

Customer_Behavior_Analysis_<Piyali_Mujumdar_&_Pranav_Jadhav>

Objective

In this case study, you will be working on E-commerce Customer Behavior Analysis using Apache Spark, a powerful distributed computing framework designed for big data processing. This assignment aims to give you hands-on experience in analyzing large-scale e-commerce datasets using PySpark. You will apply techniques learned in data analytics to clean, transform, and explore customer behavior data, drawing meaningful insights to support business decision-making. Apart from understanding how big data tools can optimize performance on a single machine and across clusters, you will develop a structured approach to analyzing customer segmentation, purchase patterns, and behavioral trends.

Loading the Datasets

In this code it initializes a Spark session, loads three CSV datasets into PySpark DataFrames, merges purchase data with survey responses using a common ID, and displays the first rows of the merged dataset.

1. Data Preparation

We loaded and merged the datasets (Amazon purchases, survey, fields) using PySpark. Data preparation ensured consistency by aligning schema, cleaning column names, and handling missing values. This step structured the dataset for further processing and analysis.

2. Data Cleaning

2.1 Handling Missing values

We identified missing values across columns and applied different strategies: dropping rows with critical nulls, filling demographic or categorical gaps with placeholders, and inputting numeric fields with 0 where appropriate. This reduced inconsistencies and improved dataset completeness.

We found some titles and categories of the product were missing, so made an assumption of none of the product was ordered from the account

Also after removal of missing values we were left with 603 null values left with the category columns, since it was negligible compared to the actual dataset size so we removed those rows from the dataset.

2.2 Feature Engineering

Perform feature engineering on the dataset to extract relevant/ create new features as required and map specific data types.

- This step focused on **feature engineering** to prepare the dataset for deeper analysis.

We first converted the "Order Date" column into a proper DateType, ensuring time-based calculations were accurate. From this, we extracted **year**, **month**, and **day** to study purchase patterns over time. Then, numeric fields like "Purchase Price Per Unit" and "Quantity" were cast into correct data types for precise computation.

Finally, we engineered a **new feature**, **Total_Value**, by multiplying unit price and quantity, representing the actual revenue per transaction. This enriched dataset provided a consistent and structured base for all further EDA and modeling.

- This step encoded categorical demographic fields into numeric values for analysis. Income ranges were mapped from low to high brackets, while gender categories were assigned codes. These mappings created new numeric columns (Income_Num, Gender_Num), enabling easier statistical analysis, clustering, and machine learning without losing demographic meaning.

2.3 Data Cleaning

Handle data cleaning techniques such as data duplication, dropping unnecessary values etc.

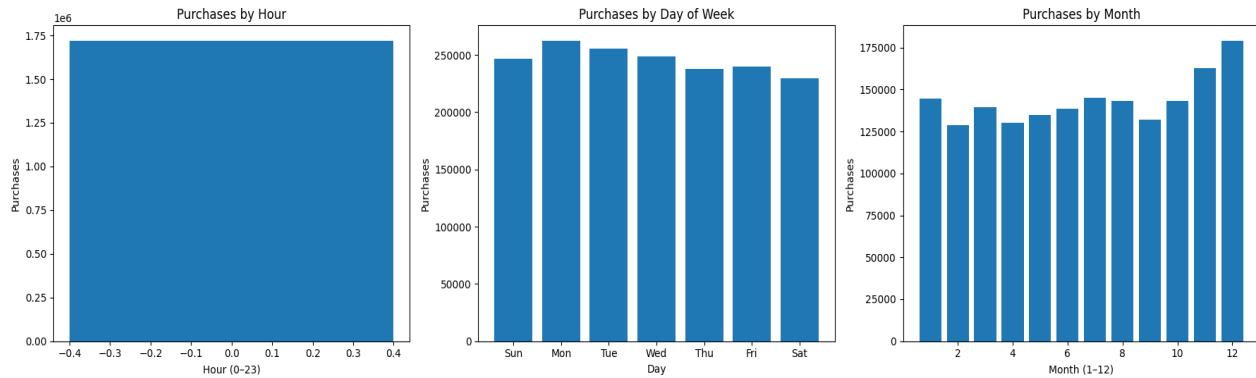
This code detects and removes duplicate records from the merged dataset. First, it calculates the number of duplicates by comparing the row count before and after applying dropDuplicates(). Then, duplicates are removed, and a second check confirms that no duplicate entries remain, ensuring cleaner and more reliable data.

3. Exploratory Data Analysis

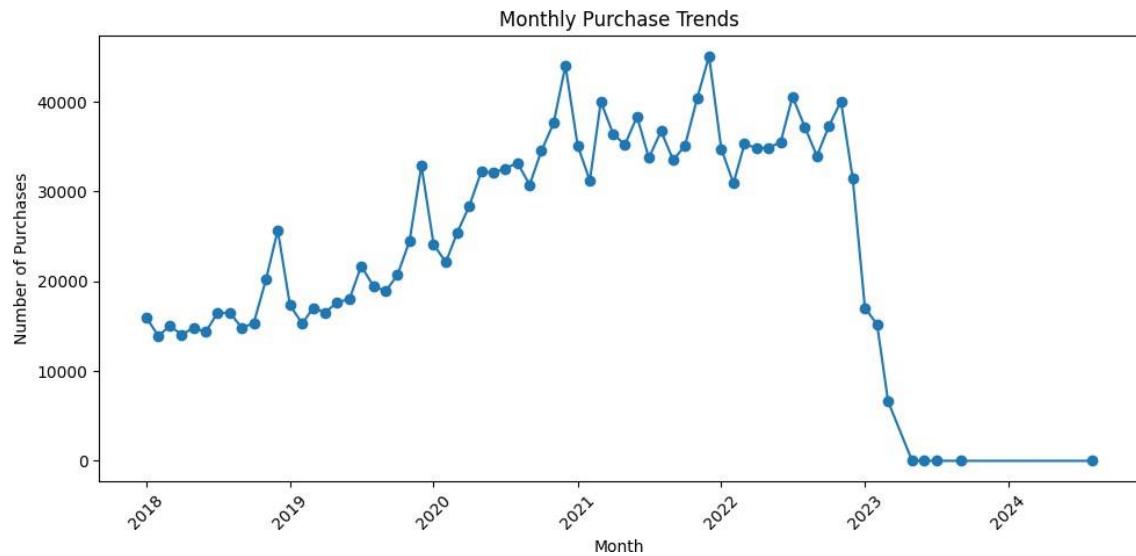
3.1 Analyse purchases by hour, day and month

We analyzed purchase frequency over time by extracting hour, day of the week, and month from order timestamps. This revealed customer shopping trends, peaks, and seasonality patterns. Bar plots and time-series plots visualized the distribution

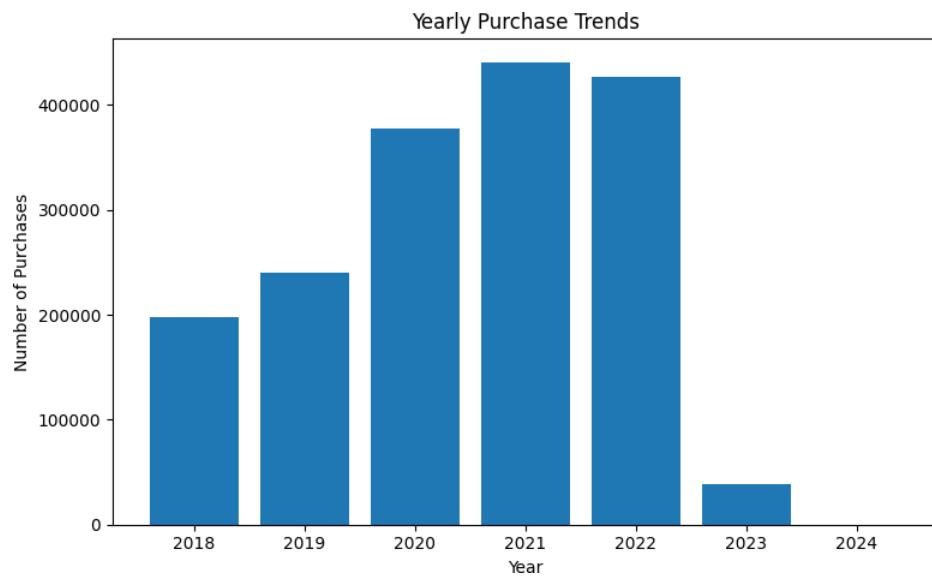
Purchase Distribution by Hour, Day, and Month



Monthly Purchase Trends

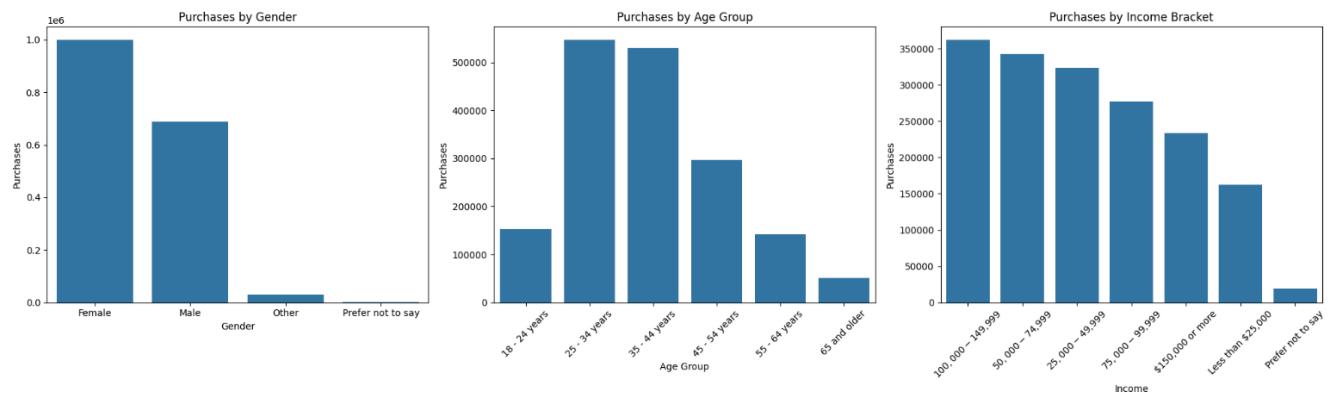


Yearly Purchase Trends



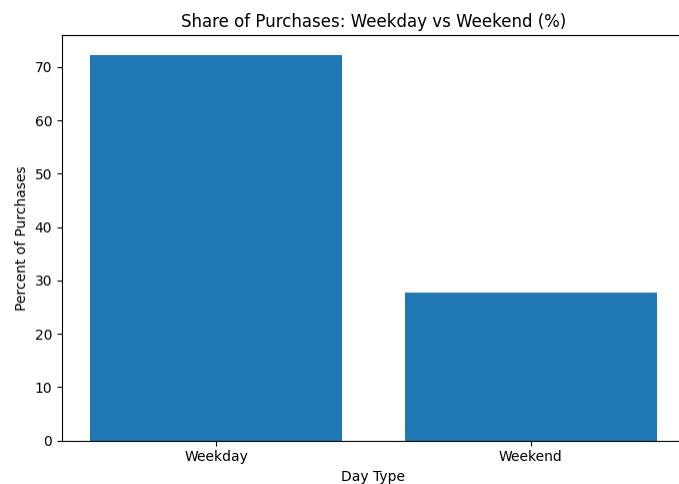
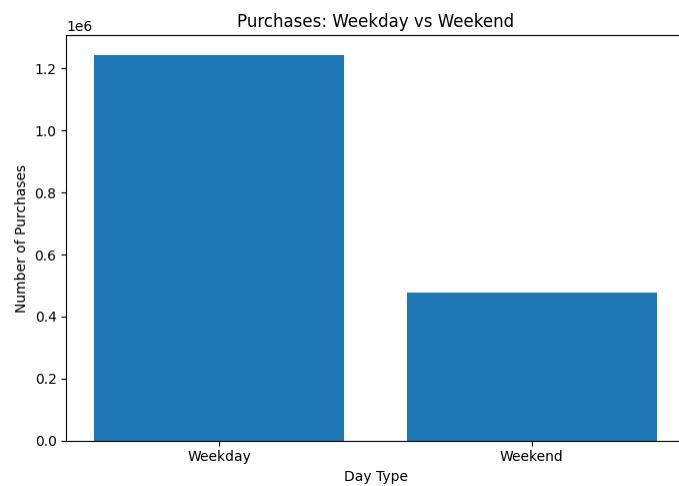
3.2 Customer Demographics vs Purchase Frequency

Demographic factors such as age, gender, income, and education were grouped and correlated with purchase frequency. This analysis highlighted which customer segments were more active buyers. Bar plots compared demographic attributes with purchase counts



3.3 Purchase behavior weekend vs weekday

Purchases were separated into weekday versus weekend transactions using date functions. We compared purchase counts and revenues, highlighting higher activity trends on specific days. Bar plots visualized differences between weekday and weekend shopping patterns

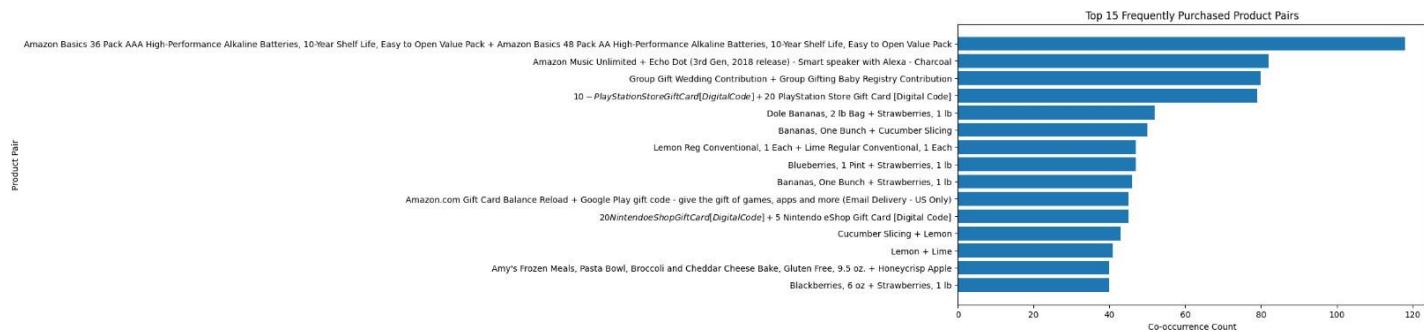


DayType Purchases Percent

DayType	Purchases	Percent
0 Weekday	1244366	72.309234
1 Weekend	476529	27.690766

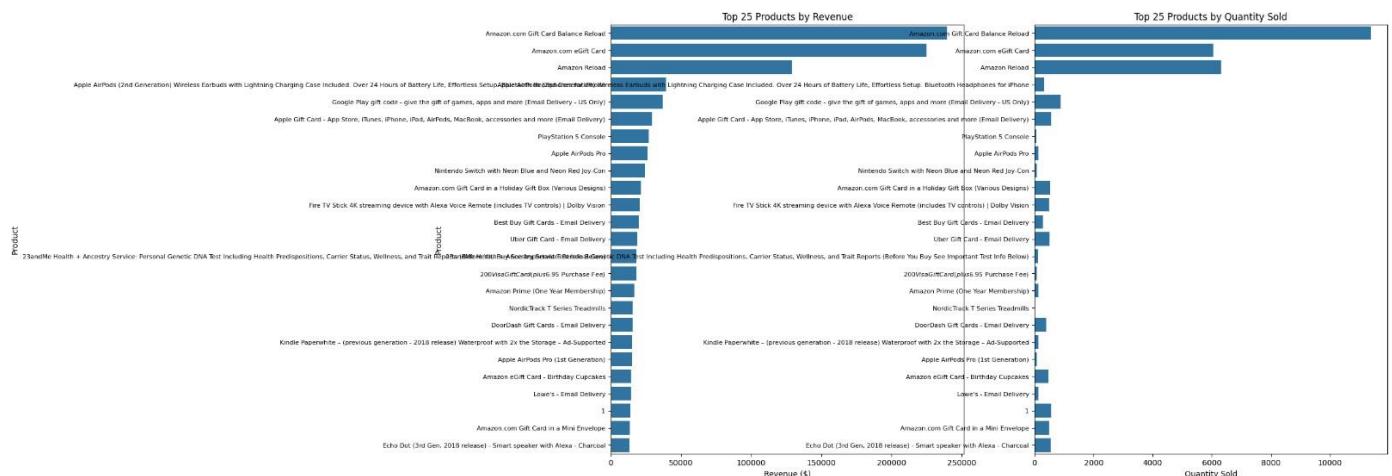
3.4 Frequently purchased product pairs

Market basket analysis grouped items purchased together by customers. Frequent item pairs were identified by co-occurrence counts, revealing associations between products. These insights are valuable for cross-selling strategies and recommendations.



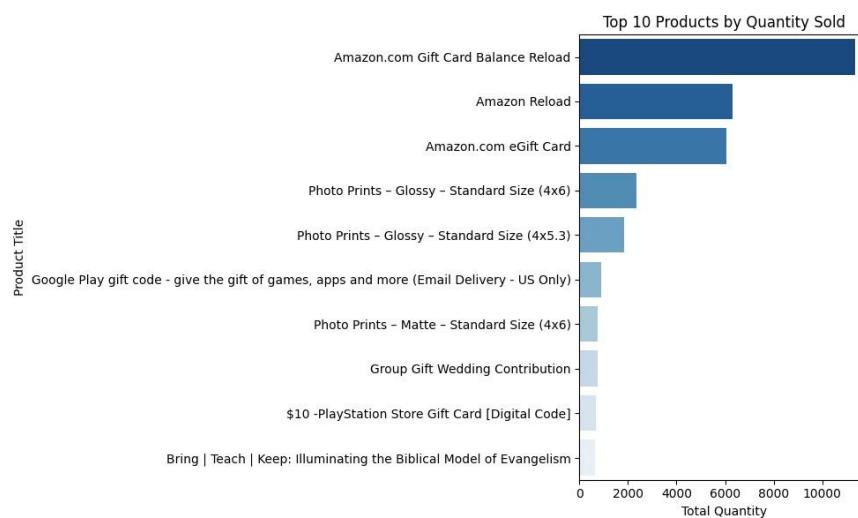
3.5 Product Performance

Products were evaluated by total revenue and purchase frequency. We identified top- performing products based on sales value and popularity. Bar plots showed contributions by revenue and quantities sold, guiding product strategy.



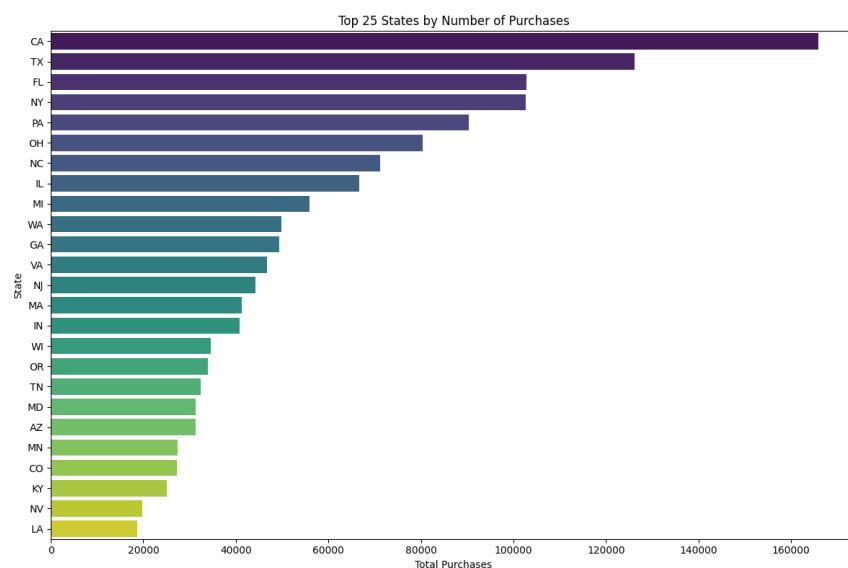
3.6 Top Products by Quantity

We aggregated purchase quantities per product and ranked them to identify the top-selling items. This helped pinpoint products in high demand. Horizontal bar plots displayed the top 10 products by sales volume.



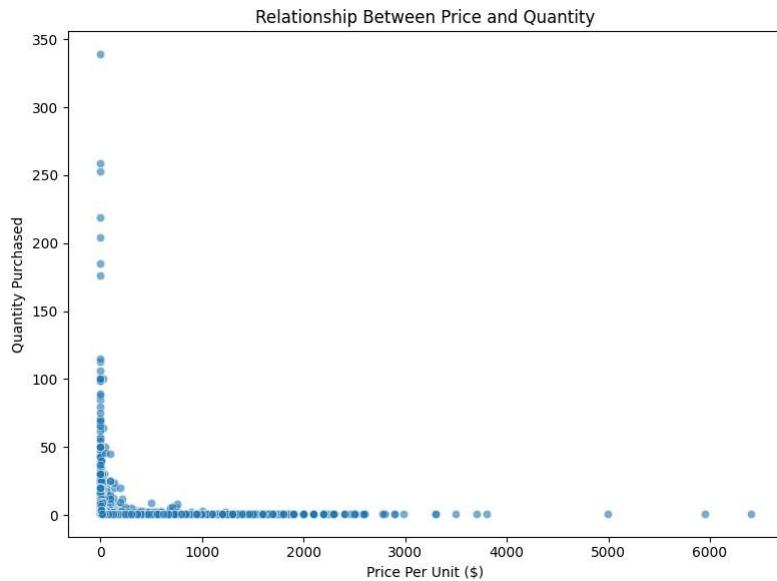
3.7 Distribution of Purchases by State

Customer purchases were grouped by state to assess geographic trends. This analysis provided insights into strong markets and potential growth regions. Bar plots highlighted states with the highest contribution to revenue.



3.8 Price vs Quantity Relationship

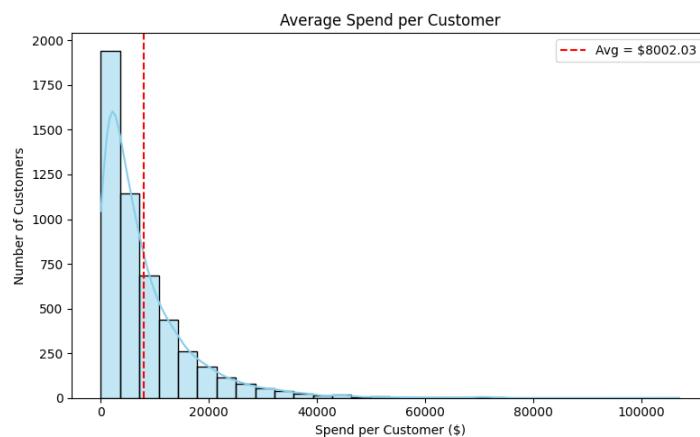
We examined correlations between product price and quantity purchased. Scatter plots visualized patterns, revealing whether higher prices reduced demand or premium products retained strong sales. This supported pricing and promotional strategies.



3.G Spending KPIs (Average Spend per Customer)

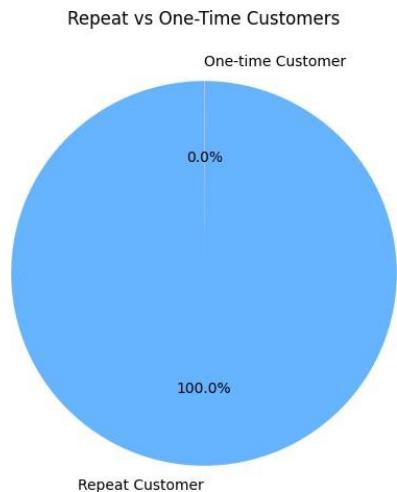
The KPI of average spend per customer was computed by dividing total spending by unique customers. This offered insights into customer value and spending capacity. KPI visuals highlighted revenue potential from existing customers.

Average Spend per Customer



Average Spend per Customer: \$8002.03

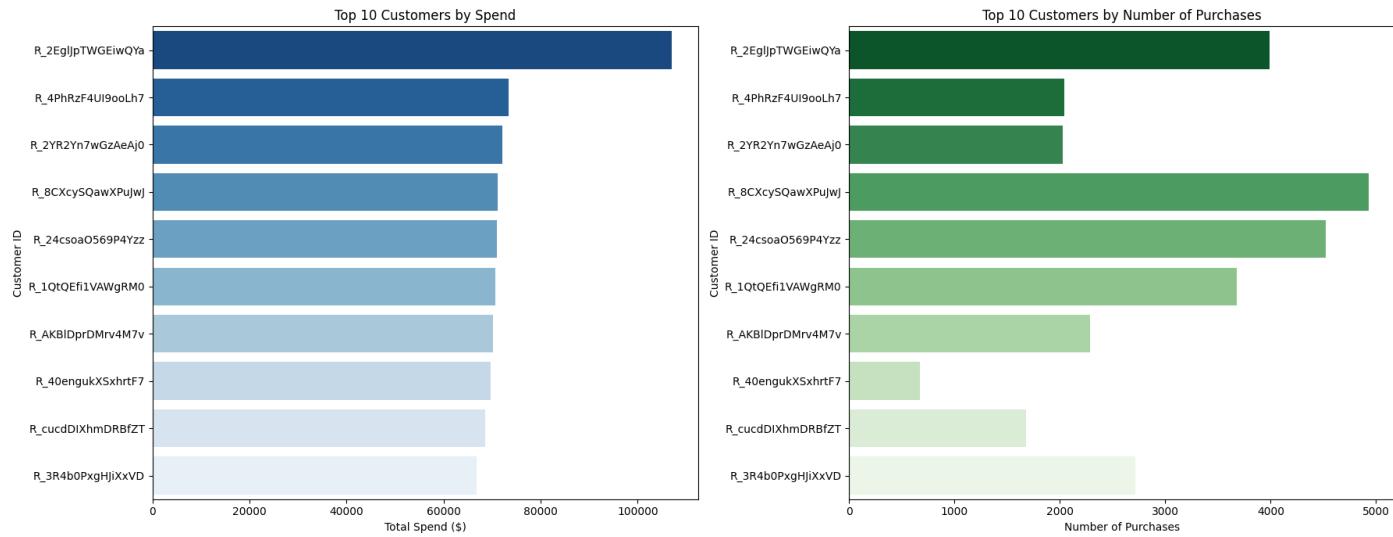
Analyse the Repeat Purchase Behavior of Customers



Purchase_Type Num_Customers

0	Repeat Customer	5019
1	One-time Customer	2

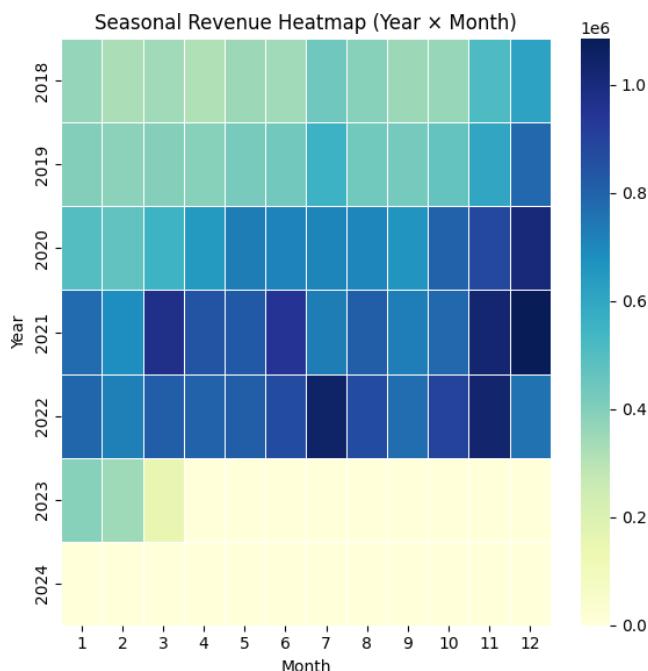
Analyse the top 10 high-engagement customers

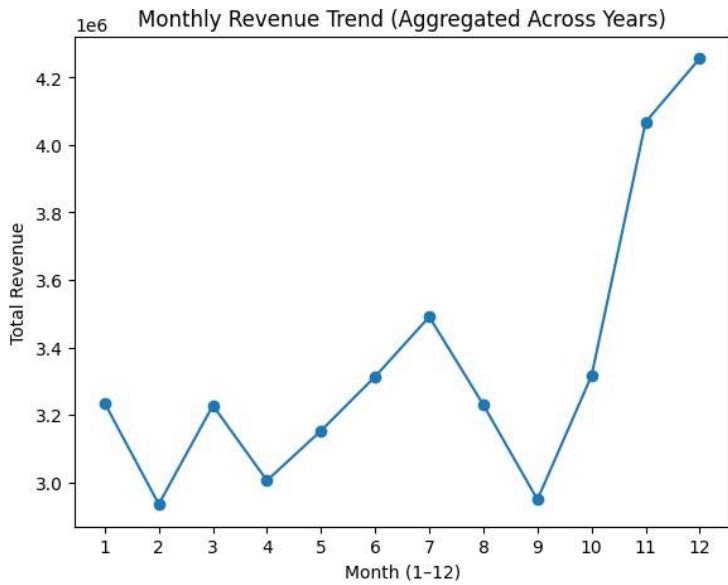


3.10 Seasonal Trends and Revenue

Impact Seasonality was analyzed by grouping sales by year and month. Trends identified peak sales seasons, informing promotional planning. Line plots of monthly revenue showed demand variations over time.

Seasonal Trends in Product Purchases and Their Impact on Revenue

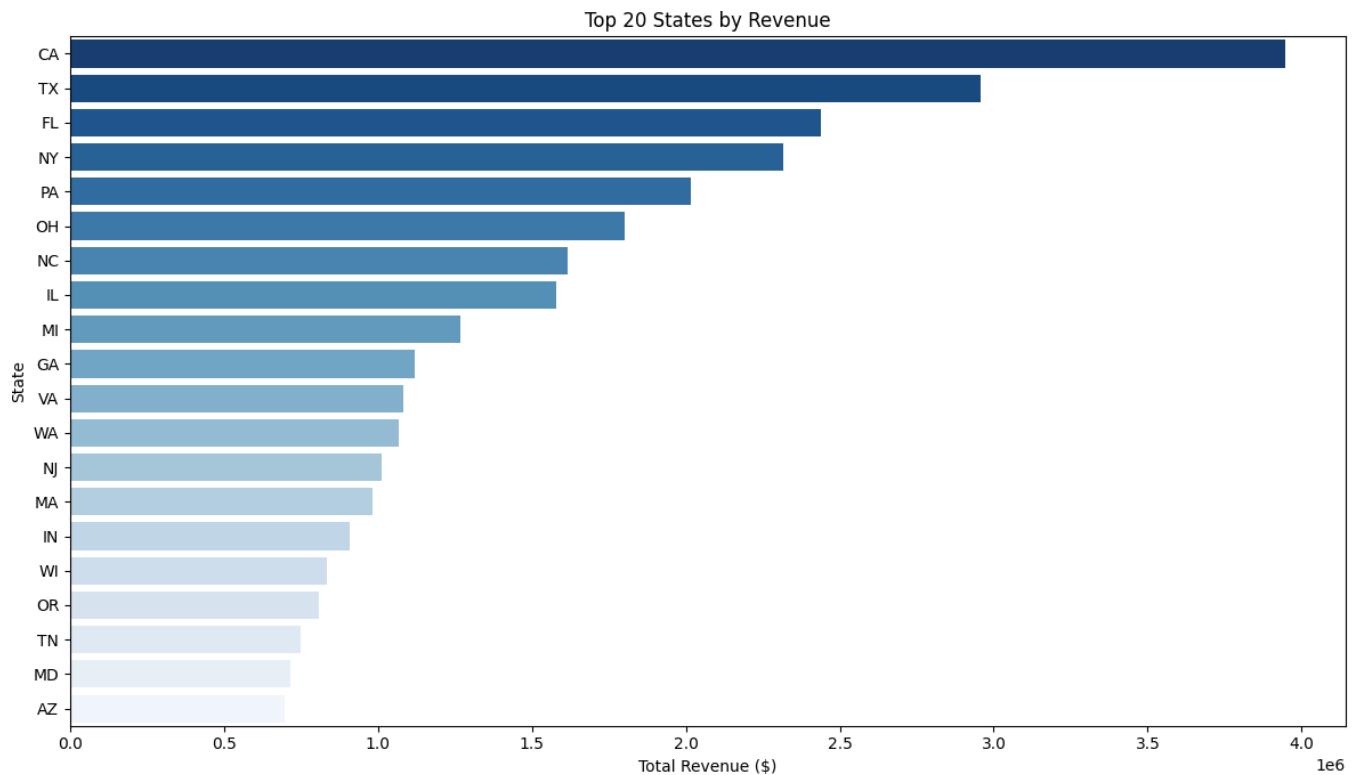




3.11 Customer Location vs Purchasing Behavior

We compared customer states against purchasing behavior to examine geographic influences. This highlighted how regional factors affect buying patterns. Plots displayed variations in spending across different states.

Relationship Between Customer Location and Purchase Behavior

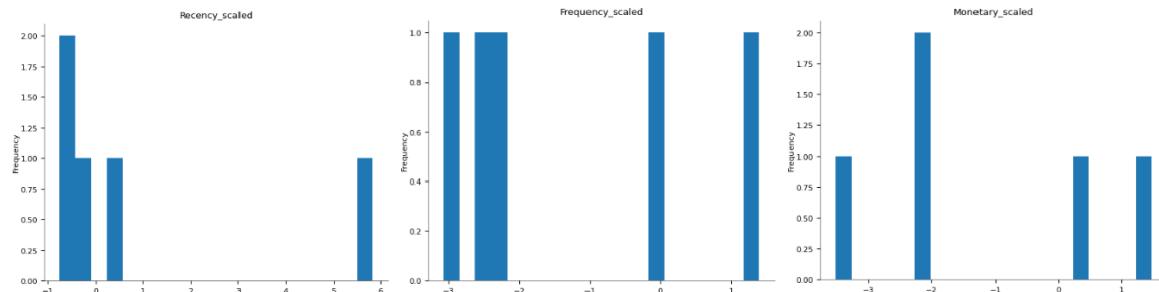


4. Customer Segmentation and Insights

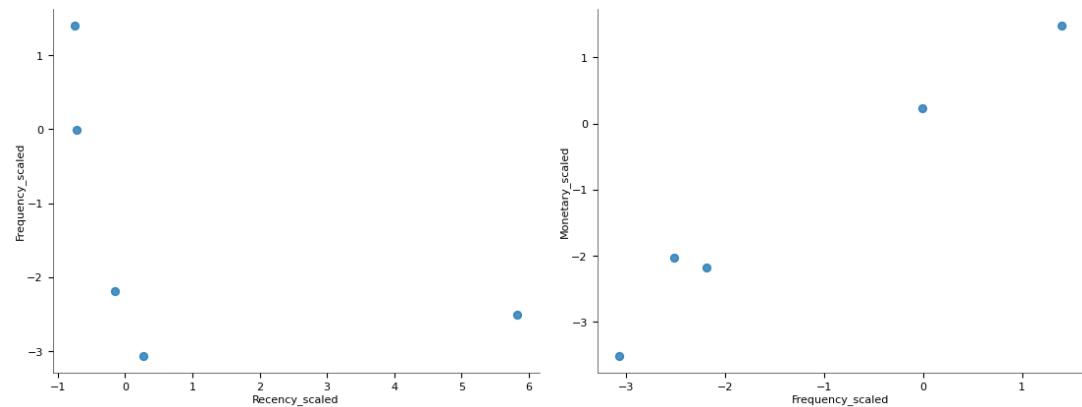
4.1 RFM Analysis

We performed Recency, Frequency, Monetary (RFM) analysis by computing days since last purchase, total purchase frequency, and total monetary spend per customer. This helped segment customers into high-value and low-value groups for targeted marketing.

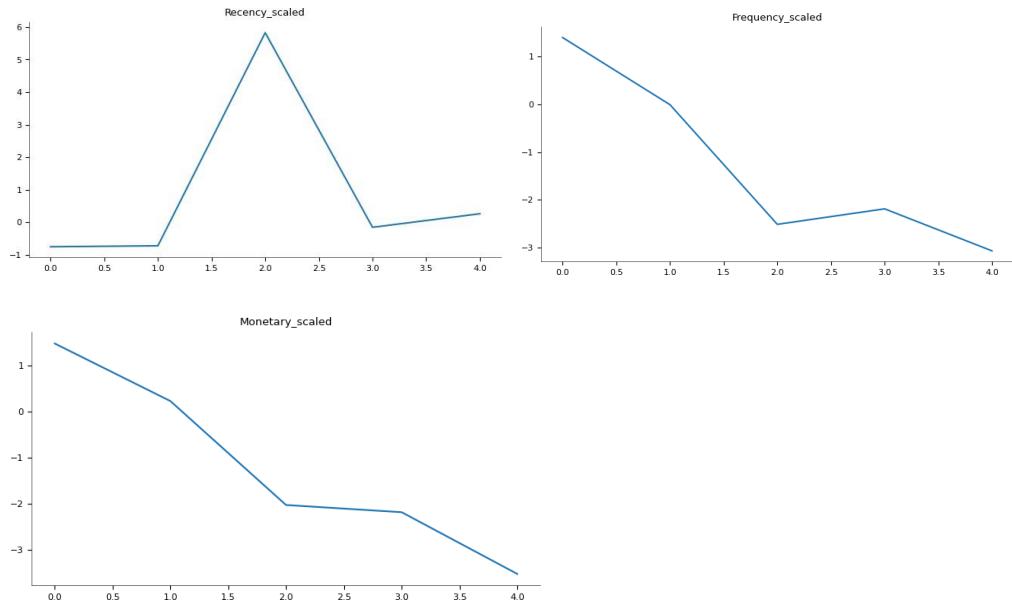
Distribution:



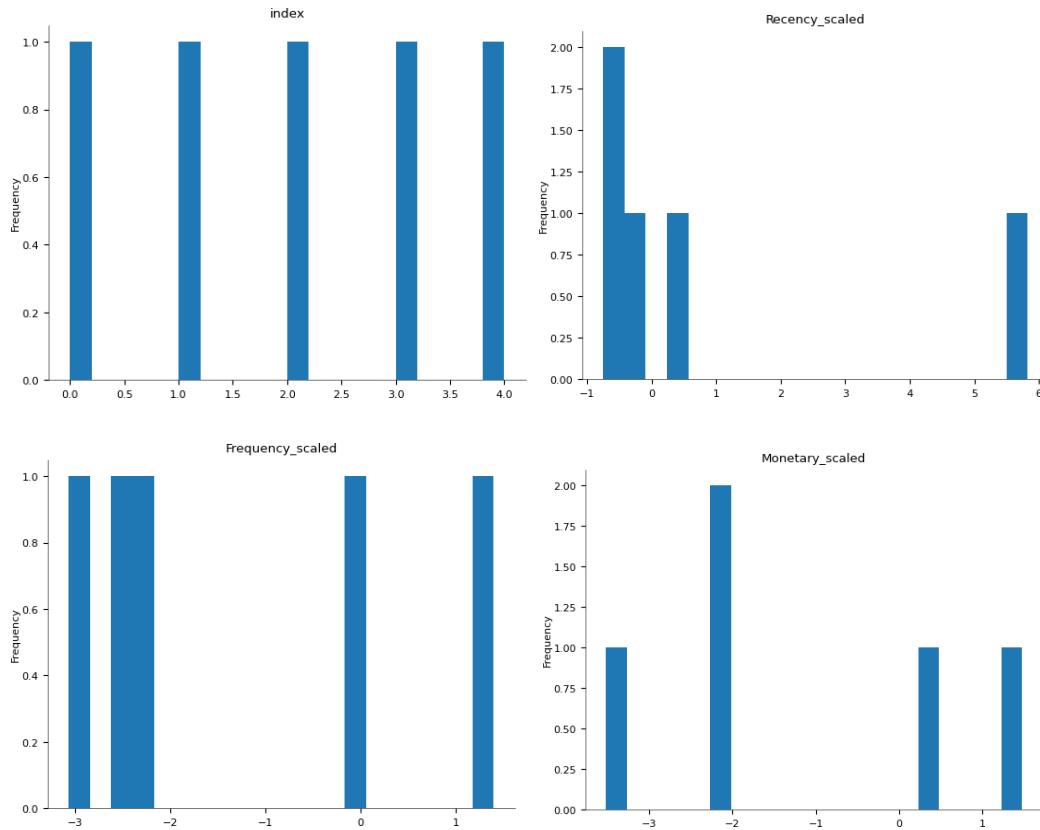
2-d distributions



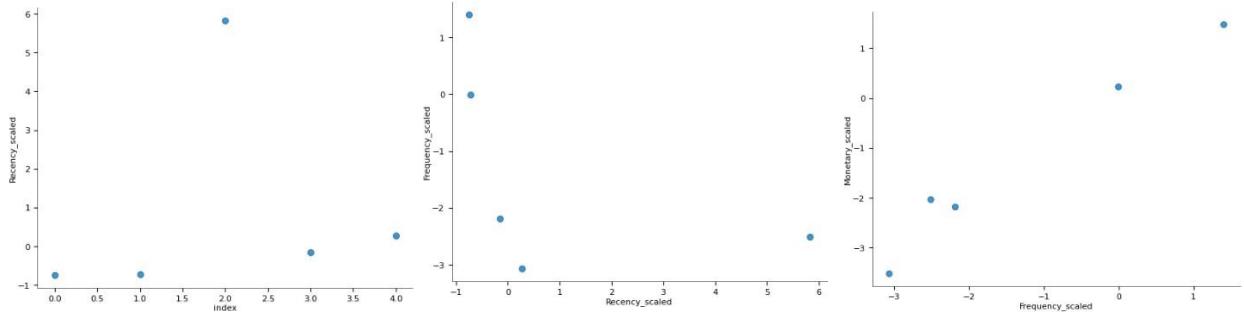
Values



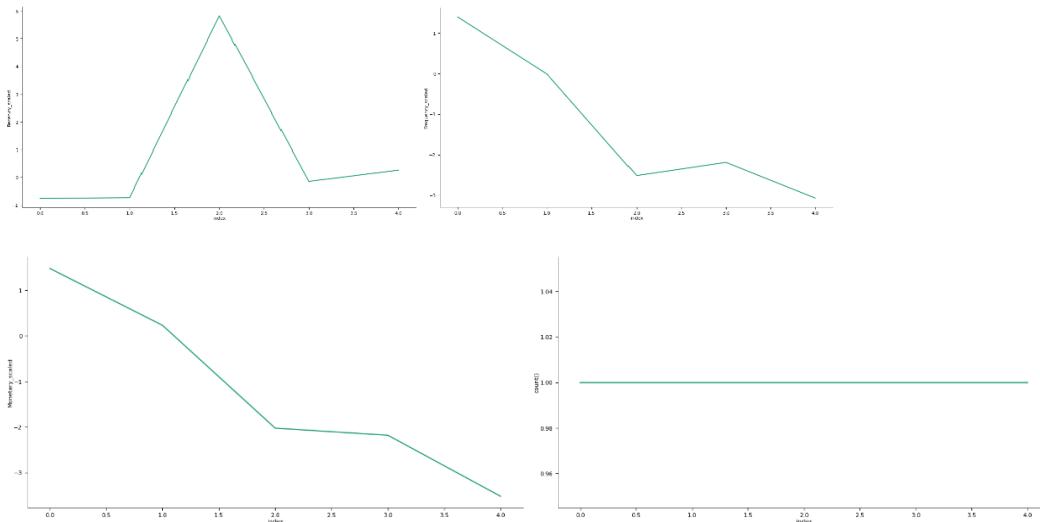
Distributions



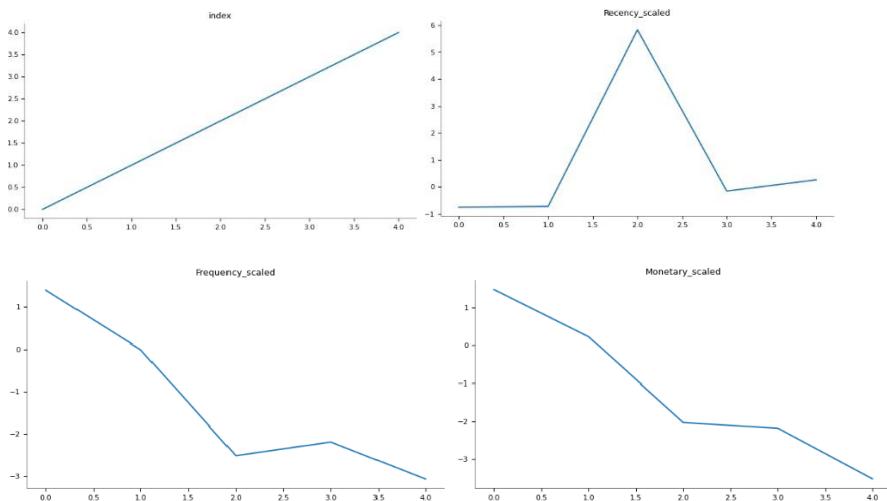
2-d distributions



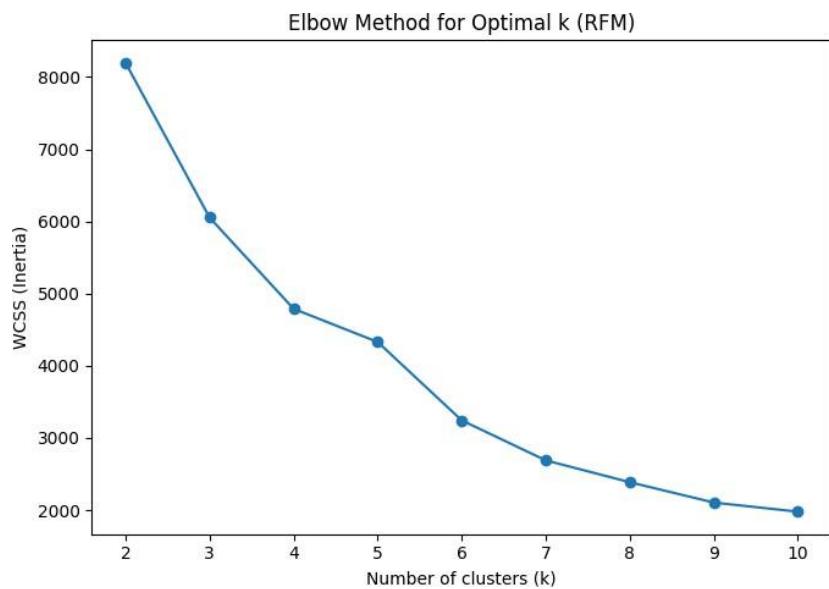
Time series



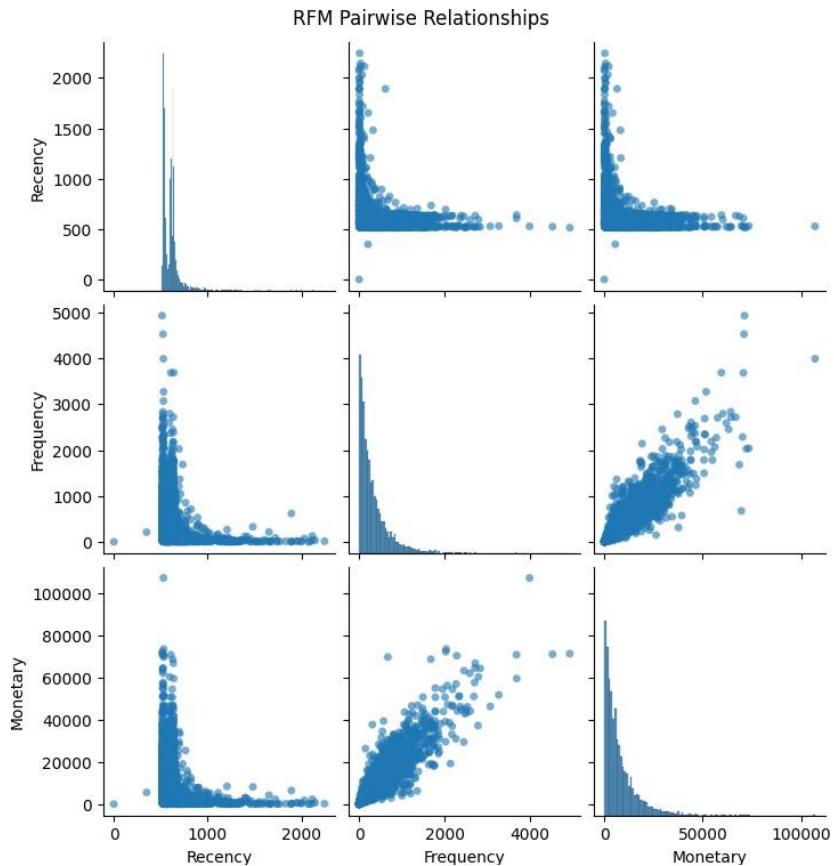
Values



Plot the elbow curve with the number of clusters on the x-axis and WCSS on the y-axis



Convert the full RFM dataset from PySpark DataFrame to Pandas DataFrame for visualisation



Behavioral Trends Analysis

Perform RFM analysis to study the behavior of customers to tailor marketing strategies

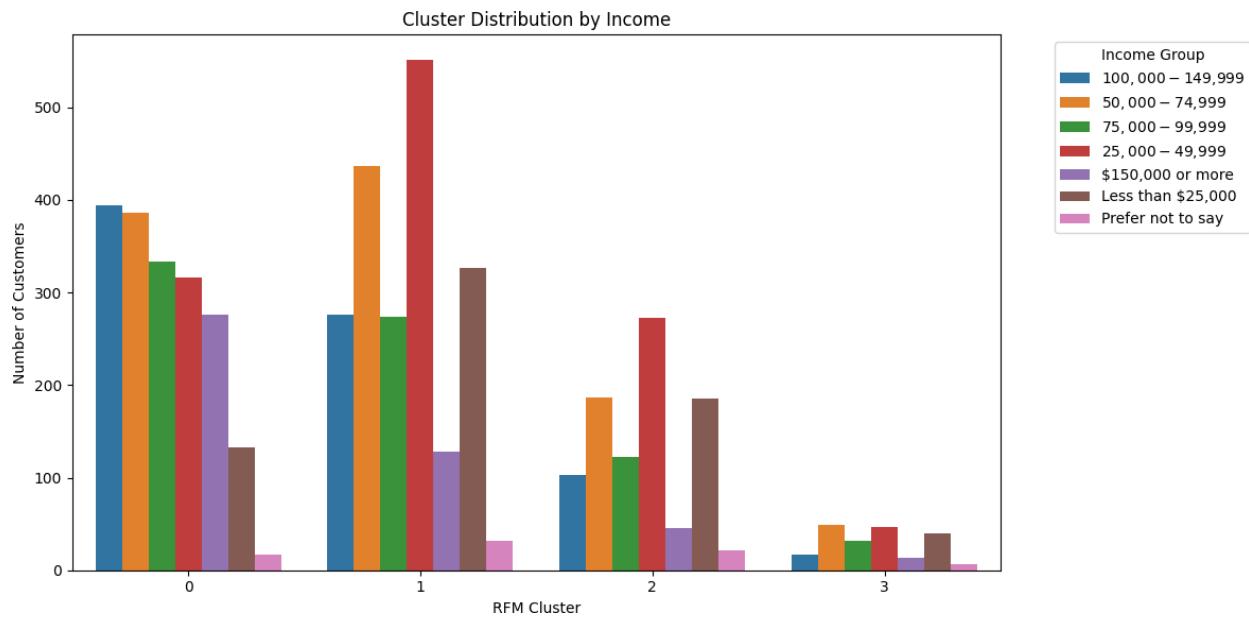
This code performs RFM (Recency, Frequency, Monetary) analysis for customer segmentation. It extracts purchase timestamps, calculates recency (days since last order), frequency (distinct products purchased), and monetary value (total spend). Data is

normalized using log transformation, converted to Pandas, and standardized with StandardScaler to prepare features for clustering models like K-Means.

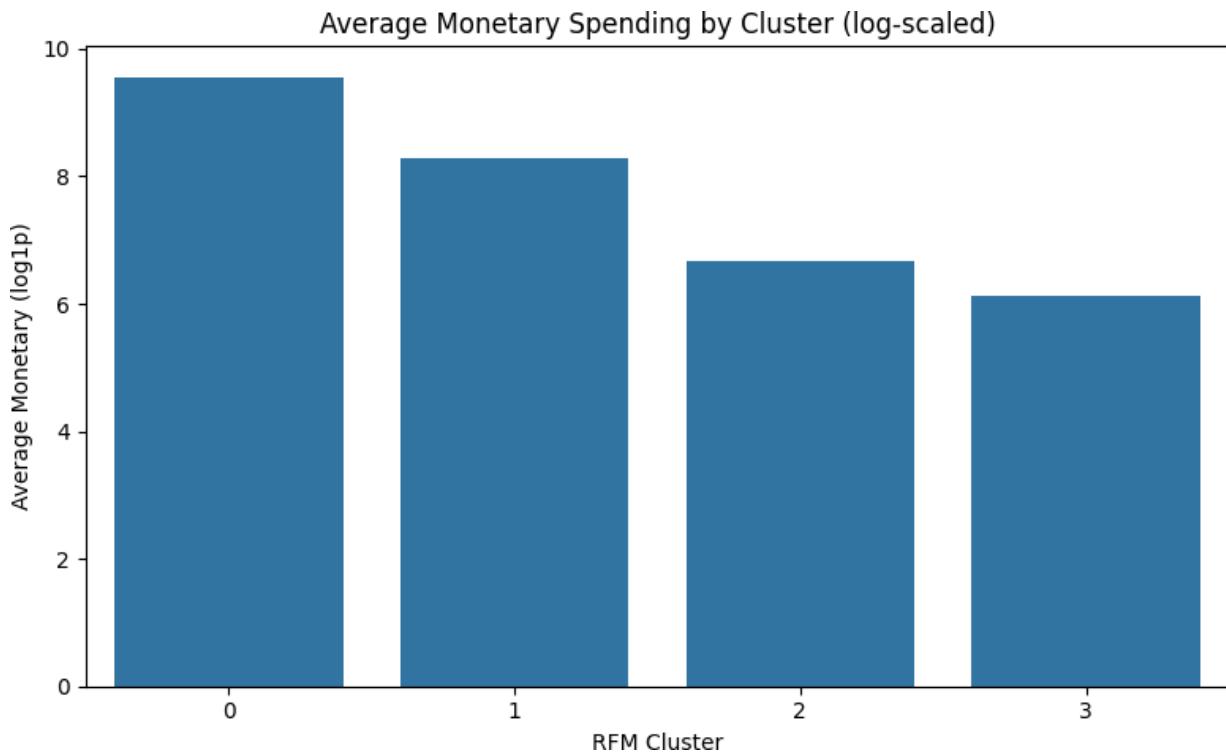
Apply K-Means clustering

This code applies **K-Means clustering** on standardized RFM features. It selects scaled inputs, fits the model with four clusters, predicts labels, and attaches them to customers. Finally, it merges results back into Spark, showing each customer's Recency, Frequency, Monetary values, and assigned cluster for segmentation insights.

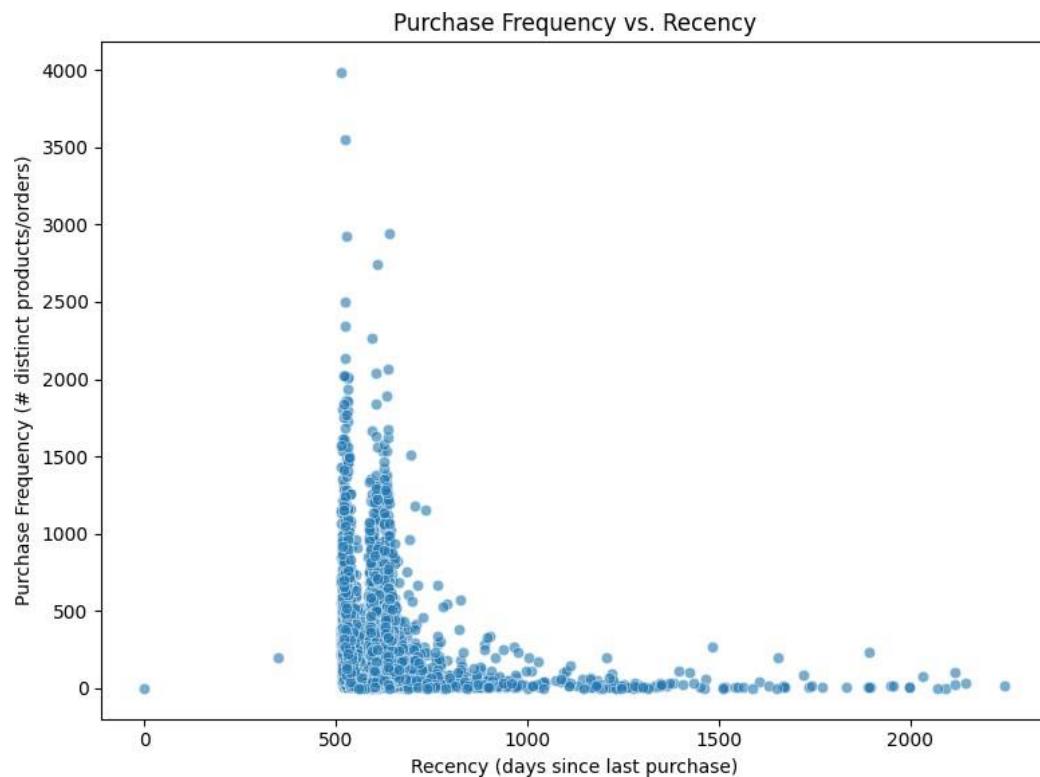
Analyse the Cluster Distribution by Income



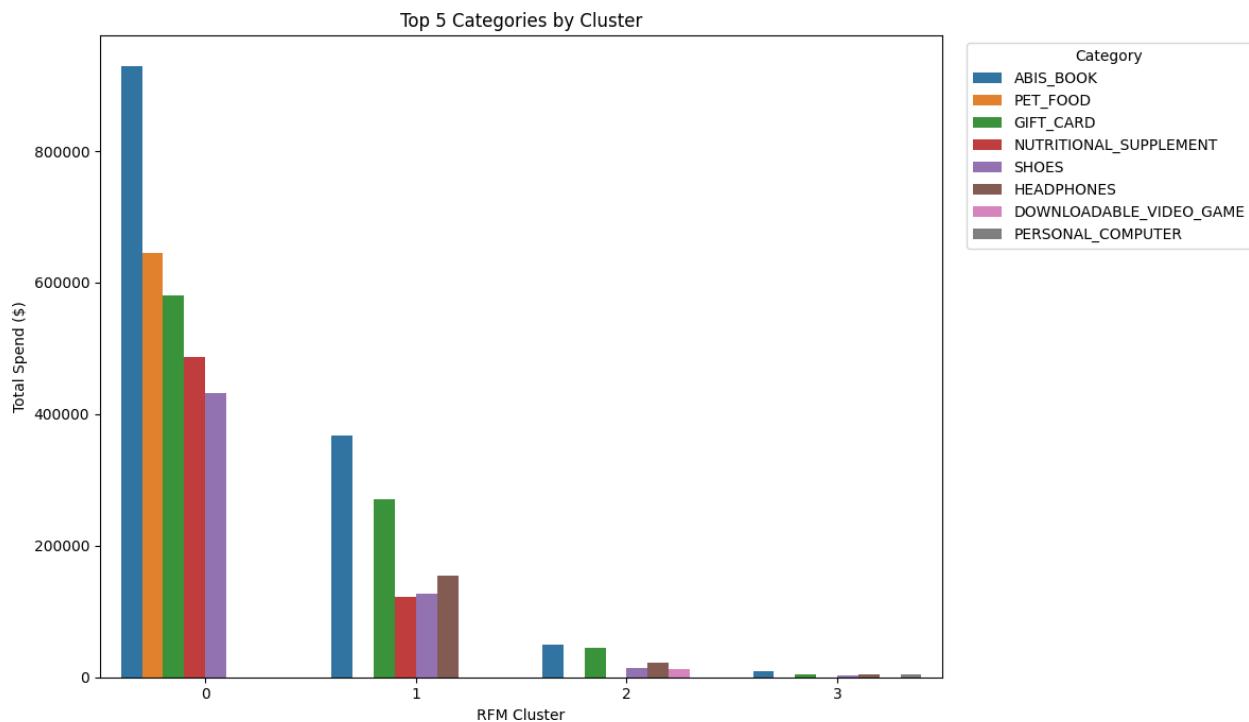
Analyse the Average Spending by Cluster



Analyse the Purchase Frequency vs. Recency



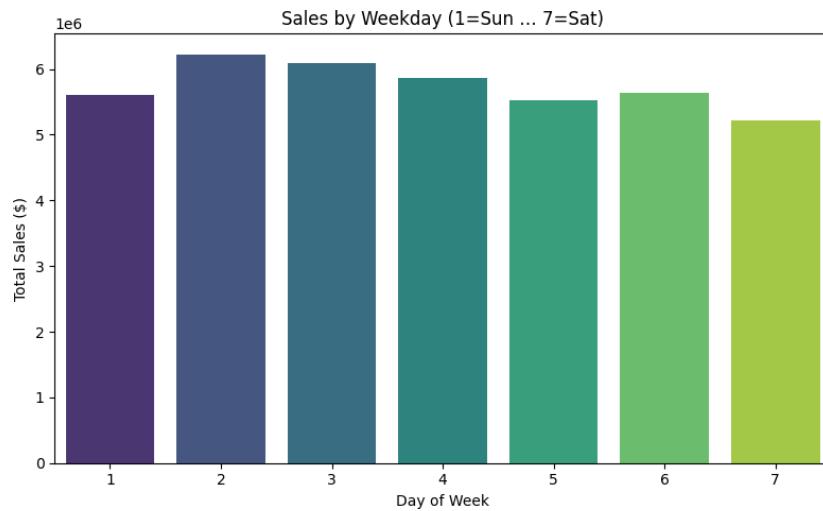
Analyse the top categories by clusters



4.2 Insights

4.2.1 Promotions by Weekday

Weekday sales distributions were calculated and plotted, revealing which days experienced high or low activity. This guided scheduling of promotions on days where they would be most effective.



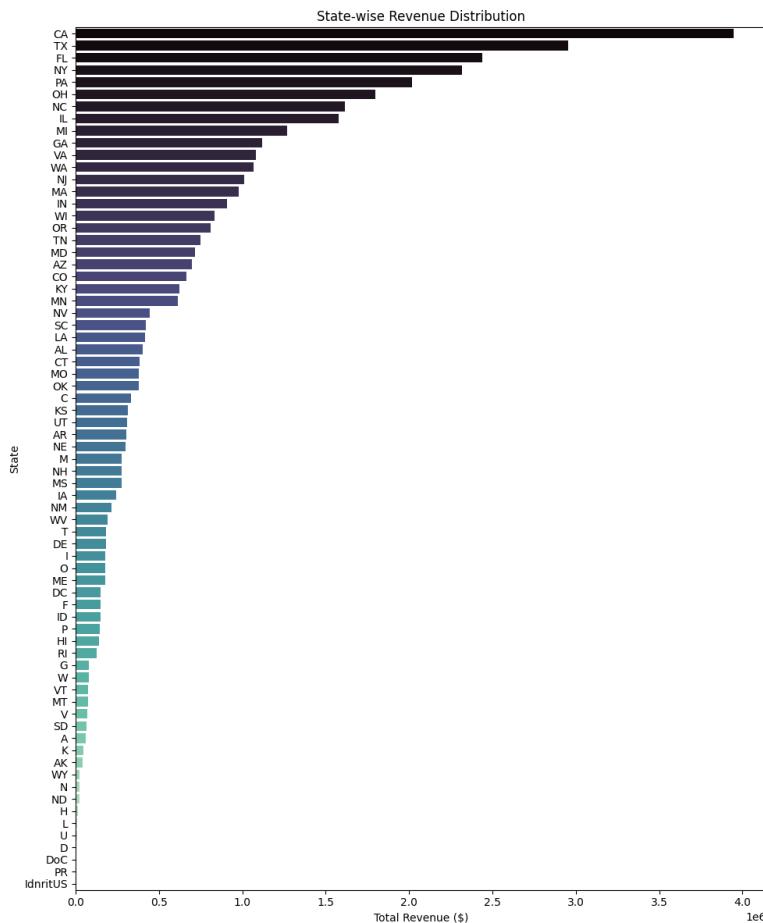
4.2.2 Top-selling Products

Products were ranked by total revenue and purchase counts. Plots highlighted the top 10 products by sales value, aiding product prioritization and promotional strategies.



4.2.3 State-wise Revenue Distribution

Revenue was aggregated by customer state to identify high-contribution regions. Bar plots showcased the top-performing states, enabling focus on high-growth areas.



4.2.4 Repeat Purchase Behavior

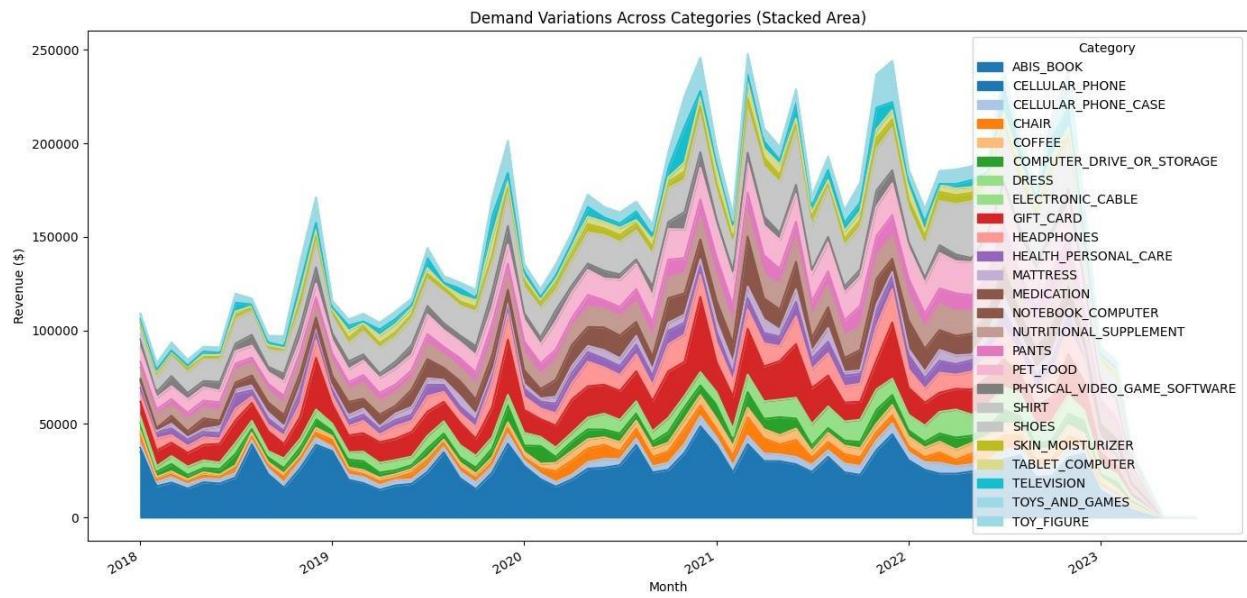
Customers with multiple purchases were identified and compared against one-time buyers. This analysis emphasized retention opportunities and loyalty program design. Histograms and counts illustrated repeat buyer behavior.

4.2.5 Fraud Detection

We flagged transactions where spending exceeded mean + 3 standard deviations. This statistical anomaly detection identified suspicious high-value orders potentially linked to fraud.

4.2.6 Demand Variations across product categories

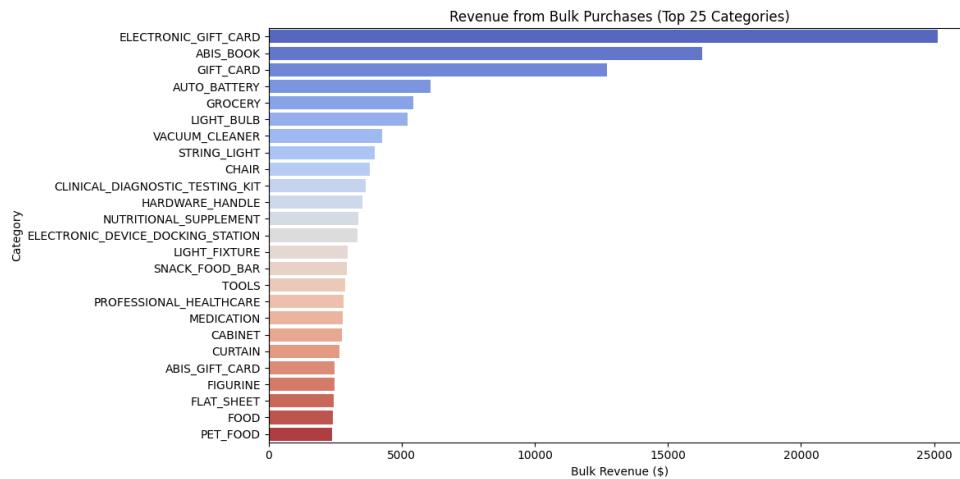
In this code we analyze **demand variations across product categories**. It calculates monthly revenue by multiplying unit price and quantity, aggregates by category and month, and selects the top 25 categories. Finally, it visualizes category-wise revenue trends over time using a stacked area chart for inventory management



insights.

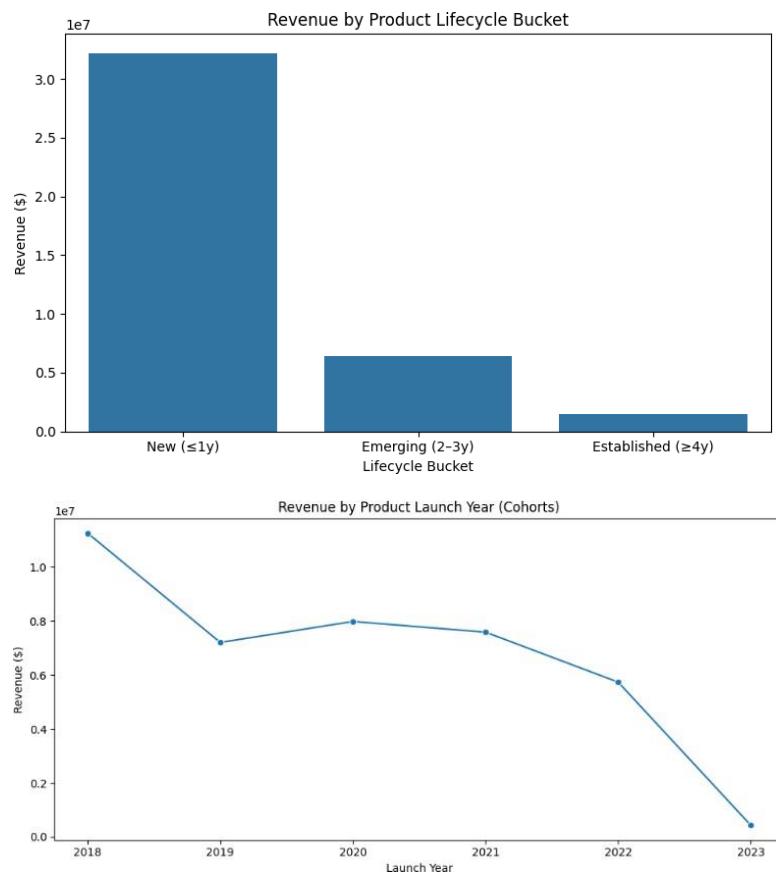
4.2.7 Bulk Purchases and Revenue

Transactions with quantities greater than 5 were classified as bulk purchases. Revenue from these was aggregated by category, helping assess supply chain pressure points. Bar plots highlighted top categories affected by bulk purchases.



4.2.8 Lifecycle Strategies

We computed launch year as the first sale per product and compared revenues of new versus established products. Line and bar plots revealed lifecycle impacts, guiding marketing and phasing strategies.



5 Conclusion

Customer behavior analysis reveals clear patterns in purchases, demographics, and product trends. Seasonal and weekday insights highlight promotion opportunities, while RFM segmentation pinpoints loyal customers. Revenue by state and category supports growth planning. Repeat purchase, fraud detection, and bulk buying analysis guide loyalty, security, and supply chain decisions effectively.

- Promotions: Midweek and seasonal campaigns can improve sales during low- demand periods.
- Segmentation: High-value clusters should be prioritized for retention and loyalty initiatives.
- Operations: Bulk purchasing and product lifecycle insights guide supply chain optimization.