trends in the MARCH-AUGUST period

Attack the April-June time period with discounts or new launches create pre-made, high-value gift bundles for the holiday season(NOVEMBER)

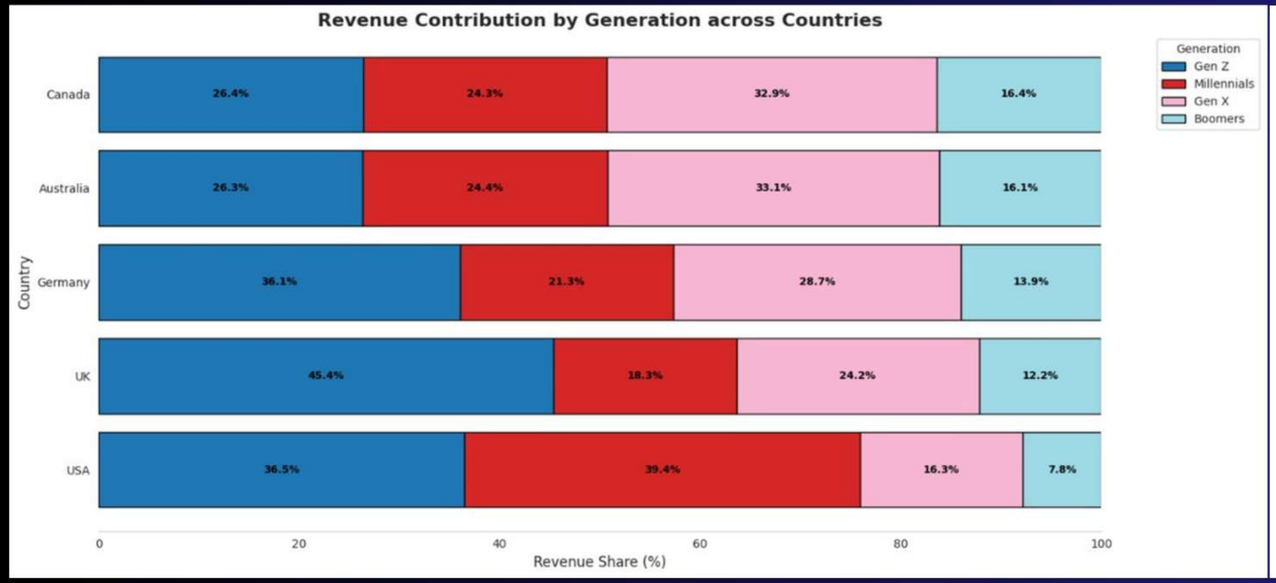
Introduce benefits specifically designed to incentivize a higher AOV for your Premium members

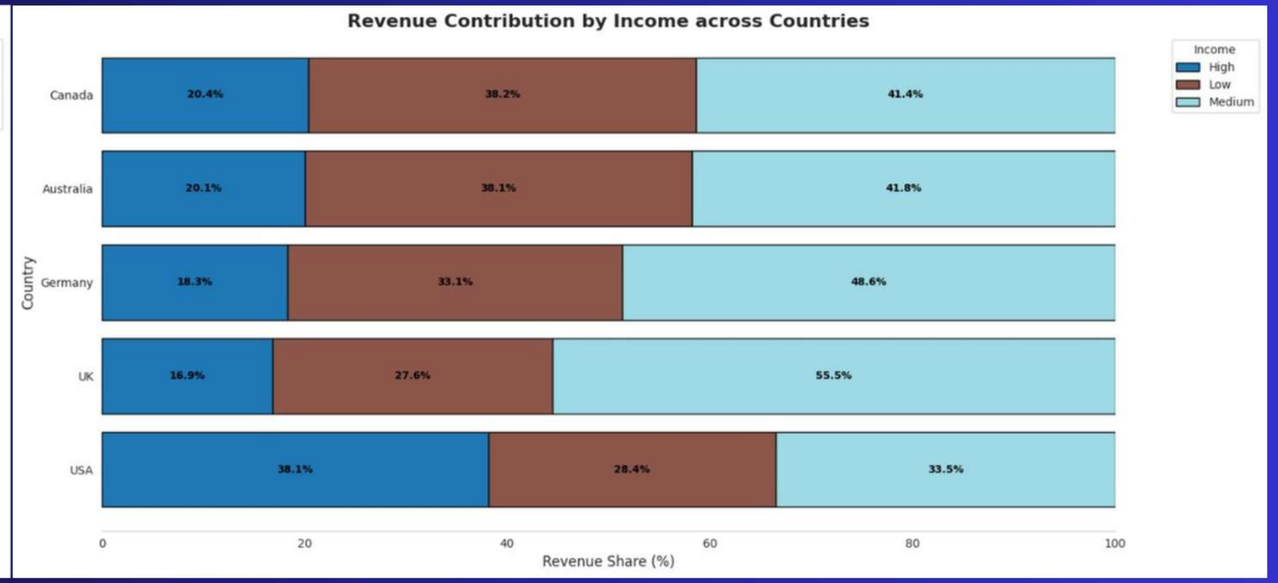
Regional Preferences by Product Type



Most states show a balanced share across categories (~18–25%).
Electronics consistently lead in many states (Ontario, England, Berlin).
Connecticut is unique with Grocery dominating (53.8%).

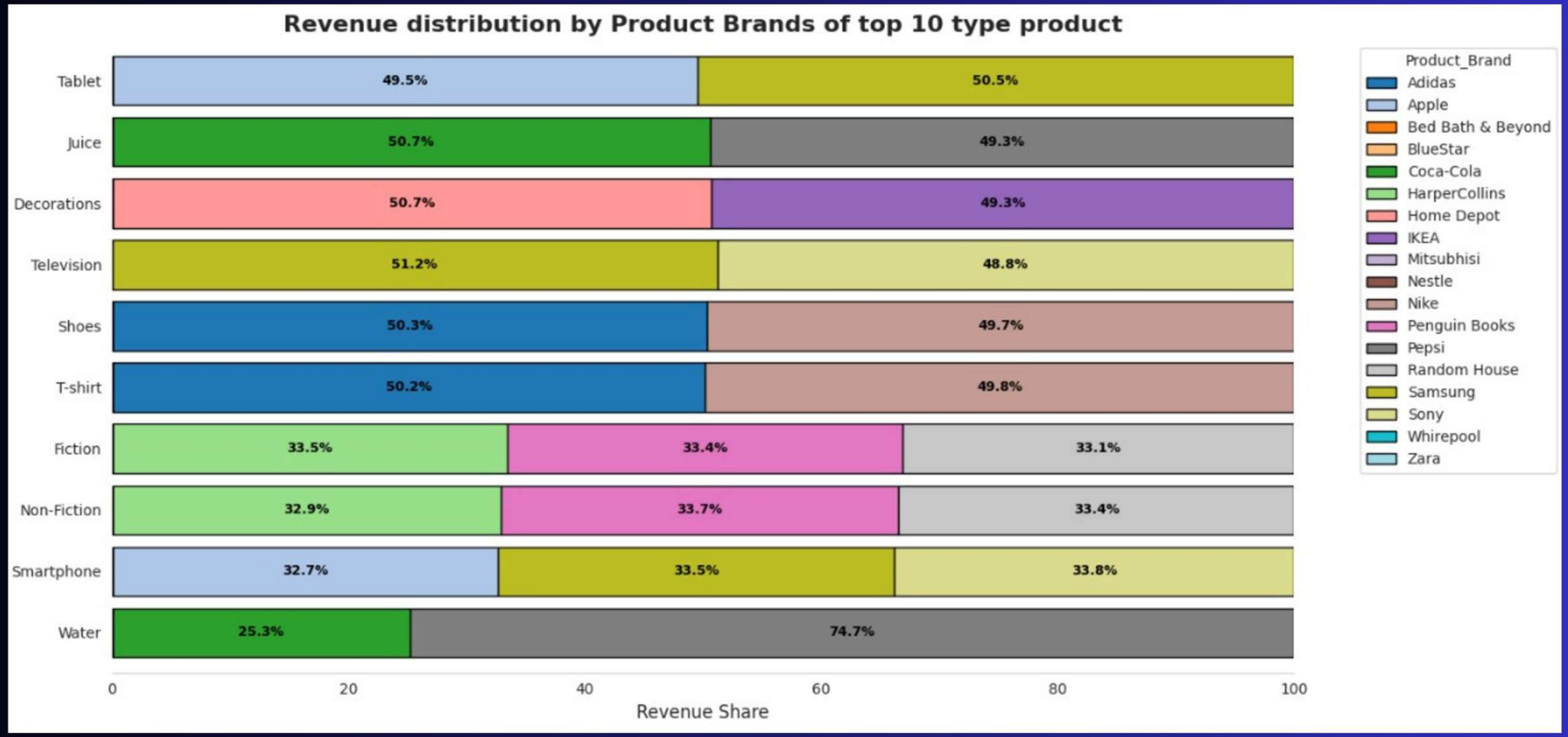
Cross-Country Revenue Patterns by Gender & Income



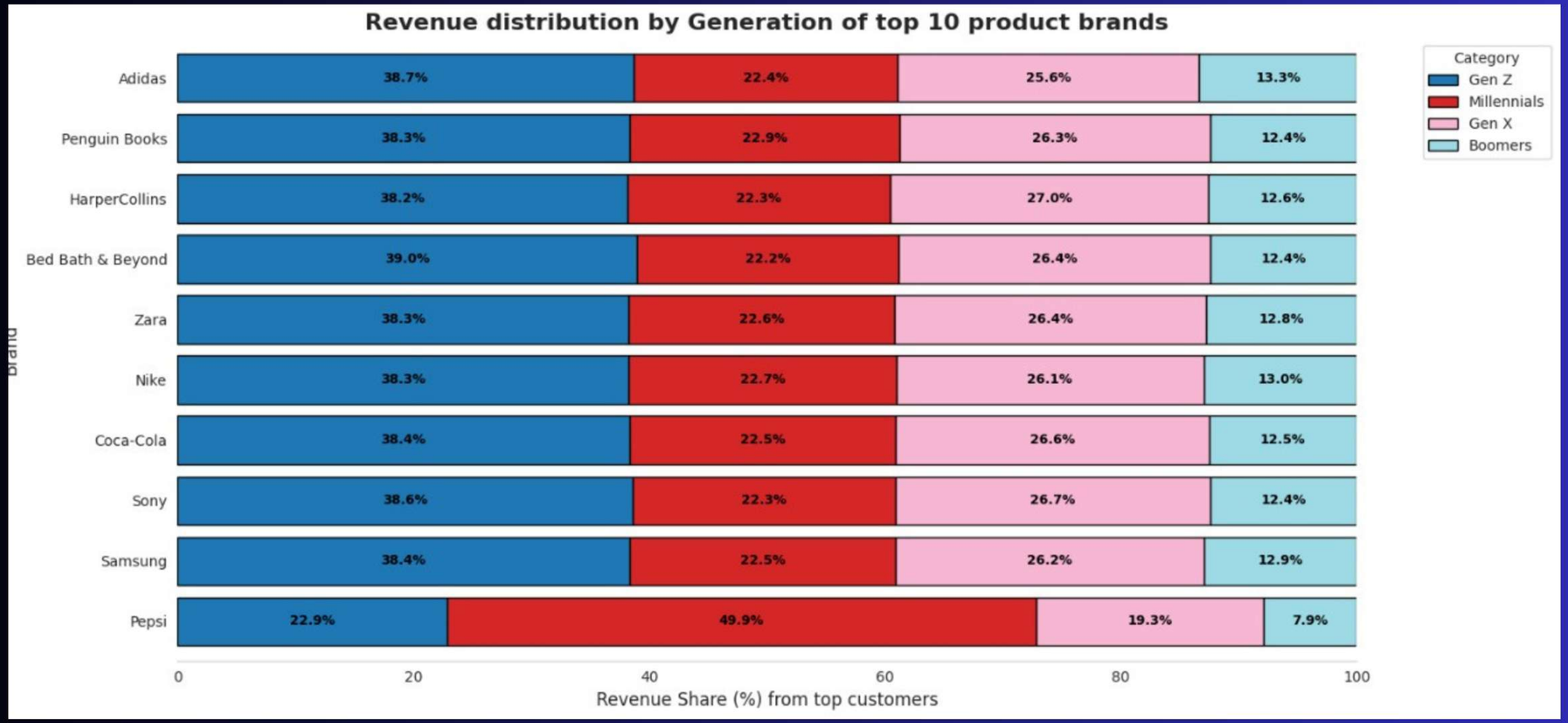


The distribution of revenue across generations varies by country.
The USA shows a higher revenue share from Millennials and Gen Z compared to other countries.
The UK has a higher revenue share from Gen Z.

The distribution of revenue across income levels varies by country.
The USA has a higher revenue share from the "High" income group.
The UK has a significantly higher revenue share from the "Medium" income group.



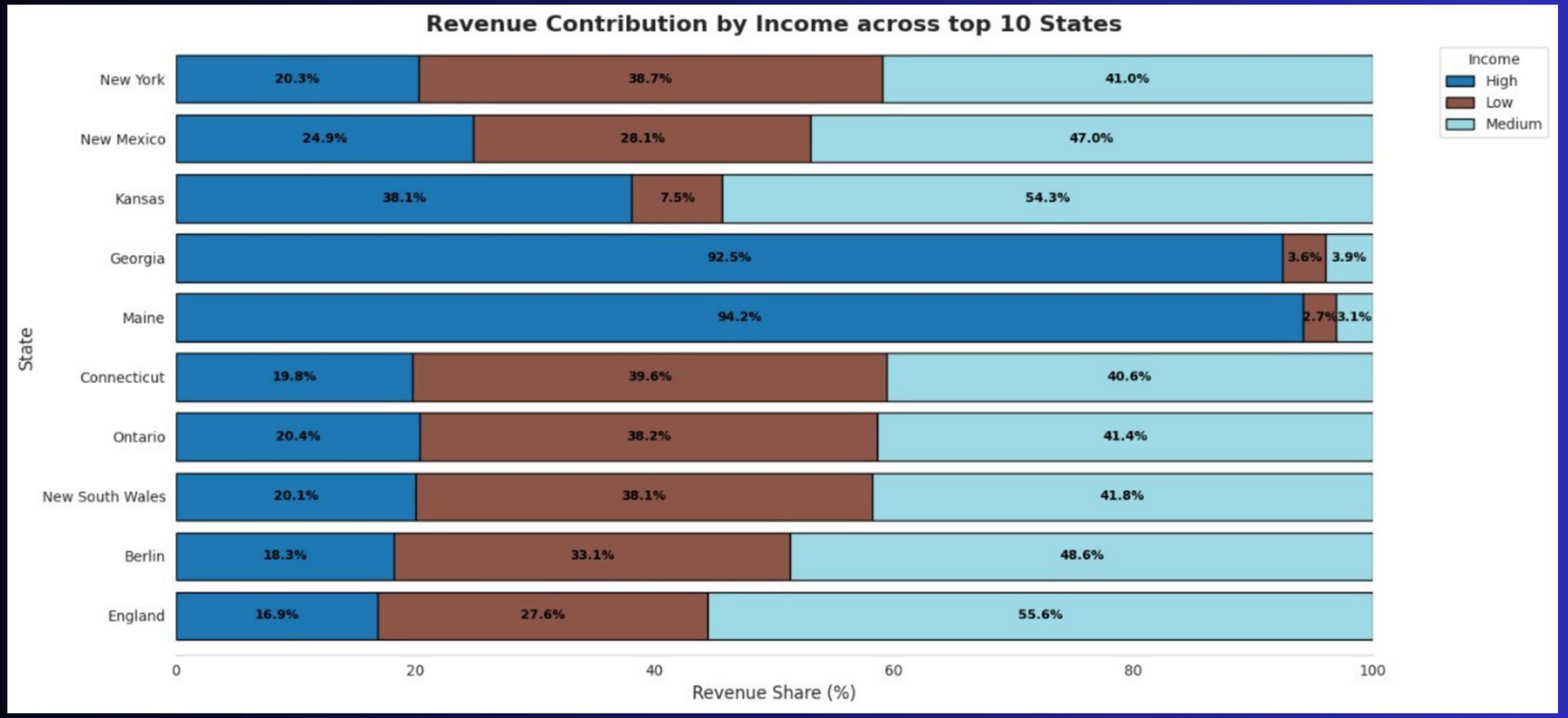
Most categories show balanced split (~50/50) across top brands.
Water is unique, with Pepsi dominating (74.7%).



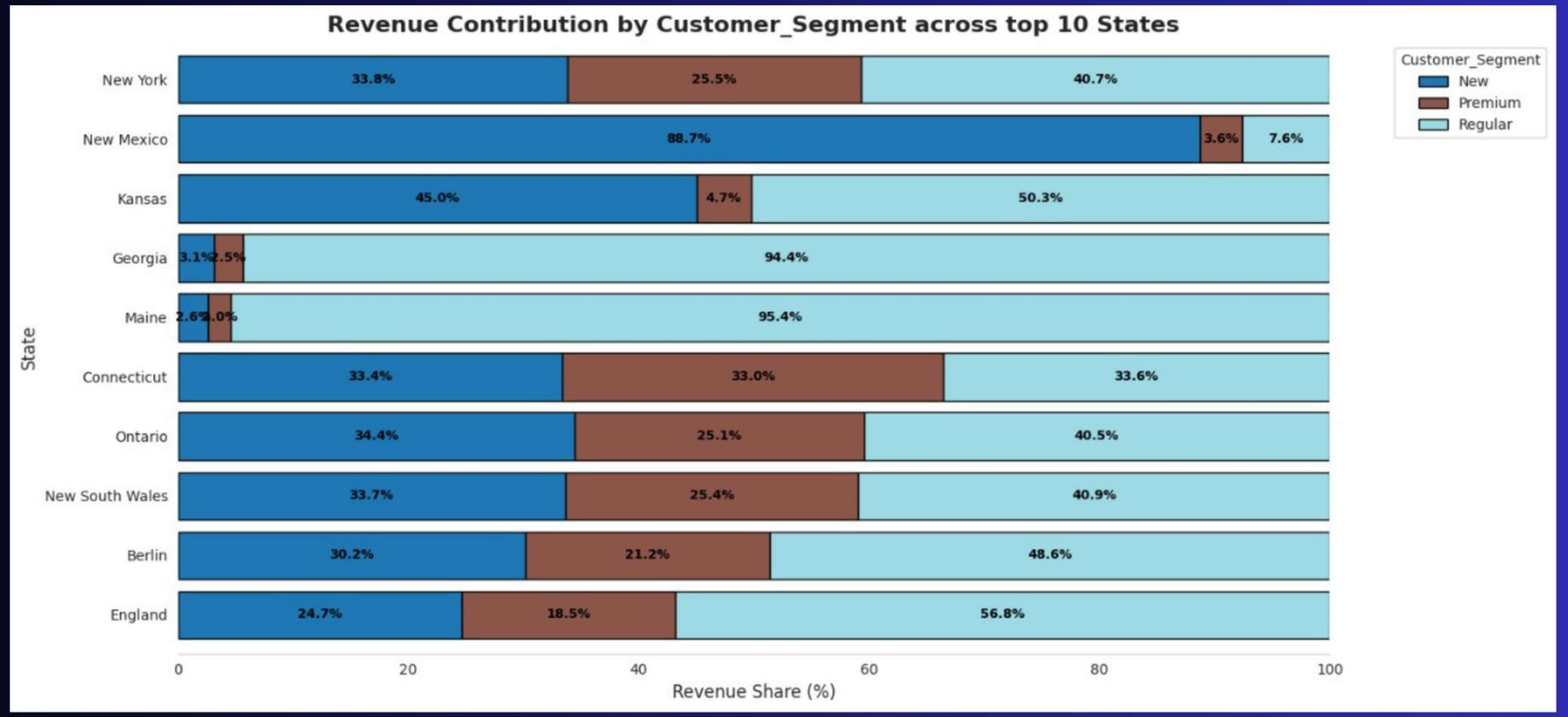
Most brands show Gen Z dominance (~38–39%).

Pepsi is an outlier, driven heavily by Millennials (49.9%).

Gen X (26%) and Boomers (12–13%) are relatively consistent across brands.



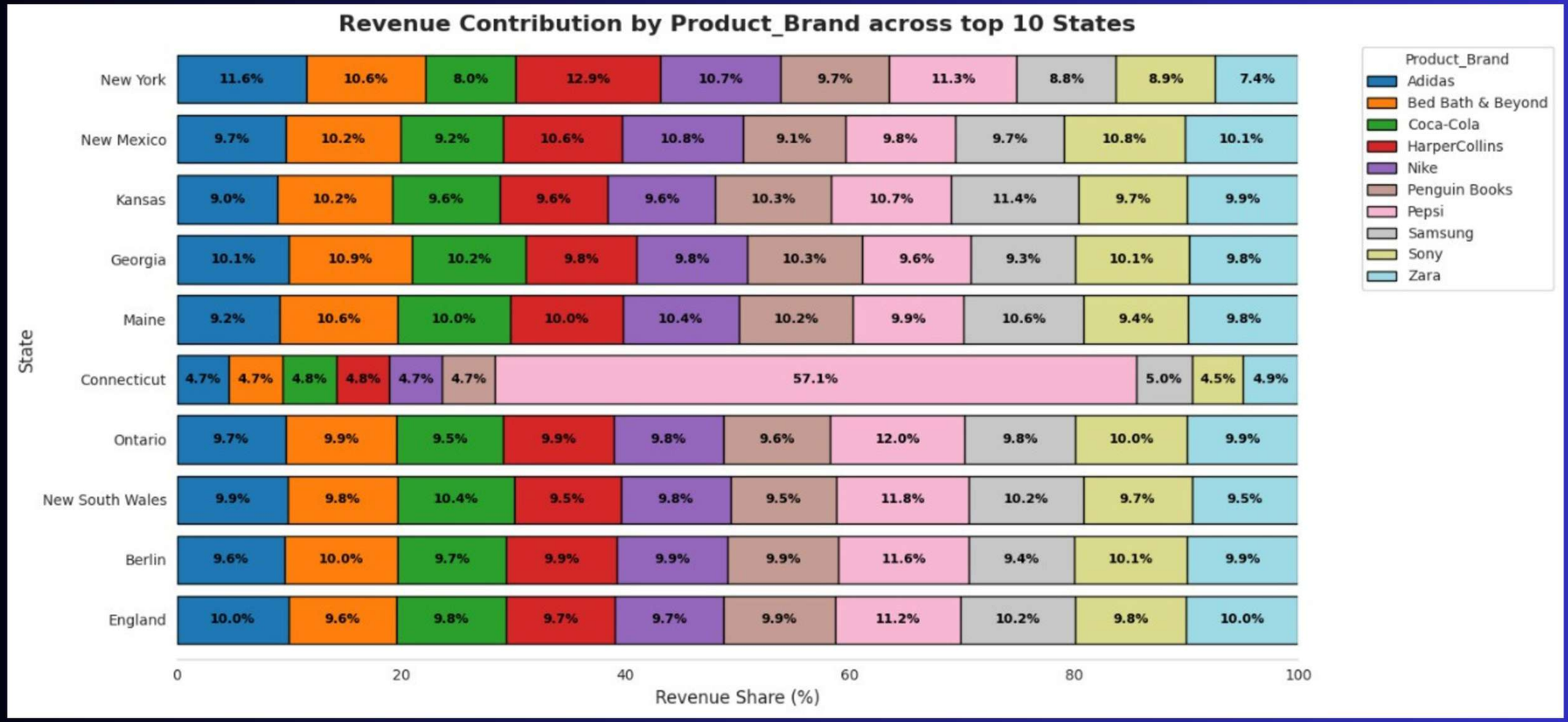
New York and Connecticut show a higher revenue share from the "High" income group.
New Mexico, Kansas, Georgia, and Maine show a very high revenue share from the "Low" income group.
England and Berlin show a higher revenue share from the "Medium" income group.



New Mexico, Georgia, and Maine show a very high revenue share from "Regular" customers.

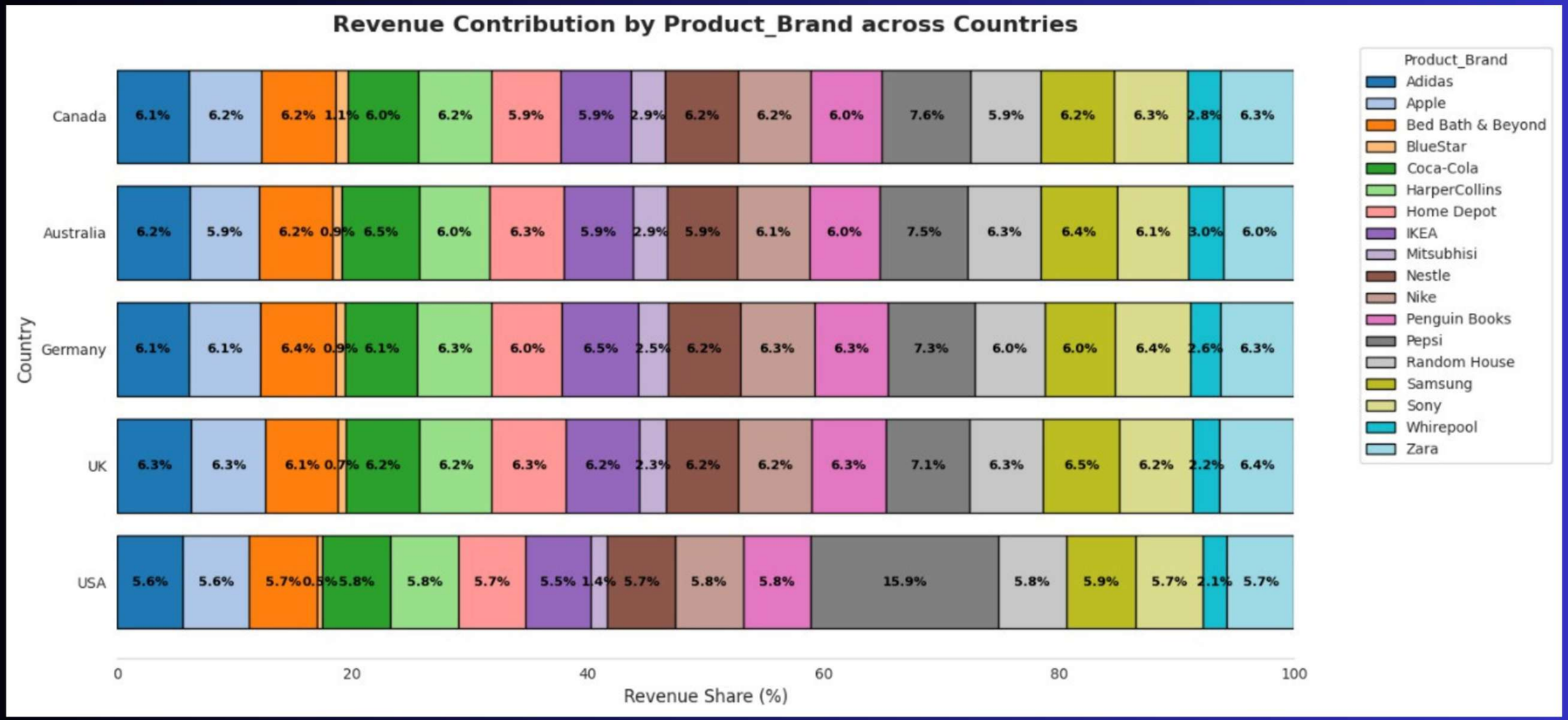
New York, Connecticut, Ontario, New South Wales, Berlin, and England show a more balanced distribution across customer segments.

Business Insight:



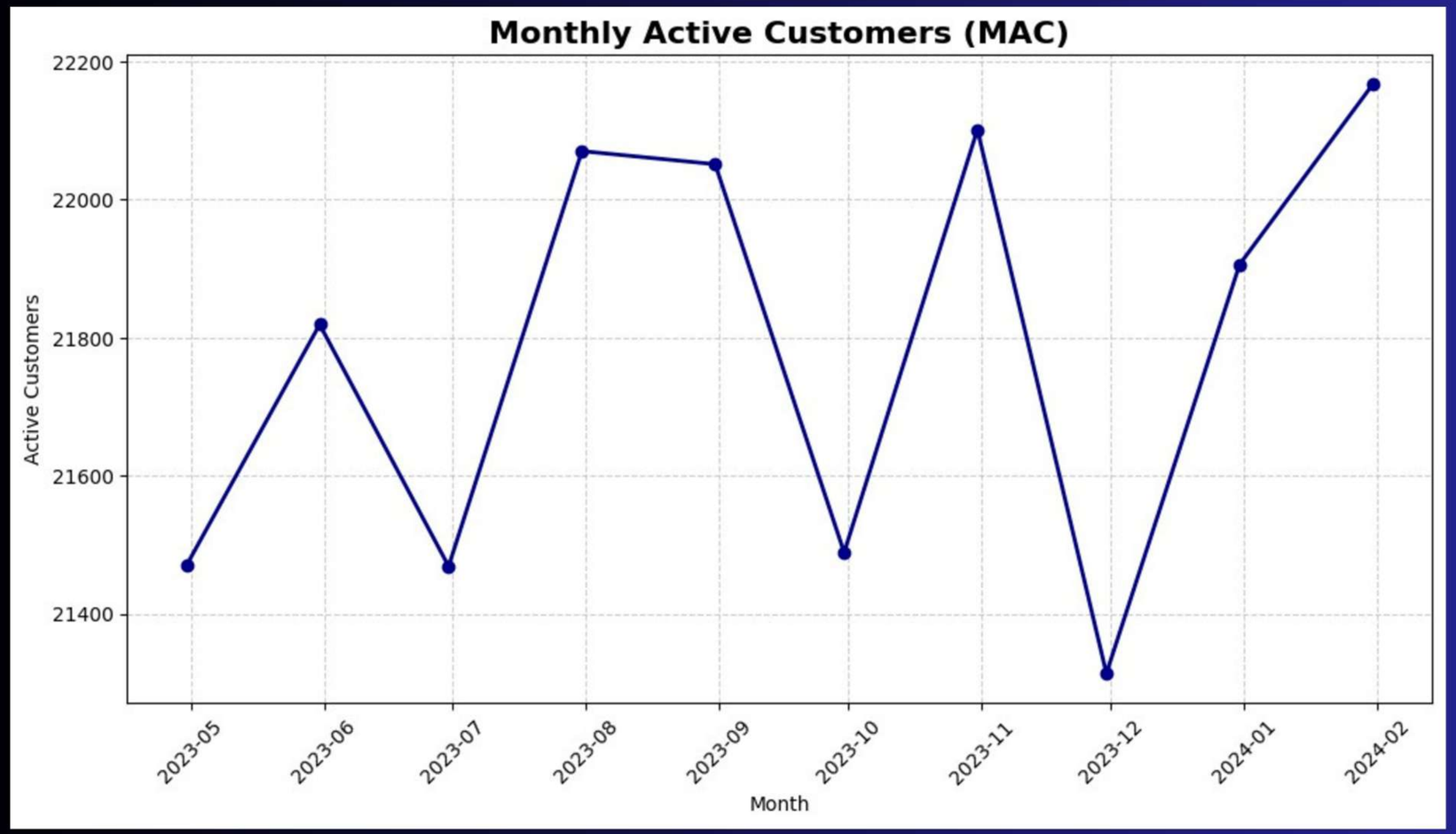
The revenue contribution of product brands is relatively consistent across the top 10 states.

Connecticut shows a significantly different distribution, with a much higher revenue share from the "Mitsubhisi" brand.



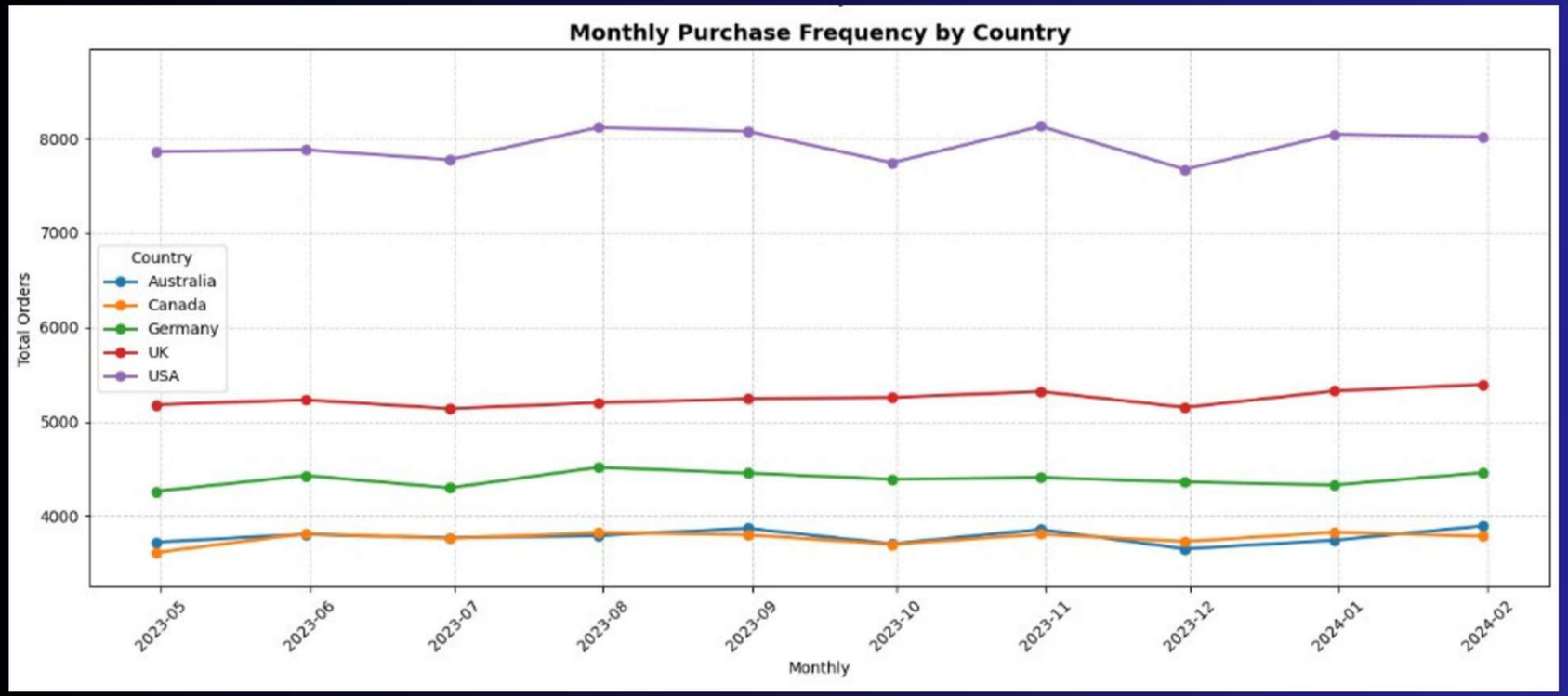
Pepsi shows a notably higher revenue share in the USA compared to other brands in that country and compared to Pepsi's performance in other countries.

Customer Engagement: Monthly Active Users



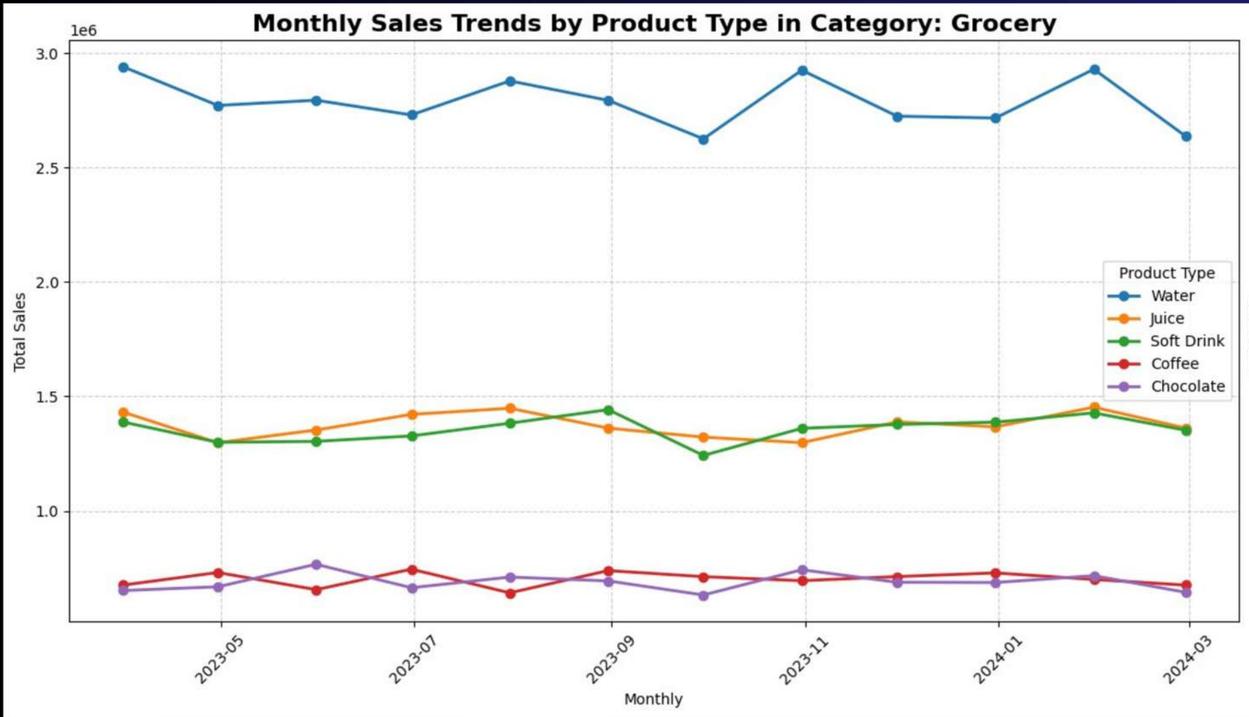
Shows fluctuations with an overall upward trend, peaking in early 2024 after seasonal dips (e.g., Dec 2023).

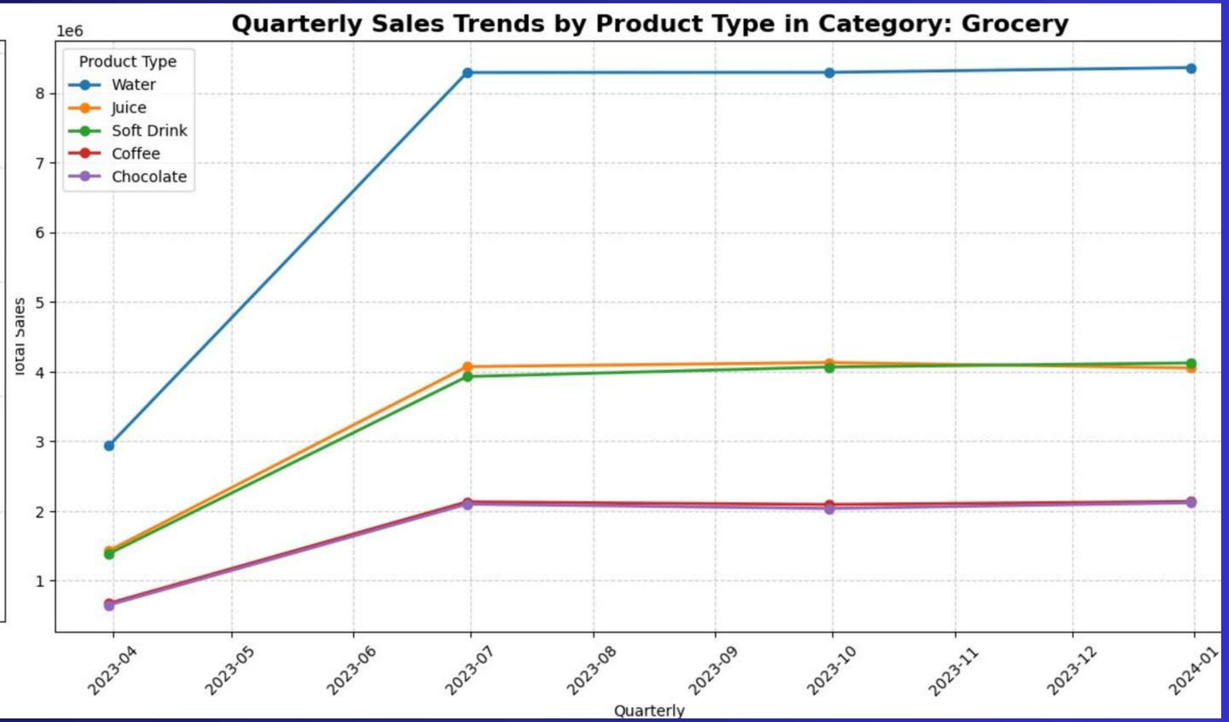
Geographic Trends in Customer Purchases



The USA consistently leads in purchase frequency, followed by the UK, while other countries (Germany, Canada, Australia) show moderate but stable activity with slight seasonal variations

Sales Dynamics of Grocery Product Types (Monthly vs Quarterly)



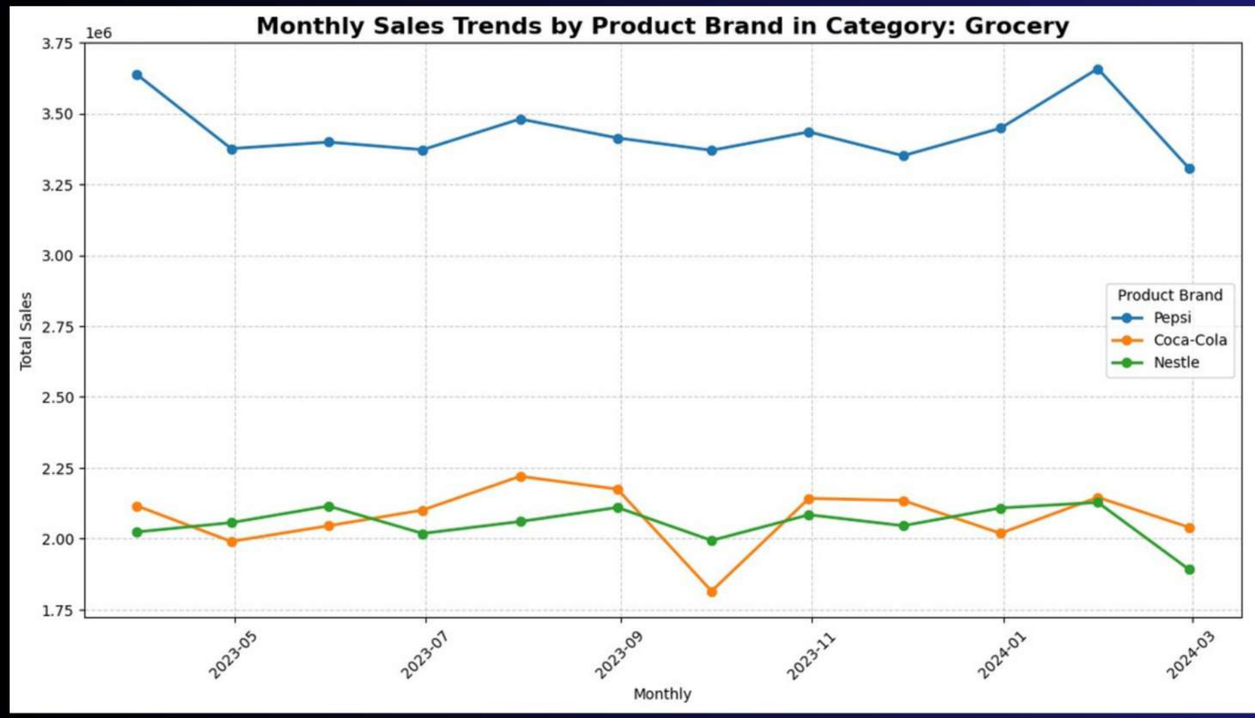


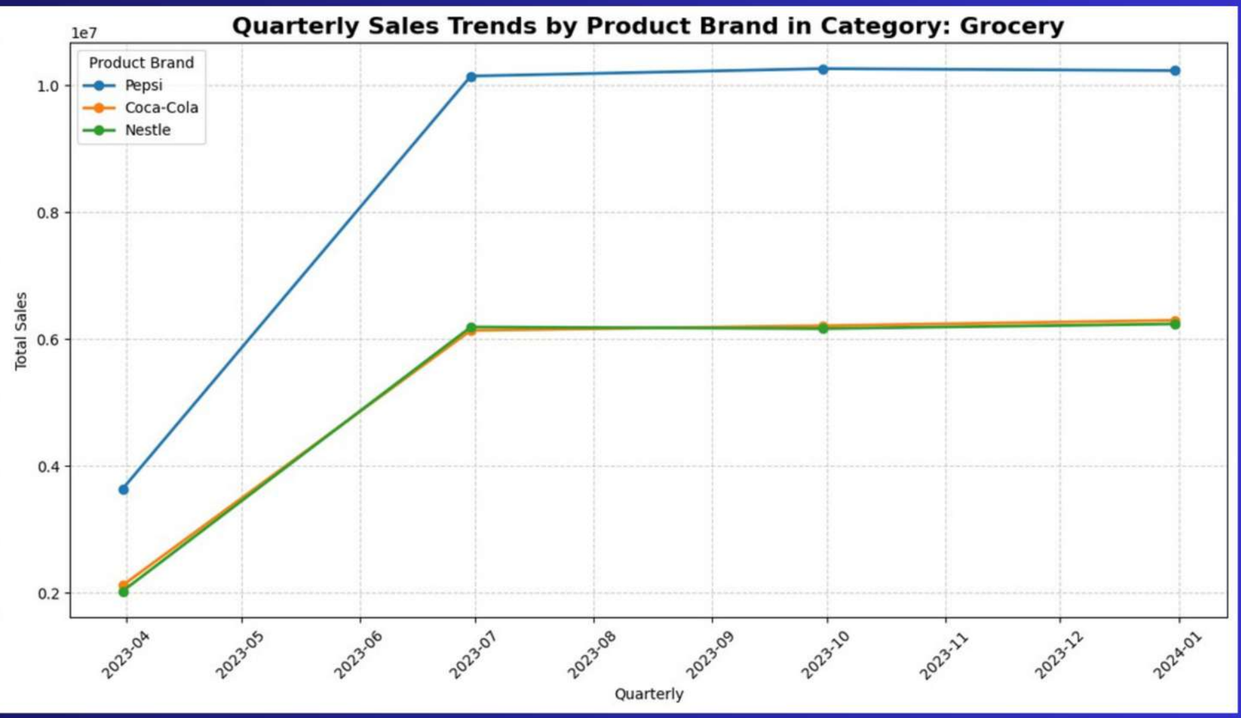
Water is the dominant product, consistently leading sales across both monthly and quarterly trends.

Juice and Soft Drinks show steady demand with mild seasonality.

Coffee and Chocolate remain niche with smaller, stable contributions.

Grocery Brand Dynamics Across Months & Quarters

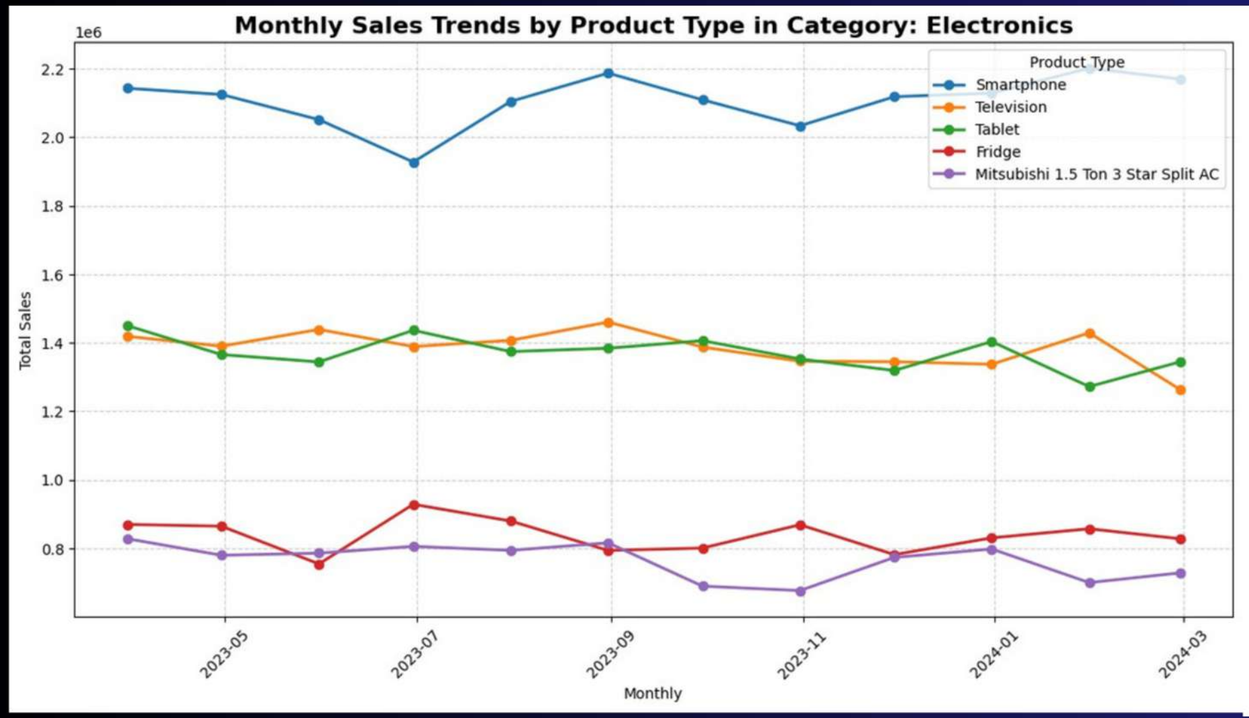


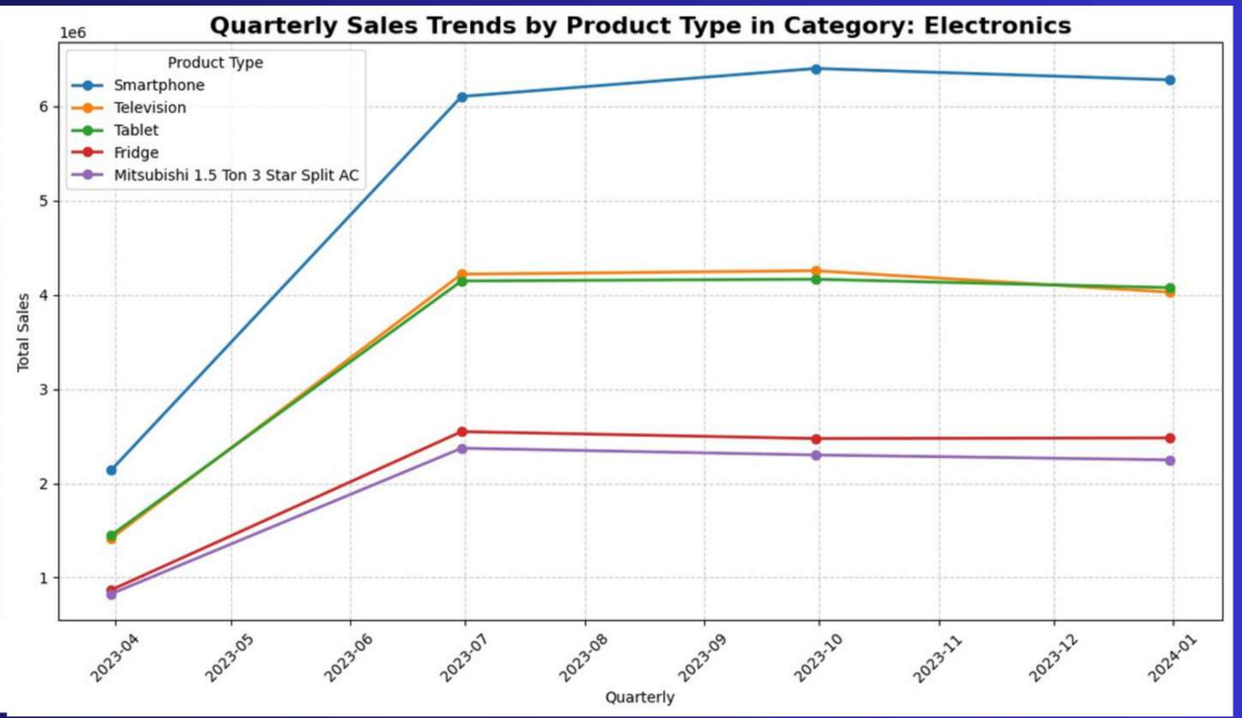


Pepsi consistently leads with strong sales, while Coca-Cola and Nestle show moderate but stable performance.

Minor fluctuations exist, but no major growth shifts across brands.

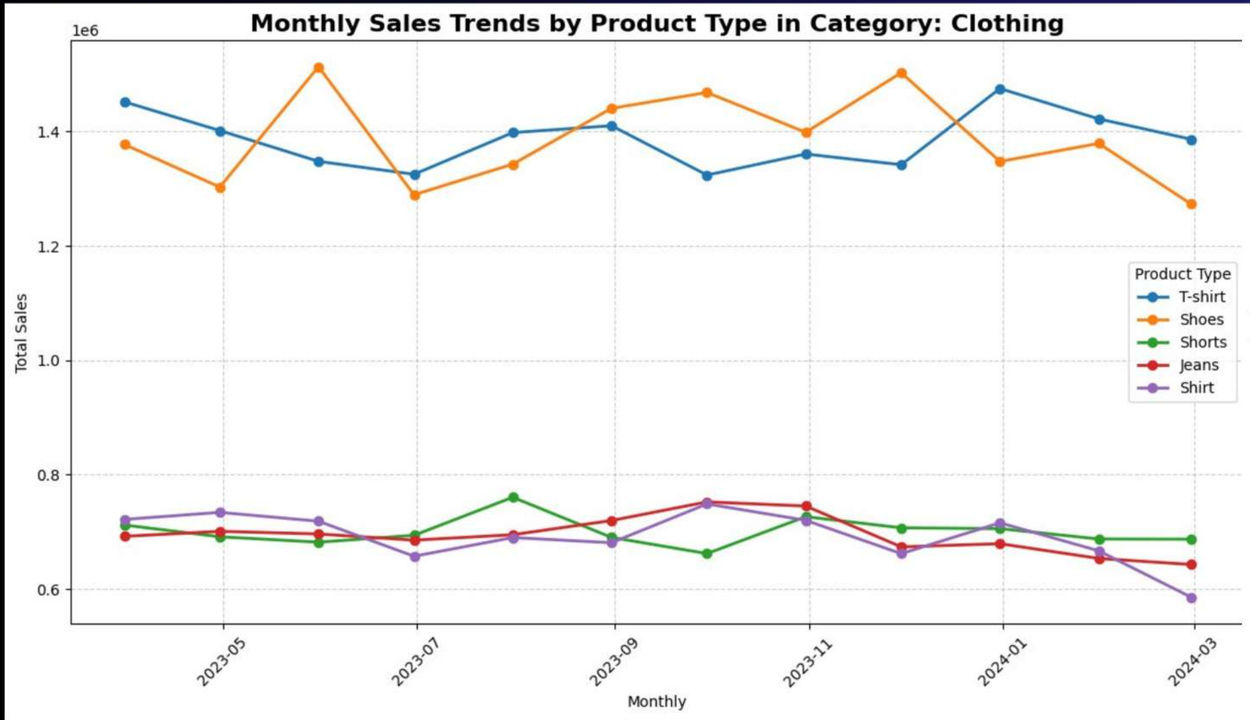
Electronics Product Type Performance Across Months & Quarters

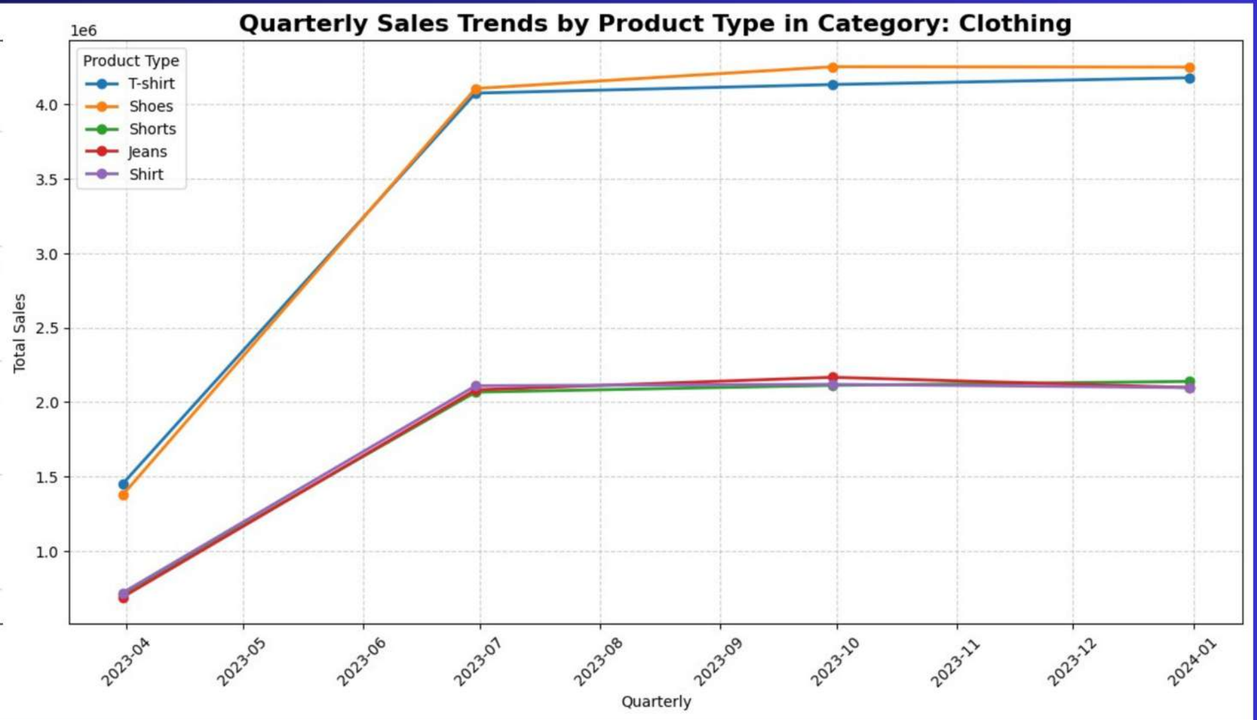




Smartphones consistently dominate sales, showing stable growth. Televisions and tablets remain mid-tier but with slight downward trends, while fridges and ACs have the lowest and most volatile sales.

Monthly & Quarterly Sales Trends – Clothing Product Types

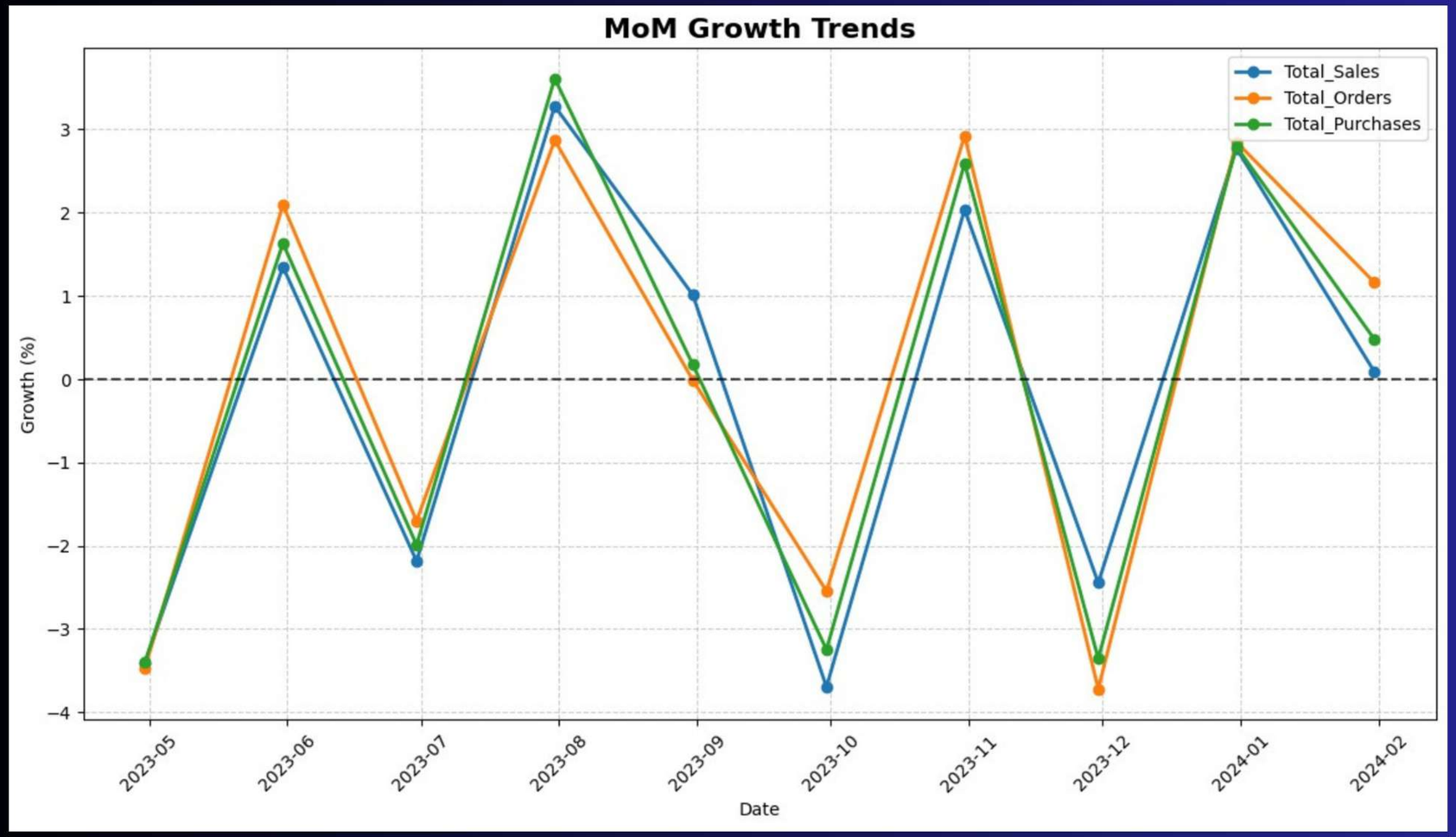




Quarterly: T-shirts and Shoes drive the bulk of sales (~4M each), showing steady growth and stability. Shorts, Jeans, and Shirts plateau around ~2M.

Monthly: T-shirts and Shoes show consistent high sales with slight fluctuations, while Jeans, Shorts, and Shirts remain secondary with modest and stable demand.

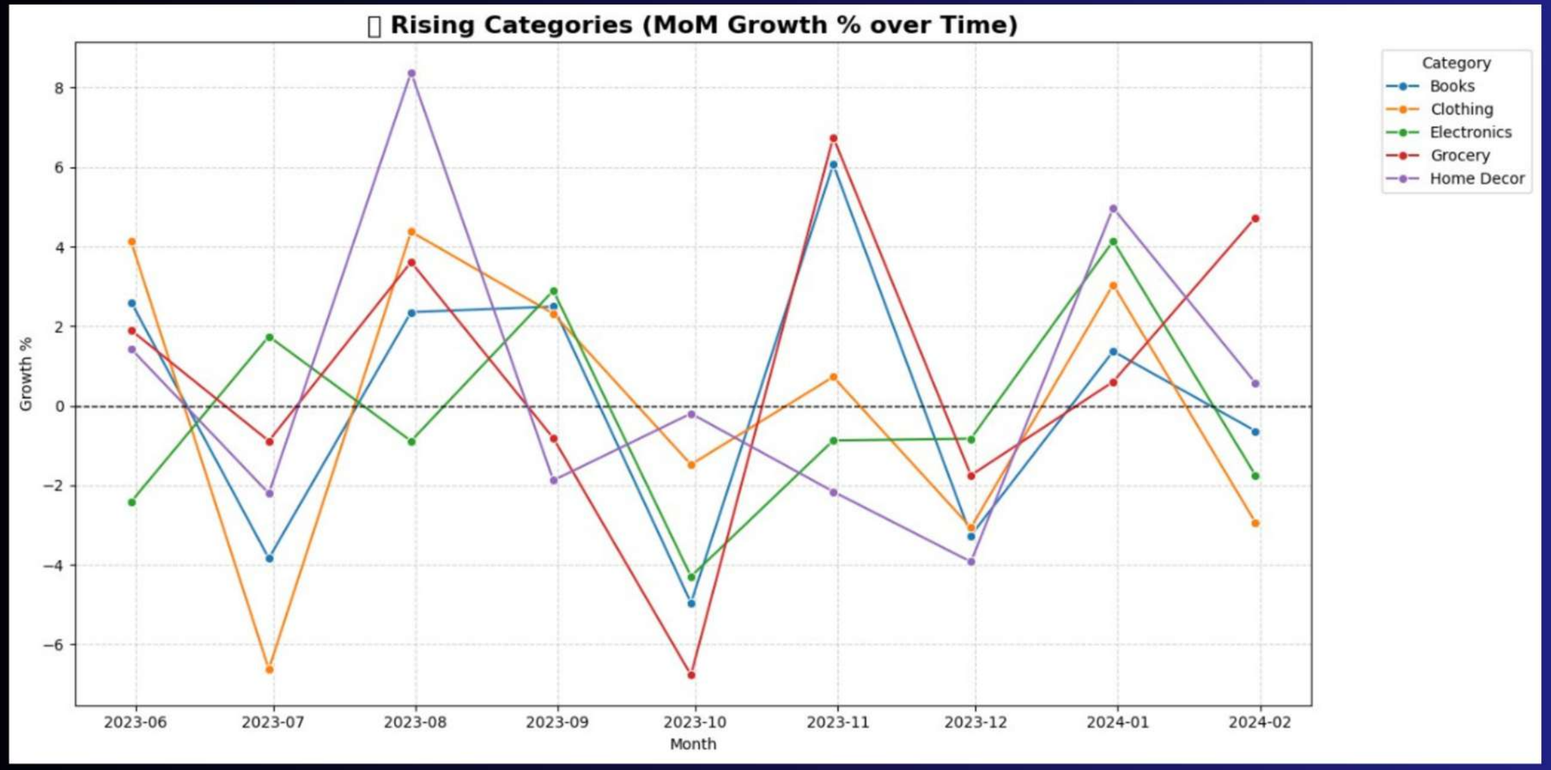
MoM Growth Insights: Tracking Sales, Orders & Purchases



The MoM growth trends show a highly cyclical pattern, with alternating positive and negative growth.

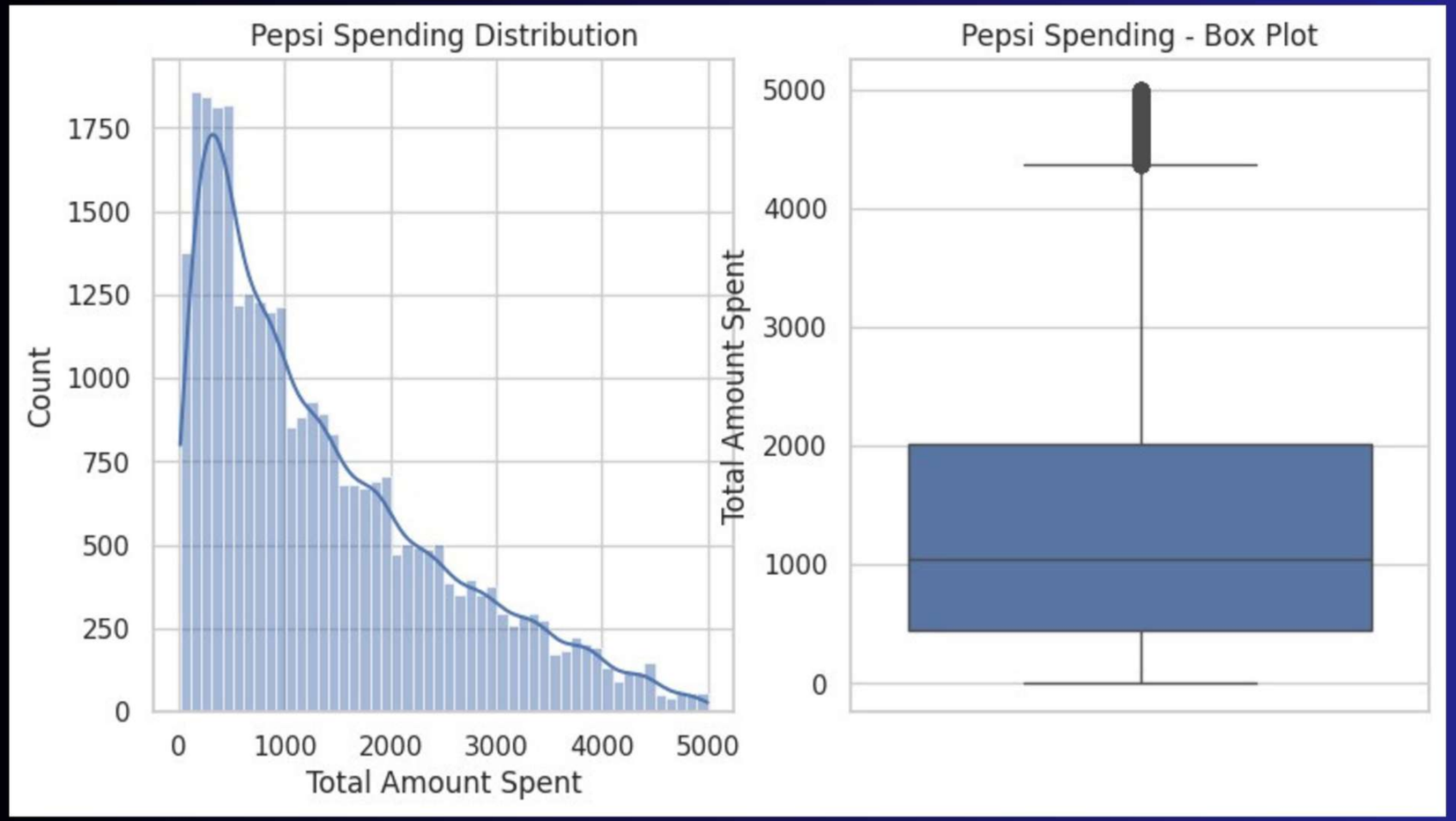
Peaks occur roughly every 2 months (e.g., Aug, Nov, Jan), followed by sharp declines. Sales, orders, and purchases move closely together, indicating demand and supply alignment.

Emerging Categories (MoM % Growth)



Growth is highly volatile across all categories, with Clothing, Grocery, and Home Décor driving sharp spikes, while Books and Electronics fluctuate moderately.

Outlier Analysis(Pepsi)



Pepsi is the #1 revenue brand, generating $27.6M (10% of total revenue).
Pepsi attracts high-value customers, with LTV 33% higher ($5,747 vs $4,332).

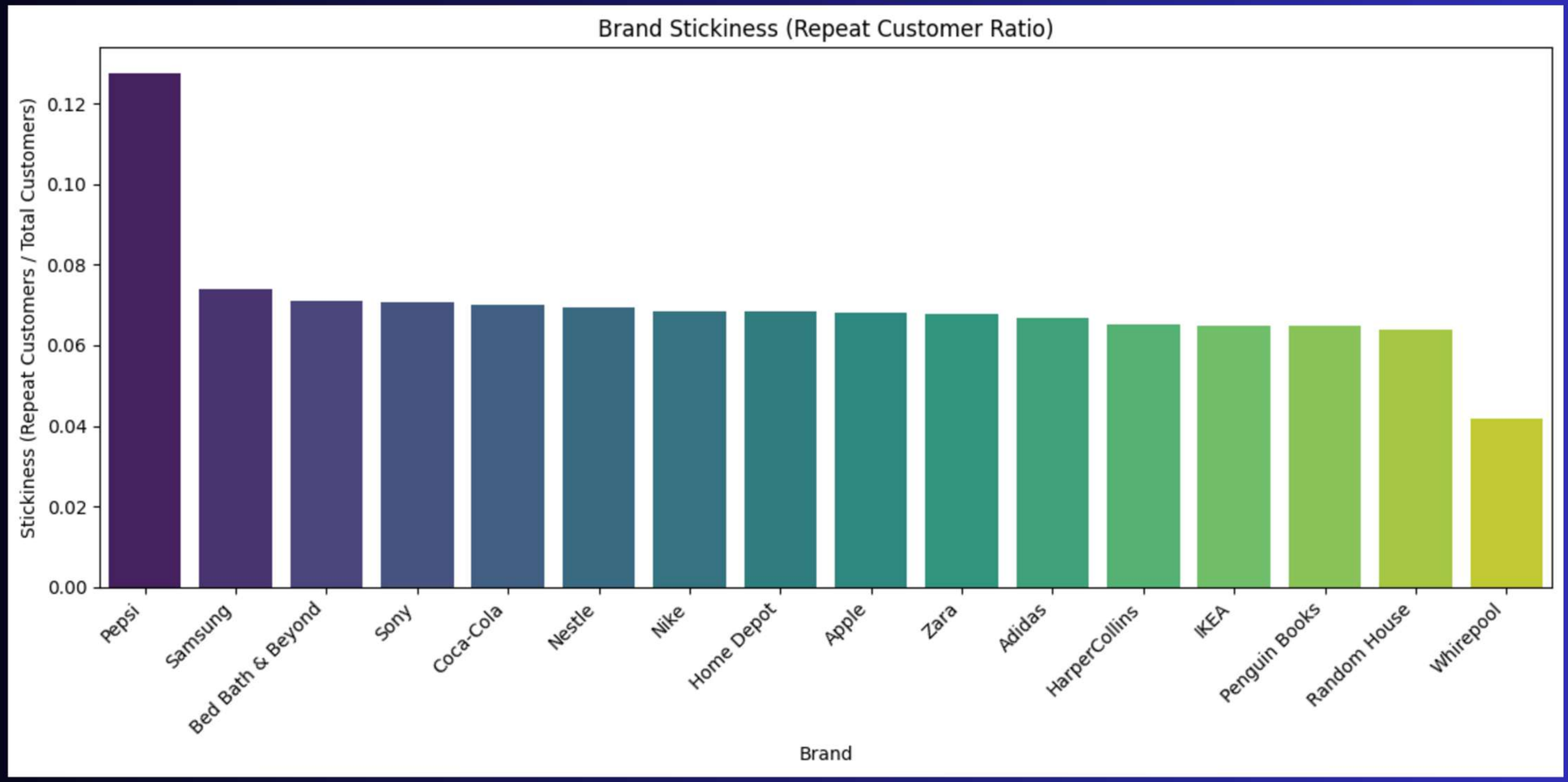
Pepsi is a staple purchase, appearing in ~30K grocery baskets with core essentials.

Pepsi has cross-generational appeal, strongest with Millennials (49.4%) and Gen Z (23.8%), but also reaches older groups.

Pepsi buyers are very loyal and also high-spending customers..

Pepsi shows geographic strength, with the US leading and solid presence in the UK, Germany, Canada, and Australia.

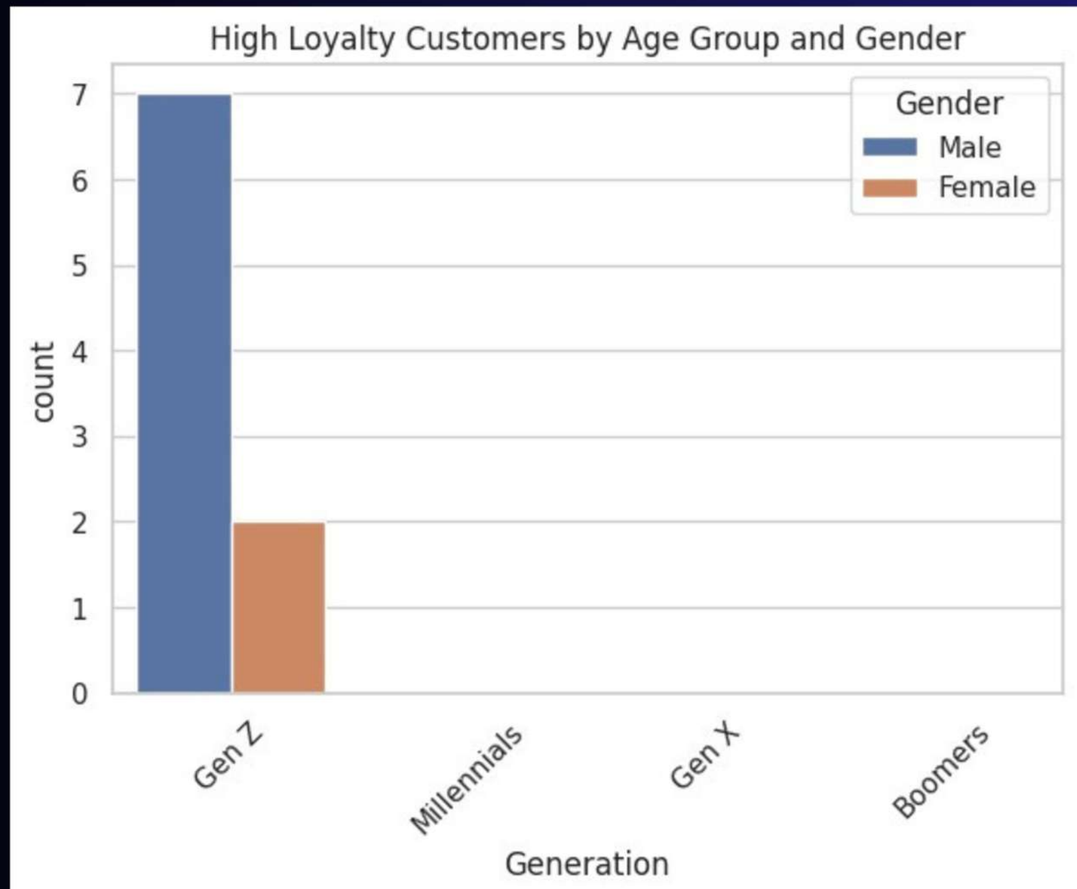
Customer Stickiness by Brand

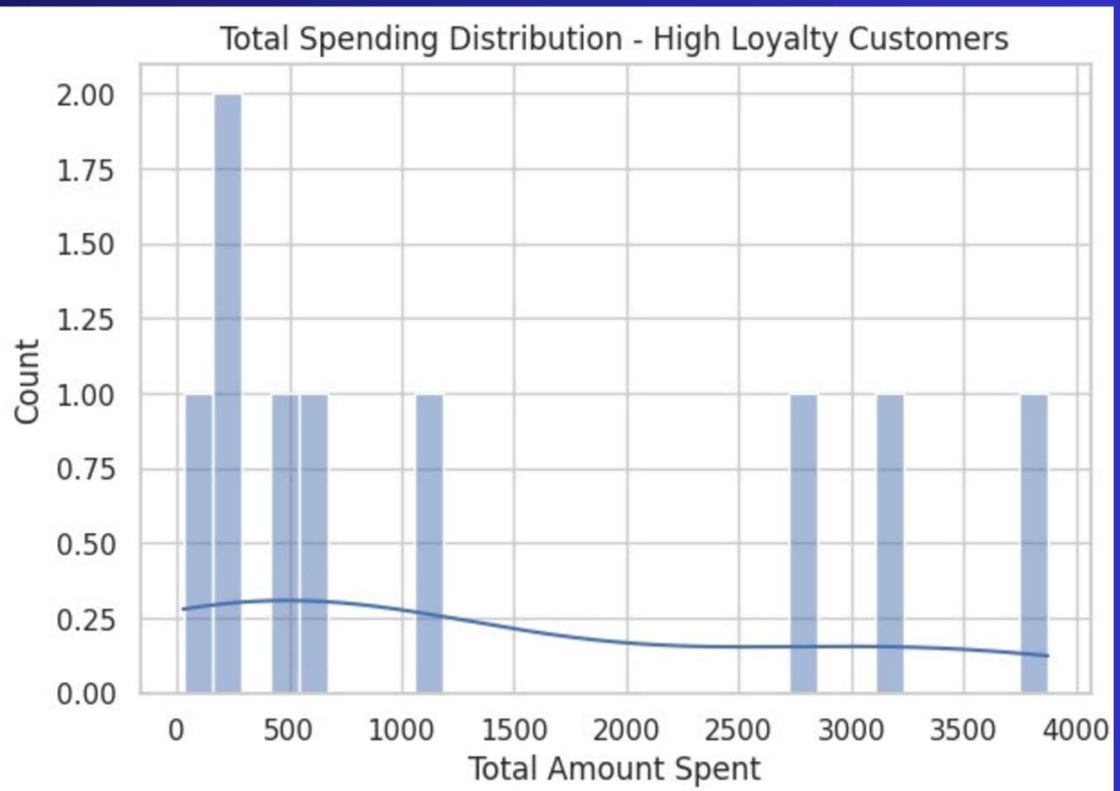


Apple's stickiness ratio is approximately 5 to 6 times higher than many other well-known and successful brands like Nike, Coca-Cola, and Sony.

A brand like Coca-Cola has near-universal recognition and massive market share but a moderate stickiness ratio. This suggests that while everyone buys Coke, they also buy many other competing beverages

Customer Loyalty





High Loyalty Customer Insights

Mostly Gen Z, skewed male.

Medium income dominates; high income = top spenders.

Business Recommendations

User-based

Drive Gen Z engagement in apparel, tech, and lifestyle, and run Millennial-focused campaigns for Pepsi (Millennials = 49.9%).

Maintain retention for Gen X & Boomers, while tailoring state-level campaigns to dominant cohorts (e.g., Millennials in New Mexico, Gen X in Maine).

Move mid-frequency buyers (3–5 purchases) into high-frequency with rewards, discounts, or VIP perks, and encourage sub-₹1000 buyers to spend more.

Focus retention on top 25% spenders and address pain points of low-rating (1–2) customers.

Build income-specific strategies: loyalty programs for medium-income, cashback/subscriptions for low-income, and exclusivity + concierge for high-income.

Strengthen the Regular customer base in states where they dominate (e.g., New Mexico, Georgia, Maine).

Business Recommendations

Location-Based

Reduce pending + processing orders to prevent revenue loss, especially in top-performing cities.

Address underperformance in Tier 1 markets like New York, Washington, and Dallas through localized strategies and competitive benchmarking.

Run tailored digital ad campaigns using local identities/values, and A/B test against national campaigns.

Optimize logistics in Hawaii, Idaho, and Oklahoma with better shipping partnerships. Prioritize high-growth states like Colorado and Washington due to stronger economies.

Localize marketing, payments, and support for international markets (e.g., Australia).

Adapt strategies to dominant income + generational cohorts in each region/state.

Investigate Connecticut’s preference for Mitsubishi to uncover demand drivers.

Business Recommendations

Time-Series Analysis

Launch pre-made, high-value gift bundles in November (holiday season).

Start exclusive promotions early in the year to build momentum for higher sales during March–August.

Run short, sharp 24–48 hour flash sales in slow weeks to create urgency and boost revenue.

Product Brands

Whirlpool should push discounts more aggressively and invest in modern, sleek designs to appeal to younger buyers.

Compete with brand-driven campaigns in markets where top brands have a balanced split (~50/50).

For Pepsi Water, leverage dominance (74.7% share) with premium pricing and loyalty pushes.

Investigate Pepsi’s strong performance in the USA to replicate success in other regions or product categories.

Business Recommendations

Product Category

Promote Electronics more aggressively in tech-oriented states.

Leverage Grocery dominance in Connecticut with bundled offers.

Use broad-based campaigns in balanced markets where no single category dominates.

Offer affordable subscriptions for essential groceries to lock in price-sensitive customers.

Create curated premium bundles for high-income customers.

Launch pre-made gift bundles during festive and holiday seasons.

Thank You

Sukrit Mukherjee

https://github.com/SukritM2004

Debayudh Khag

https://github.com/Debayudh337

Anindita Saha

https://github.com/aninditasaha6041

Soumyajit Paul

http://github.com/SoumyajitPaul-git