Report for Customer Behaviour Analysis

1.Data Preparation:The data set comprises three CSV files containing transactional and demographic data. Preprocessing ensures consistency and readiness for analysis.

A) Load the CVS using S3 bucket of AWS Apache Spark,

- amazon_purchases_path = "s3a://customeranalysis123/amazon-purchases.csv"
- survey_path = "s3a://customeranalysis123/survey.csv"
- fields_path = "s3a://customeranalysis123/fields.csv"

B) Run all the packages for importing the libraries

- Print the Schema for Amazon_purchase and Survey DataSets and merge the data on the basis of common column(reponse_id)

2.Data Cleaning

- Fixing columns
- Handling missing values
- Handling outliers
- Feature engineering.

Output we got : Number of Duplicates: 11516
Number of Duplicates After Cleaning: 0

- Cleaned data in saved in S3 Bucket for further analysis.

3.Exploratory Data Analysis

4.Analyse purchases by hour, day and month

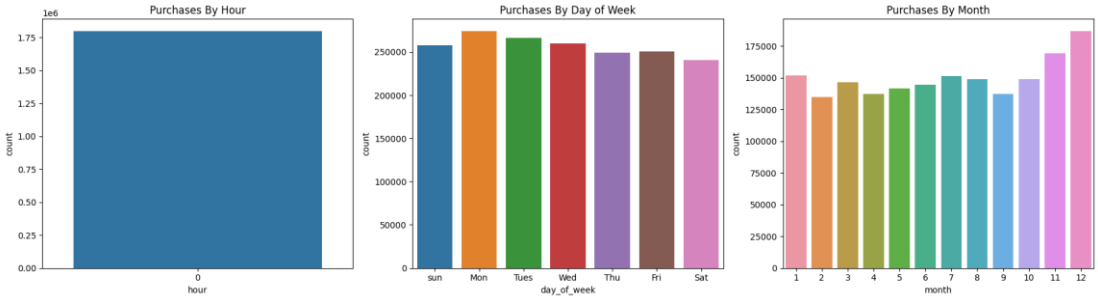


Fig a) month_count.png

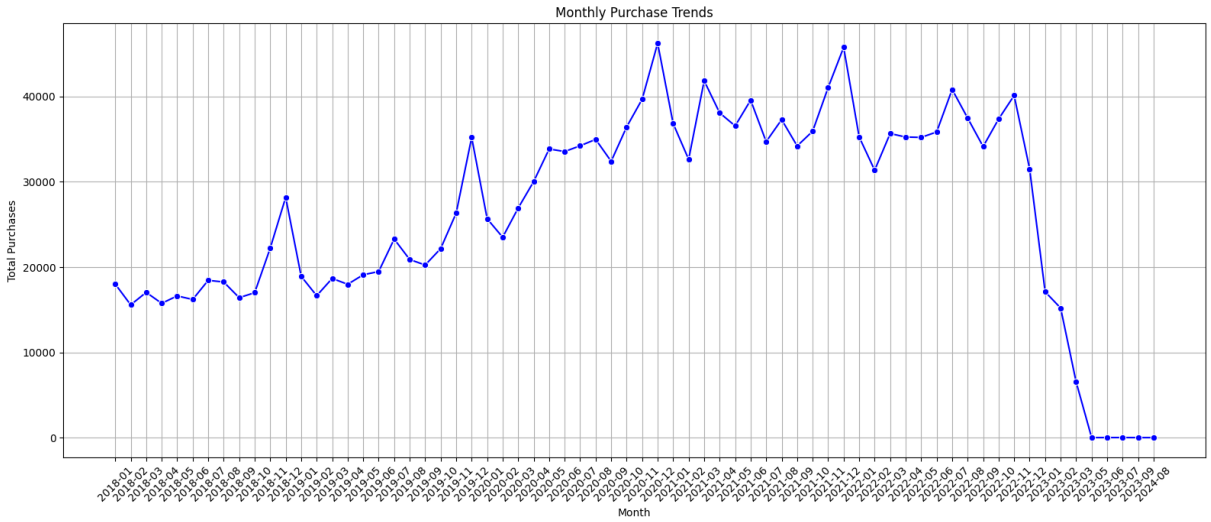


Fig b) monthly_trend.png

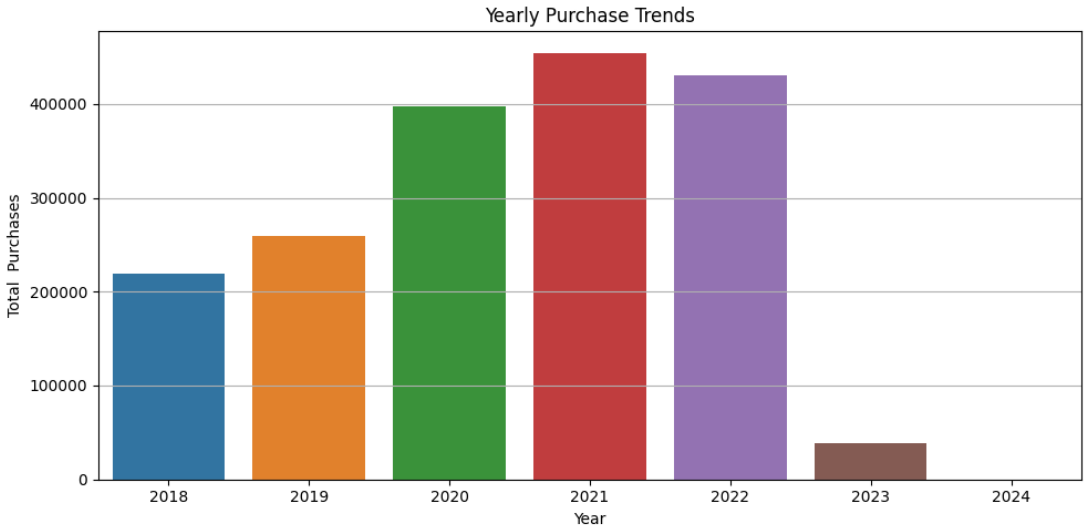


Fig c) yearly_trend.png

5.Analyse the trends between the customer deographics and the purchase frequency(By income, age, gender)

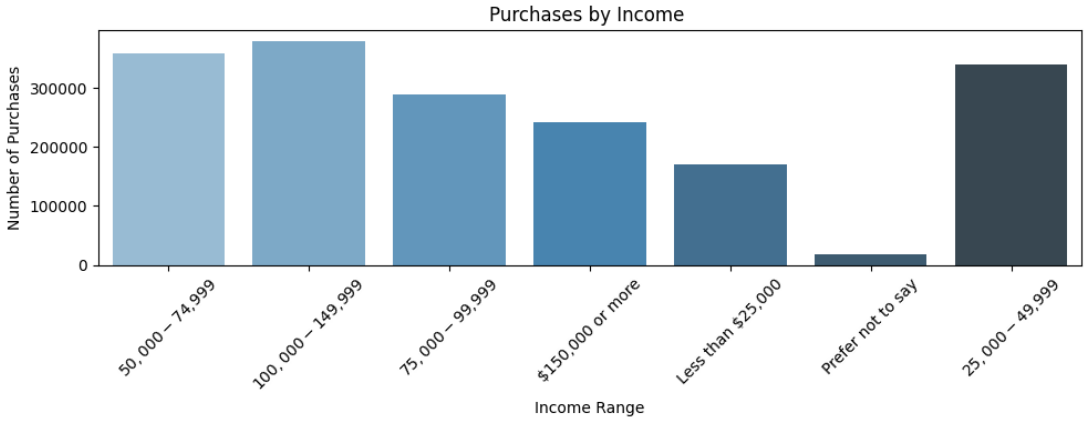
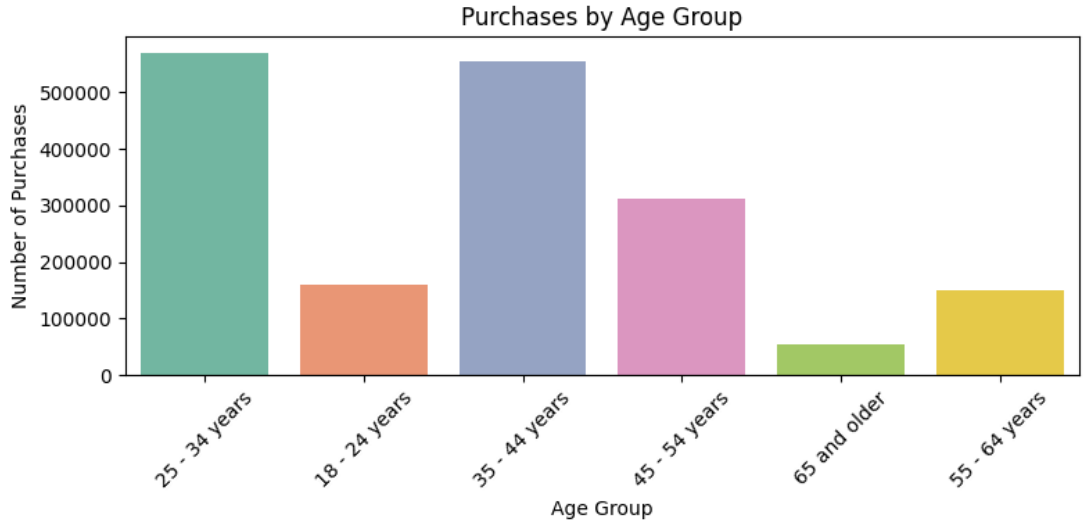


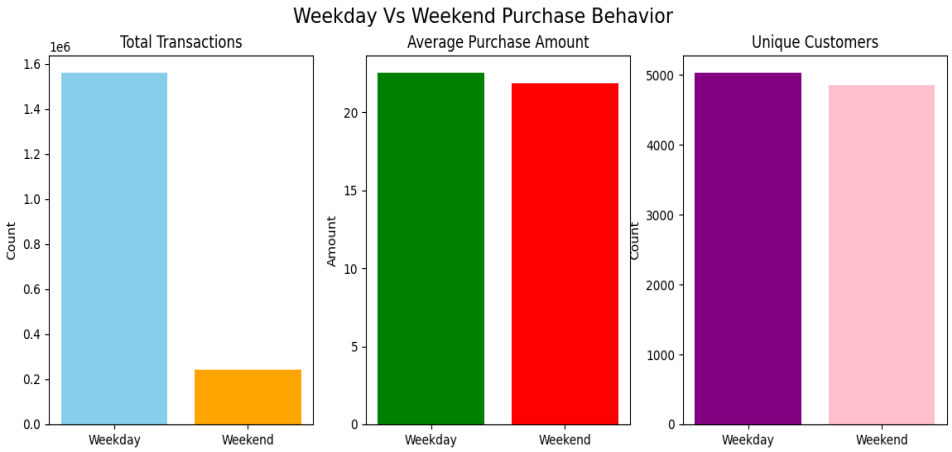
Fig a) income_trend.png



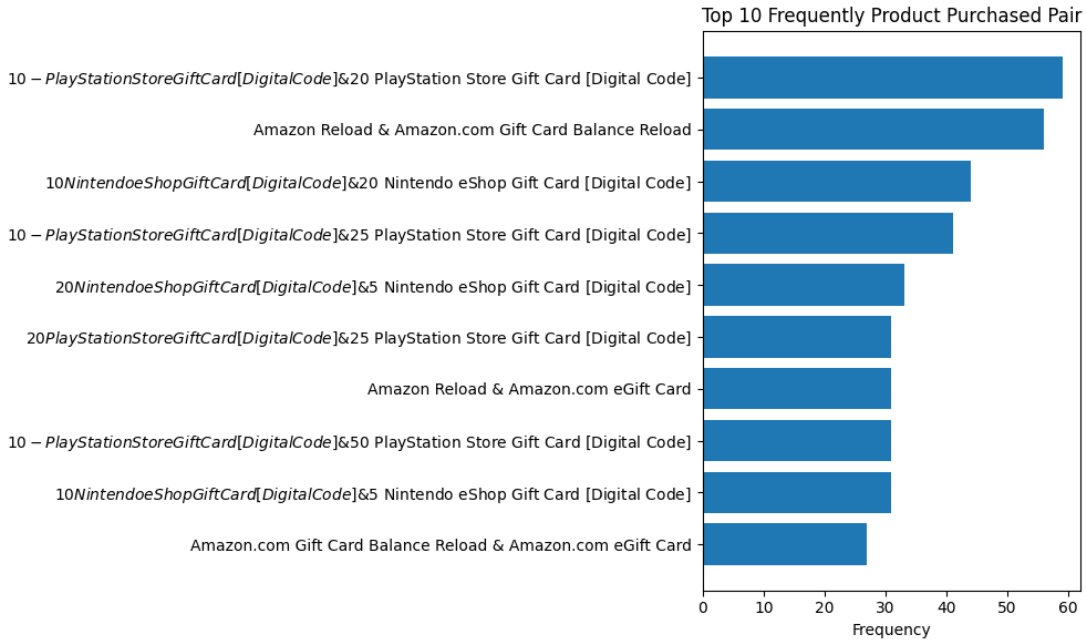
Figb)age_trend.png

Fig c) gender_trend.png

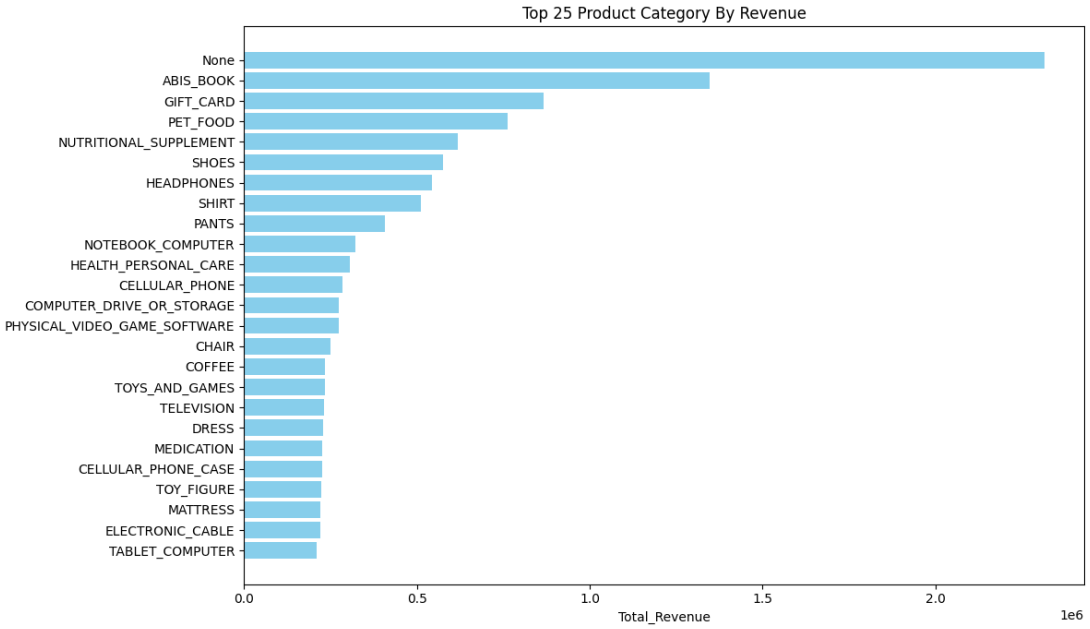
6.Compare the purchase behavior of customer's on weekdays vs. weekends.



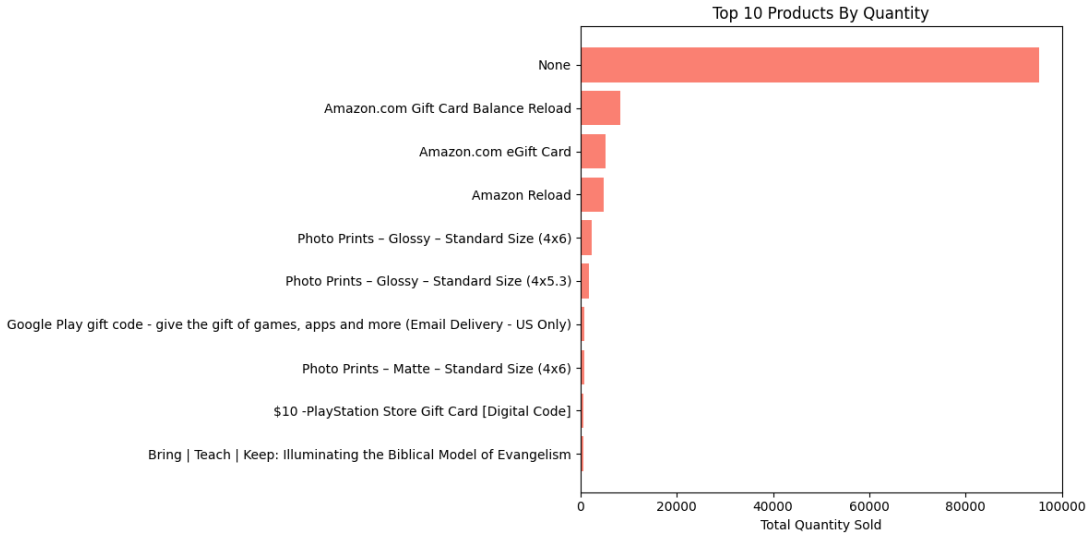
7.Frequently purchased product pairs



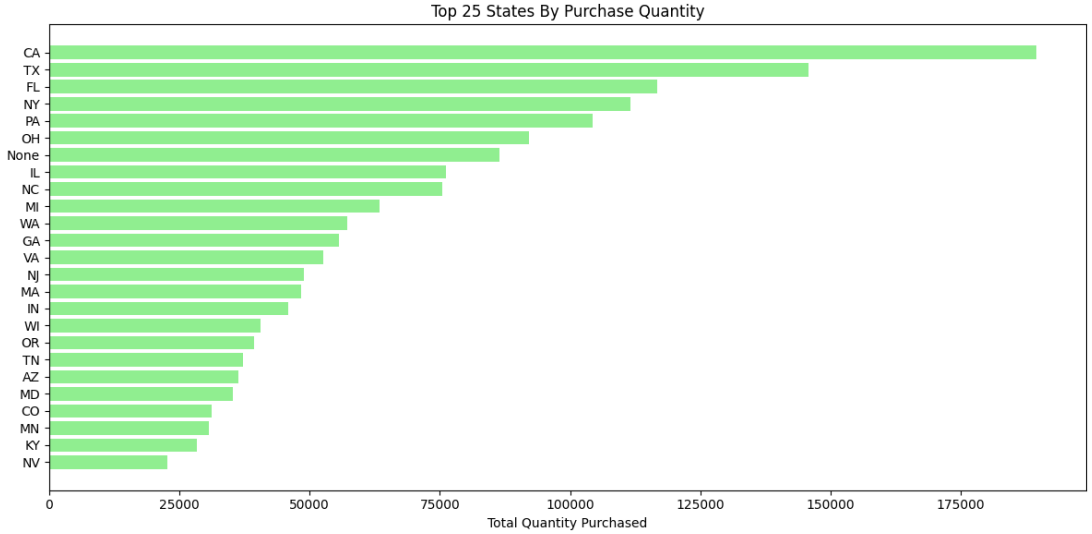
8.Examined Product Performance



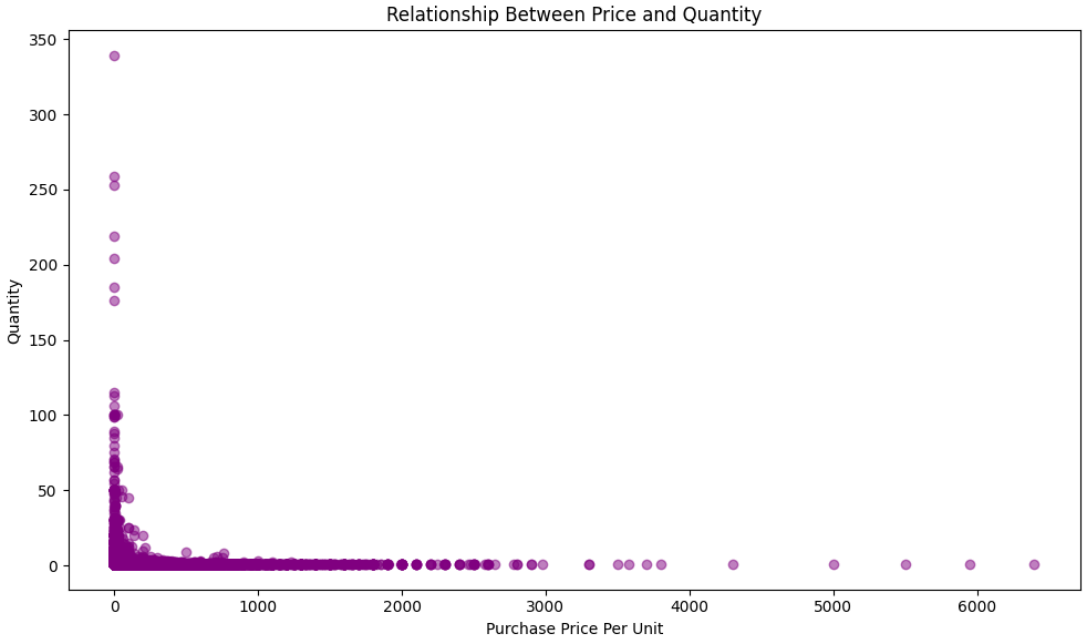
9.Top products by quantity



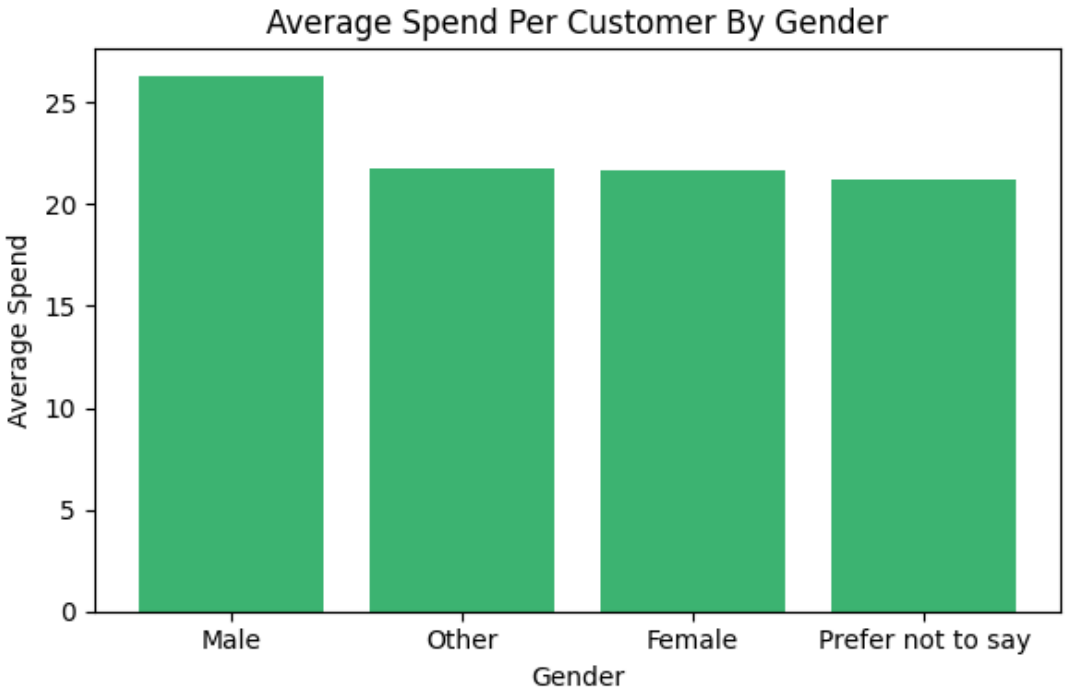
10.Distribution of Purchases by State



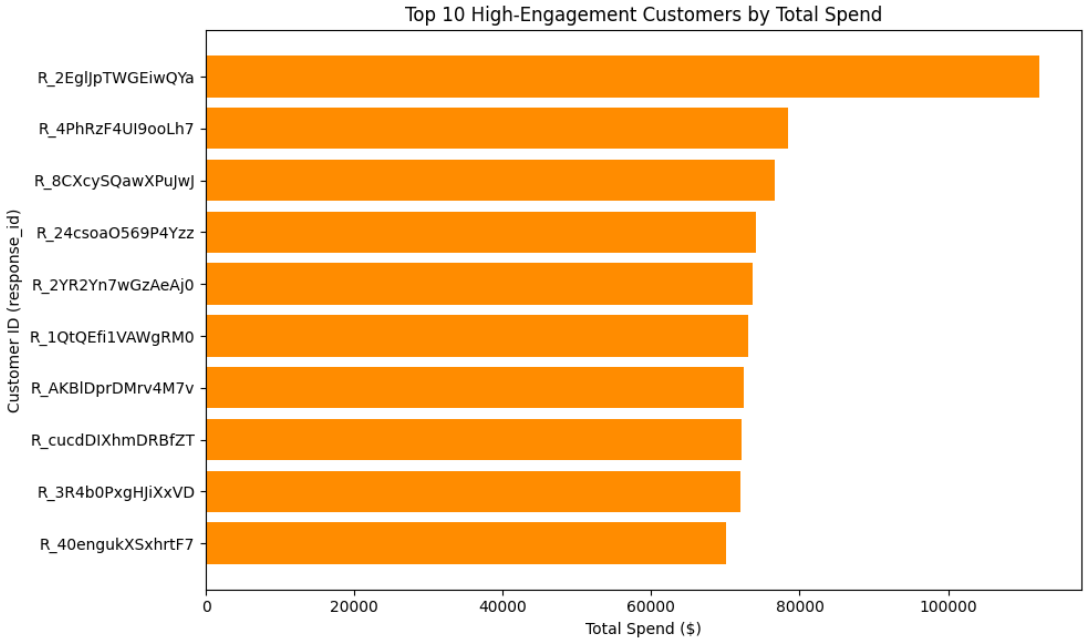
11.Price vs Product Quantity



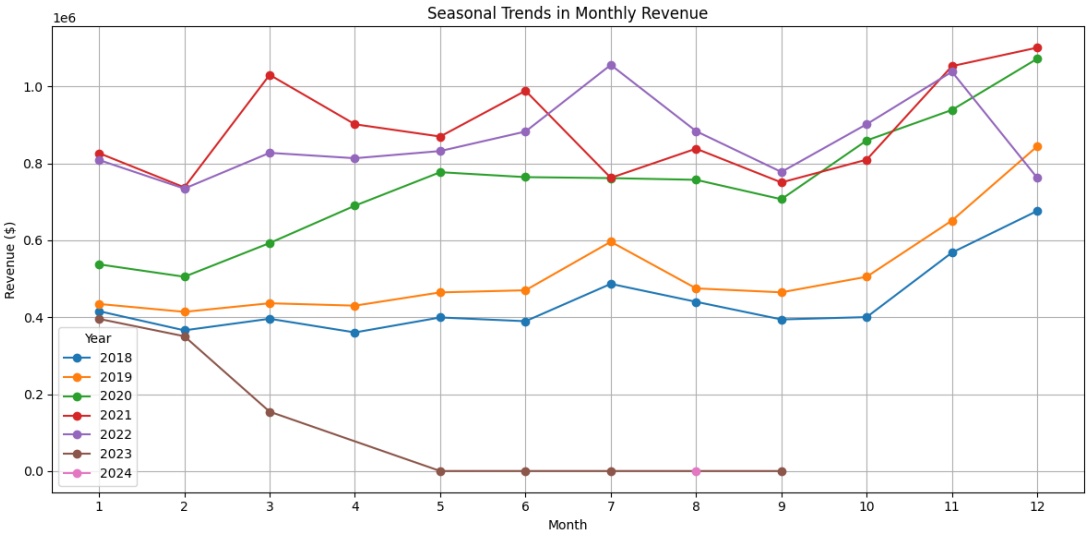
Analyse the spending KPIs : A popular KPI is average spend per customer.



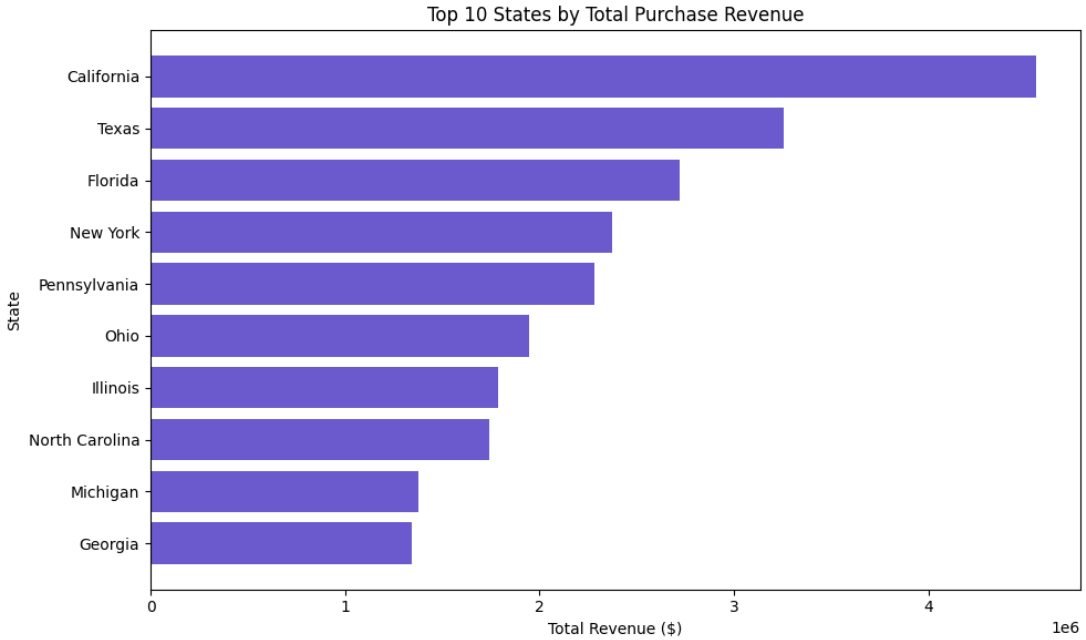
12.Analyse the top 10 high-engagement customers



Seasonal trends in product purchases and their impact on revenues



Customer location vs purchasing behavior



We required to Prepare data for RFM Analysis , Scaled the RFM data , the output is "Recency_scaled", "Frequency_scaled", "Monetary_scaled".

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

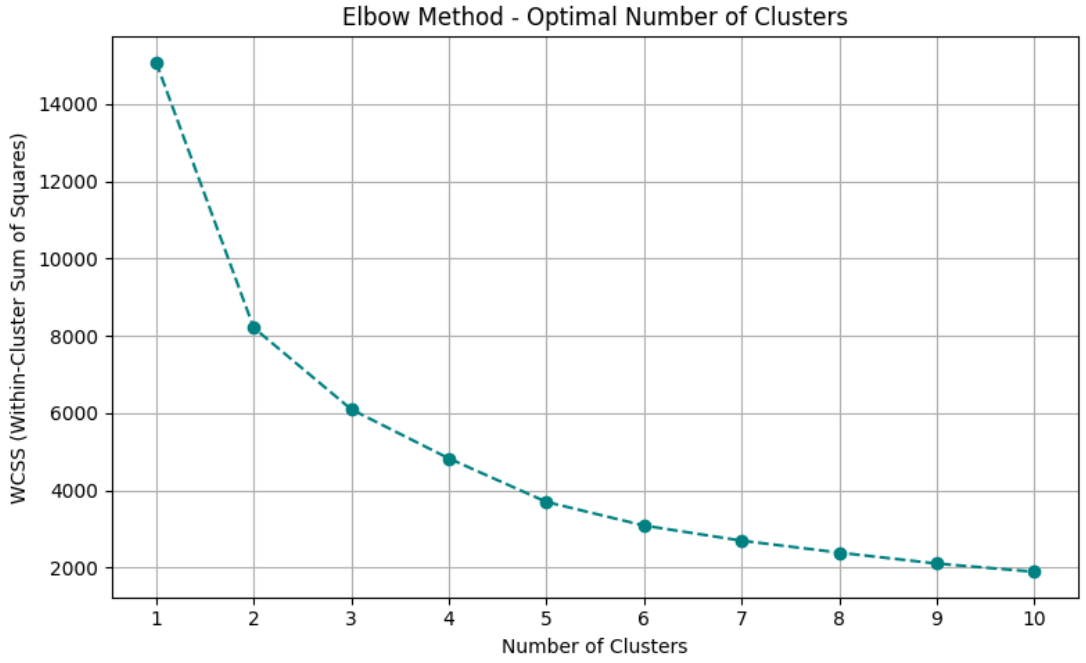
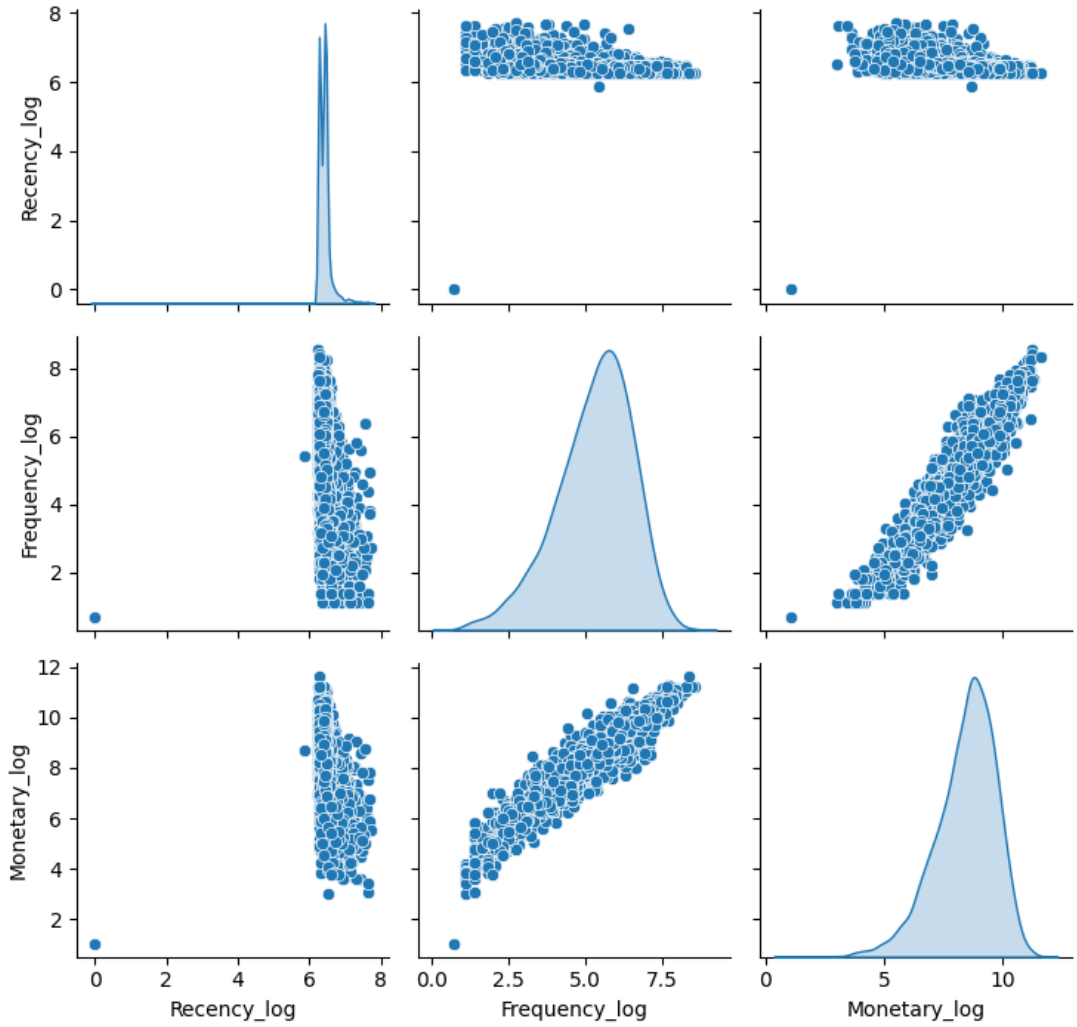


Fig : elbow_curve_rfm.png

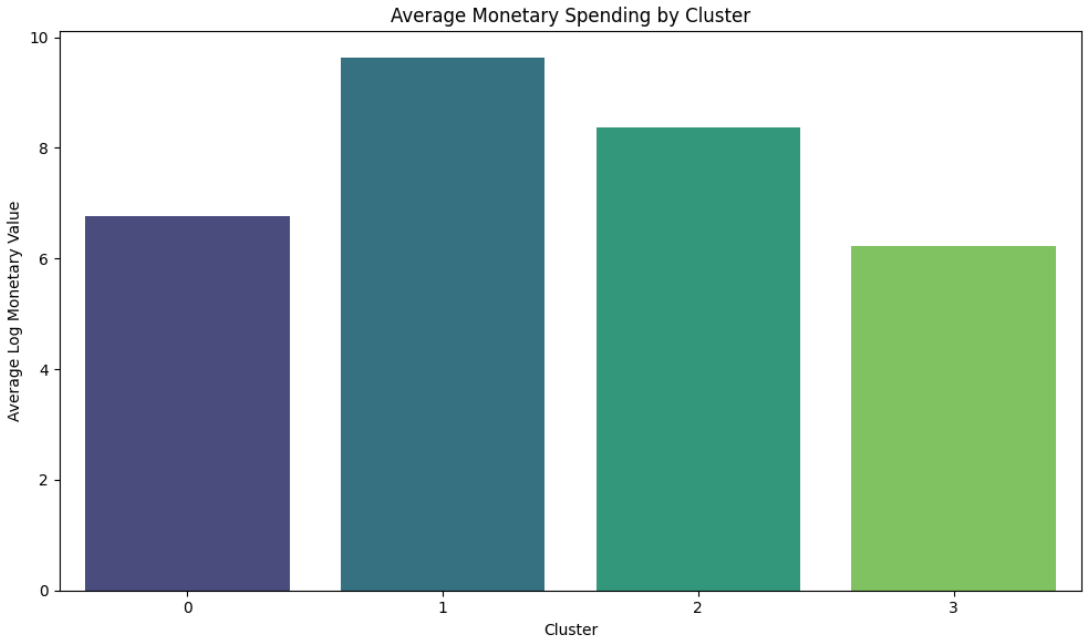
13.Fit the K-Means model using the optimal number of clusters obtained after understanding the elblow plot.

14.Generate a pair-plot to visualize the relationships between the numeric RFM columns

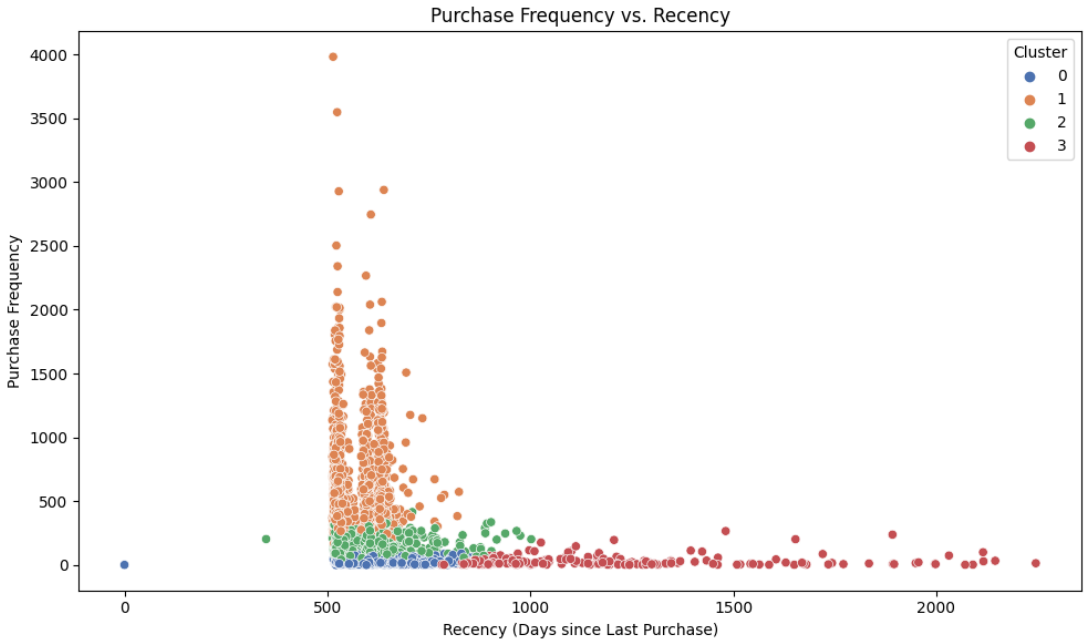


15.Behavioral Trends Analysis: Perform RFM analysis to study the behavior of customers to tailor marketing strategies and futher apply K-means clustering at rfm_scaled_data.

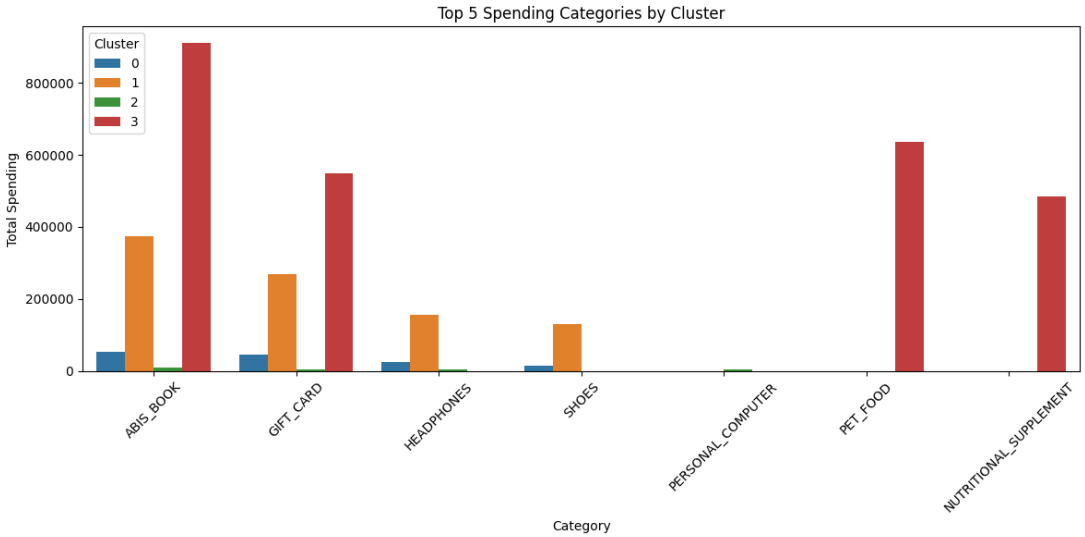
16.Analyse the Cluster Distribution by Income



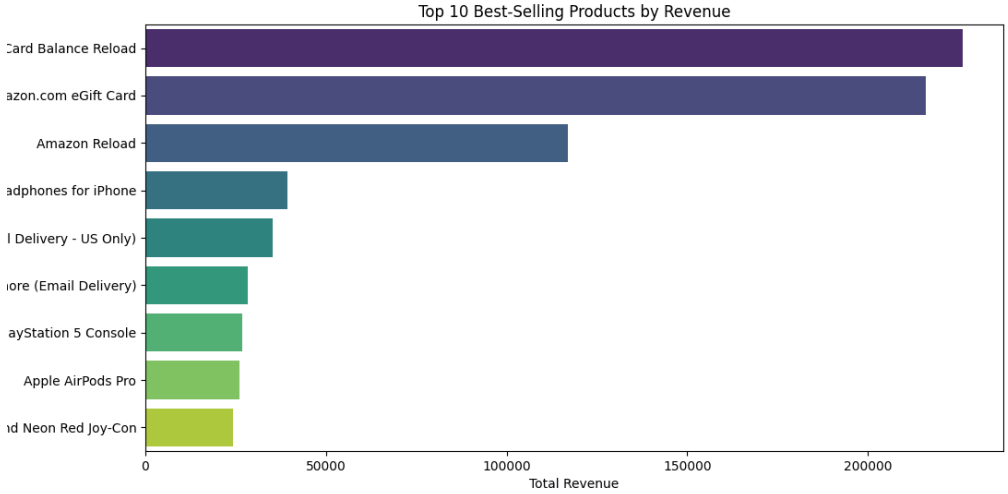
17.Analyse the Purchase Frequency vs. Recency



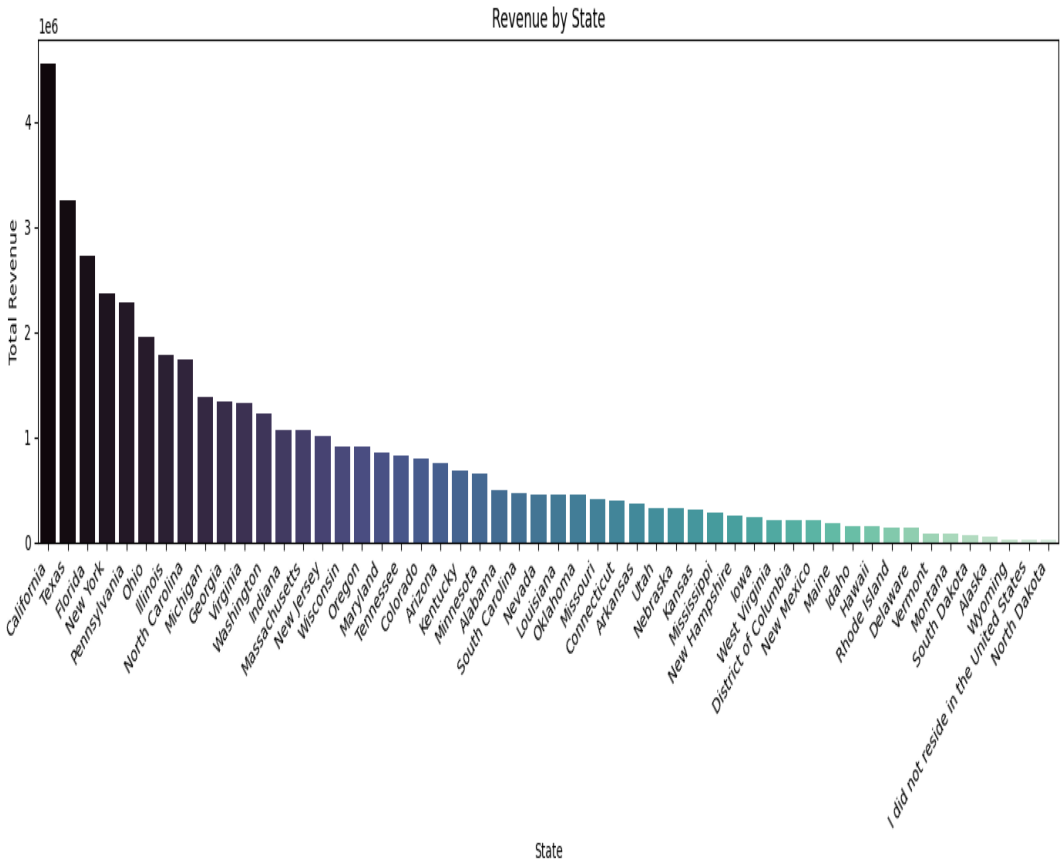
18.Analyse the top categories by clusters



19.Top 10 products by revenue



20.Revenue By State



21.Examined repeat purchase behavior to enhance retention initiatives: Data for Repeated customer and their purchase quantity

22.Flagging Potential Fraud : Data for suspicious transactions.

23.Demand Variations across product categories :Performed inventory management by monitoring demand variations across product categories.

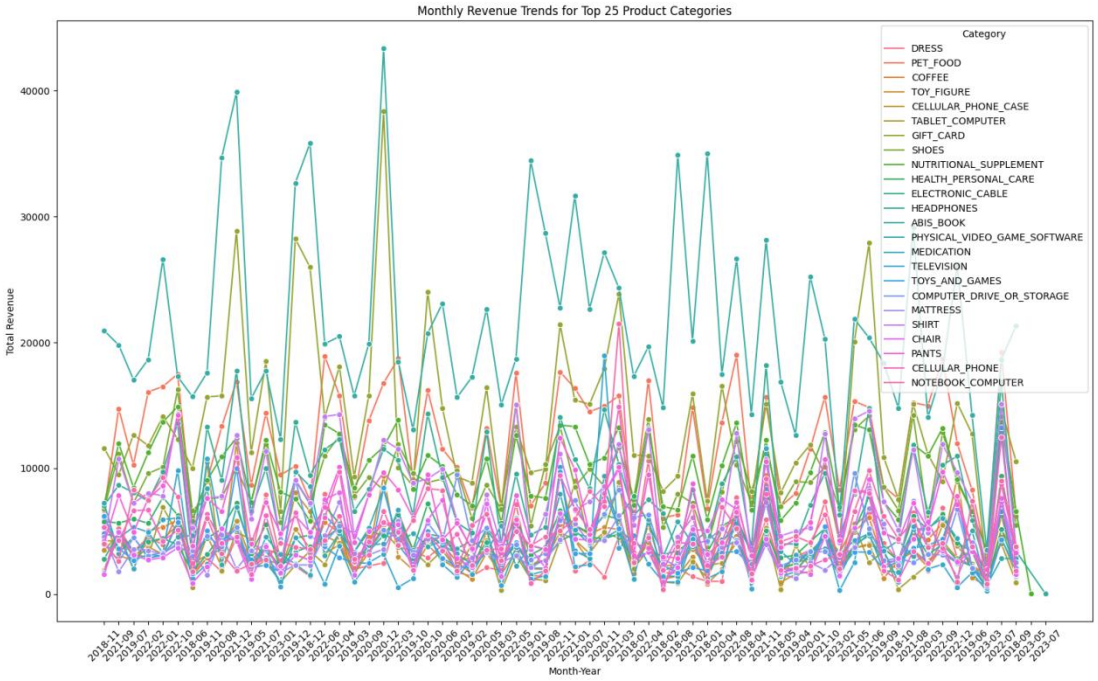
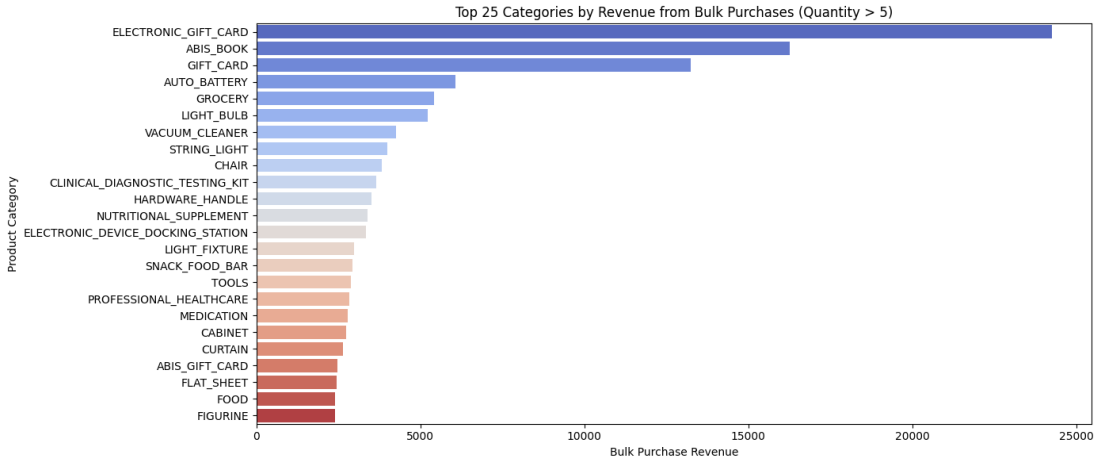
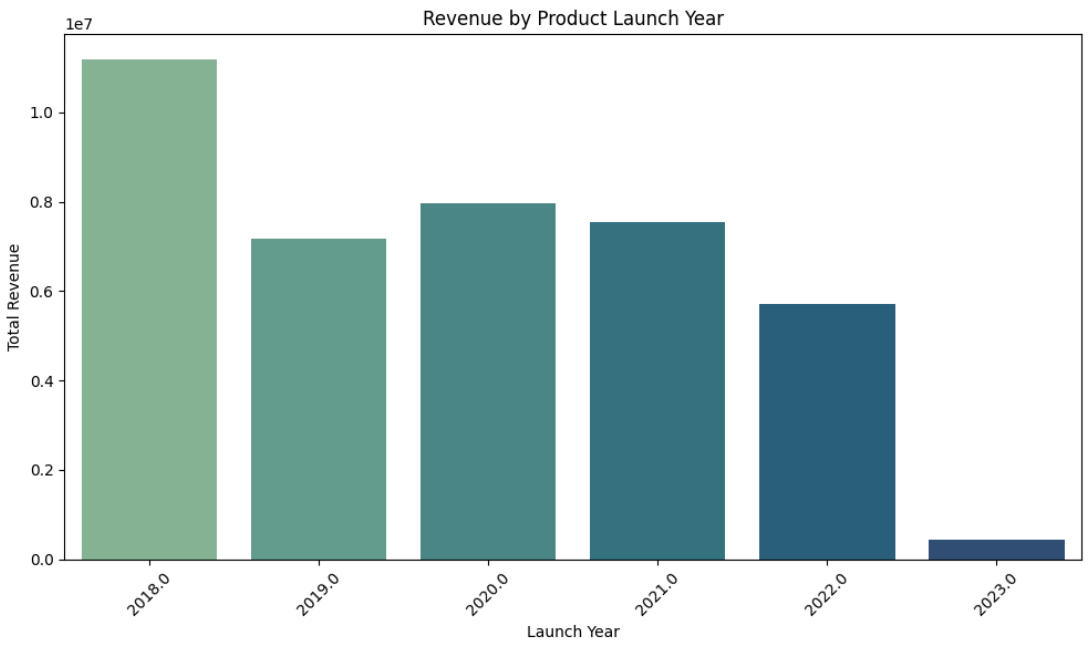


Fig : Top_25_Category_Trends.png

24.Analyse the impact of how bulk purchasing behavior affects revenue and the overall supply chain operations.



25.Compared new and established products to inform and compare lifecycle strategies to make informed decisions.



26.Conclusion: In this project, we conducted an in-depth analysis of customer purchase behavior using PySpark on a large-scale e-commerce dataset. The key objectives were to uncover trends in revenue, product performance, state-wise demand, and potential fraud indicators. Below are the major findings and insights:

a) Weekly Sales Trends: Schedule marketing campaigns and promotions around Sundays and Mondays to maximize customer engagement.

b) Top-Selling Products :Focus inventory and marketing efforts on these top performers for better ROI.

c) State Wise Revenue Distribution:Invest in targeted ads and logistics in high-performing states; explore growth strategies in underperforming regions.

d) Bulk Behaviour Purchase:ailor bulk discount campaigns for top bulk-buying categories to optimize inventory turnover.

e) Repeat Purchase Analysis: Develop retention initiatives like loyalty programs or email follow-ups to convert one-time buyers into repeat customers(Repeat customer data useful for analysis)

f) Suspicious Transaction Detection:Investigate flagged transactions further to distinguish between genuine bulk purchases and fraud attempts.

g) Category Wise Demand Trends:Use this data for inventory planning and seasonal promotions to match demand cycles.

h) Product Lifecycle Analysis: Maintain support for high-performing old products while strategically investing in launch and promotion of new items.

So, This analysis provides valuable direction for improving,Marketing strategies (based on time and region),Inventory management (through category trends and bulk behavior),Customer retention,and fraud prevention.