Customer Shopping (Latest Trends) Dataset

An Exploratory Data Analysis (EDA) Report.

BY

AMIYA SALIM

INTRODUCTION

This report presents an in-depth exploratory data analysis (EDA) of the **Customer Shopping (Latest Trends) Dataset**, which provides valuable insights into consumer behavior in the retail sector. The dataset encompasses a wide array of transactional, demographic, and behavioral data points, including detailed transaction information (such as purchase dates, transaction values, product categories, and payment methods), customer demographics (such as age group, gender, location, and loyalty status), and shopping behavior (including frequency of purchases, seasonal trends, and average spending per transaction).

The main goal of this analysis is to uncover trends, patterns, and correlations within the data that can enhance our understanding of how various factors influence consumer purchasing behavior. By analyzing purchasing habits across different customer segments, this report aims to identify key drivers of sales, the most popular product categories, and how factors like age, location, and payment methods impact overall shopping trends.

The report begins with a rigorous data cleaning process to ensure the accuracy and integrity of the dataset, handling missing values and addressing any discrepancies. Feature engineering follows to create new data columns that facilitate deeper insights into consumer behavior, such as categorizing products by seasonality or calculating the average spend per transaction for different customer groups. Various statistical and visualization techniques, such as histograms, bar charts, and heatmaps, are applied to reveal trends in product category popularity, payment methods, and customer demographics.

Ultimately, the findings from this analysis provide actionable insights that can inform retail strategy development, customer segmentation, and sales forecasting. By understanding the nuances of consumer shopping behavior, retailers can tailor their product offerings, marketing strategies, and customer engagement efforts to enhance customer satisfaction, optimize sales, and improve long-term business performance. This analysis also lays the foundation for further predictive modeling and recommendation system development in the retail industry, with the potential to drive even more personalized and targeted marketing initiatives.

DATASET DETAILS

- Number of Records: 3,900
- Number of Features: 19
- Memory Usage: (453.25 kB)

Key Features

- **Customer Information**
  - **Customer ID**: Unique identifier for customers.
  - **Age**: Customer's age in years.
  - **Gender**: Male/Female.

  - **Location**: Geographic location of the customer.
  - **Size**: Size of the purchased item (e.g., S, M, L, XL).
  - **Subscription Status**: Indicates whether the customer has an active subscription (Yes/No).
  - **Color**: Color preference of the customer for purchased items.

- **Transaction Details**
  - **Item Purchased**: Name of the purchased item.
  - **Category**: Product category.
  - **Purchase Amount (USD)**: Transaction value in dollars.
  - **Payment Method**: Method used for payment (e.g., Credit Card, PayPal).
  - **Preferred Payment Method**: Customer's preferred payment method.
  - **Season**: Time of year (e.g., Winter, Summer).
- **Behavioral Insights**
  - **Frequency of Purchases**: How often the customer makes purchases.
  - **Previous Purchases**: Total previous transactions.
  - **Review Rating**: Customer's rating (1.0–5.0).
- **Promotional Insights**
  - **Discount Applied**: Indicates if a discount was used.
  - **Promo Code Used**: Tracks promo code usage.
- **Shipping Information**
  - **Shipping Type**: Delivery method chosen.

DATA COLLECTION

Libraries such as pandas**,** numpy, and matplotlib are imported for data manipulation and visualization. After setting up the environment, the dataset is read into a DataFrame using the read_csv() function.

DATA CLEANING

1. **Checking for Duplicates** :

   An analysis was conducted to detect duplicate records in the dataset. The result confirmed that there are no duplicate entries, as the count returned is zero. This ensures that the dataset contains only unique records.

2. Missing Value Handling :

An analysis of this dataset revealed that there are zero null values, indicating that the dataset is complete with no missing data.

Customer ID	0
Age	0
Gender	0
Item Purchased	0
Category	0
Purchase Amount (USD)	0
Location	0
Size	0
Color	0
Season	0
Review Rating	0
Subscription Status	0
Payment Method	0
Shipping Type	0
Discount Applied	0
Promo Code Used	0
Previous Purchases	0
Preferred Payment Method	0
Frequency of Purchases	0

3. Outlier Detection :

Outlier detection was performed using boxplots for all numerical columns in the dataset. The analysis revealed that no outliers were present, as all data points fell within the expected range.

4. Data Type Conversion :

The data types of all columns were reviewed and found to be appropriately assigned.

Feature Engineering

Feature engineering involves creating new features or modifying existing ones to improve model performance. It enhances the dataset's relevance and helps machine learning models make more accurate predictions.

Age Group Binning:

The Age column was binned into categories like "Young," "Middle-Aged," and "Older," enabling a demographic analysis based on age segments.

- Process: By assigning each customer to a specific age group, a new "Age Group" column was created to categorize customers based on their age. This helps in analyzing patterns and trends more effectively across different age segments.

- Purpose: The goal of age group binning is to better understand and analyze customer behavior by segmenting them into age categories. This helps in tailoring marketing efforts and product offerings to meet the specific needs of each group, improving customer targeting and sales strategies.

Purchase Frequency Binning :

The purpose is to group customers' purchase frequencies into categories (Regular, Occasional, Rare Purchases) for more focused analysis and strategy development.

- Process:
  1) **Mapping Frequencies:** Using the purchase_group_map, we categorize purchase frequencies into three groups.
     - Regular Purchases (Weekly, Fortnightly, Bi-Weekly)
     - Occasional Purchases (Monthly, Quarterly, Every 3 Months)
     - Rare Purchases (Annually).
  2) **Applying the Mapping:** The .map() function adds a new Frequency Group column to the dataframe based on these categories.

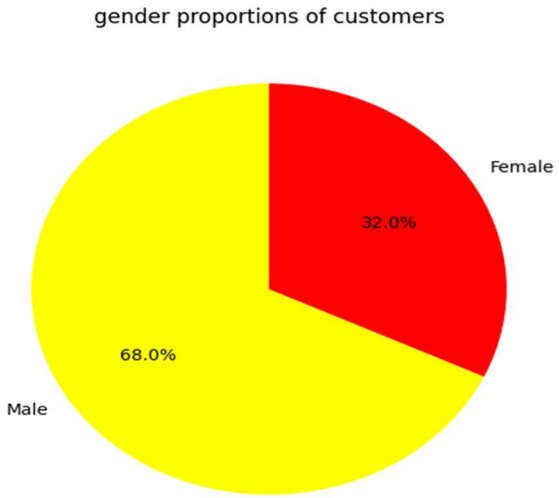
- Purpose : This simplifies analysis, helps in targeting specific customer groups, and provides actionable insights for marketing strategies and customer retention.

DATA ANALYSIS

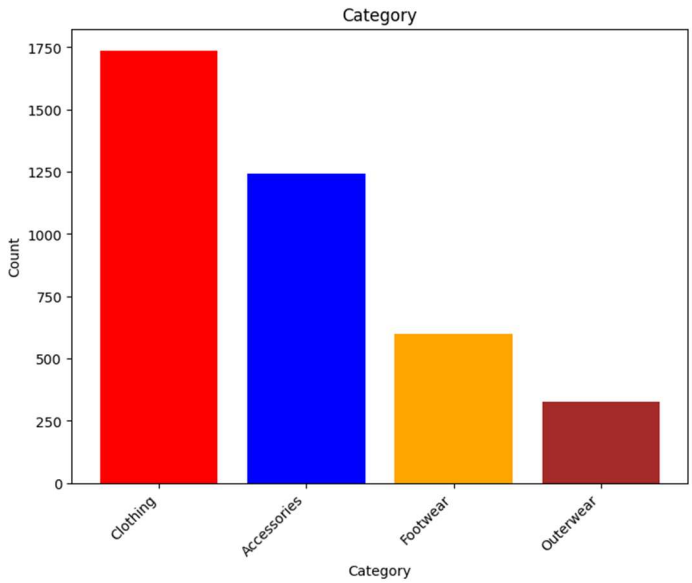
Customer Demographics

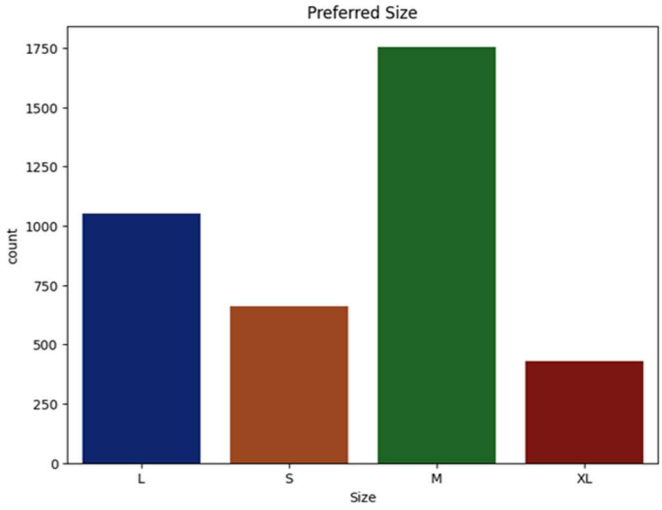
i. Gender proportions of customers

We can see from the below distribution that Male demographic seems to be the majority with a 68% of the total customers, while Female represent the remaining 32%.



ii. Category Available & Preferred Size



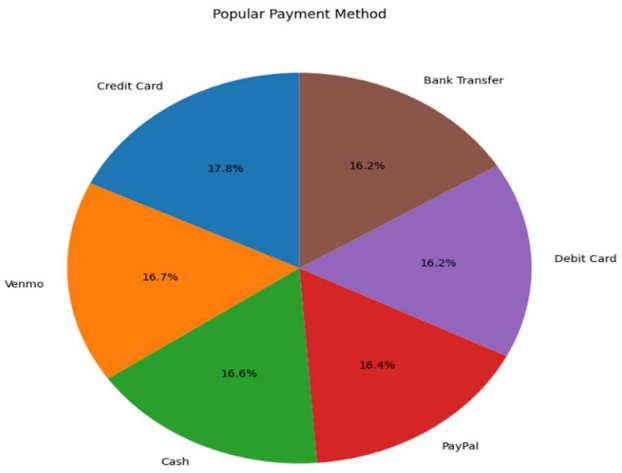


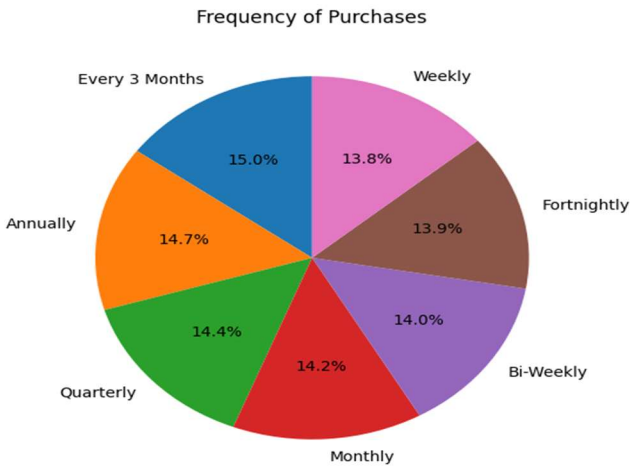
Insight:-

Category: -Clothing itmes are the most purchased ones by a big margin, followed by Accessories, Footwear, and Outerwear.

Size:- A large portion of the customers order Medium sized products, suggesting people just choose the middle ground as not to accidently purchasing too big or small items.

iii. Popular Payment Method & Frequency of Purchases



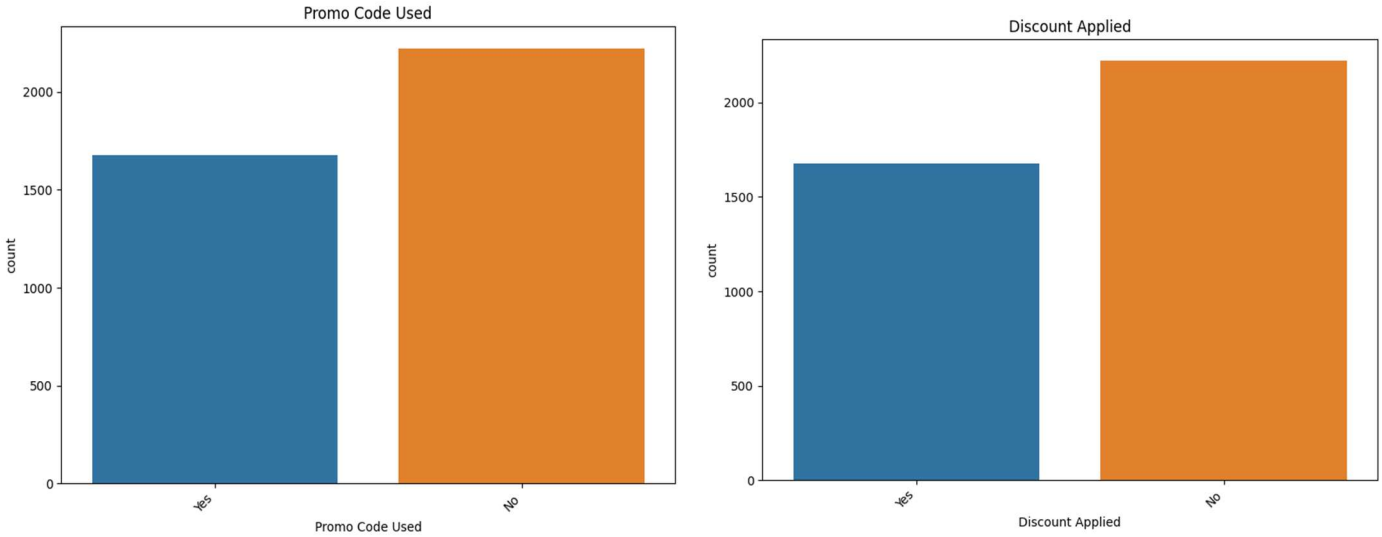


Insight:-

Payment Method :- Majority of the customers use Credit Card for payment, 17.8% while the distribution for the rest are around 16%.

Frequency of Purchases :- Every unique instance of purchasing time period has same distribution, with Every 3 Months instance having highest distribution by a very small margin of 15% while the rest being at around 14%

iv. Discount Applied & Promo Code Used



Insight:-

Discount Applied: Looking at the distribution most people purchase the products with no discount.

Promo Code: Just like discount, most people don't use promo codes or they don't have them.

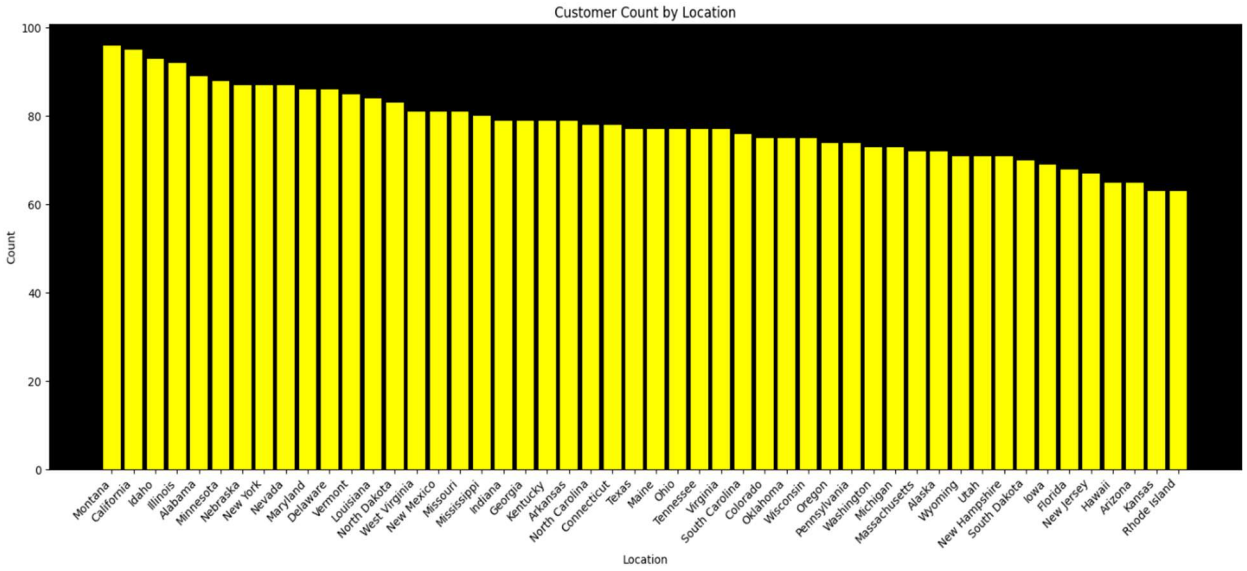
v. Customer most Purchased, Location and Preferred Color

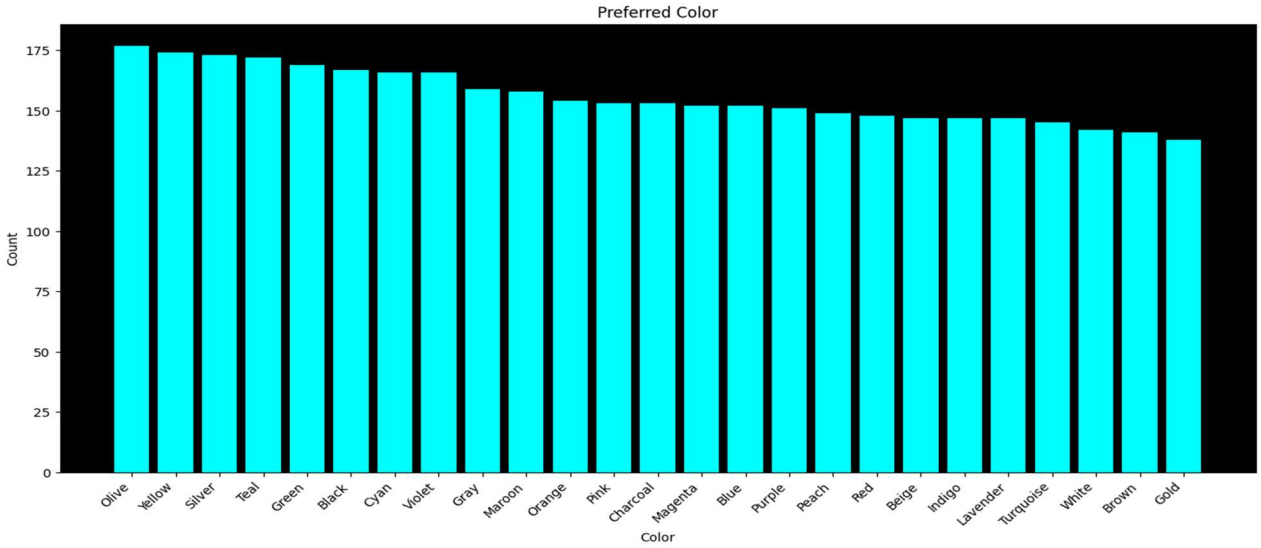
Insight:-

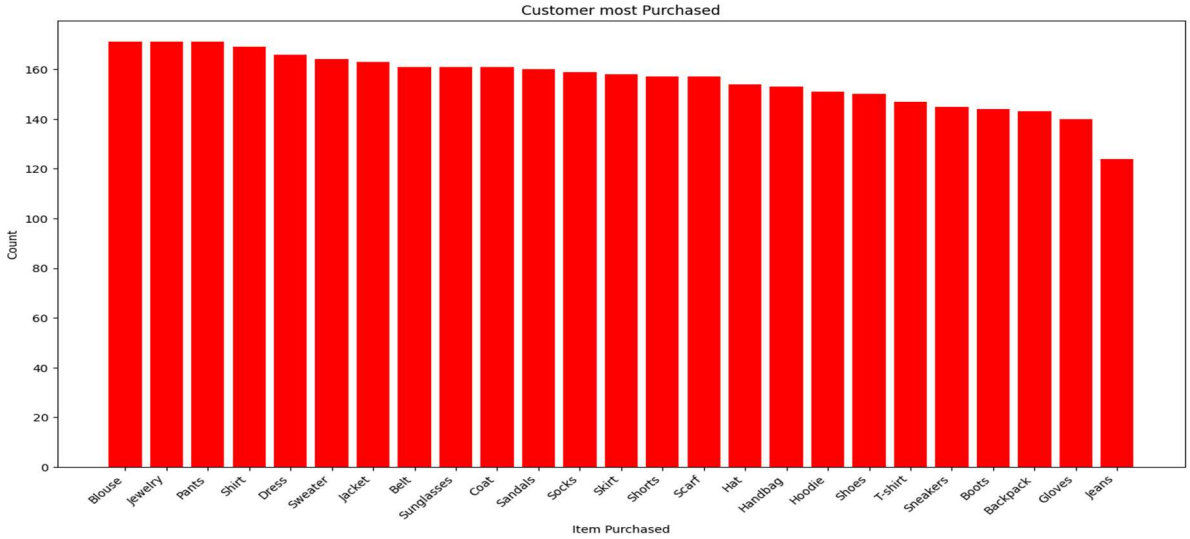
Color Preferences: Olive, Yellow, and Silver are the most popular colors, while Gold and Brown are least preferred.

Location Insights: Montana, California, and Idaho lead in purchases, with Kansas and Rhode Island having the fewest.

Top Items: Blouse, Jewelry, and Pants are the top items, each with 171 purchases, followed by Shirt, Dress, and Sweater

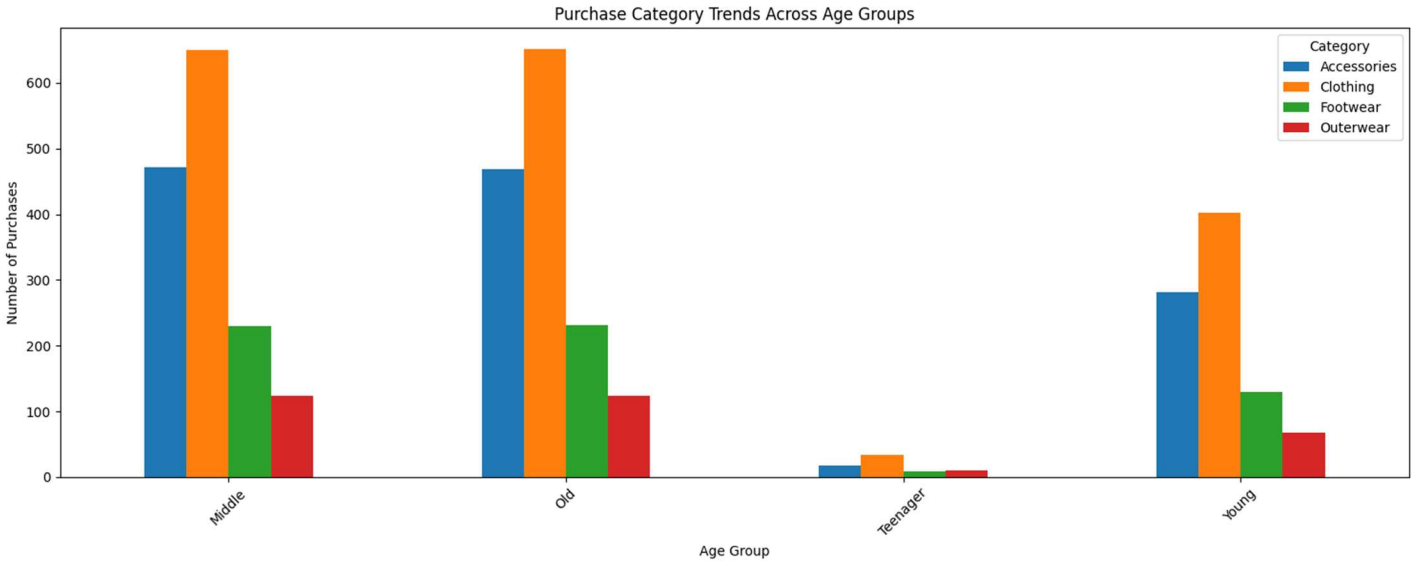

Customer Count by Location
Count
Location
Montana
California
Idaho
Illinois
Alabama
Minnesota
Nebraska
New York
Nevada
Maryland
Delaware
Vermont
Louisiana
North Dakota
West Virginia
New Mexico
Missouri
Mississippi
Indiana
Georgia
Kentucky
Arkansas
North Carolina
Connecticut
Texas
Maine
Ohio
Tennessee
Virginia
South Carolina
Colorado
Oklahoma
Wisconsin
Oregon
Pennsylvania
Washington
Michigan
Massachusetts
Alaska
Wyoming
Utah
New Hampshire
South Dakota
Iowa
Florida
New Jersey
Hawaii
Arizona
Kansas
Rhode Island


Preferred Color
Count
Color
Olive
Yellow
Silver
Teal
Green
Black
Cyan
Violet
Gray
Maroon
Orange
Pink
Charcoal
Magenta
Blue
Purple
Peach
Red
Beige
Indigo
Lavender
Turquoise
White
Brown
Gold


Customer most Purchased
Count
Item Purchased
Blouse
Jewelry
Pants
Shirt
Dress
Sweater
Jacket
Belt
Sunglasses
Coat
Sandals
Socks
Skirt
Shorts
Scarf
Hat
Handbag
Hoodie
Shoes
T-shirt
Sneakers
Boots
Backpack
Gloves
Jeans

Detailed Analysis

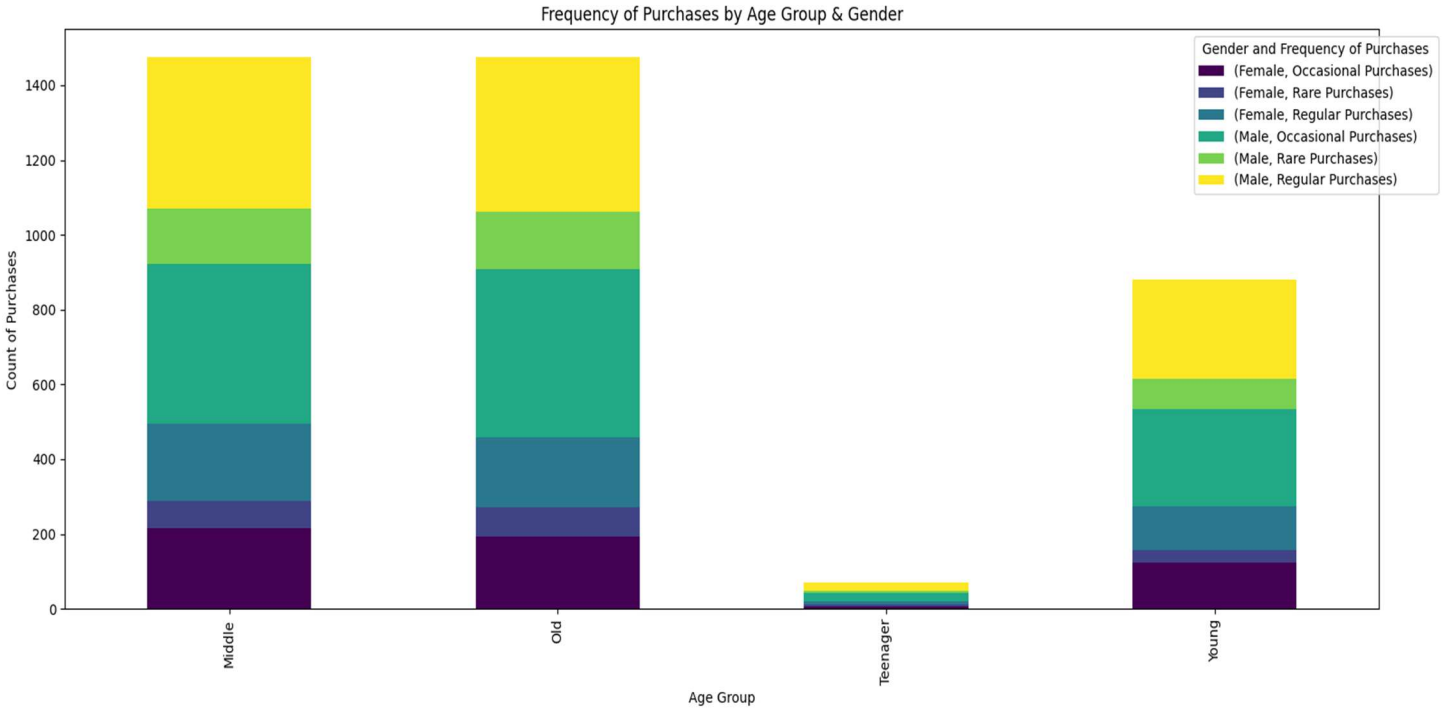
vi. Purchase Category Trends Across Age Groups



Insight:-

Middle-aged and older adults dominate purchases, with clothing being the most popular category. Accessories and footwear follow, while outerwear sees minimal demand. Teenagers contribute the least, and young adults focus mainly on clothing, suggesting marketing efforts should target older age groups with a focus on clothing.

vii. Frequency of Purchases by Age Group & Gender



Insight:-

Key insight:

- Purchase frequency increases with age, with males generally making more purchases than females. Males dominate in Regular and Occasional Purchases, especially in Middle and Old groups. Females are highly engaged in Regular Purchases, notably in the Middle and Old groups. Teenagers show minimal activity, while Young and Middle age groups are the most active, offering prime marketing opportunities.

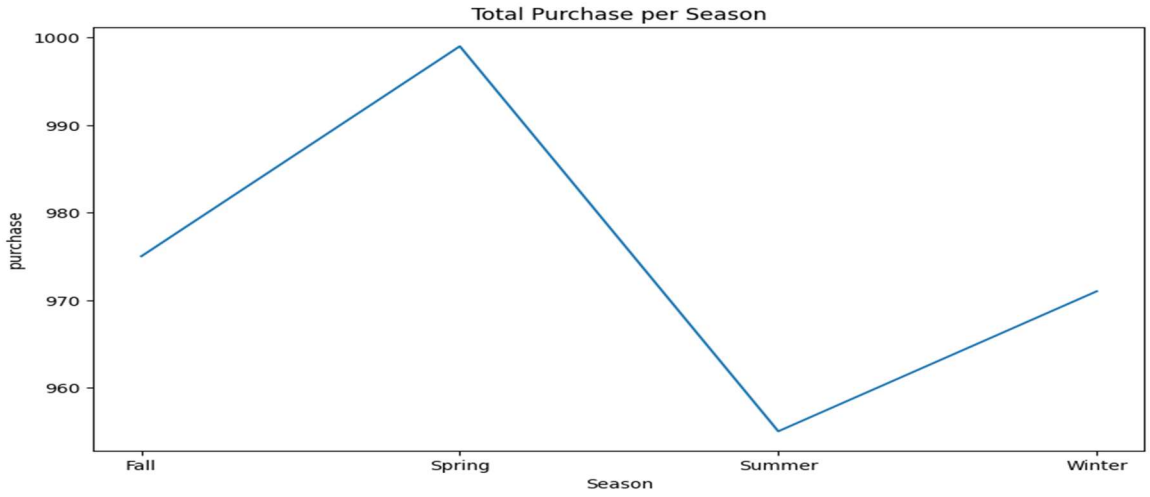
Middle Age (31-40): Purchase frequency is high for both genders, with males leading.

Old Age (41-50): Significant increase in purchases for both genders, with males remaining dominant.

Teenagers (10-20): Low purchase frequency for both males and females.

Young Age (21-30): Purchase frequency rises for both genders, with males in the lead.

viii. Total Purchase per Season



Insight:-

Fall: *Moderate purchase activity.*

Spring: *Highest purchase frequency.*

Summer: *Lowest purchase activity.*

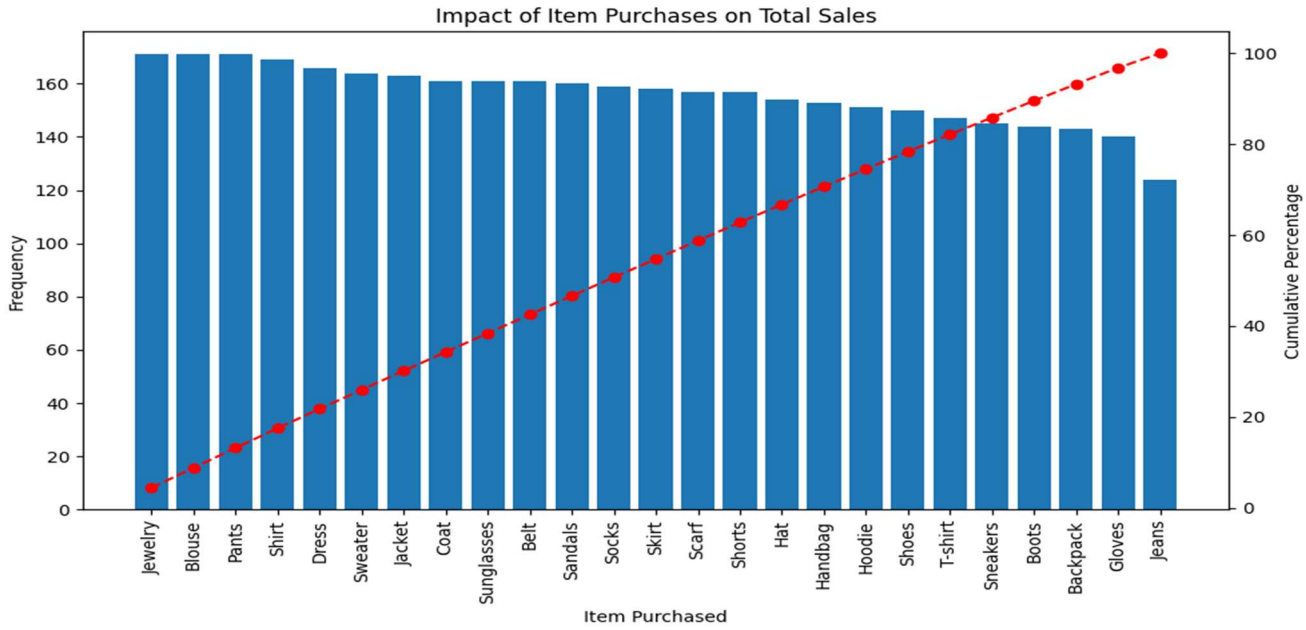
Winter: *Moderate to high purchase frequency, recovering from the summer dip*.

ix. Impact of Item Purchases on Total Sales

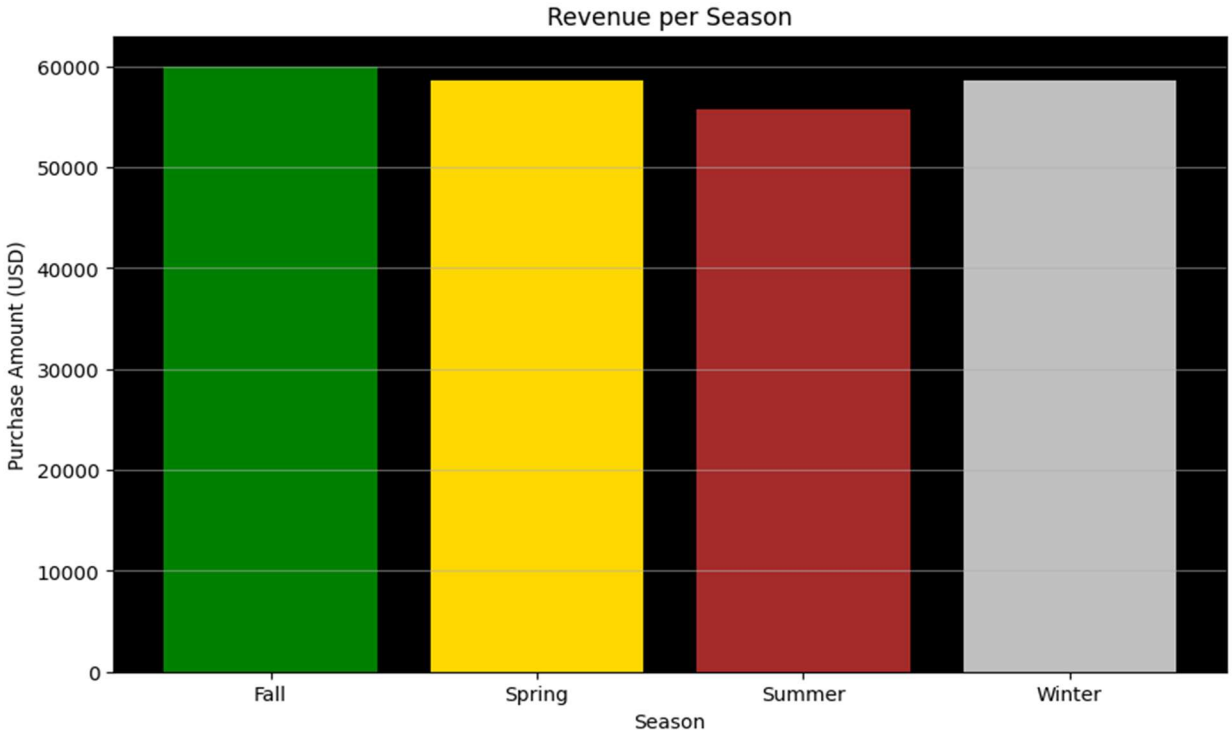
Insight:-

A few items, like "Blouse," "Pants," and "Jewelry," make up most of the total purchases, while many others have low sales, forming a "long tail." The cumulative percentage shows that top-

performing items dominate sales, with diminishing returns as you move down the list. Businesses can focus on promoting best-sellers to boost sales effectively.



x. Revenue per Season



Insight:-

Overall: *Revenue remains relatively consistent across all seasons.*

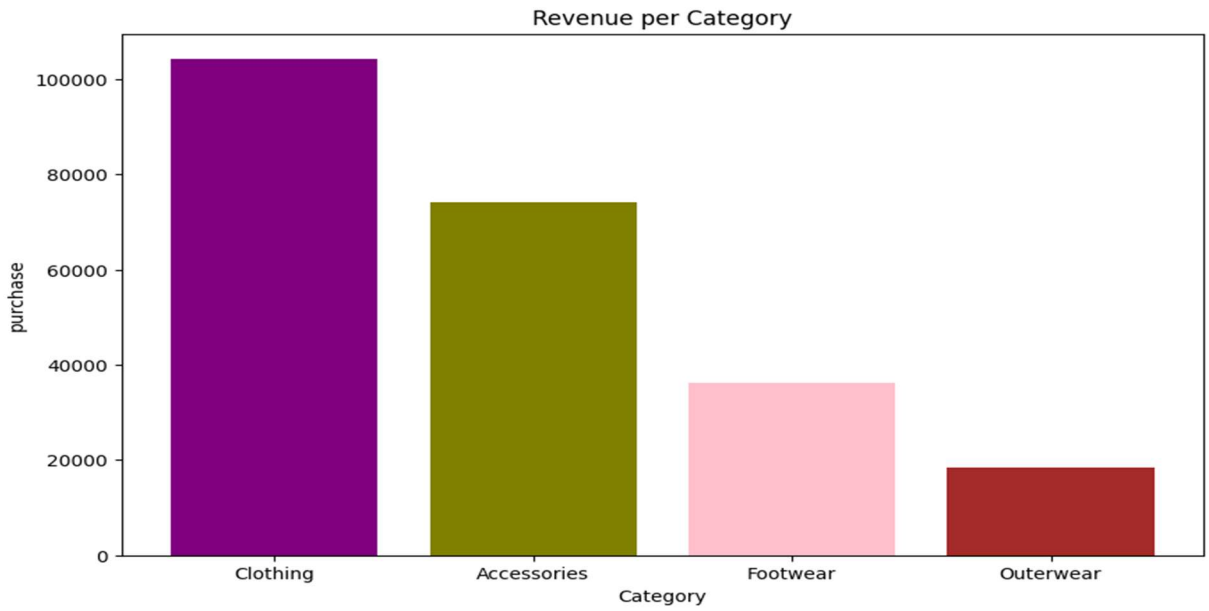
Fall: *Highest revenue.*

Spring: *Second highest revenue.*

Winter: *Third highest revenue.*

Summer: *Lowest revenue*

xi. Revenue per Category



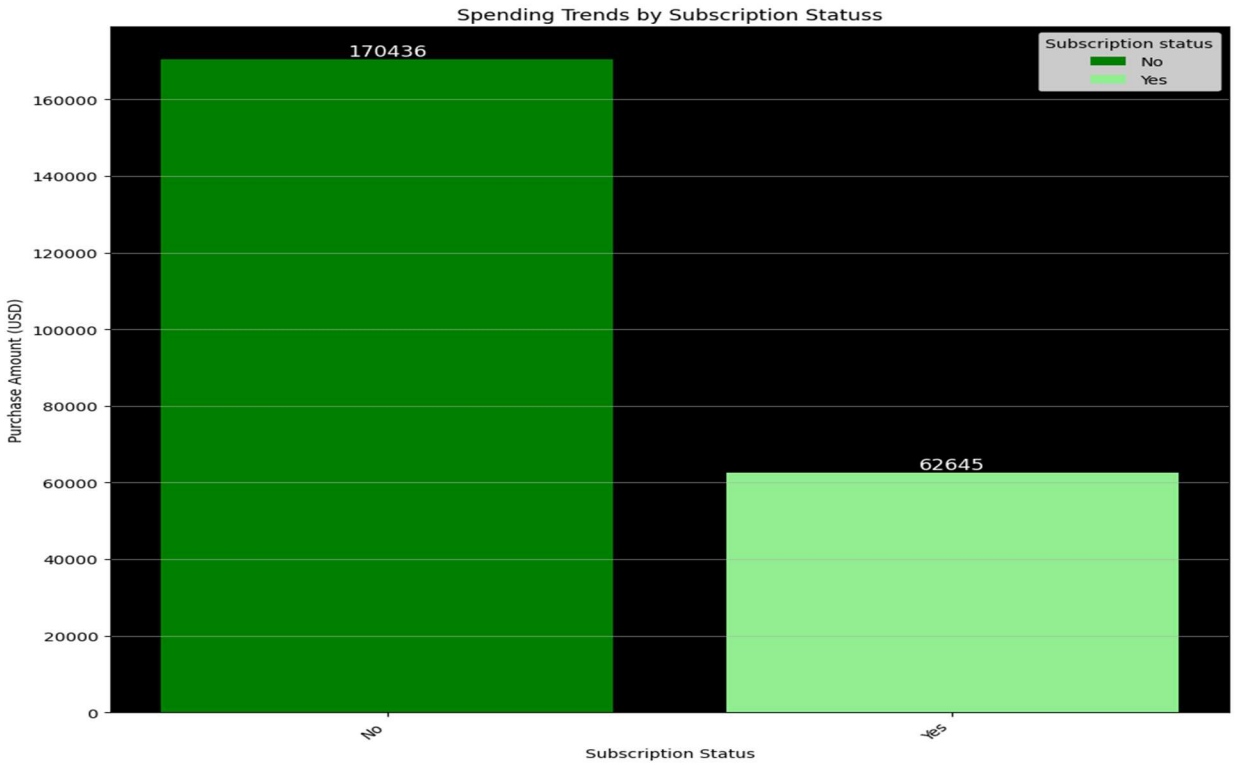
Insight:-

Clothing generates the highest revenue, followed by Accessories and Footwear. Outerwear has the lowest revenue.

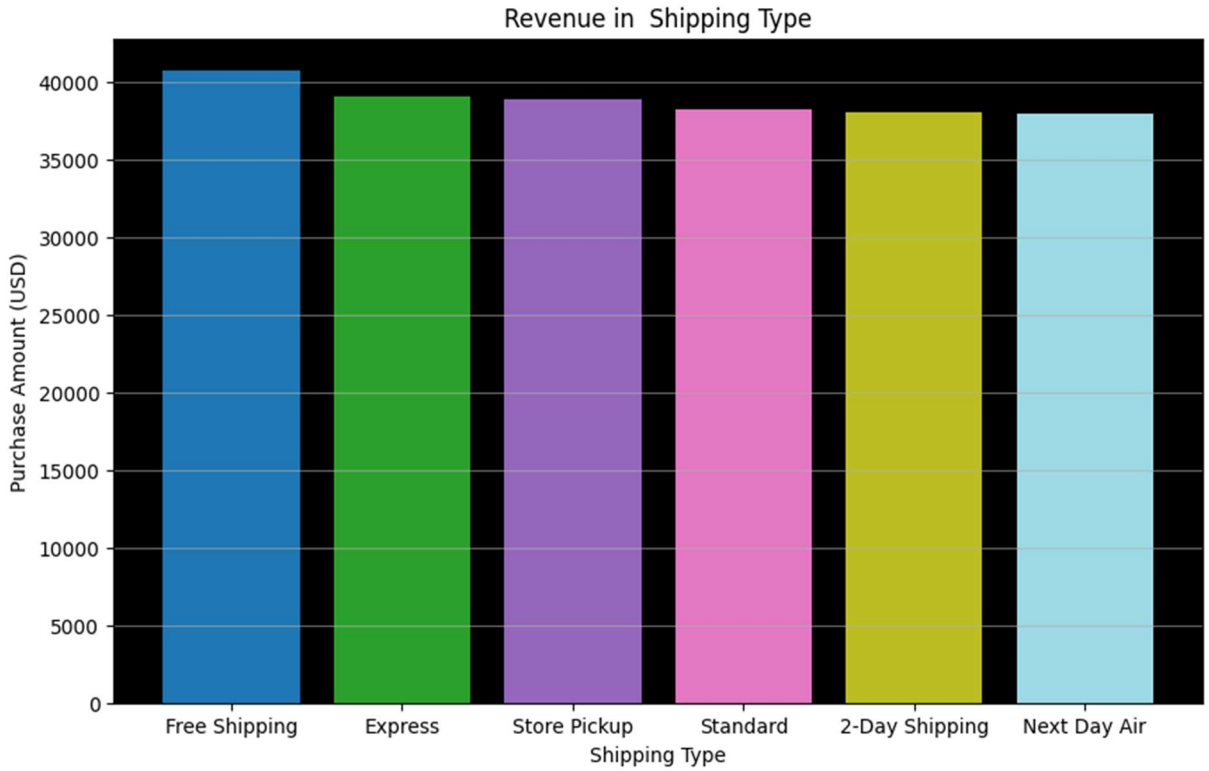
xii. Purchase Amount Analysis by Subscription Status

Insight:-

Non-subscribers have a slightly higher average purchase amount ($59.87) compared to subscribers ($59.49). However, the total purchase amount for non-subscribers is significantly higher ($170,436 vs. $62,645), indicating that non-subscribers contribute more to overall sales despite a smaller average spend. This suggests that increasing subscriptions could help drive repeat purchases while maintaining strong overall revenue.



xiii. Revenue in Shipping Type



Insight:-

The analysis shows that shipping methods like express shipping lead to higher spending, indicating customers are willing to pay more for faster delivery. Free shipping results in comparatively lower spending. They also boosts purchases, suggesting that promotions offering free shipping can drive sales mean.

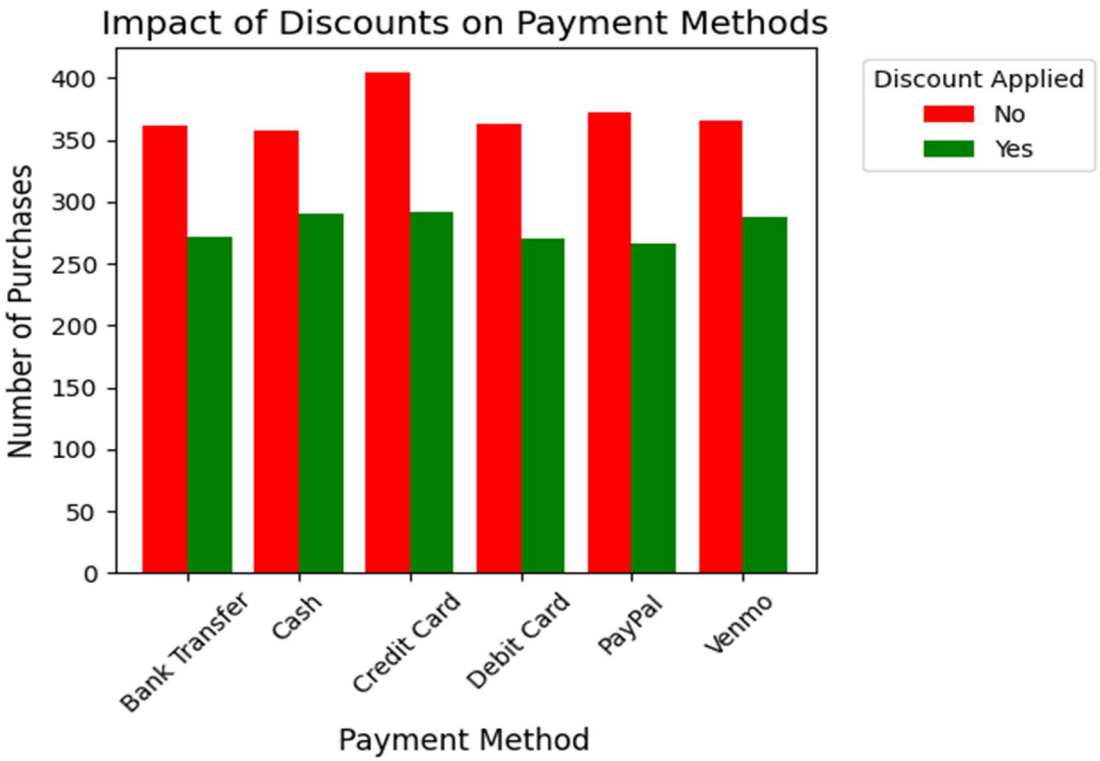
1. Free Shipping leads in total purchase amount, suggesting that customers are likely to make more purchases when free shipping is offered.

2. Express Shipping comes in second, indicating that customers are willing to pay for faster delivery.

3. Store Pickup and Standard shipping are closely aligned, showing moderate spending with these options.

4. 2-Day Shipping and Next Day Air have the lowest total purchase amounts, suggesting these options might be less popular or have fewer customers choosing them

xiv. Impact of Discounts on Payment Method



Insight:-

The analysis shows that while most customers prefer full-price purchases across all payment methods, a notable portion still opts for discounts. Cash, Venmo, and PayPal users show a moderate preference for discounts, whereas Credit Card and Bank Transfer users tend to buy at full price.

1. Bank Transfer: 361 non-discounted and 271 discounted purchases, showing a preference for full-price purchases.

2. Cash: 358 non-discounted and 290 discounted purchases, with a moderate inclination toward discounts.

3. Credit Card: 404 non-discounted and 292 discounted purchases, indicating high overall purchases but a lower percentage of discounts.

4. Debit Card: 363 non-discounted and 270 discounted purchases, with a preference for non-discounted items.

5. PayPal: 372 total, with 266 discounted purchases, showing a slightly higher preference for discounts.

6. Venmo: 365 non-discounted and 288 discounted purchases, indicating a moderate balance between both.

Conclusion

The analysis of consumer purchasing behavior from the dataset offers a comprehensive understanding of key trends and actionable insights for retail optimization. Males tend to make more purchases across all age groups, particularly in the 31-50 range, while females show a balanced distribution between regular and occasional purchases. Age-wise, the purchase frequency increases with age, with older age groups showing higher engagement in purchases, though males consistently lead in total purchase volumes. Payment methods such as Venmo, PayPal, and Credit Card significantly influence purchasing behavior, with discounts enhancing engagement, especially when paired with these methods. Popular items like Blouses, Sweaters, and Jeans dominate the sales, while certain regions, including Montana and California, show higher engagement, indicating potential for region-specific marketing. Analyzing shipping types reveals preferences for quicker delivery options such as Express and Next Day Air, particularly for higher-priced items like Handbags and Jackets, which can guide logistics strategies. Additionally, reviewing subscription status indicates that subscribers tend to make more frequent purchases, further supporting the need for tailored subscription-based offers. The long tail of low-frequency items suggests opportunities for cross-selling or targeted promotions to boost sales in underperforming categories. These findings highlight the importance of personalized marketing, strategic discount offerings, region-based campaigns, and product-specific optimizations to drive higher sales, enhance customer loyalty, and maximize retail revenue.