CUSTOMER SEGMENTATION

Devanaga Saputra

1

2

3

4

5

6

7

8

9

BUSINESS UNDERSTANDING

The Superstore Sales dataset provides comprehensive sales data for multiple products sold by a retail superstore. This dataset is suitable for exploring various aspects of sales analysis, including trend analysis, product performance, geographical segmentation, and consumer behavior. This dataset can be accessed at the link https://www.kaggle.com/datasets/aditisaxena20/superstore-sales-dataset/code

BUSINESS QUESTION

- Who are our most profitable customers?
- What are the distinct characteristics of each customer segment?
- Which products or services are preferred by different segments?

3. Data Understanding

1
2
3
4
5
6
7
8
9

- order_id: A unique identifier for each order.
- order_date: The date when the order was placed.
- ship_date: The date when the order was shipped.
- ship_mode: The method used for shipping the order.
- customer_name: The name of the customer who placed the order.
- segment: The market segment to which the customer belongs.
- state: The state where the order was placed.
- country: The country where the order was placed.
- market: The market in which the order was placed.
- region: The region in which the order was placed.
- product_id: A unique identifier for each product.
- category: The broad category to which the product belongs.
- sub_category: A more specific category within the broader category.
- product_name: The name of the product.
- sales: The total sales amount for the order.
- quantity: The quantity of products ordered.
- discount: The discount applied to the order.
- profit: The profit generated from the order.
- shipping_cost: The cost associated with shipping the order.
- order_priority: The priority level of the order.
- year: The year in which the order was placed.

4. Data Peparation

1
2
3
4
5
6
7
8
9

Program Language: Python

Packages:

Numpy

Pandas

Matplotlib

Seaborn

Plotly

Sklearn

Yellowbrick

Tabulate

Dateutil

4.1. Data Cleansing

This data has no missing values and also no duplicate values.

4.2. Dataset Overview

Inferences of Summary Statistics

- Quantity: The average quantity per order is around 3.48 items, with a standard deviation of approximately 2.28. The quantity ranges from 1 to 14 items.
- Discount: On average, there is a discount of about 14.29% per order, with a standard deviation of approximately 21.23%. The discount varies from 0% to 85%.
- Profit: The mean profit per order is roughly $28.64, with a considerable standard deviation of approximately $174.42. Profit ranges from - $6599.98 to a gain of $8399.98.
- Shipping Cost: The average shipping cost per order is around $26.38, with a standard deviation of approximately $57.30. Shipping costs range from $0.00 to $933.57.

1
2
3
4
5
6
7
8
9

Inferences of Summary Statistics

- order_id: There are 25,035 unique order IDs, with "CA-2014-100111" being the most frequent, occurring 14 times.
- ship_mode: There are 4 unique shipping modes, with "Standard Class" being the most frequent, occurring 30,775 times.
- customer_name: There are 795 unique customer names, with "Muhammed Yedwab" being the most frequent, occurring 108 times.
- segment: There are 3 unique market segments, with "Consumer" being the most frequent, occurring 26,518 times.
- country: There are 147 unique countries, with "United States" being the most frequent, occurring 9,994 times.
- market: There are 7 unique markets, with "APAC" (Asia-Pacific) being the most frequent, occurring 11,002 times.
- region: There are 13 unique regions, with "Central" being the most frequent, occurring 11,117 times.
- product_id: There are 10,292 unique product IDs, with "OFF-AR-10003651" being the most frequent, occurring 35 times.
- category: There are 3 unique categories, with "Office Supplies" being the most frequent, occurring 31,273 times.
- sub_category: There are 17 unique sub-categories, with "Binders" being the most frequent, occurring 6,152 times.
- product_name: There are 3,788 unique product names, with "Staples" being the most frequent, occurring 227 times.
- order_priority: There are 4 unique order priorities, with "Medium" being the most frequent, occurring 29,433 times.

4.3. Data Transformation


```
: df['profit'].describe()
```

```
: count    51290.000000
  mean        28.641740
  std        174.424113
  min      -6599.978000
  25%          0.000000
  50%          9.240000
  75%         36.810000
  max       8399.976000
  Name: profit, dtype: float64
```

Negative values are found in the profit variable, which is inconsistent with the meaning of profit itself. Therefore, the treatment for this case is to remove data with profit values < 0.

4.4. Feature Engineering

RFM stands for Recency, Frequency, and Monetary Value, and it is a method used in marketing analysis to segment customers based on their transaction history:

- **Recency (R): Refers to how recently a customer has made a purchase. It is usually measured in terms of the number of days since the last purchase. Customers who have made a purchase more recently are considered to be more valuable.**
  - **Days_Since_Last_Purchase**
- **Frequency (F): Indicates how often a customer makes a purchase within a specific period, such as a month or a year. Customers who make purchases more frequently are generally considered more valuable.**
  - **Total_Transactions**
  - **Total_Products_Purchased**
  - **Unique_Products_Purchased**
- **Monetary Value (M): Represents the total amount of money a customer has spent on purchases. This metric helps identify high-value customers who contribute significantly to revenue.**
  - **Total_Profit**
  - **Average_Transaction_Profit**

Behavior features capture the actions and interactions of customers with a business. These features provide insights into the purchasing patterns, preferences, and engagement levels of customers.

- **Day_Of_Week**
- **Is_SC**
- **Buy_OS**
- **Number_Using_Discount**
- **Discount_Rate**

4.4. Feature Engineering

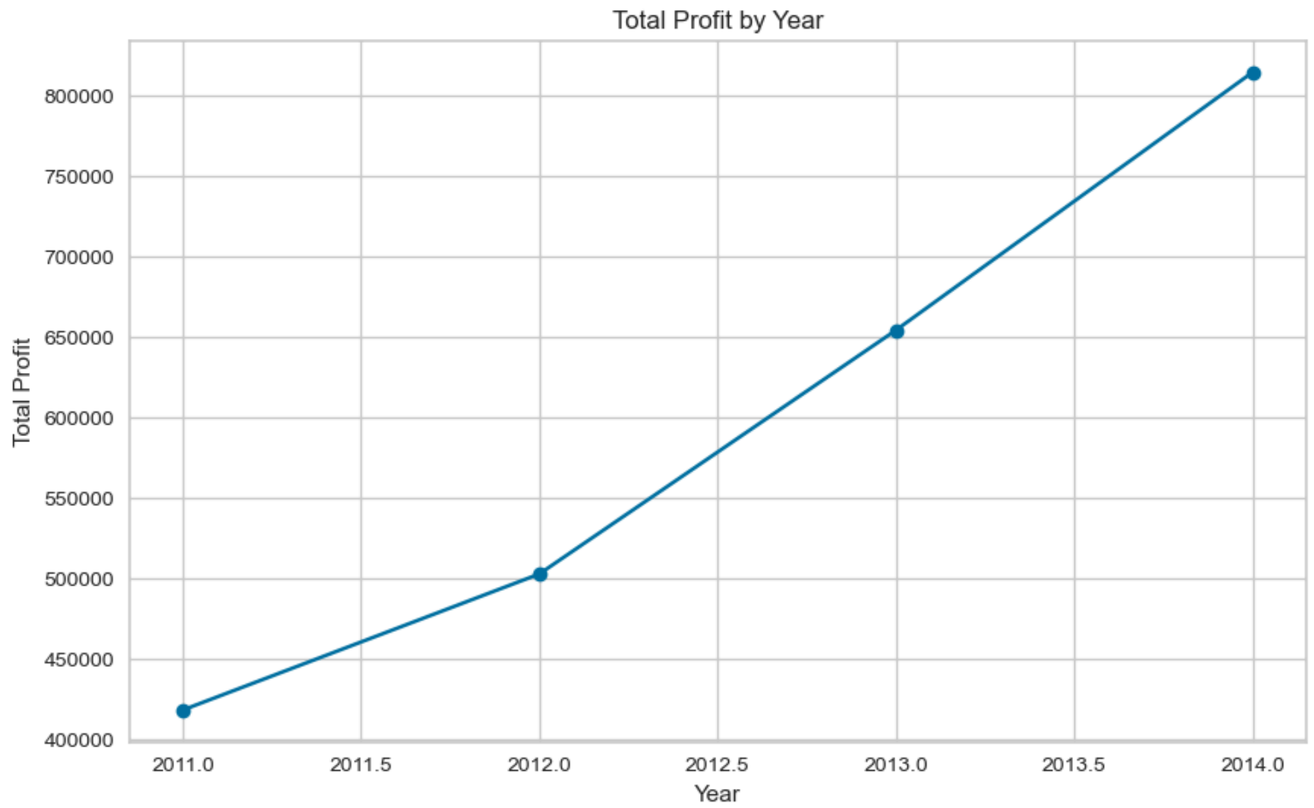
Geographic features refer to the location-based attributes associated with customer data. These features provide insights into where customers are located and how geographic factors influence their purchasing behavior.

- **In_Americas**

Profit trend features analyze the changes in profitability over time. These features help in understanding how profit margins and overall profitability evolve, often in relation to other variables like sales, costs, and market conditions.

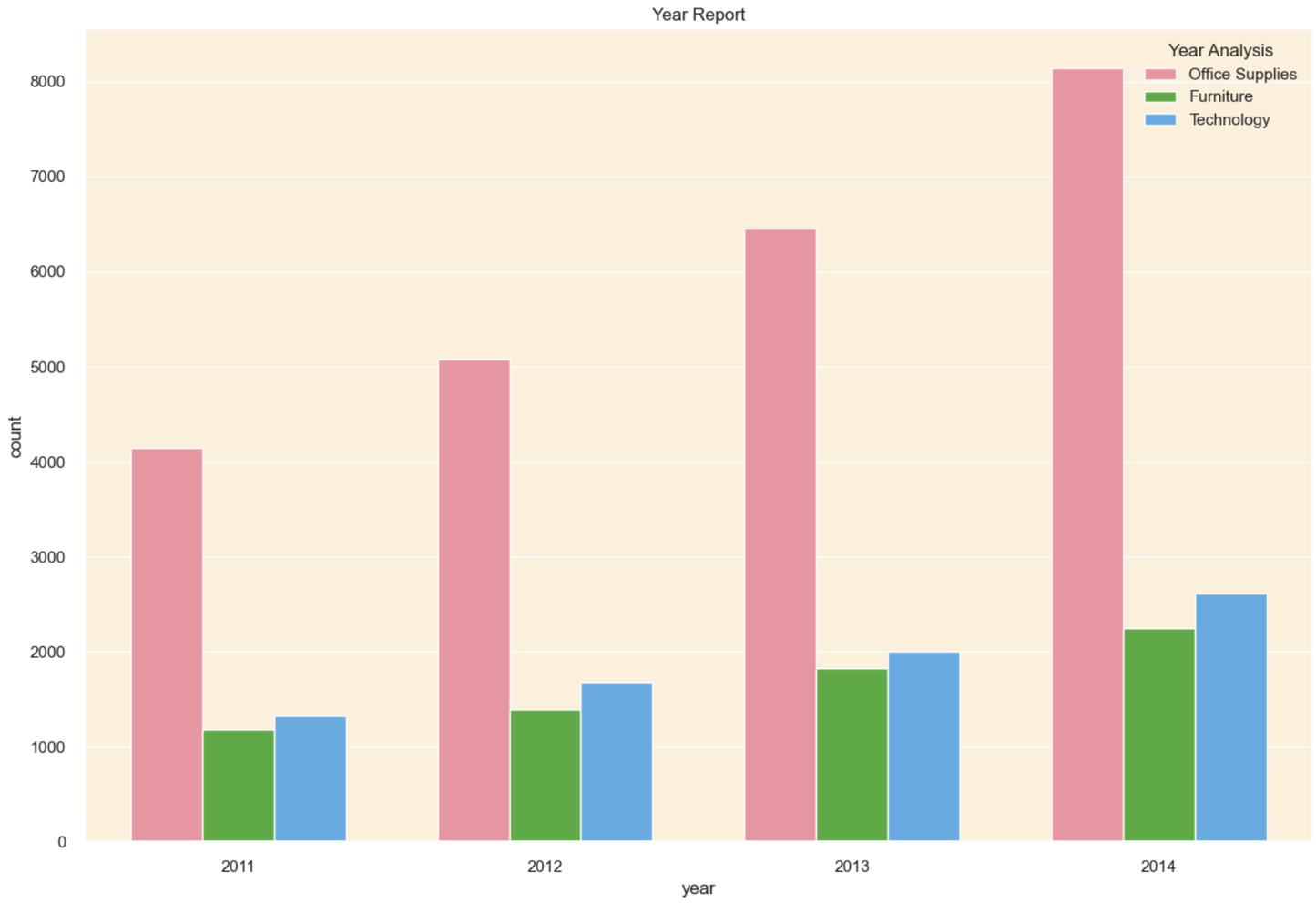
- **Monthly_Profit_Mean**
- **Monthly_Profit_Std**
- **Spending_Trend**

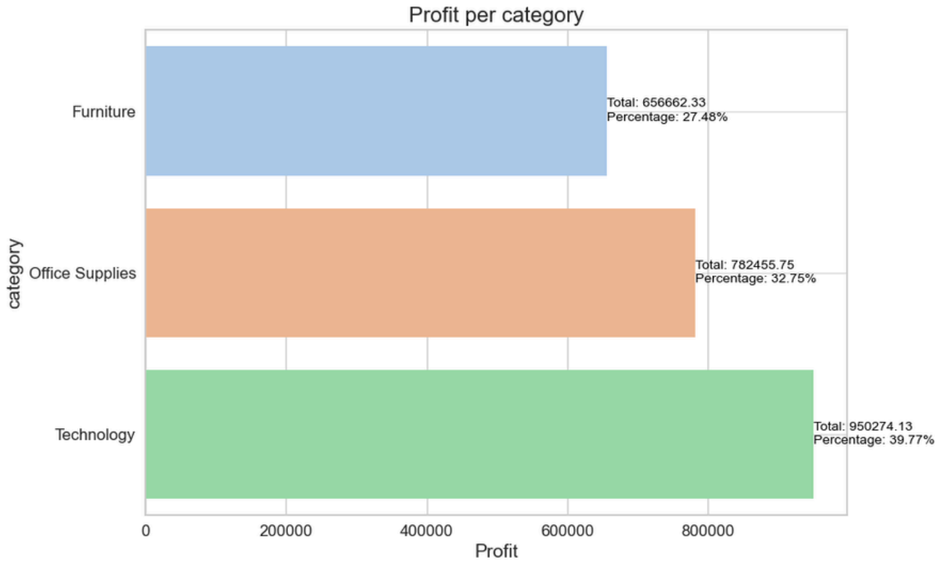
5. EDA

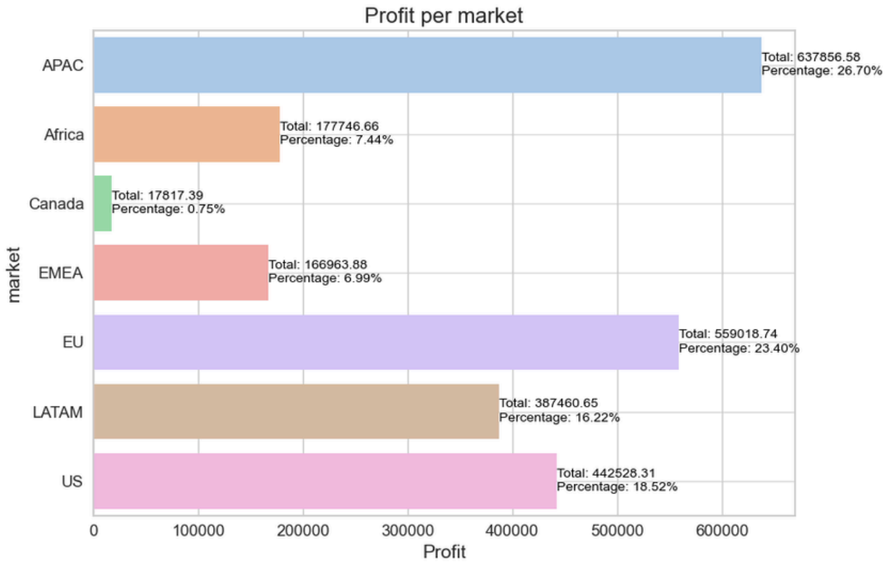


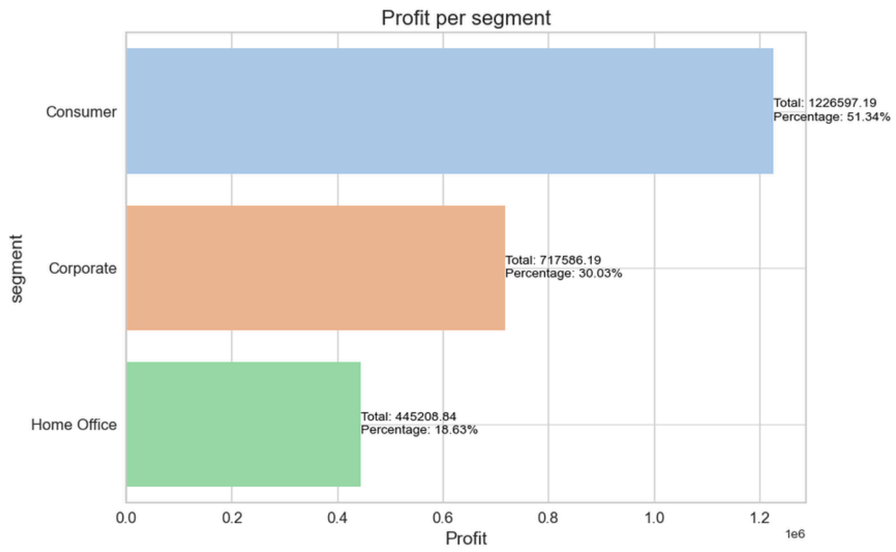
It appears that the profit has consistently increased year by year, especially noticeable from 2012 to 2014, indicating significant growth in profitability over this period.

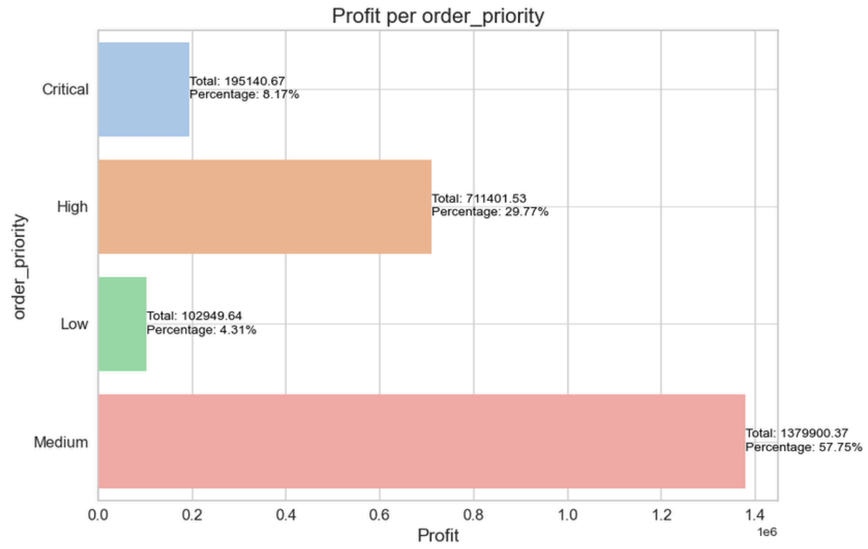
- **Increase in the number of businesses and offices: This period may have witnessed growth in the number of businesses and offices, which require office supplies such as pens, paper, erasers, and others.**
- **Development of office technology: Despite technological advancements, there is still a significant need for traditional office supplies. However, there have also been advancements in office technology that may have influenced the interest in office supplies.**



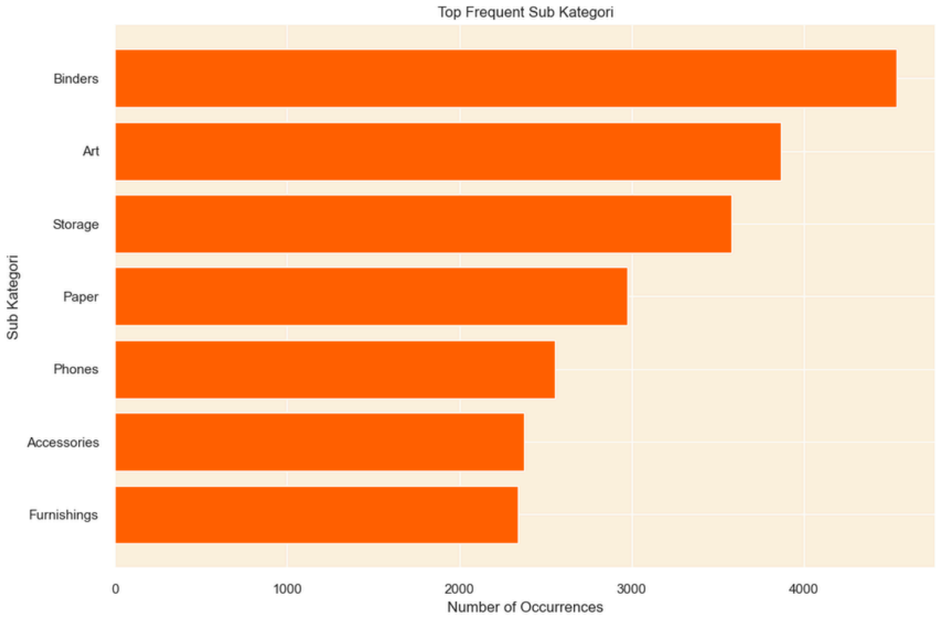


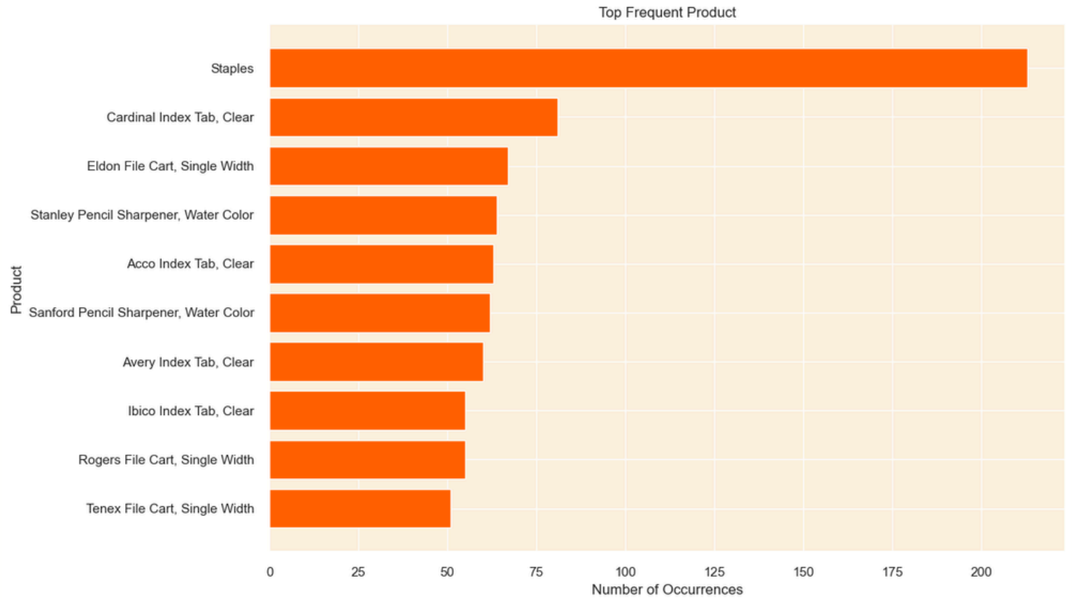




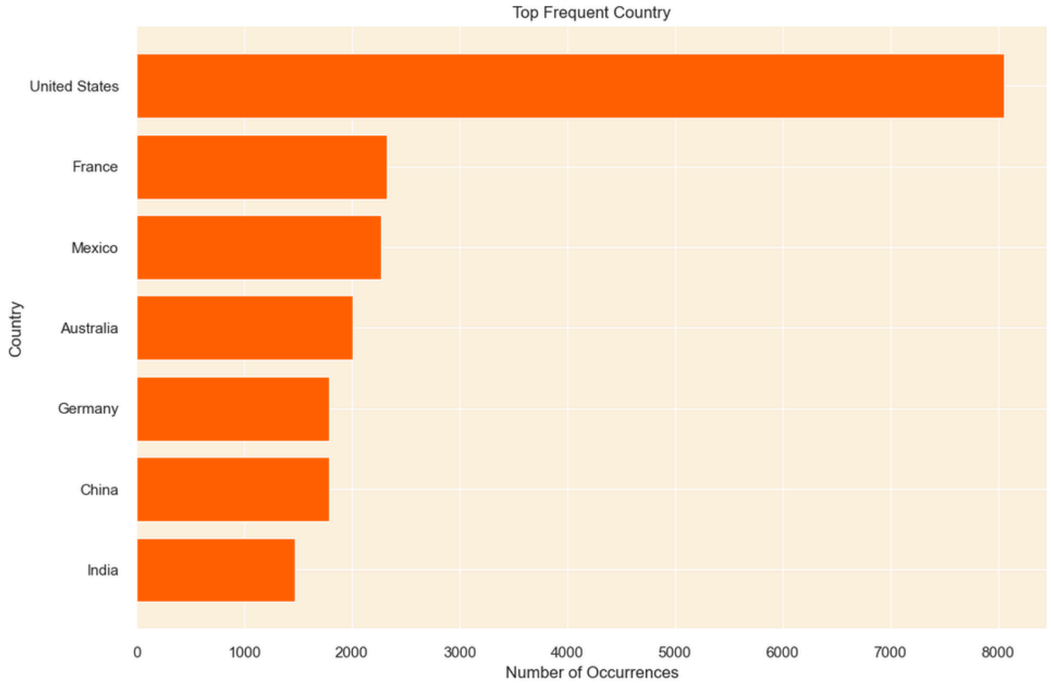


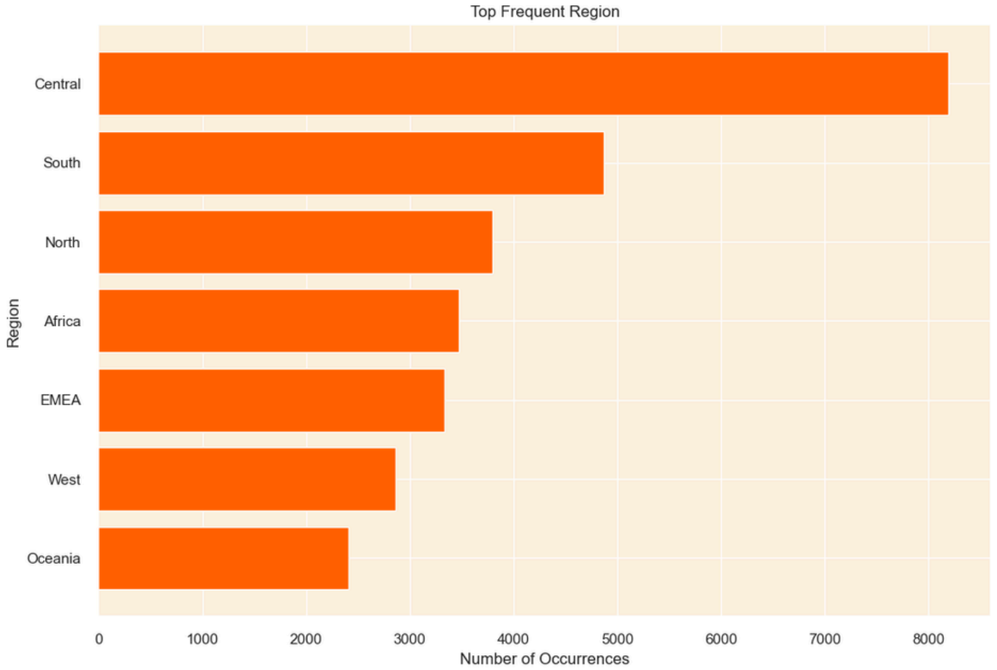
- **The technology category products generate more profit compared to office supplies, which was previously known to be the category with the highest sales volume. This indicates a significant difference in the cost price versus the selling price of the products.**
- **The majority of profits based on market come from APAC, despite the previous knowledge that the US had the highest number of orders. This suggests that technology purchases in APAC far exceed those in the US. It's also noted that the profit per item generated from office supplies is not very high.**
- **In terms of segment-based profit, the consumer segment, or purchases for personal rather than corporate or home office use, yields the highest profit. This further indicates that consumer purchases dominate in the technology category.**
- **Regarding order priority, orders with a medium priority level generate the most profit, followed by high priority. This is because the quantity of medium priority orders is significantly higher than that of high and critical priority orders. This also suggests that the higher the priority of the order, the higher the profit per transaction.**

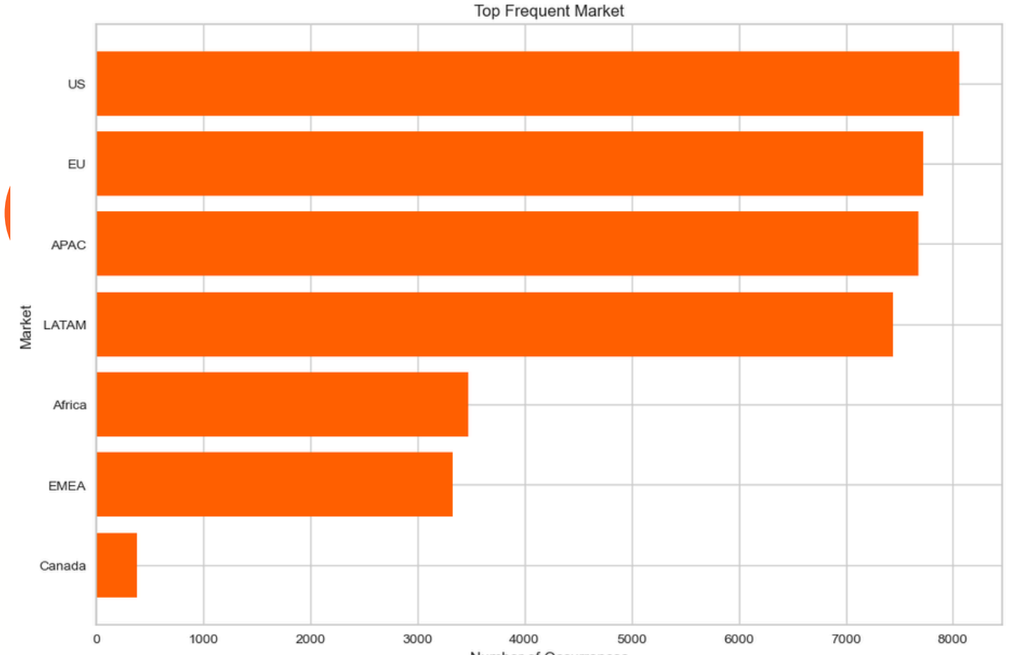




The binder sub-category has become the top choice for customers to purchase in the store, with a significant difference compared to other sub-categories where purchases are much lower. The most frequently purchased product by customers is Staples. The purchase of Staples products is much higher compared to other products.

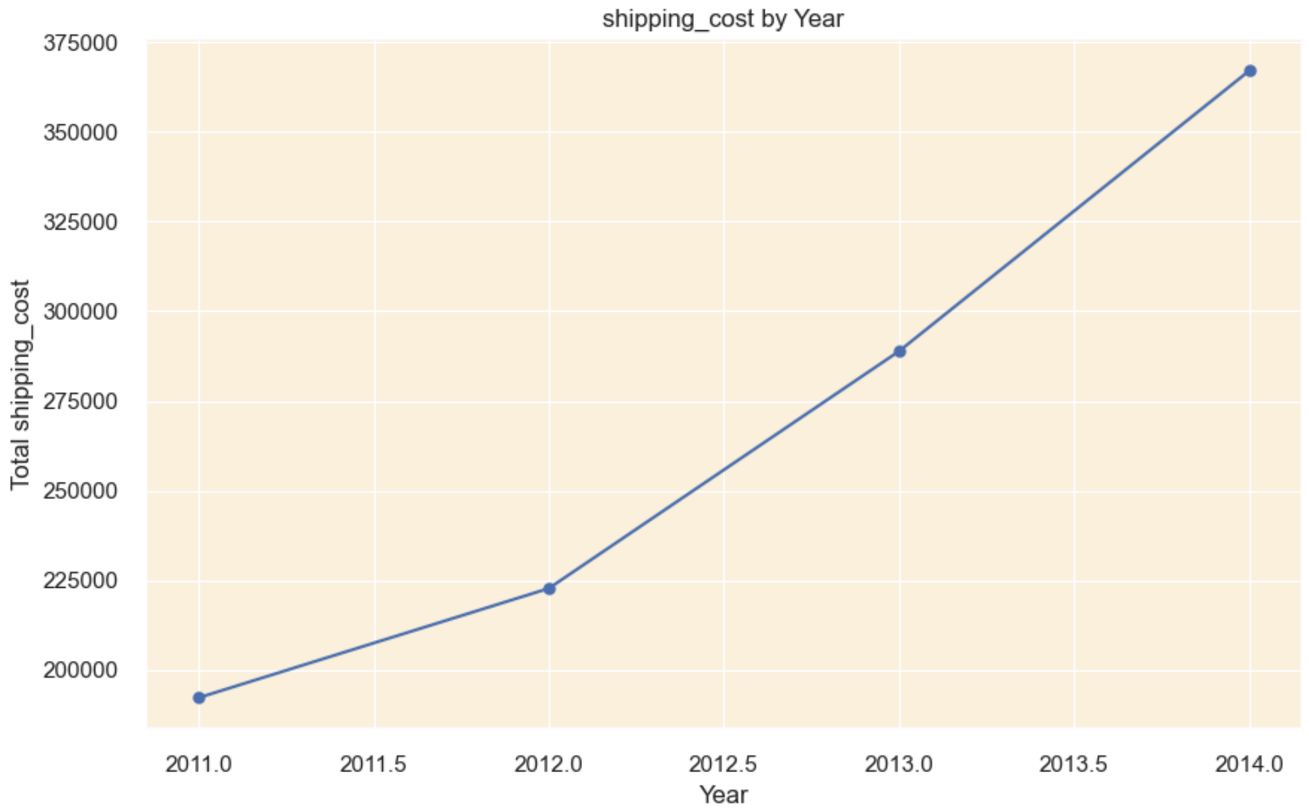


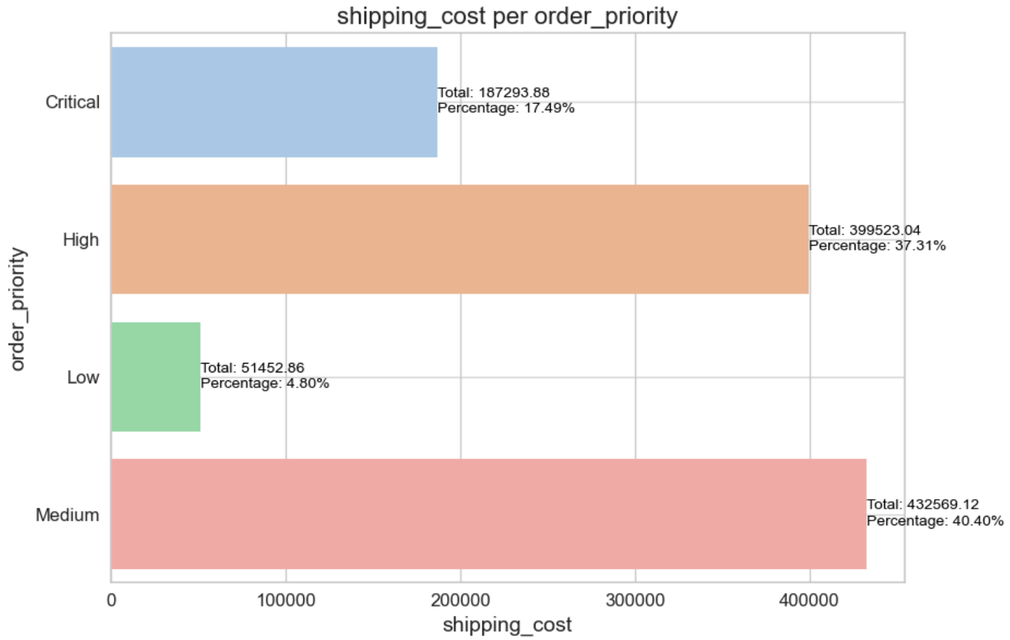


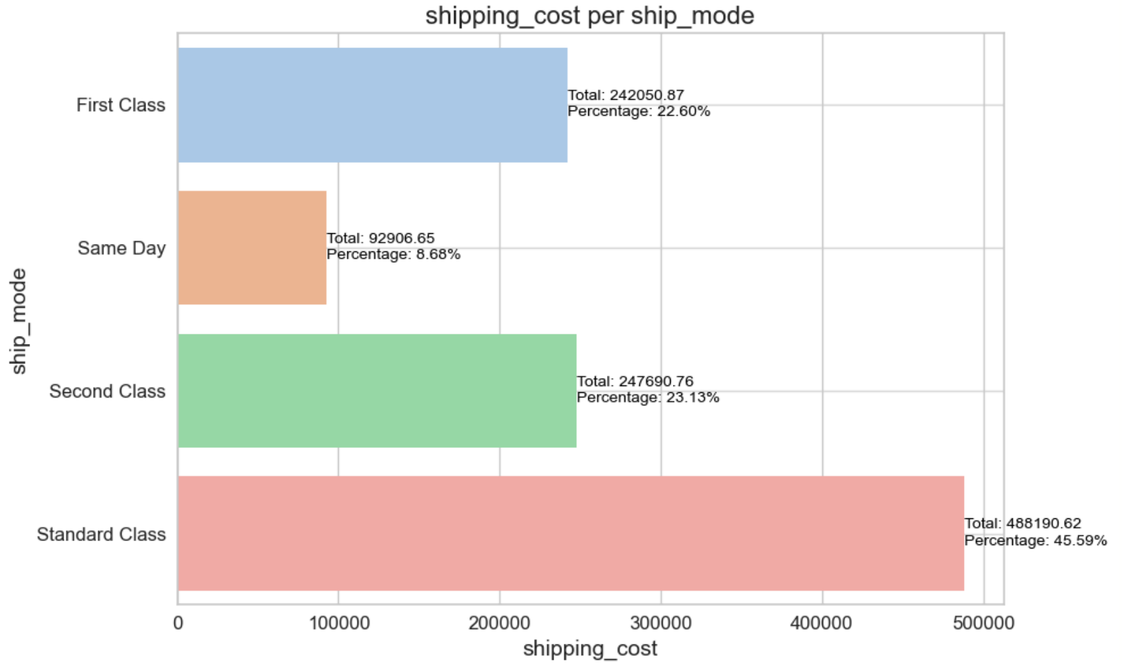


The United States is the country with the highest frequency of shipments. Among the states, Central America has the highest frequency of shipments. However, from the total sales in the US, EU, APAC, and LATAM regions, the frequency is almost equally distributed.

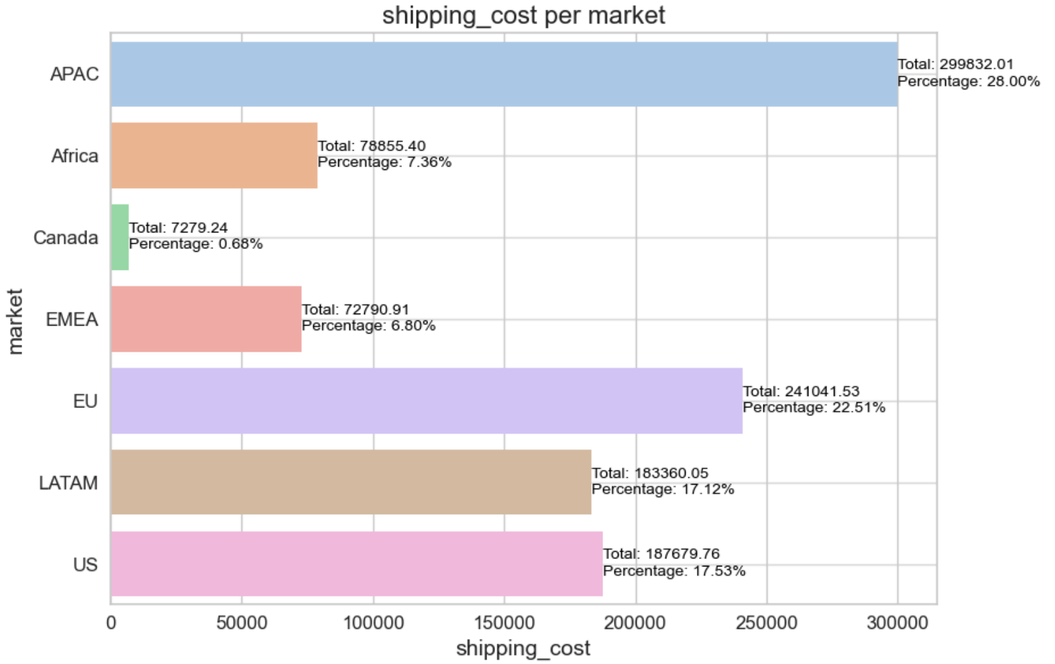
The increase in shipping costs is attributed to the increase in the number of purchases during those years. Similar to profit, shipping costs began to increase significantly from 2012 to 2014.

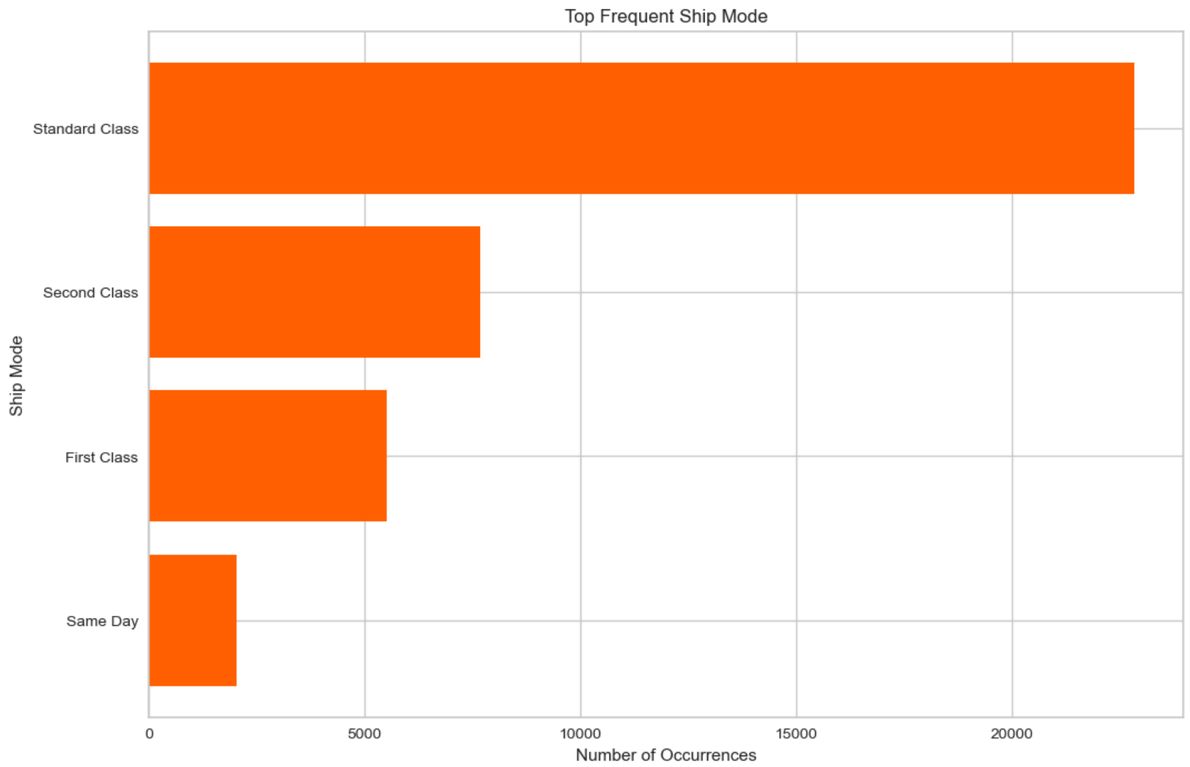


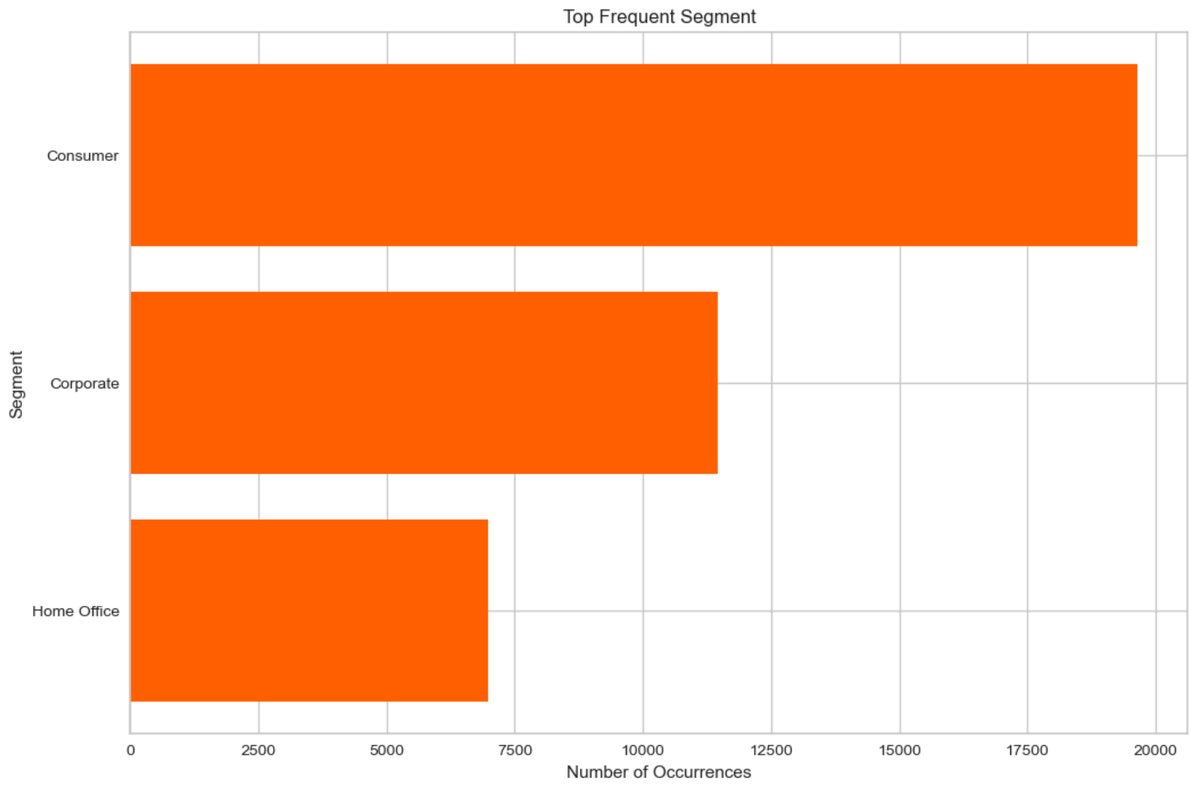




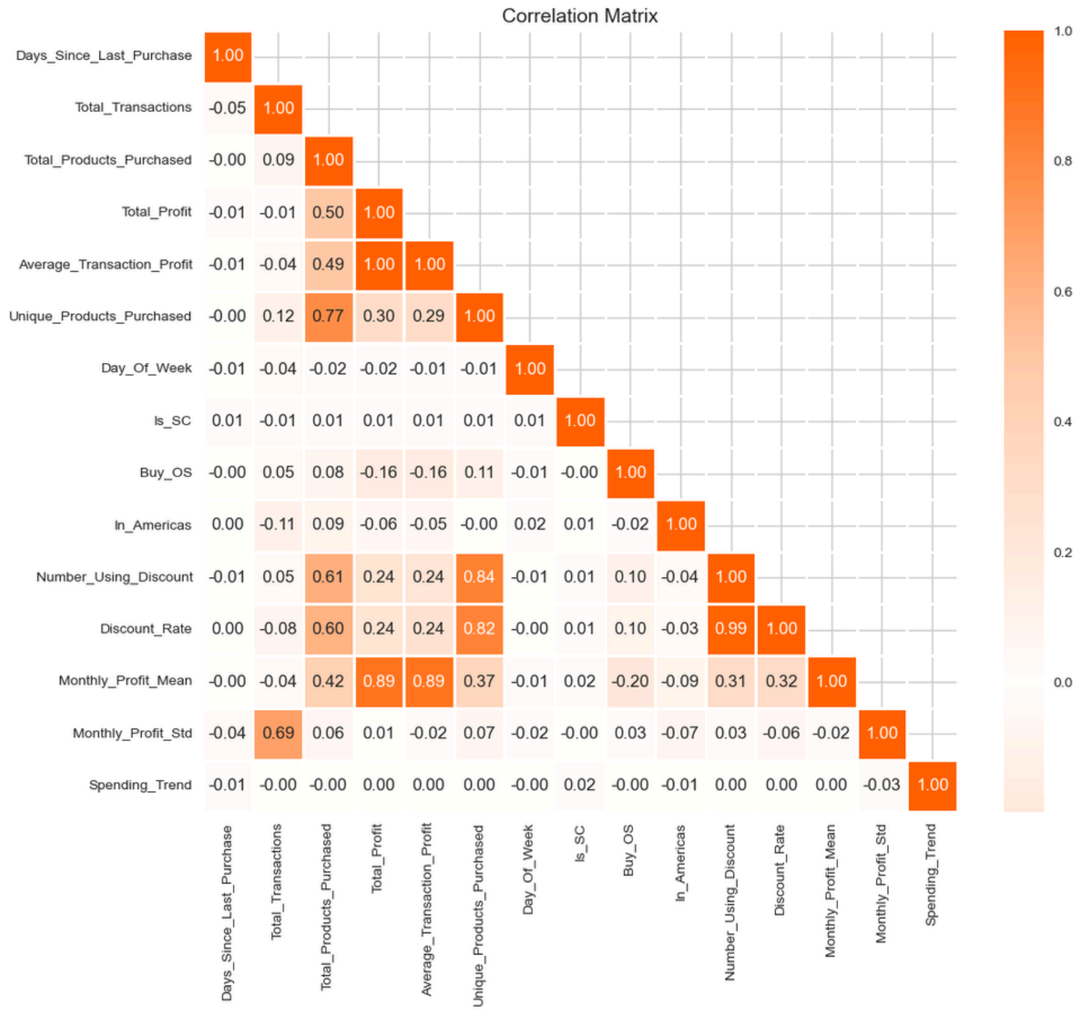
- **The shipping cost based on priority is predominantly attributed to medium priority orders. However, when examining the profit generated from these priorities, it's evident that they are highly profitable, as seen from the significant amount of profit relative to the shipping cost incurred. The pricing strategy for products with medium priority appears to be very effective. Further exploration is needed for high and critical priorities to optimize product pricing for better profitability.**
- **The shipping cost to APAC appears to be notably high, despite APAC not being the highest in terms of transaction volume. This indicates that shipping goods from the center to APAC is quite distant.**
- **The shipping cost for the standard class is the highest among the shipping modes.**







Standard class is the primary choice for customers as the most selected shipping mode because, in addition to being affordable, this shipping mode has a wide delivery range. The largest customer segment is consumers, indicating that many customers make transactions through the retail system. This also suggests that consumers are the most likely to make the most technology purchases.



Monthly_Profit_Mean, Total_Profit, and Average_Transaction_Profit, Unique_Products_Purchased, Number_Using_Discount, and Discount_Rate, Total_Profit,Average_Transaction_Profit have high correlations. This indicates that the data to be clustered experiences multicollinearity. Therefore, before processing the data into segments, PCA is applied to reduce dimensionality and address multicollinearity. Multicollinearity affects the clustering results, and dimensionality reduction ensures that the process does not take too long to run.

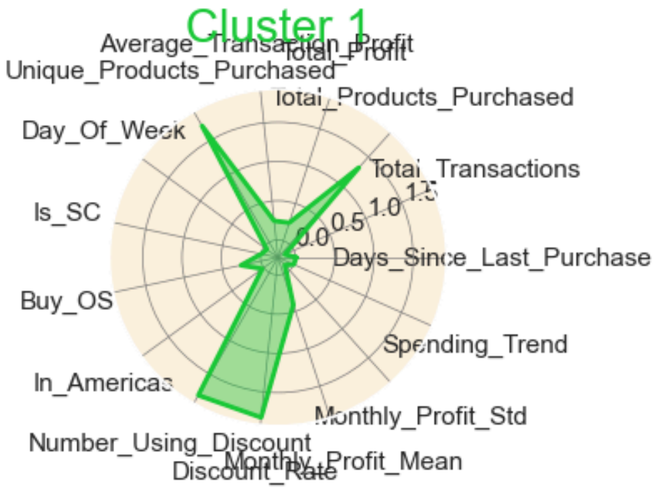
6. Cluster Analysis

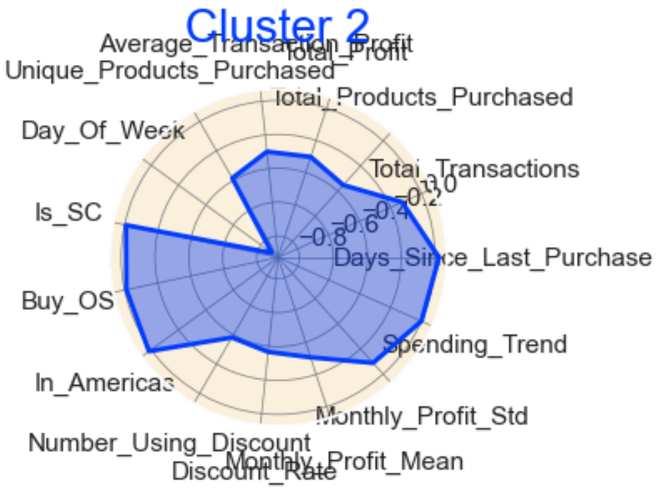
The variables used in forming the clusters are:

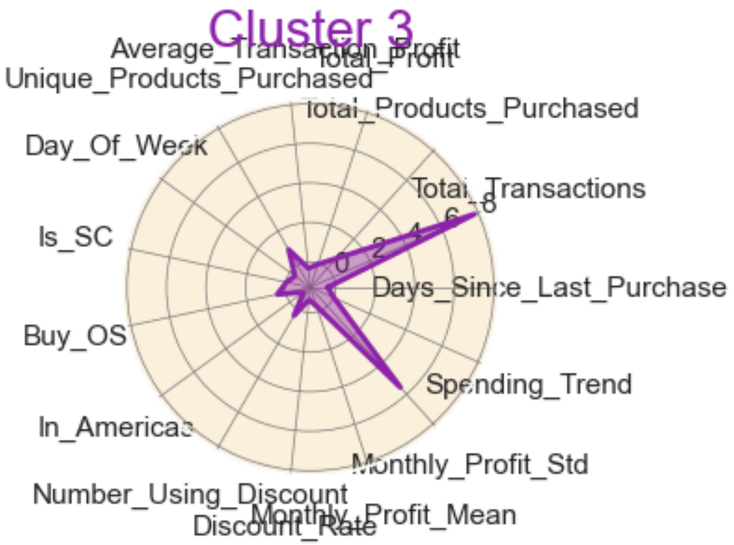
- **Days_Since_Last_Purchase: Number of days since the customer's last purchase.**
- **Total_Transactions: Total number of transactions made by the customer.**
- **Total_Products_Purchased: Sum of all products purchased by the customer.**
- **Unique_Products_Purchased: Number of distinct products purchased by the customer.**
- **Total_Profit: Total profit generated from the customer's transactions.**
- **Average_Transaction_Profit: Average profit earned per transaction from the customer.**
- **Day_Of_Week: Day of the week when the customer made purchases.**
- **Is_SC (SC=Standard Class): Indicates if the customer used Standard Class shipping.**
- **Buy_OS (Office_Supplies): Indicates if the customer purchased office supplies.**
- **Number_Using_Discount: Number of times the customer used a discount.**
- **Discount_Rate: Average discount rate applied to the customer's purchases.**
- **In_Americas: Indicates if the customer is located in the Americas.**
- **Monthly_Profit_Mean: Average monthly profit from the customer's transactions.**
- **Monthly_Profit_Std: Standard deviation of monthly profit from the customer's transactions.**
- **Spending_Trend: Pattern of the customer's spending over time.**

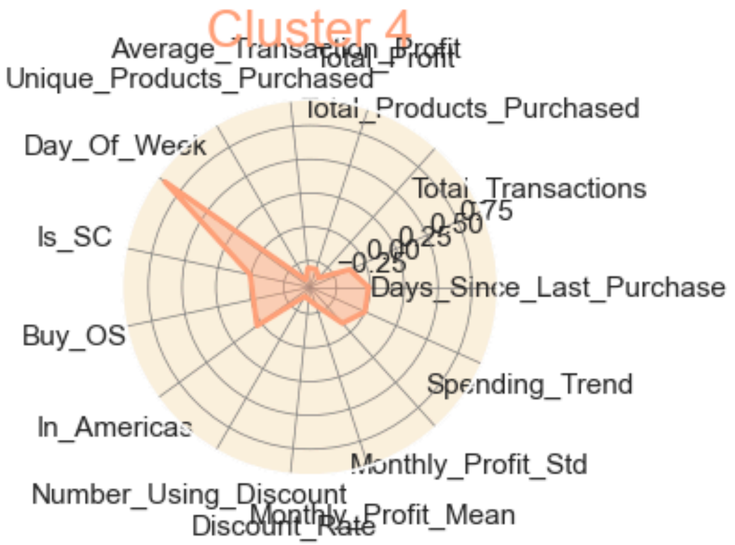
Clustering using data that has been processed through PCA (due to multicollinearity and high dimensionality). The clustering method used is K-Means Clustering.


Cluster 0
Average_Transaction_Profit
Total_Profit
Unique_Products_Purchased
Total_Products_Purchased
Day_Of_Week
Total_Transactions
Is_SC
Days_Since_Last_Purchase
Buy_OS
Spending_Trend
In_Americas
Monthly_Profit_Std
Number_Using_Discount
Monthly_Profit_Mean
Discount_Rate


Cluster 1
Average_Transaction_Profit
Total_Profit
Unique_Products_Purchased
Total_Products_Purchased
Day_Of_Week
Total_Transactions
Is_SC
Days_Since_Last_Purchase
Buy_OS
Spending_Trend
In_Americas
Monthly_Profit_Std
Number_Using_Discount
Monthly_Profit_Mean
Discount_Rate


Cluster 2
Average_Transaction_Profit
Total_Profit
Unique_Products_Purchased
Total_Products_Purchased
Day_Of_Week
Total_Transactions
Is_SC
Days_Since_Last_Purchase
Buy_OS
Spending_Trend
In_Americas
Monthly_Profit_Std
Number_Using_Discount
Monthly_Profit_Mean
Discount_Rate


Cluster 3
Average_Transaction_Profit
Total_Profit
Unique_Products_Purchased
Total_Products_Purchased
Day_Of_Week
Total_Transactions
Is_SC
Days_Since_Last_Purchase
Buy_OS
Spending_Trend
In_Americas
Monthly_Profit_Std
Number_Using_Discount
Monthly_Profit_Mean
Discount_Rate


Cluster 4
Average_Transaction_Profit
Total_Profit
Unique_Products_Purchased
Total_Products_Purchased
Day_Of_Week
Total_Transactions
Is_SC
Days_Since_Last_Purchase
Buy_OS
Spending_Trend
In_Americas
Monthly_Profit_Std
Number_Using_Discount
Monthly_Profit_Mean
Discount_Rate

Cluster 0: Monthly Shopper

Characteristics:

- **Moderate amount of shopping, neither too much nor too little**
- **Generates significant profit**
- **Regularly shops almost every month**

Cluster 1: Discount Shopper

Characteristics:

- **Always purchases different items**
- **Buys a considerable number of products**
- **Consistently uses discounts in every purchase**
- **Purchase at the beginning of the week.**

Cluster 2: Office Supplies Subscriber

Characteristics:

- **Consistently buys office supplies**
- **Shipping is done using standard class mode**
- **Purchases are sent to the US or LATAM**
- **Frequently makes purchases**

Cluster 3: Same Purchase Pattern

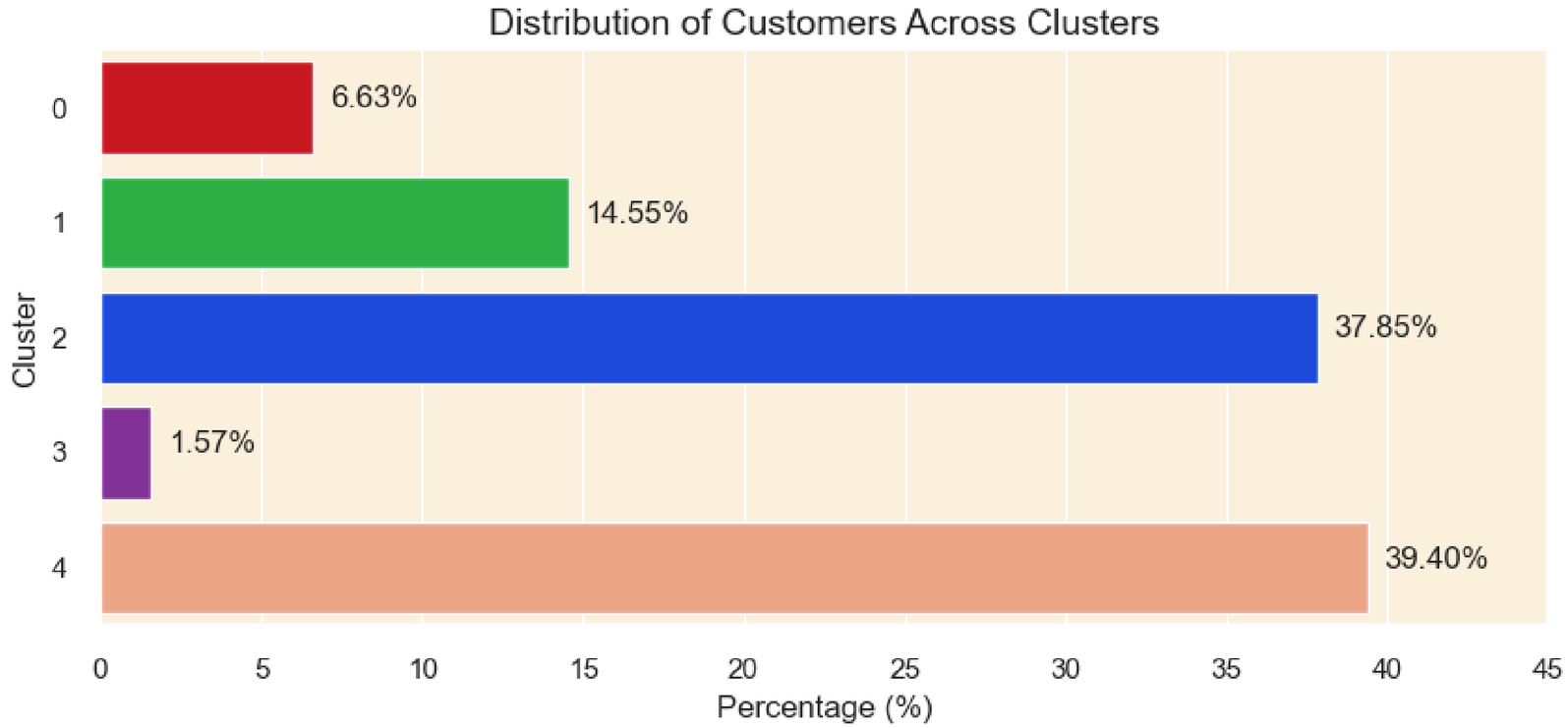
Characteristics:

- **Likes to transact**
- **Makes purchases almost every month**
- **Not located in the American continent**

Cluster 4: Holiday Shopper

Characteristics:

- **Purchases are made on weekends**
- **Never uses discounts**



The best distribution of clusters is in clusters 1, 2, and 4. Therefore, the ideal target for marketing would be those who exhibit the characteristics of these clusters.

63]:

order_id	Rec1_product_id	Rec1_product_name	Rec2_product_id	Rec2_product_name	Rec3_product_id	Rec3_product_name
UZ-2014-6400	OFF-BI-10001253	Acco Binder Covers, Recycled	OFF-AR-10000594	Binney & Smith Highlighters, Water Color	OFF-FA-10002803	Accos Staples, Metal
IN-2013-18805	OFF-BI-10001524	GBC Premium Transparent Covers with Diagonal L...	FUR-TA-10001095	Chromcraft Round Conference Tables	OFF-EN-10003850	GlobeWeis Clasp Envelope, Security-Tint
ES-2011-3911616	OFF-BI-10001294	Fellowes Binding Cases	OFF-AR-10002783	Stanley Pencil Sharpener, Water Color	TEC-PH-10004071	PayAnywhere Card Reader
PL-2013-8560	OFF-AR-10001228	Stanley Markers, Water Color	FUR-CH-10003733	Hon Steel Folding Chair, Set of Two	OFF-BI-10003883	Acco Binder, Economy
ES-2014-5263546	OFF-AR-10001228	Stanley Markers, Water Color	FUR-CH-10003733	Hon Steel Folding Chair, Set of Two	OFF-BI-10003883	Acco Binder, Economy
CA-2012-151841	OFF-BI-10001524	GBC Premium Transparent Covers with Diagonal L...	FUR-TA-10001095	Chromcraft Round Conference Tables	OFF-EN-10003850	GlobeWeis Clasp Envelope, Security-Tint
ES-2012-2393484	OFF-BI-10001524	GBC Premium Transparent Covers with Diagonal L...	FUR-TA-10001095	Chromcraft Round Conference Tables	OFF-EN-10003850	GlobeWeis Clasp Envelope, Security-Tint
ID-2014-77864	OFF-BI-10001524	GBC Premium Transparent Covers with Diagonal L...	FUR-TA-10001095	Chromcraft Round Conference Tables	OFF-EN-10003850	GlobeWeis Clasp Envelope, Security-Tint
CG-2011-3350	OFF-BI-10001294	Fellowes Binding Cases	OFF-AR-10002783	Stanley Pencil Sharpener, Water Color	TEC-PH-10004071	PayAnywhere Card Reader
CA-2013-133340	OFF-BI-10001524	GBC Premium Transparent Covers with Diagonal L...	FUR-TA-10001095	Chromcraft Round Conference Tables	OFF-EN-10003850	GlobeWeis Clasp Envelope, Security-Tint

1
2
3
4
5
6
7
8
9

Thank you

Contact Details

Phone : *+6281546743628*

Address : *Jakarta BArat*

Email : *devanagas7@gmail.com*