Logistics Performance Analysis and Delay Prediction and Consumer Segmentation

1. Introduction

1.1 Project Overview

In today's competitive consumer market, efficient logistics are vital for maintaining customer satisfaction and business success. Timely product deliveries boost customer loyalty, while delays can lead to dissatisfaction, decreased profits, and operational challenges.

This project analyzes logistics performance in consumer product deliveries, focusing on identifying factors contributing to delays, such as shipping modes, product categories, and regions. By predicting delays, businesses can optimize supply chain management, minimize disruptions, and improve overall performance.

1.2 Objectives

The primary objectives of this analysis are:

- **Gain insights** into factors affecting delivery performance, such as regions, shipping modes, product categories, and customer segments.
- **Identify patterns** in sales, profit, and customer behavior across different regions, product categories, and shipping methods.
- **Develop a predictive model** to classify orders into Delayed, On Time, or Early Arrival using logistics, geographical, and transactional data.
- **Provide actionable recommendations** to reduce delays and enhance profitability in logistics operations.
- **Create a Power BI dashboard** for real-time visualization of key metrics and performance insights.

1.3 Tools and Technologies Used

- **Python**: For data analysis, feature engineering, and machine learning (with libraries such as Pandas, NumPy, and Scikit-learn).
- **Data Visualization**: Seaborn and Matplotlib for visualizing trends, distributions, and relationships in the data.
- **Machine Learning Algorithms**: Classification models including Decision Trees, Random Forest, and Gradient Boosting for delay prediction.
- **Jupyter Notebook**: For interactive analysis and reporting.
- **SSMS (SQL Server)**: For efficient data storage, retrieval, and management.
- **Power BI**: For creating a dynamic dashboard to visualize key metrics such as order volume, sales, profitability, and delay trends.

.

2. Data Collection and Preprocessing

2.1 Data Source

The data used for this analysis comes from Kaggle's dataset, titled *incom2024_delay_example_dataset*. This dataset contains real-world logistics data, including product delivery information and delivery labels indicating whether the shipment was delayed, on time, or arrived early.

LINK: https://www.kaggle.com/datasets/pushpitkamboj/logistics-data-containing-real-world-data

2.2 Data Description

The dataset contains a wide variety of features such as:

- **Customer and Order Information**: Customer ID, order date, shipping date, product name, etc.
- **Geographic Information**: Customer city, state, and location (latitude, longitude).
- **Logistics and Shipping Information**: Order item quantity, shipping mode, and department.
- **Target Variable**: The label indicating delivery status (delay, on time, or early arrival). The dataset has no null values and has been cleaned and prepared for analysis. Unique categories were identified, especially for categorical variables.

2.3 Data Cleaning

Data pre-processing steps include:

- **Dropped unnecessary columns**: Columns like department_id and category_id were dropped.
- **Mapped categorical variables**: Transformed category_name and department_name into numeric values using dictionary mapping.
- **Date formatting**: Converted date columns (e.g., order_date, shipping_date) into datetime objects.
- **Calculated new features**: Derived order_to_shipping_days to calculate the number of days between ordering and shipping.
- **Created a new UUID for orders**: A unique order_id was generated for each record.

2.4 Feature Engineering

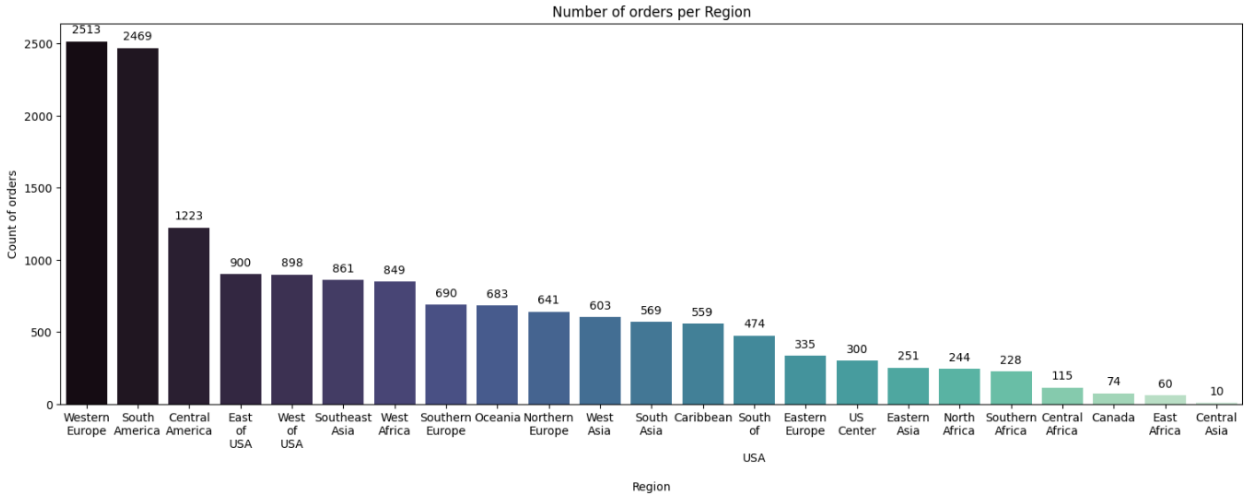
Several new features were created to enrich the dataset:

 Date-related features: Extracted day, month, and year from the order_date and shipping_date columns to capture temporal patterns.
 Shipping delay calculation: Created the order_to_shipping_days feature to measure the time lag between order and shipment.
 Delay label transformation: Transformed the numeric delay label into categories (e.g., Delayed, On Time, Early Arrival) for classification modeling.

3. Exploratory Data Analysis (EDA)

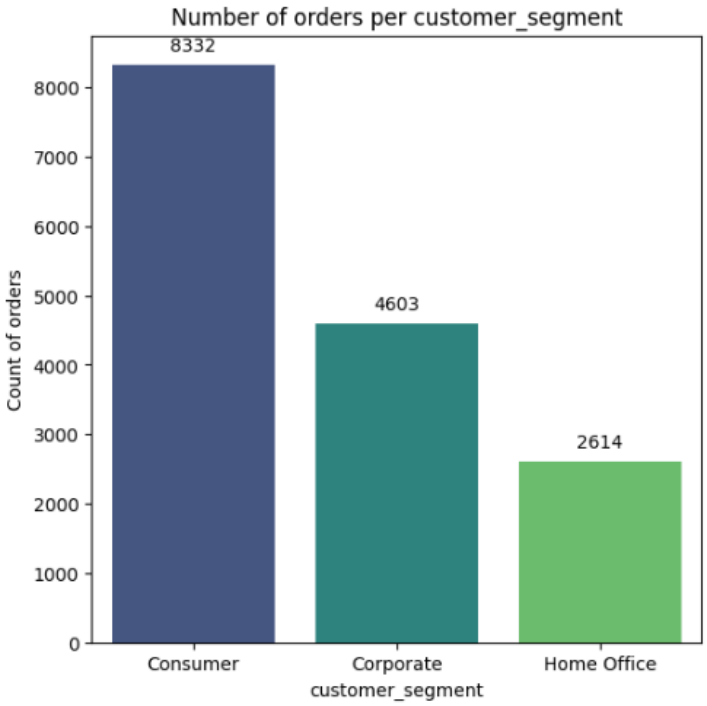
❖ **Order Volume Analysis:**

- Which regions have the highest number of orders?



Ans) Western Europe has the highest number of records with 2513 orders

- How does the number of orders vary by customer segments (e.g., Consumer, Corporate, etc.)?

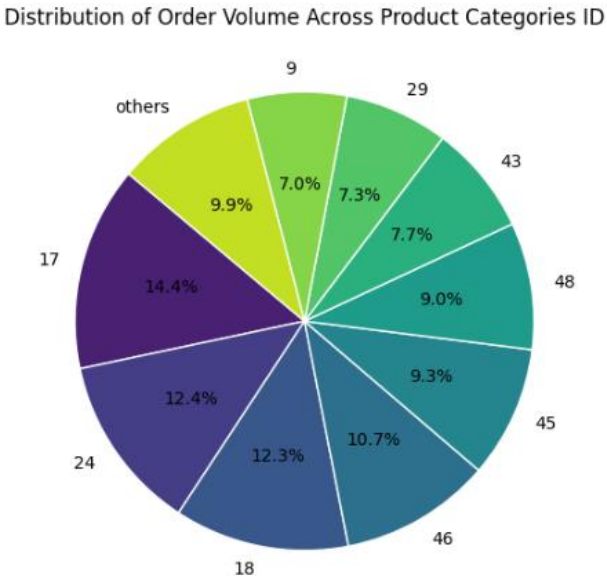


Ans) Consumer Segment has highest number of orders 8332 followed by Corporate 4603 and Home Office 2614

- How does the number of orders differ by shipping mode (e.g., Standard, Express)?

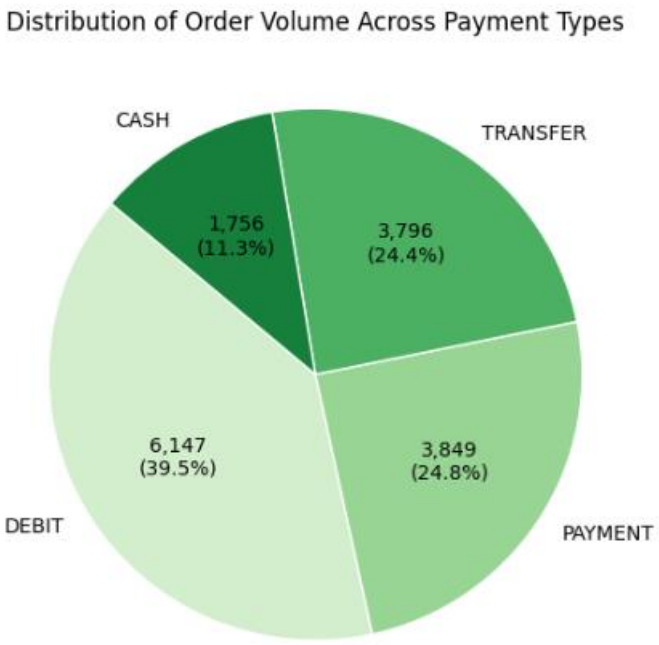
Ans) Standard Class has the highest 9116 orders followed by Second Class 3283 First Class 2391 and lastly same day with 759 orders

- What is the distribution of order volume across different product categories?



Ans) Category ID 17 has the highest distribution followed by Category ID 2 and 18 and 46. In Others category there are several category id each having a minute share. Cumulative value of each is 9.9 %

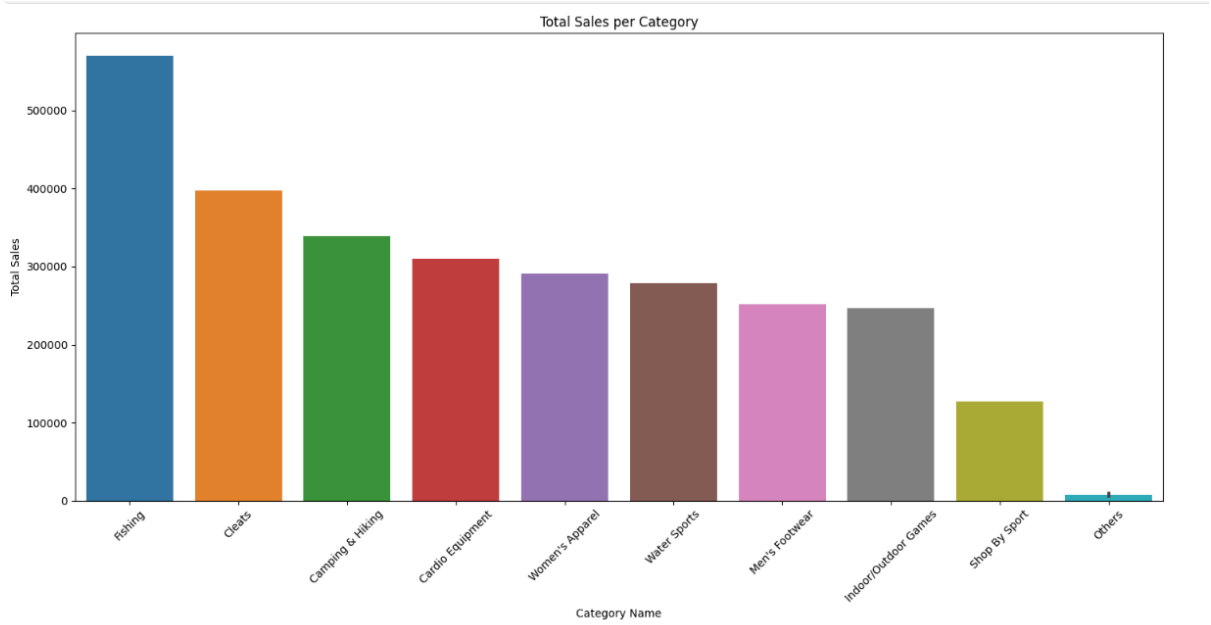
- Are there any noticeable trends in order volume based on payment types?

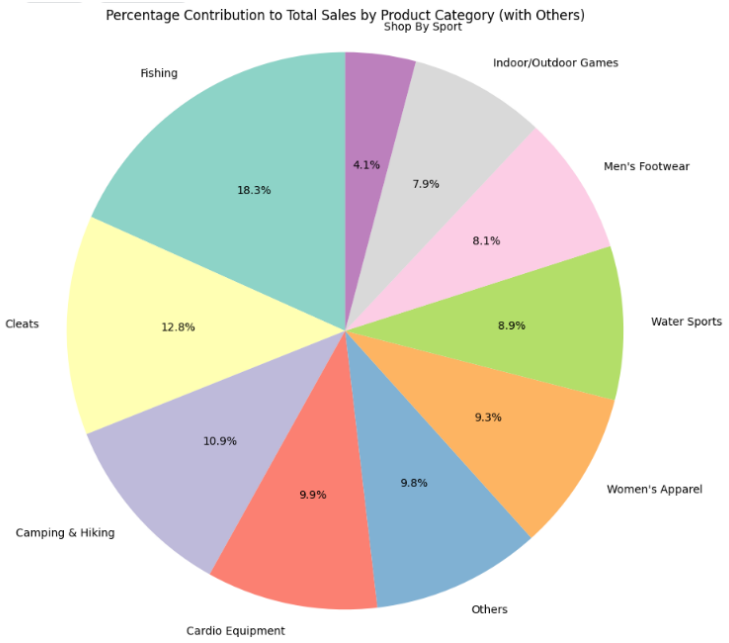


Ans) Debit has highest order percentage of 39.5% followed by Payment with 24.8% and Transfer with 24.4%

❖ **Sales Performance:**

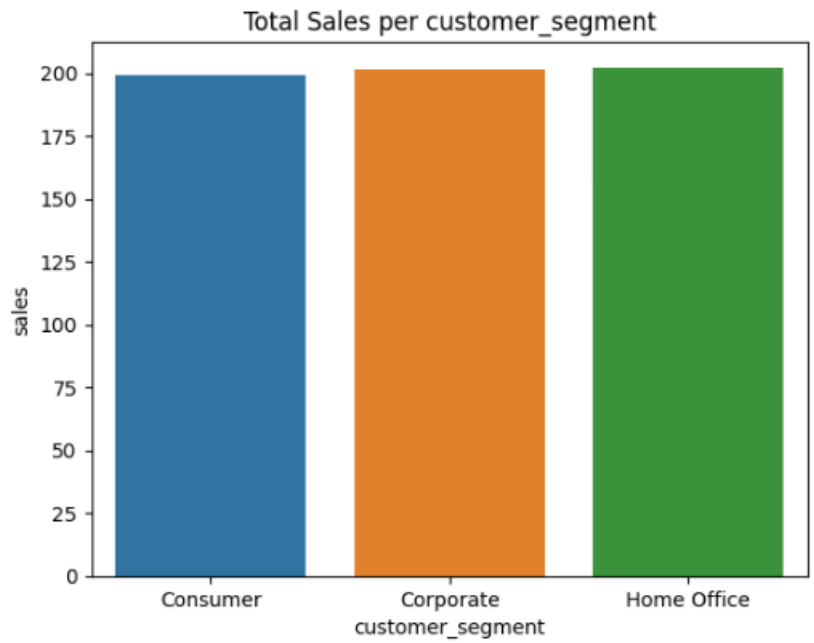
- Which product categories contribute the most to total sales?
- **Ans)** Fishing Category has highest contribution in sales followed by Cleats and Camping & Hiking





Ans) Fishing Category has highest contribution in sales followed by Cleats and Camping & Hiking

- What is the average sales per customer segment?

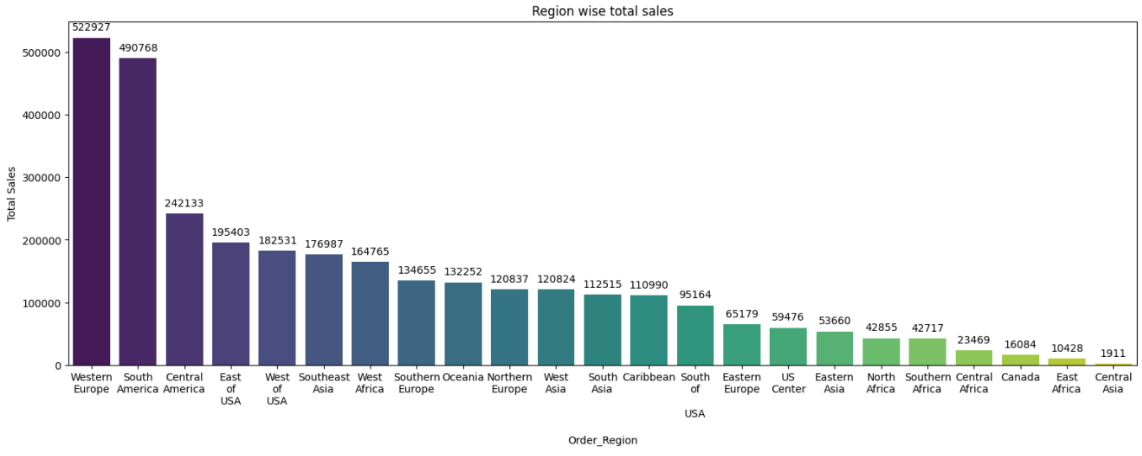


[165...

	customer_segment	sales
0	Consumer	199.340673
1	Corporate	201.780166
2	Home Office	202.311569

Ans) Average sales is similar for all customer segments nearly 200 units

- How do sales vary across different regions and countries?

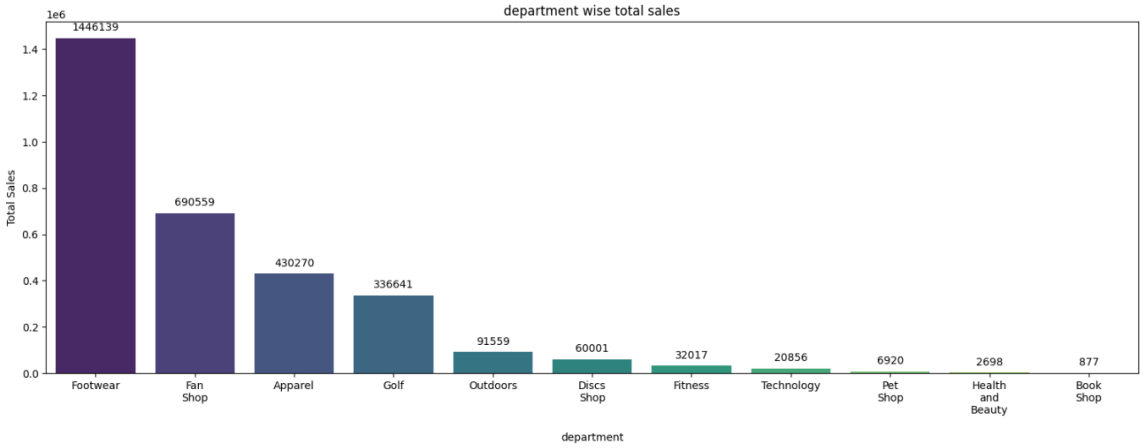


Ans) Europe and American countries have most amount of sales followed by Africa and Asia

- What are the top-performing markets in terms of total sales?

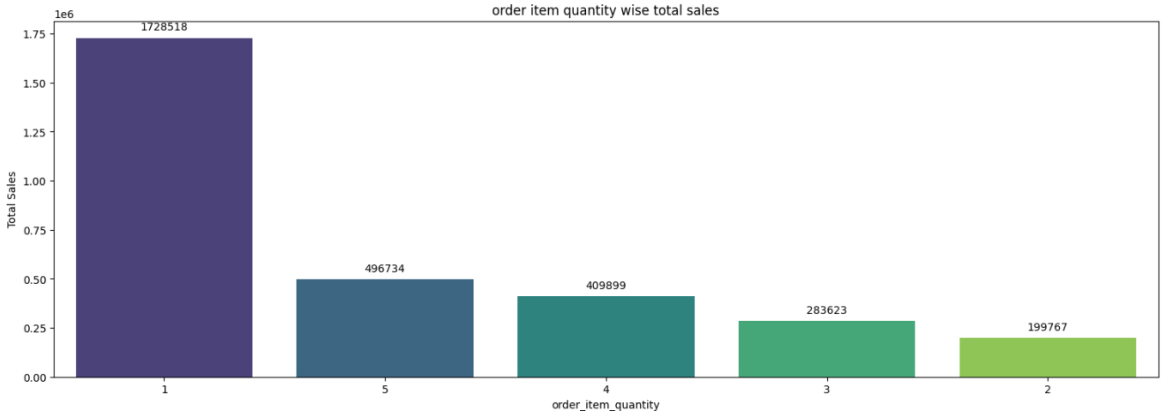
Ans) Europe and American countries have most amount of sales followed by Africa and Asia

- How does the sales distribution vary by department (e.g., Footwear, Fan Shop)?



Ans) Taking into account the number of sales Footwear has highest number of sales followed by Fan shop , Apparel, Golf etc

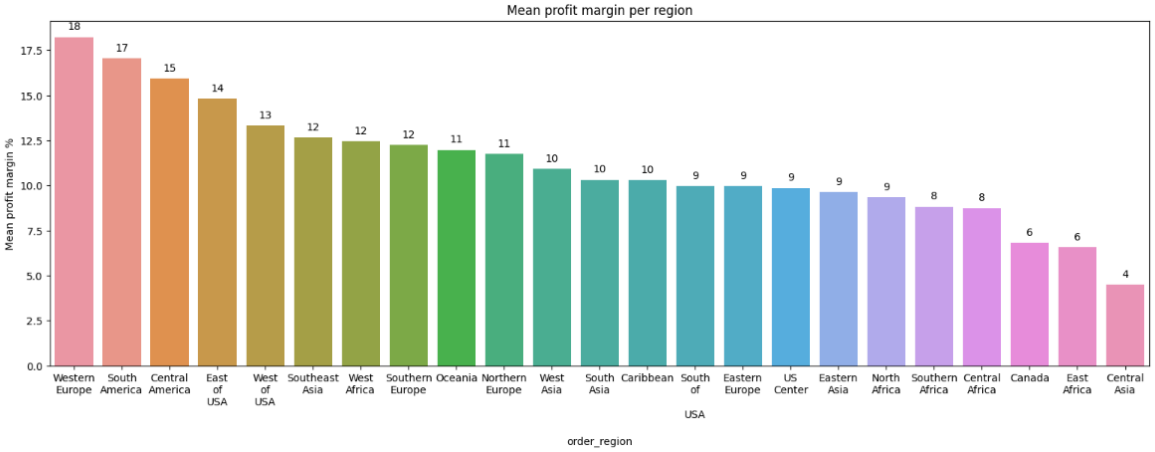
- How does the number of items ordered per order affect total sales?



Ans) Order quantity 1 has the highest number of sales followed by order count 5 ,4,3,2.

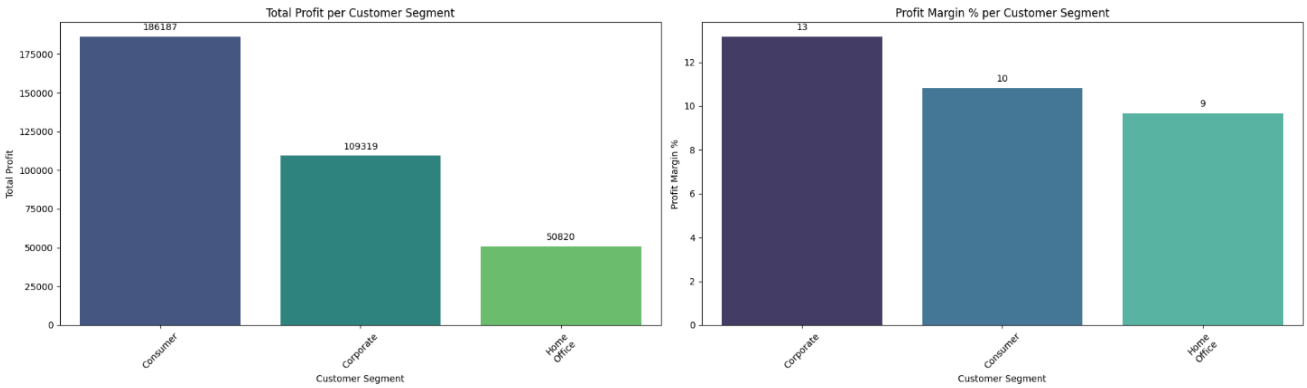
❖ Profitability Analysis

- Which regions or countries have the highest profit margins?



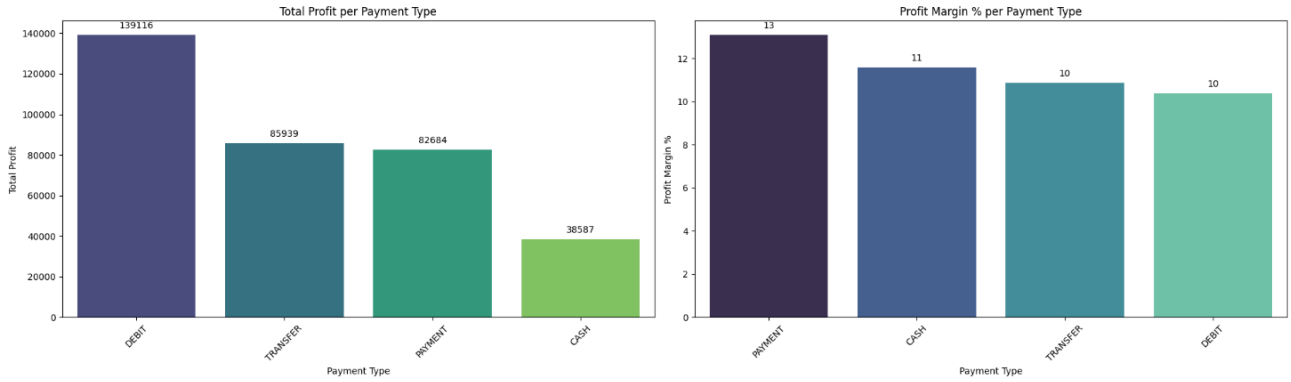
Ans) Europe and US regions have the highest profit margin.

- How does profitability vary by customer segment?



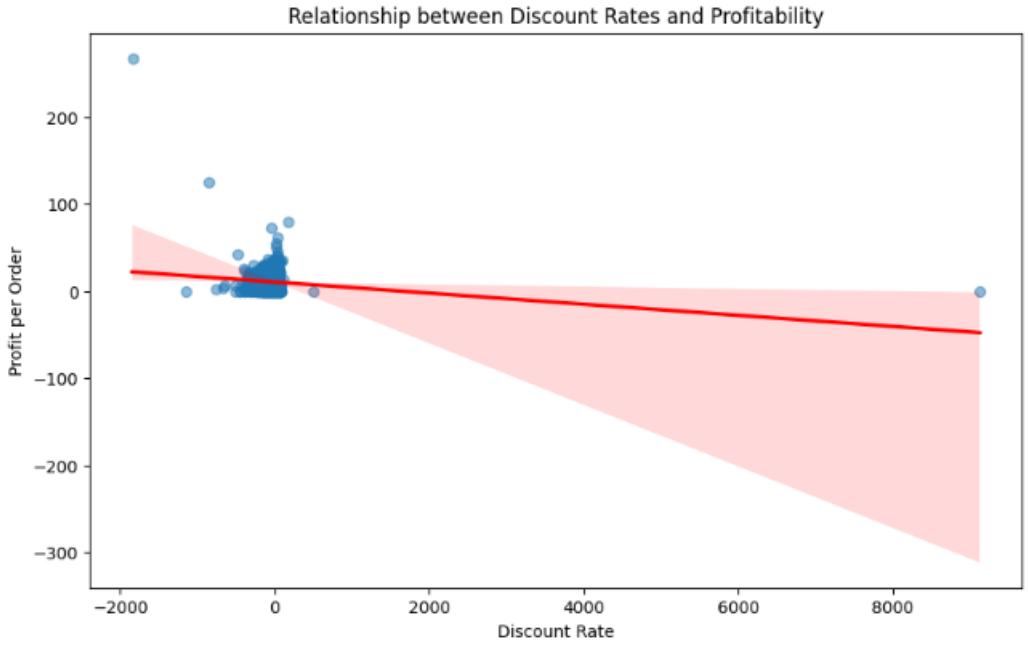
Ans) Consumer Segment has produced the highest profit followed by Corporate and Home Office. Profit margins of Corporate is higher than consumer and Home Office

- Which payment types are associated with the highest profitability?



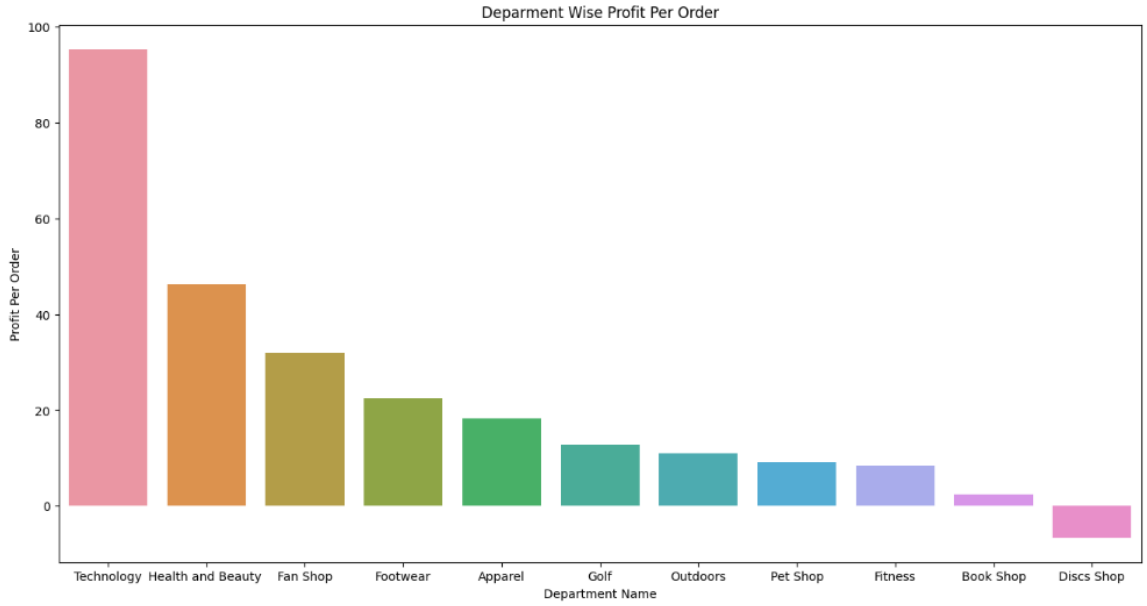
Ans) Debt has provided the most profit followed by transfer, Payment and cast. Whereas payment has the highest profit margin closely followed by Cash transfer and Debit

- What is the relationship between discount rates and profitability? Do higher discounts reduce profitability significantly?



Ans) Yes , there is a relationship declining relationship between higher discounts and lower profits

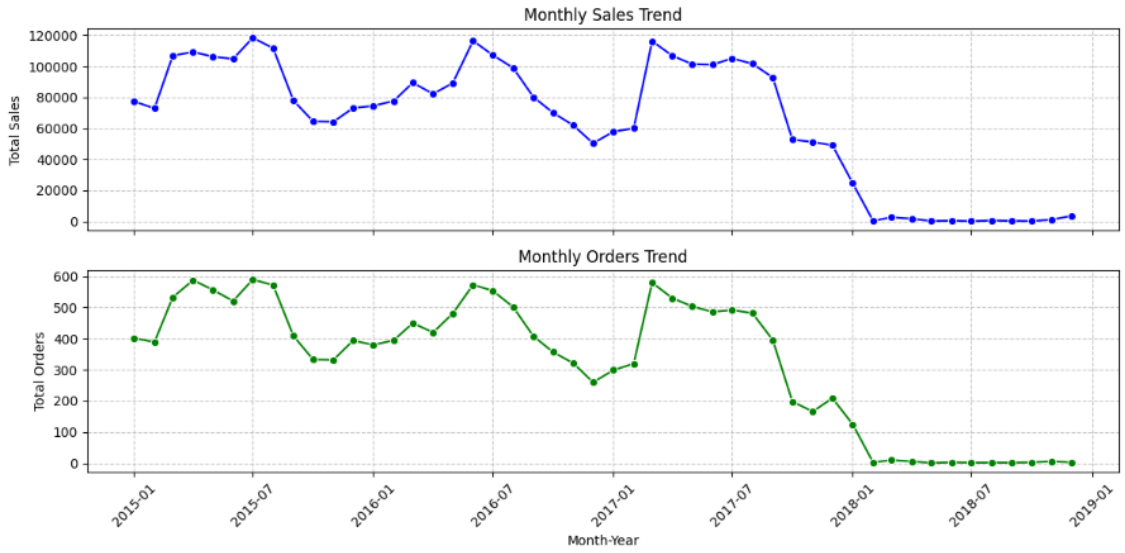
- How does the profit per order vary across different departments?



Ans) Profit per order is highest for Technology , then Health_and_beauty at half profit per order than technology then FanShop Footwear closely follow in profit per order. Book Shop has negligible profit per order and Disc shop has negative profit per order.

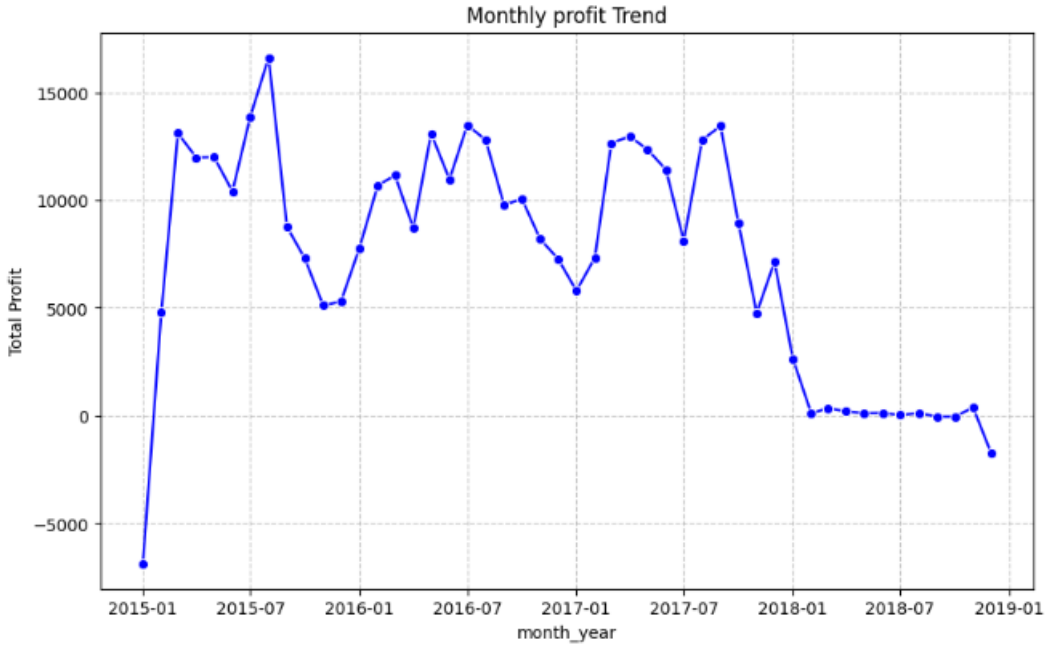
❖ **Order and Sales Trends Over Time:**

- Are there seasonal patterns in the number of orders or sales?



Ans) Every Year first 6 months have high sales and order count then in the last 6 months sales and order drop compared to first 6 monthf from 2015 to 2017. In 2018 to 2019 these has been huge drop of sales and order count

- How does profitability fluctuate throughout the year or across months?



Ans) Similarly profit is high in first 6 months and low in last 6months from 2015 to 2017 . In the year 2018 to 2019 profits have reduced drastically closely touching loss.

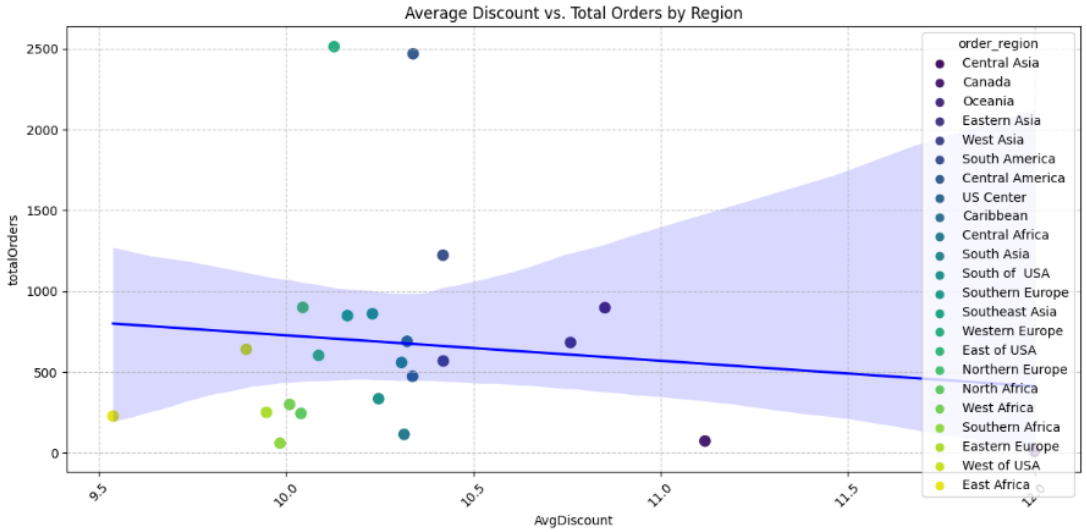
❖ **Customer Behaviour Insights:**

- Which customer countries or regions are associated with repeat orders?

```
number of Repeat Orders-->58
number of unique Orders-->15491
```

Ans) Neglible amount of orders are repeated – 58 out of 15000

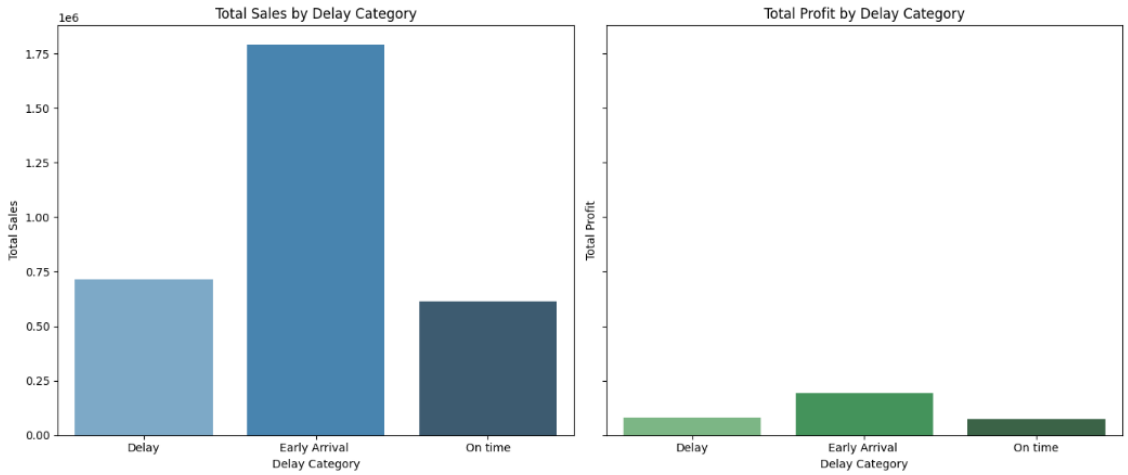
- How do customers from different countries or segments respond to discounts (e.g., Do they place more orders with discounts)?



Ans) There is a very slight decrement of order if discounts are increased . But Discounts don't increase number of order except in Europe and USA

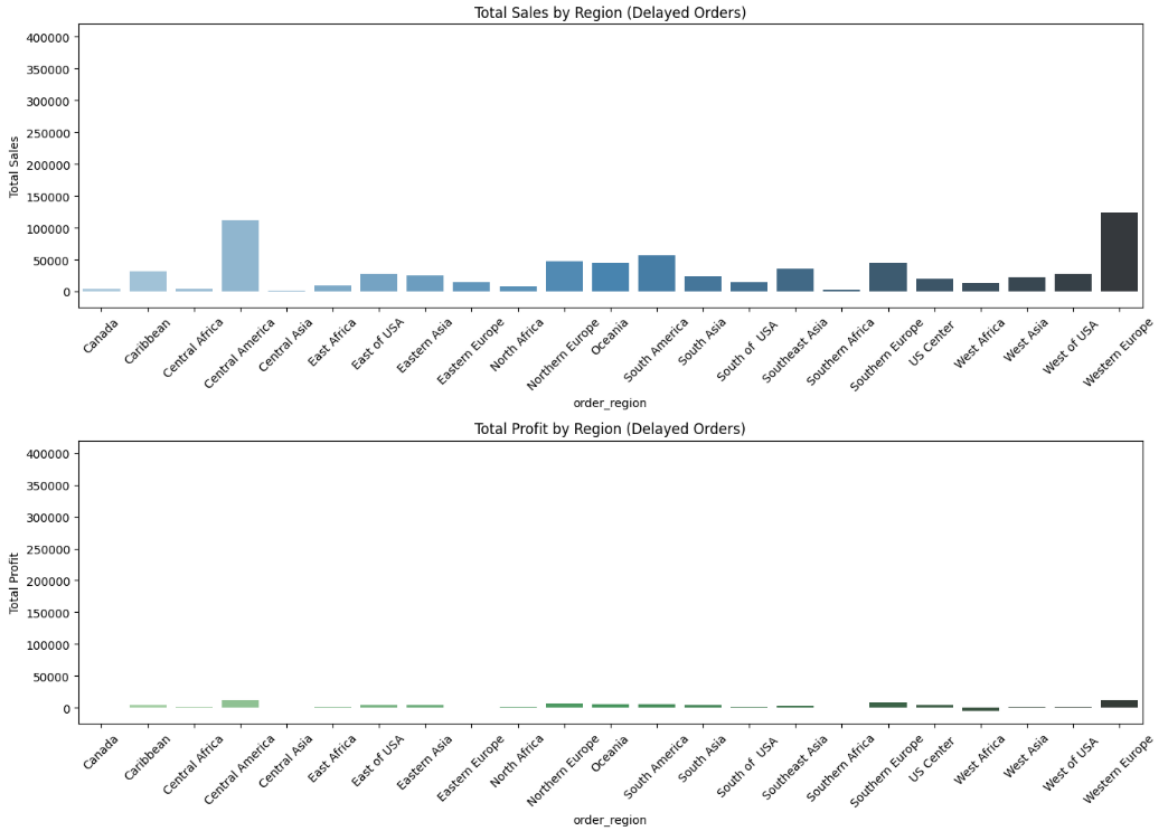
❖ **Order Delay and Impact on Profitability:**

- Sales and profit by delay category

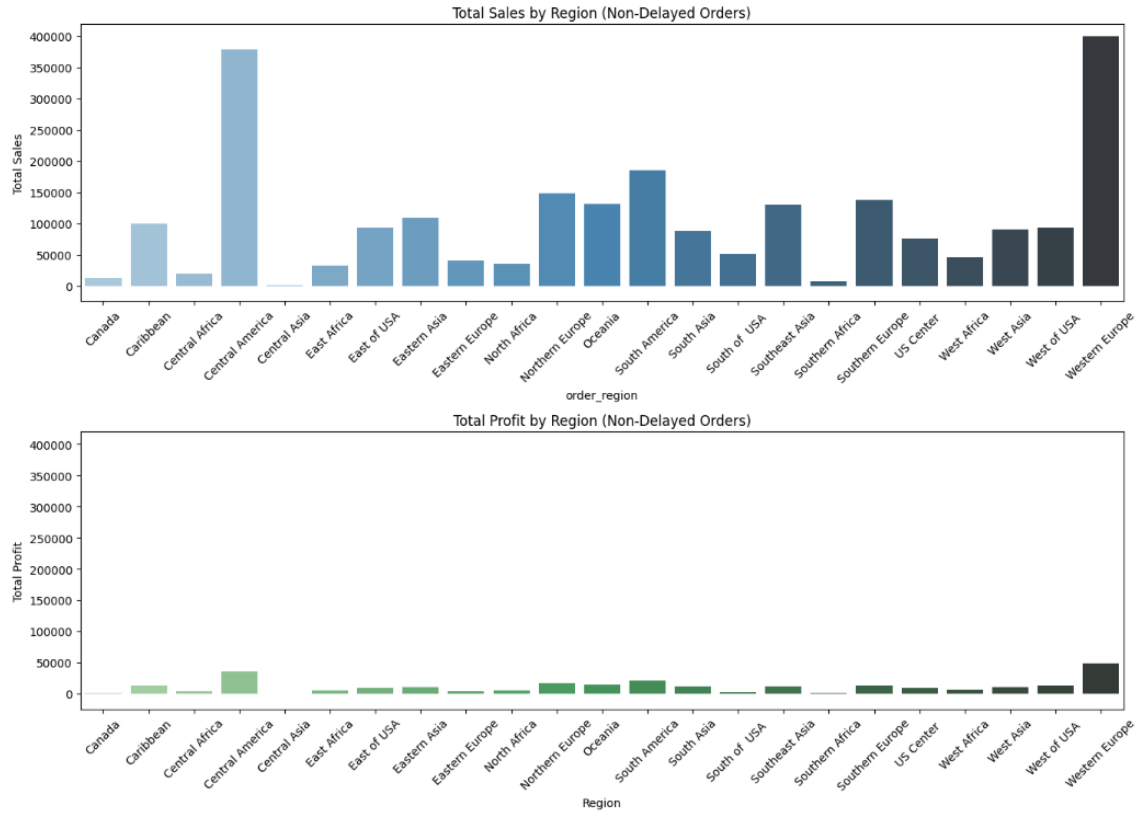


Ans) Early Arrivers are in most sales and profit. Follwed by Delay sales and profit lastly we have on time sales and profit

- Is there a significant difference in sales or profit for delayed versus on-time orders?



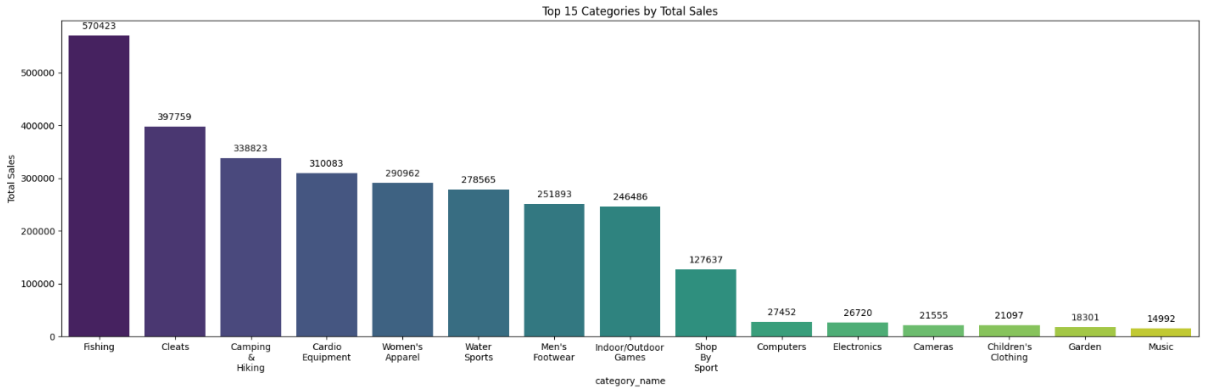
Ans) Western Europe and Central Usa has highest number of delayed orders followed by Asia and Africa

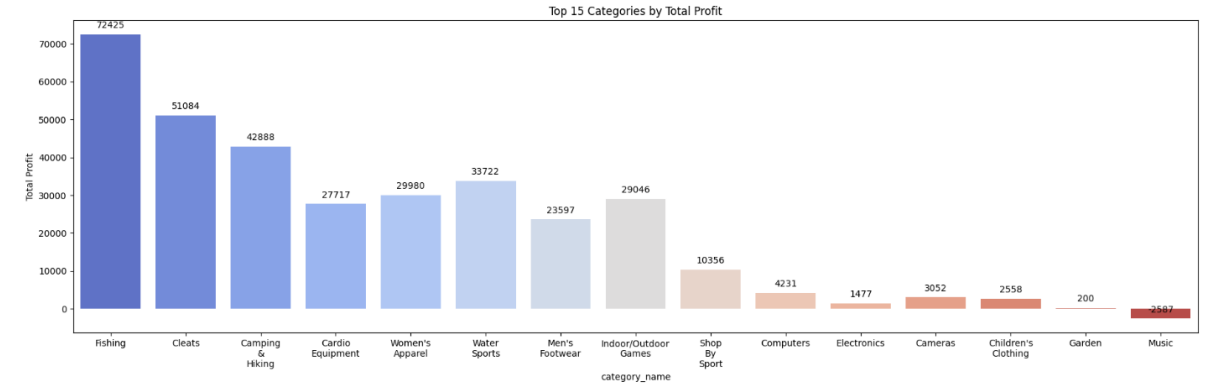


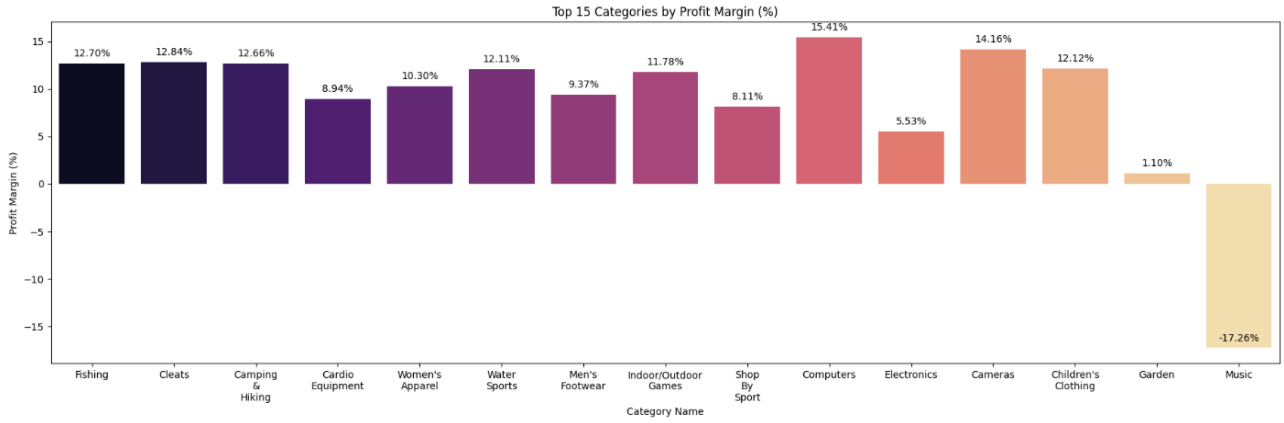
Ans) Similar trends is seen in non delayed orders

❖ Product-Specific Performance:

- Which specific products (e.g., by product name or category) have the highest sales and profit margins?

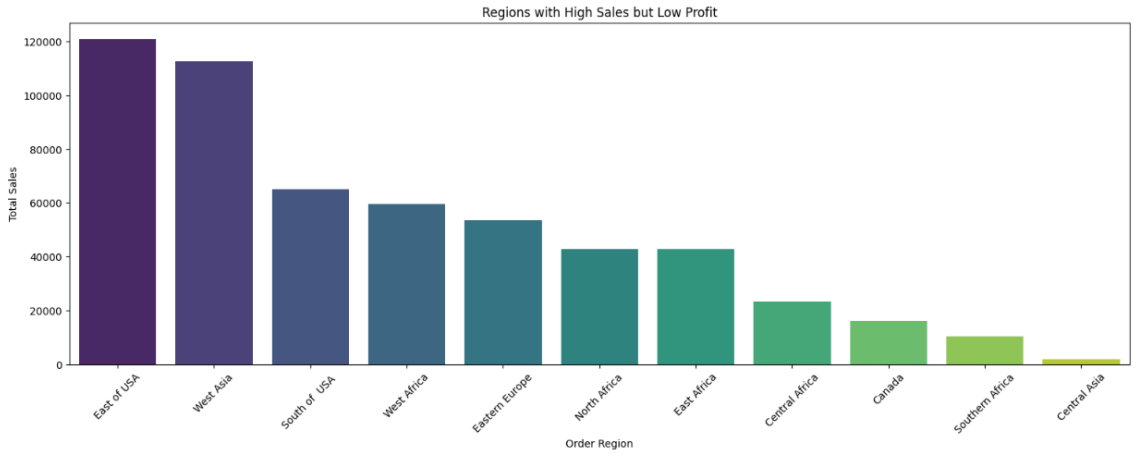






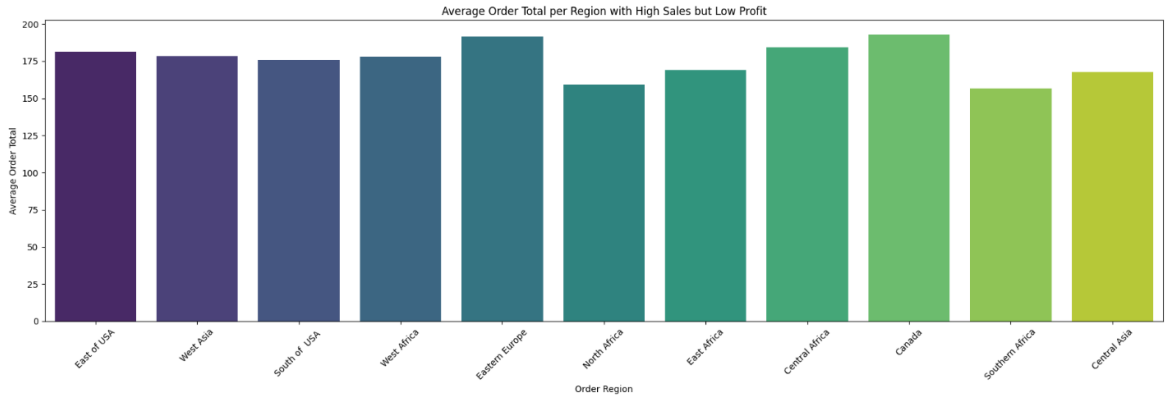
Ans) Fishing Category has highest contribution in sales followed by Cleats and Camping & Hiking Music has a loss impact.

❖ **Geographical Insights:**



Ans) East of USA South of USA , West asia, West Africa Eastern Europe etc has high sales but low profit

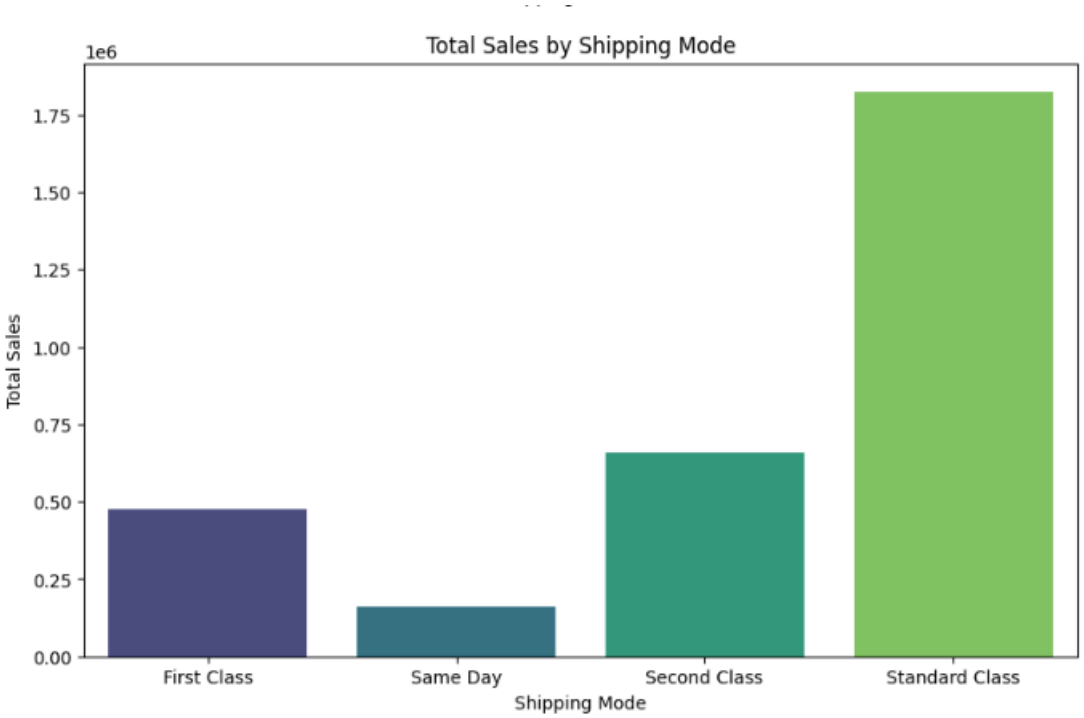
Ans) Average Discounts at high sales and low profit regions is similar. Some parts of Central Asia and Southern are not giving profit as USA and Asia but getting similar discounts



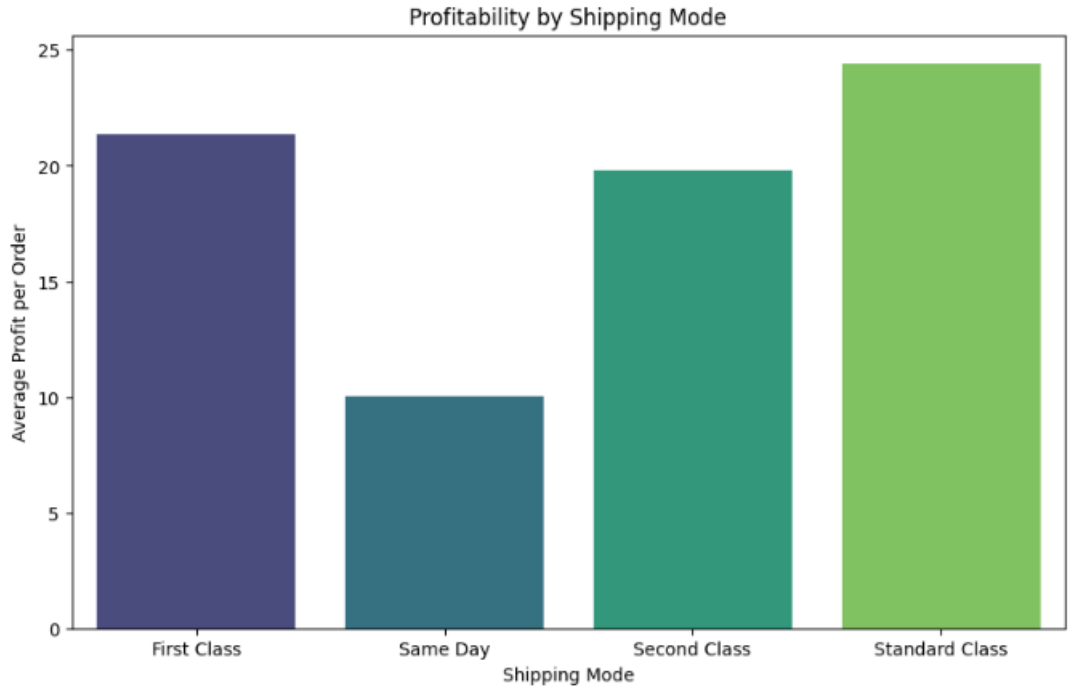
Ans) Number of orders is similar for all these regions

❖ **Sales and Profitability by Shipping Mode:**

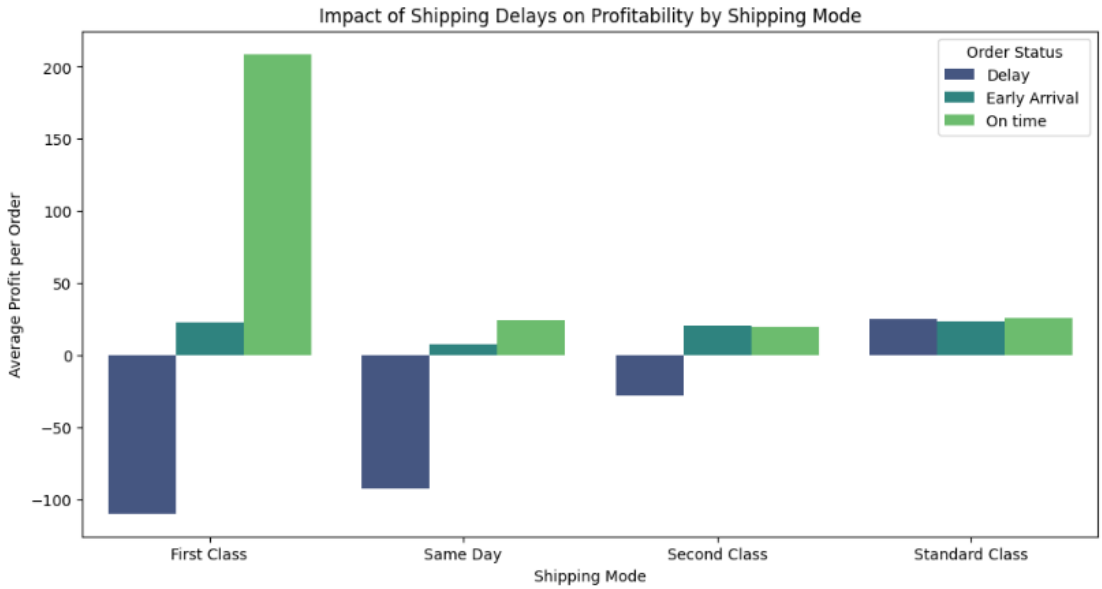
- How does profitability vary across different shipping modes (e.g., Standard, Express)?



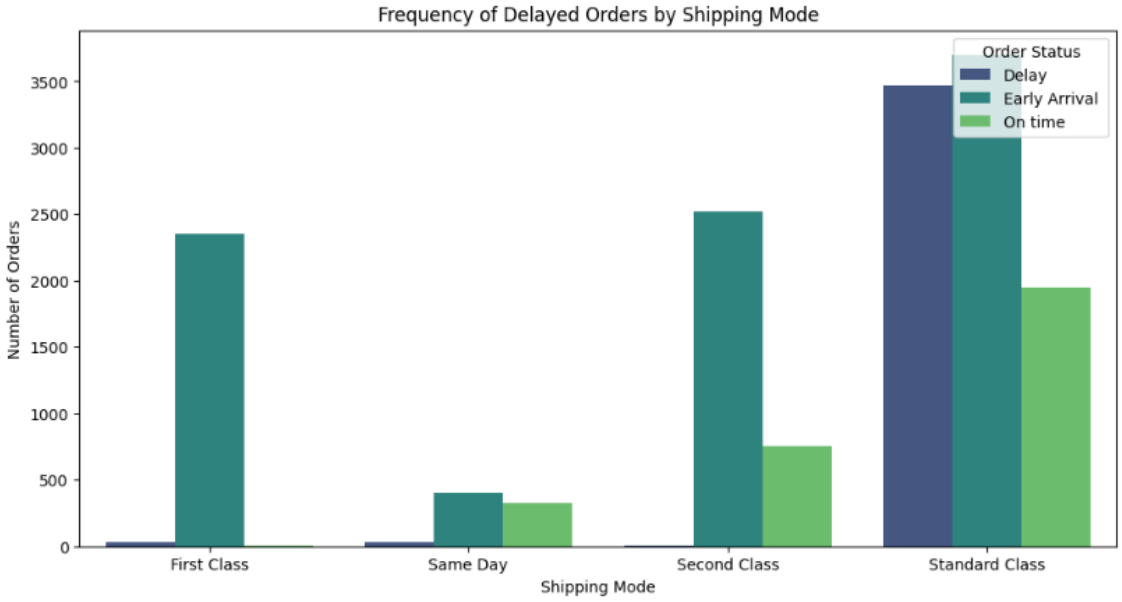
Ans) Standard Class has much higher sales compared to Second_class, Same_Day, First_Class



Ans) Standard Class has much higher profits compared to Second_class, Same_Day, First_Class. Second Class And first class has similar to profits and close to standard class. Same Day is lower profits compared to others



Ans) For first Class and Same Day Delay orders are creating high loss ,followed by second Class. Lastly we have Standard Class which is in profits even in delay orders



Ans) Highest frequency of Delay orders with standard class whereas First Class Same Day Second Class have negligible Delay orders

4. Inferences

Order Volume Insights:

- Western Europe has the highest number of orders, especially in the consumer segment.
- The Standard Class shipping method is most popular, but there is a noticeable distribution of orders across various product categories, with Category ID 17 being dominant.
- Debit payments account for the majority of orders.

Sales Performance:

- The Fishing category contributes significantly to overall sales, with Europe and the Americas driving most of the revenue.
- Sales are relatively consistent across customer segments, but regions like Europe and the Americas dominate in total sales.
- Footwear is the top-performing department in terms of sales.

Profitability Analysis:

- Europe and the U.S. are the regions with the highest profit margins.
- The consumer segment yields the most profit, while higher discount rates generally reduce profitability.
- Technology has the highest profit per order among departments.

Order and Sales Trends Over Time:

- There is a clear seasonal trend where sales and orders peak in the first half of the year (2015-2017) before declining. However, 2018-2019 shows a significant drop in sales and profitability overall.

Customer Behaviour:

- Repeat orders are minimal, and discount rates do not significantly impact order volumes except in Europe and the U.S.

Order Delay and Impact on Profitability:

- Delayed orders significantly impact sales and profits, especially in regions like Western Europe and Central USA.

Product-Specific Performance:

- Specific categories such as Fishing, Cleats, and Camping & Hiking contribute heavily to sales, whereas music-related products show losses.

Geographical Insights and Sales and Profitability by Shipping Mode:

- Regions with high sales (e.g., Eastern Europe, Western Africa, and parts of the U.S.) may struggle with lower profit margins despite having similar discount levels to more profitable regions.

- Standard Class is most profitable even when accounting for delayed orders, while Same Day shipping shows losses with high delay frequency.

5.Key Decisions

 Optimize Inventory in Western Europe: Given the region’s high order volume, inventory management should prioritize Western Europe.

 Target Consumer Segment: With the highest order and profit volume, marketing and promotional efforts should focus on the Consumer segment.

 Shipping Mode Optimization: Encourage customers to use Standard Class to maximize profitability while reducing the negative financial impact of delayed orders in Same Day and First-Class shipping.

 Reconsider Discount Strategies: The current discounting structure does not significantly boost sales in most regions and tends to reduce profitability. Fine-tuning discount strategies in key markets may yield better financial results.

 Improve Delayed Order Management: Implement processes to reduce delays, especially in Western Europe and Central USA, as delays in these regions are severely impacting profitability.

 Focus on Product Categories with Higher Profit Margins: Strengthen marketing efforts around profitable categories like Fishing, Cleats, and Technology, while re-evaluating the performance of loss-making categories like music-related products.

6.Modeling

To predict delivery status (Delayed, On Time, Early Arrival), several classification models were developed and evaluated. The models included Decision Tree, Random Forest, AdaBoost, Gradient Boosting, and Stacking classifiers. Below are the details of the models and their performance.

6.1 Data Preparation

The features used for modeling included a mix of customer information, order details, and logistics data, such as:

- Customer segment
- Shipping mode
- Product category
- Geographic region
- Order-to-shipping days
- Payment type

The dataset was split into training (75%) and testing (25%) sets using train_test_split() with a random state of 1 to ensure reproducibility.

6.2 Model Performance

Decision Tree Classifier
The Decision Tree model was trained with class balancing and minimal complexity pruning (ccp_alpha=0.001). The results show that the model captures patterns with moderate precision and recall, though it slightly overfits the training data.

- Train Accuracy: 68.25%
- Test Accuracy: 69.14%
- Precision and recall were higher for delayed orders but lower for on-time deliveries.

Random Forest Classifier
With 100 trees and similar class balancing, the Random Forest model provided stable results. Although its accuracy was slightly better than the Decision Tree, the model demonstrated better generalization due to its ensemble nature.

- Train Accuracy: 68.60%
- Test Accuracy: 69.29%
- Precision for delayed orders was stronger than for early or on-time orders.

AdaBoost Classifier
The AdaBoost model with 50 estimators and the SAMME algorithm performed similarly to the Random Forest, showing good generalization and moderate accuracy for both training and test sets.

- Train Accuracy: 68.60%
- Test Accuracy: 69.29%
- The model had balanced precision and recall across all delay categories.

Gradient Boosting Classifier
The Gradient Boosting model performed slightly better than previous models, showing improved precision for delayed orders. The model could handle the complexity of the relationships between features better.

- Train Accuracy: 70.04%
- Test Accuracy: 69.67%
- It consistently classified delayed orders with higher precision.

Stacking Classifier
A Stacking Classifier was built using Random Forest, Support Vector Classifier (SVC), and Gaussian Naive Bayes as base estimators, with Logistic Regression as the final estimator. This model captured more complex interactions between features, providing slightly improved accuracy.

- Train Accuracy: 70.23%
- Test Accuracy: 69.82%
- The precision for all categories, especially delayed and early arrivals, improved compared to other models.

6.3 Optuna Optimization for Gradient Boosting Classifier:

In this analysis, the **Gradient Boosting Classifier** was optimized using **Optuna**, an efficient hyper para meter optimization framework. The best set of hyper parameters found through Optuna are:

- **Learning Rate**: 0.0147
- **Number of Estimators**: 342
- **Max Depth**: 8
- **Min Samples Split**: 48
- **Min Samples Leaf**: 3
- **Subsample**: 0.773

Train Classification Matrix (Optimized Gradient Boosting Classifier)

Metric	Class 0	Class 1	Overall/Macro
Precision	0.63	0.87	0.75
Recall	0.87	0.63	0.75
F1-Score	0.73	0.73	0.73
Support	4913	6748	11661

Test Classification Matrix (Optimized Gradient Boosting Classifier)

Metric	Class 0	Class 1	Overall/Macro
Precision	0.61	0.82	0.71
Recall	0.83	0.60	0.71
F1-Score	0.70	0.69	0.70
Support	1660	2228	3888

6.4 Summary of Model Results

All models performed comparably, with the Stacking and Gradient Boosting classifiers showing slightly better performance. Overall, the models were able to predict delays and on-time orders with moderate accuracy, but the prediction of early arrivals proved more challenging. The Random Forest, AdaBoost, and Stacking classifiers were more robust in handling the classification of delayed orders, which is critical for reducing the negative impact of late shipments.

7.Customer Segmentation

In this analysis, both **K-Means** and **Agglomerative Clustering** algorithms were applied to segment customers based on various attributes. The dimensionality of the data was reduced using **PCA (Principal Component Analysis)** for visualization purposes, and the clusters were analysed to compare the performance of the two algorithms.

Key Observations:

 PCA Visualization:

- The scatter plots generated from the PCA-transformed data for both **K-Means** and **Agglomerative Clustering** showed very similar cluster structures. This indicates that both algorithms identified similar patterns in the data.
- The clusters are well-defined in 2D space, but due to the dimensionality reduction, finer details of the clustering behavior might not be fully visible.

 Cluster Analysis:

- The attribute means for each cluster, derived from both K-Means and Agglomerative Clustering, were also quite similar. This suggests that both methods grouped the customers in a comparable way, despite their different clustering approaches.

 Silhouette Score:

- The **K-Means** algorithm achieved a silhouette score of **0.56**, while **Agglomerative Clustering** achieved a score of **0.54**. Both scores are quite close, indicating that both clustering methods are performing at similar levels.
- A silhouette score closer to **1** indicates well-separated, dense clusters. In this case, the scores are moderate, implying that while the clusters are reasonably well-formed, there is some overlap between them.

 DBSCAN:

- **DBSCAN** was tested as an alternative algorithm but resulted in a large number of points being assigned to the cluster labeled **-1** (noise). This suggests that DBSCAN struggles with the data due to its skewness and high variability. Unlike K-Means and Agglomerative Clustering, DBSCAN is highly sensitive to data distribution and density, which likely caused the clustering issues.

8. Conclusions

The analysis revealed several key factors that influence the logistics performance of consumer products, especially regarding delayed orders. Models like Random Forest, AdaBoost, and Gradient Boosting provided reasonable accuracy in predicting order delays based on features like shipping mode, region, and customer segment.

Key insights from the modelling process include:

1. **Shipping Mode Impact:** Orders shipped via Standard Class tend to perform well even when delays occur, whereas Same Day and First Class shipments lead to significant losses when delayed.
2. **Regional Influence:** Delays are particularly prevalent in regions like Western Europe and Central USA, making these areas prime targets for logistic optimizations.
3. **Profit and Delay Correlation:** Early arrivals are associated with higher profitability, while delayed shipments negatively impact both sales and profit margins.

9.Power Bi Dashboard (In progress)

10.Links