Optimizing Delivery Process and Increasing Sales through Personalized Marketing: Target Corporation Case Study

A final-term report for the BDM capstone project

Submitted by

Name: Kamal Kishor Chaurasiya

Roll number: 21f2000804

IITM Online BS Degree program,

Indian Institute of Technology, Madras, Chennai,

Tamil Nadu, India, 600036.

Contents

1. Executive Summary and Title 1
2. Proof of Originality of the Data Collected 1
3. Meta Data and Descriptive Statistics 2
4. Detailed explanation of analysis process/ method 3
    a. Data pre-processing, feature engineering and Basic Analysis 3
    b. Order distribution across different statuses and states 4
    c. Actual vs Estimated Delivery Time Analysis 4
    d. Correlation Analysis: Number of Orders vs Number of Sellers 4
    e. Delayed Orders Analysis 5
    f. Customer Purchase Frequency Analysis 5
    g. State-wise Customer Distribution by Purchase Frequency and Total Count 5
    h. Product Category Demand Distribution 6
    i. Product Affinity Analysis 6
    j. Customer Segmentation Analysis Using K-Means Clustering 7
5. Results and findings 8
6. Interpretation of results and recommendation 17

References

Collab Notebook Link

1. Executive Summary

The report presents a comprehensive analysis of the dataset from Target's operations in the United States to identify key issues and develop strategies to address the challenges faced by the organization. The business is a B2C (Business-to-Consumer) company operating in the retail segment, primarily focusing on delivering products directly to end consumers.

The company has been facing several issues in the operations, particularly with optimizing the delivery process and increasing sales, which have impacted customer satisfaction and overall revenue growth. This project utilizes MS Excel and Python to uncover insights and develop strategies to solve these issues.

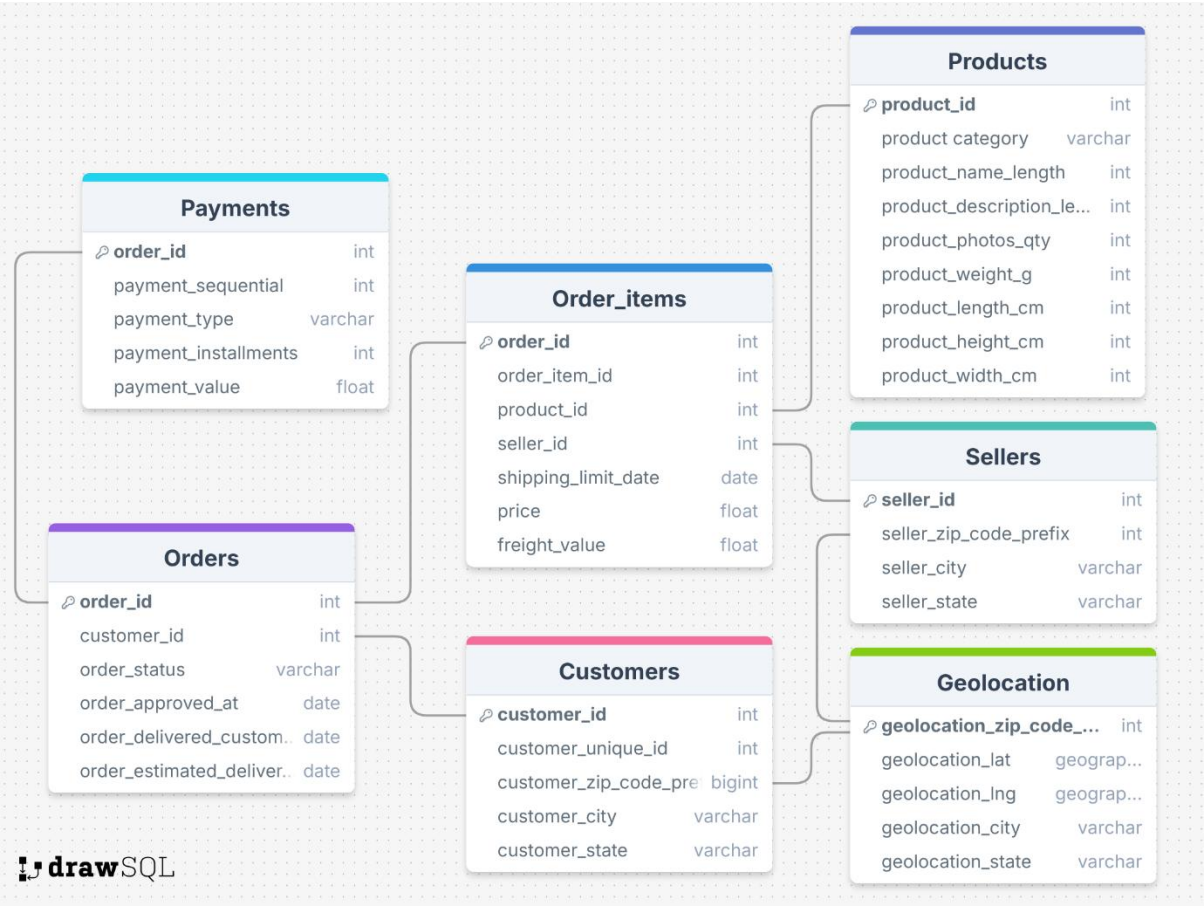
As only 8.11% of the orders were delivered late, the logistics process is largely efficient. However, there is still room for improvement. Through correlation analysis and additional insights obtained via data visualizations, I provided recommendations to further streamline the delivery process. These optimizations aim to improve operational efficiency, enhance customer satisfaction, and ultimately drive an increase in sales and revenue. In addition, the project includes customer segmentation based on recency, purchase frequency, and monetary value, enabling targeted marketing strategies. Product affinity analysis identifies complementary products for bundling, enhancing sales potential. Also, a product preferences analysis highlights the top product categories, while analysis of purchase frequency focuses on identifying key customers to drive sales growth.

By implementing the insights and recommendations derived from the analysis of this project, the company is set to gain a significant competitive advantage in the retail market. The actionable strategies proposed are expected to directly address operational challenges. The anticipated outcomes from these data-driven approaches will not only enhance customer satisfaction but also drive higher sales and improve overall revenue growth.

2. Proof of Originality of the Data Collected

The data used in this project has been collected from the publicly available Target Dataset on Kaggle, created by the user devarajv88. The dataset can be accessed via the following link: https://www.kaggle.com/datasets/devarajv88/target-dataset.

3. Meta Data and Descriptive Statistics



Descriptive Statistics of Datasets

- **Customer Dataset**

  Total Rows: 99,441 | Unique Customers: 96,096 | Cities: 4,119 | States: 27

- **Orders Dataset**

  Total Rows: 99,441 | Delivered Orders: 96,478 | Avg. Actual Delivery: 12 days 13 hrs | Avg. Estimated Delivery: 23 days 18 hrs

  Brief description of order status:

  - *Approved*: The order has been reviewed and approved for processing.
  - *Cancelled*: The order has been cancelled by the customer.
  - *Created*: The order has been placed and recorded in the system.
  - *Delivered*: The order has been successfully delivered to the customer.
  - *Invoiced*: An invoice has been generated and issued for the order.
  - *Processing*: The order is currently being prepared or handled for shipping.
  - *Shipped*: The order has been dispatched and is in transit to the customer.
  - *Unavailable*: The order could not be fulfilled due to stock or logistical issues.

- **Order Items Dataset**
  Total Rows: 112,650 | Products: 32,951 | Avg. Freight Value: 20 | Total Revenue: 13,591,643.7
- **Product Dataset**
  Total Rows/ Products: 32,951 | Product Categories: 73
- **Sellers Dataset**
  Total Rows/ Sellers: 3,095 | Seller Cities: 611 | Seller States: 23
- **Geolocation Dataset**
  Total Rows: 1,000,163 | Cities: 8,011 | States: 27
- **Payments Dataset**
  Total Rows: 103,886 | Payment Methods: 5 | Most Common Payment Mode: Credit Card | Avg. Payment Value: 154.1

4. Detailed explanation of analysis process/ method

The Target Corporation project analysed operational data to uncover key trends, identify challenges, and provide actionable recommendations for optimizing logistics and boosting sales. By addressing pain points and leveraging data-driven strategies, the project aimed to enhance efficiency and drive revenue growth. Below is an explanation of the analysis process and methodology employed:

4.1 Data pre-processing, Feature engineering and Basic Analysis

The data pre-processing phase was critical to ensuring the dataset's reliability and consistency for further analysis. The data, initially stored in separate csv files, was joined to create a unified dataset using python Pandas library. This integration enabled a comprehensive analysis by combining relevant information across multiple dimensions.

Since the data was obtained from a secondary source, it was mostly clean and consistent. However, to streamline the analysis, irrelevant columns were removed to reduce noise, missing values were handled appropriately, and new features were created through feature engineering to capture more nuanced patterns and insights from the data.

A quick data summary was generated using Pandas to gain an initial understanding of the dataset, including its structure, key statistics, and potential inconsistencies, while visualizations provided a clear understanding of the dataset's overall structure. Basic descriptive statistics such as mean, median and others were calculated to understand the data

distribution. The relationships between key variables, such as delivery times, customer locations, and payment methods, were examined using correlation and scatter plots.

The data pre-processing and basic analysis laid a solid foundation for deeper insights by ensuring the dataset was clean, structured, and enriched with meaningful features. This step was crucial in driving the subsequent stages of analysis and delivering actionable results.

4.2 Order Distribution by State and Status

Initially, an order status analysis was conducted to examine the distribution of orders across different statuses, including approved, cancelled, created, delivered, invoiced, processing, shipped, and unavailable. The data was grouped based on status and then percentage was calculated (*Fig 5.1.2*). The analysis revealed that approximately **97%** of the orders were successfully delivered, while only **3%** remained undelivered.

Additionally, a bar chart (*Fig 5.1.2*) was plotted to depict the state-wise distribution of orders using plotly in python. This insight highlights the overall efficiency of the order fulfilment process, while also indicating areas where delivery optimization efforts can be focused to further reduce undelivered orders.

4.3 Actual vs Estimated Delivery Time Analysis

The aim of the Actual vs Estimated Delivery Time Analysis was to compare the actual average delivery time with the average estimated delivery time. The dates were converted to datetime format, and the averages were calculated using pandas to plot a bar chart (*Fig 5.2.1*) and derive meaningful insights from the data.

4.4 Correlation Analysis: Number of Orders vs Number of Sellers

The correlation analysis between the number of sellers and the number of orders aimed to evaluate the relationship between these two variables. The following steps were taken to perform the analysis:

- **Data Gathering**: Data on the number of sellers and the number of orders per state was collected and aligned to create a comprehensive dataset.
- **Data Visualization**: A scatter plot was created using Plotly (*Fig 5.3.1*) to visually inspect the relationship between the number of sellers and the number of orders per state. This helped identify any visible patterns or trends.
- **Correlation Calculation**: The Pearson correlation coefficient was computed to quantify the strength and direction of the relationship between the two variables. This

statistical measure provided a numeric value to understand whether the variables are positively or negatively correlated, and the strength of that correlation.

4.5 Delayed Orders Analysis

In the Delayed Orders Analysis, the goal was to evaluate the factors contributing to delays in order delivery. I first did some descriptive statistics calculation to find total delayed orders, its percentage, average actual delivery time, average estimated time, average approval time and average shipping time. Next, I conducted a test to analyse the distribution of delayed orders based on their geographic movement. Specifically, I examined how many delayed orders were shipped either within the same state or across different states and zip codes (*Fig 5.4.1*). This helped to identify any regional or logistical factors contributing to the delays, providing a clearer understanding of the impact of location on delivery performance.

A scatter plot was plotted between shipping delay time and time taken to deliver (*Fig 5.4.2*) to check how they are associated. A moderate but positive correlation was observed indicating presence of other factors leading to delay in order delivery. To study other factors, I analysed the correlation between various variables, including overall delivery time, transit time, shipping delay, approval delay, and delivery time accuracy. A correlation heatmap (*Fig 5.4.3*) was then plotted to visually represent these relationships and gain deeper insights into the factors influencing delivery delays.

4.6 Customer Purchase Frequency Analysis

The Purchase Frequency Analysis was conducted to understand customer buying behaviour, focusing on how often customers make purchases and how this correlates with the total spending. This analysis helps to identify patterns in customer engagement, highlight high-value customers, and optimize marketing efforts.

The analysis was performed by segmenting customers based on the number of times they made purchases (purchase frequency). For each frequency group (e.g., 1-time buyers, 2-time buyers, etc.), the total amount spent percentage, average order value, and the number of customers in each category were calculated and plotted using MS Excel (Fig 5.5.1).

4.7 State-wise Customer Distribution by Purchase Frequency and Total Count

The Location-wise Customer Distribution Analysis was performed to examine the distribution of customers based on their purchase frequency across different states.

This analysis was done by aggregating the data based on the number of purchases made by customers in each state. The total number of customers and the count of 1-time, 2-time, 3-time, and so on, buyers were computed for each state.

After organizing this data, a chart was created in MS Excel to visualize the distribution (*Fig 5.5.2*). This visualization helps in identifying patterns such as states with higher frequencies of repeat buyers and those with a larger proportion of one-time buyers, offering valuable insights for targeted strategies in different regions.

4.8 Product Category Demand Distribution

The Product Preference Analysis was conducted to gain valuable insights into the distribution of customer preferences across different product categories. The goal was to identify the most popular categories and understand customer demand patterns, which can directly inform inventory management, marketing strategies, and sales initiatives to maximize revenue.

To achieve this, a Product Category vs Order count plot was created using Plotly (*Fig 5.6.1*). The data was first aggregated by counting the number of orders in each product category, providing a clear picture of customer preferences. Categories with a low order count were grouped into an "Others" category for simplicity.

This visualization enabled a detailed understanding of which product categories were performing the best, revealing opportunities to focus on high-demand areas while optimizing offerings in underperforming categories.

4.9 Product Affinity Analysis

To conduct the Product Affinity Analysis for orders with two or more items, the goal was to identify product combinations frequently bought together, which can be leveraged to improve cross-selling, product bundling, and marketing strategies. The analysis was performed using the mlxtend library, a powerful tool for market basket analysis.

- The data was put in a Pandas DataFrame and then encoded into binary values (1 for the presence of a product in the order, 0 for absence). This transformed the data into a binary matrix suitable for association rule mining.
- Using the apriori algorithm from the mlxtend.frequent_patterns library, the binary matrix was processed to find frequent itemsets.

- The association_rules() function was then used to generate rules from the frequent itemsets. The lift metric was used to evaluate the strength of the relationships, with rules that showed high lift indicating strong associations between products.

4.10 Customer Segmentation Analysis Using K-Means Clustering

The customer segmentation analysis using K-means clustering was performed with the following steps:

- The data was loaded and a rank for each customer was calculated based on the order of their purchases, allowing us to focus on the most recent transactions.
- For each customer, three variables were calculated – recency: the number of days since their most recent purchase, relative to the earliest purchase date in the dataset, frequency: the total number of orders placed and Monetary Value: the total payment value.
- Then boxplots were generated for Recency, Frequency, and Monetary value to detect outliers and entries with absolute z-scores greater than 3 were removed, ensuring only valid data remained for clustering. Also, data was standardized using StandardScaler to ensure each feature (Recency, Frequency, Monetary Value) had a mean of 0 and standard deviation of 1.
- To find the optimal number of clusters, the Elbow method was employed and Silhouette Score was used to evaluate the clustering performance.
- Finally, a k-means clustering algorithm was trained with 4 clusters and silhouette score of 0.45, indicating a moderate level of clustering quality.

This process allowed for the segmentation of customers into meaningful four groups (*Fig 5.8.1*) based on their buying behaviour, enabling targeted marketing, personalized offers, and strategic business decisions.

5. Results and findings

5.1 Order Distribution by State and Status

order_status	approved	canceled	created	delivered	invoiced	processing	shipped	unavailable
Order count	2	625	5	96478	314	301	1107	609
Percentage	0.002	0.629	0.005	97.02	0.316	0.303	1.113	0.612

Fig 5.1.1 Distribution of Orders by Status

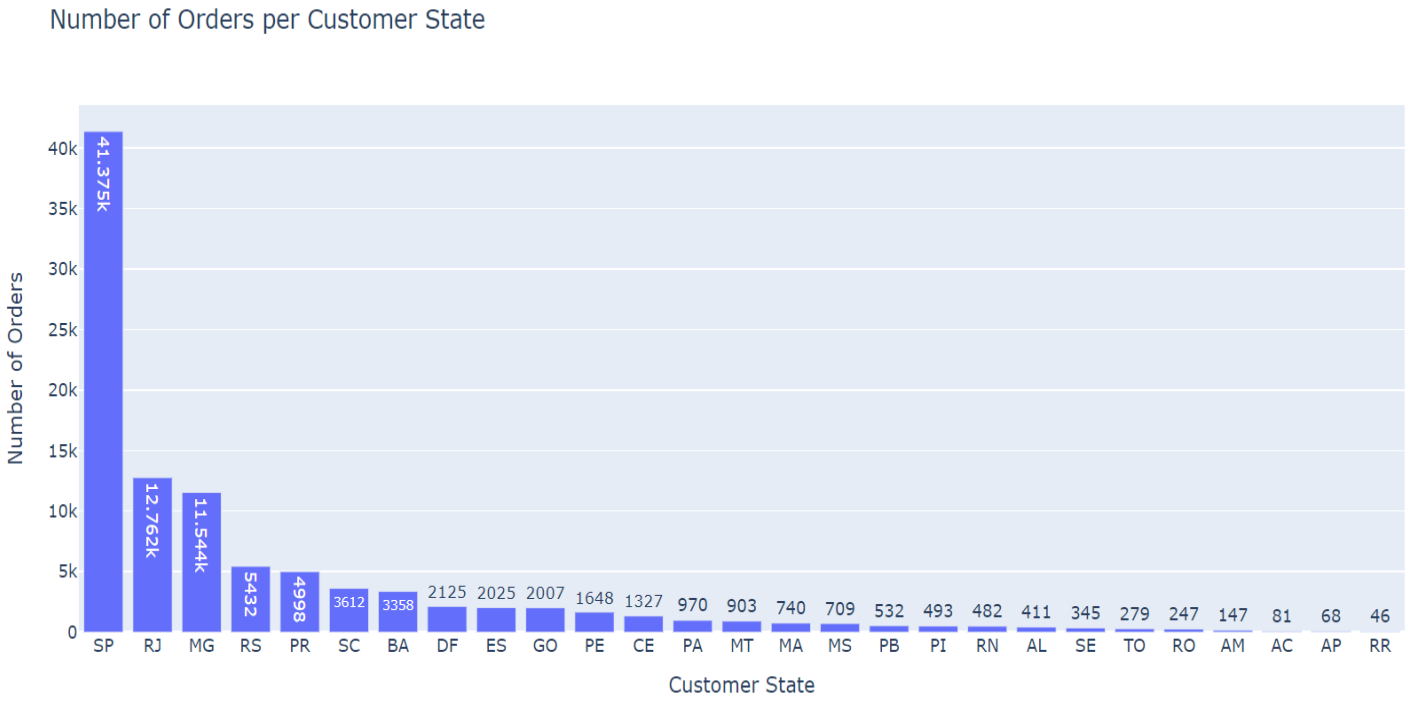


Fig 5.1.2 Distribution of Orders by State

- A significant majority of the orders (**97.02%**) are marked as delivered. This indicates the business is efficiently fulfilling most of its orders. However, the relatively lower percentages in shipped, invoiced and processing may indicate opportunities to enhance shipping logistics and order processing times.
- Orders are highly concentrated in **SP, RJ**, and **MG**, while northern and less urbanized states show minimal order volumes. The business can explore growth opportunities in underrepresented regions through targeted marketing, improved logistics and new partnerships.
- The high concentration in the above states may be attributed to larger urban populations and well-developed infrastructure, leading to increased demand and the presence of high number of sellers cater to a wide range of customer needs efficiently.

5.2 Actual vs Estimated Delivery Time Analysis

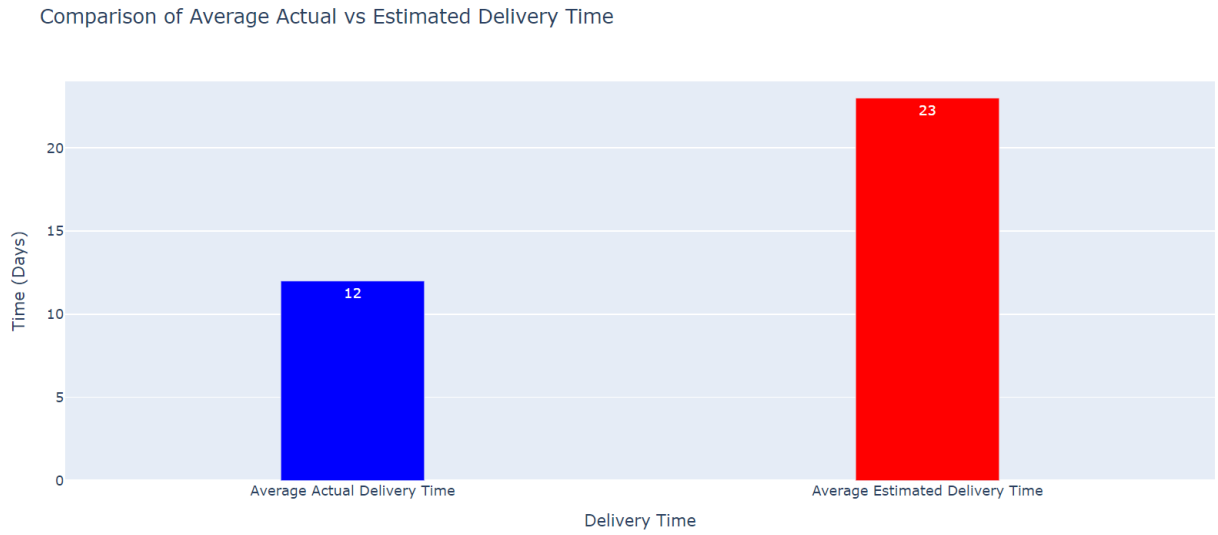


Fig 5.2.1 Average actual and estimated delivery time

- The average actual delivery time is **12 days and 13 hours**, significantly lower than the average estimated delivery time of **23 days and 17 hours**. This indicates overly conservative estimations, which could mislead customers into expecting delayed deliveries.

5.3 Correlation Analysis

The correlation analysis between the number of sellers and order count revealed a strong positive correlation (**0.83)**. This indicates that as the number of sellers increases, the order count also rises significantly.

Fig 5.3.1 Correlation between Number of Orders and Number of Sellers each state

5.4 Delayed Orders Analysis

Out of a total of **96,478** delivered orders, **7,826** orders were delivered late, resulting in a late delivery percentage of **8.11%** which suggests the logistics process is largely efficient. However, there is still room for improvement to minimize delays further.

	late_delivered_order_count
with_same_customer_seller_zip_code	2
with_different_customer_seller_zip_code	7824

Fig 5.4.1 Late Delivered Orders: Comparison of Same Zip Code vs. Different Zip Code Deliveries

- The data indicates that nearly all late deliveries (99.97%) occur when the customer and seller are in different zip codes. This suggests that inter-regional deliveries are the primary cause of delays, likely due to longer transit distances and inefficient routing.
- The lateness in inter-regional deliveries may be attributed to geographic distances.

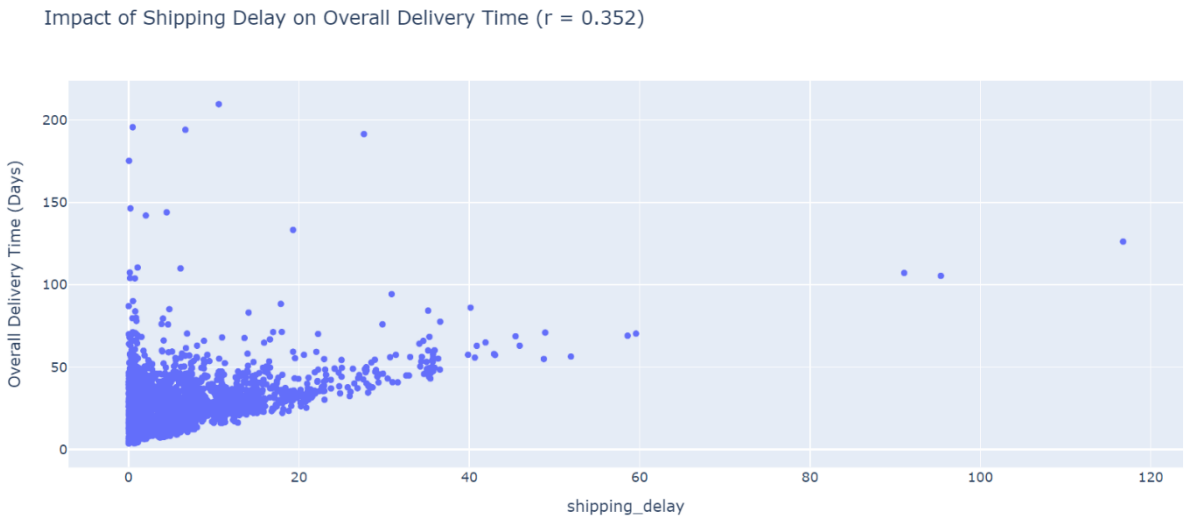


Fig 5.4.2 Moderate correlation between shipping delay (>0) and delivery time

A moderate positive correlation of **0.35** suggests that shipping delays have some influence on the outcome, but other factors also contribute significantly.

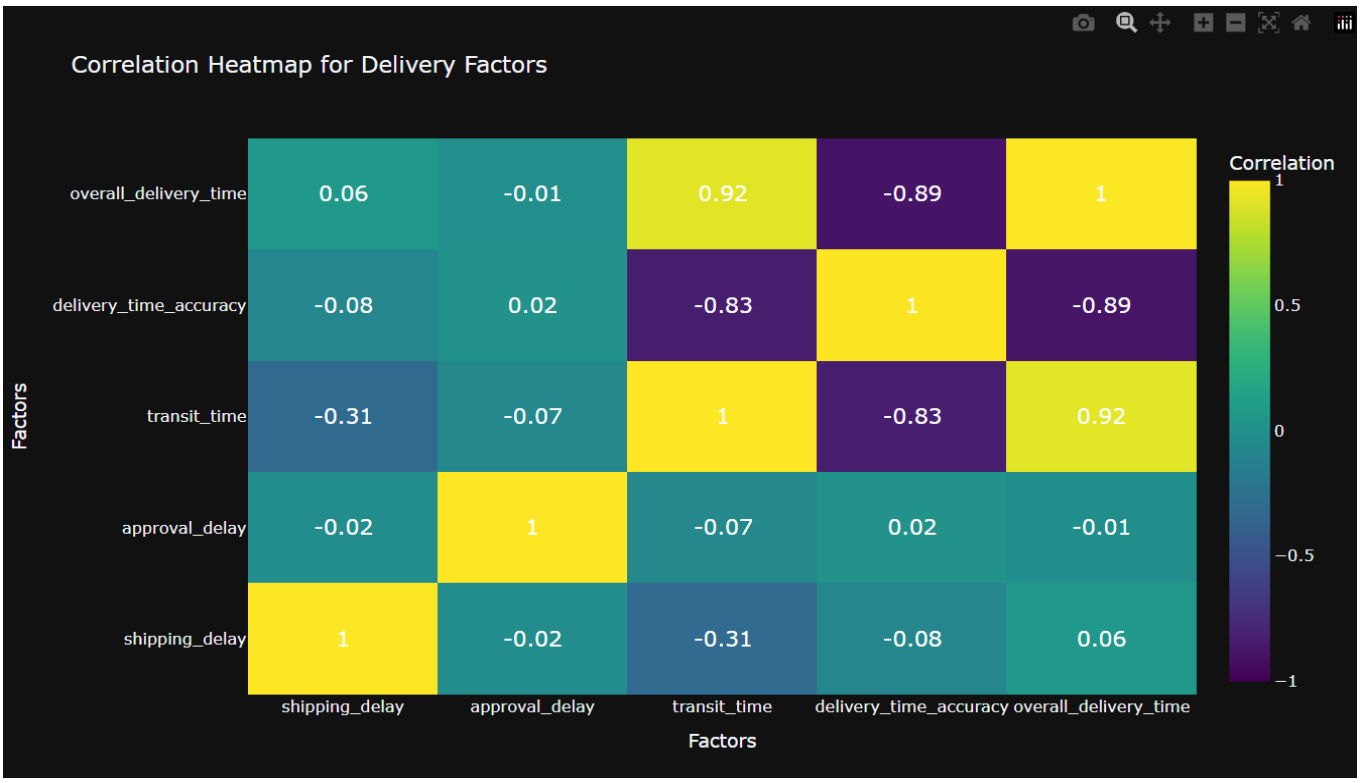


Fig 5.4.3 Correlation between delivery factors

- The strong positive correlation of **0.92** suggests that delays during transportation are a key contributor to late deliveries. This implies that transportation logistics are a major factor affecting the timeliness of deliveries.
- The strong negative correlation of **-0.83** indicates that estimated delivery times are consistently overestimated, meaning customers are expecting longer delivery windows than what is actually required. This can affect customer satisfaction and trust.

The pie chart illustrates the State-Wise Delay Percentage Breakdown, highlighting that the highest delays are observed in SP (**30.5%**) and RJ (**21.3%**). These states also account for the highest number of orders, which could be a significant contributing factor to the delays.

A significant number of states, including **RS, SC, PR**, and **ES**, along with several others, exhibit a delay percentage of less than **5%**.

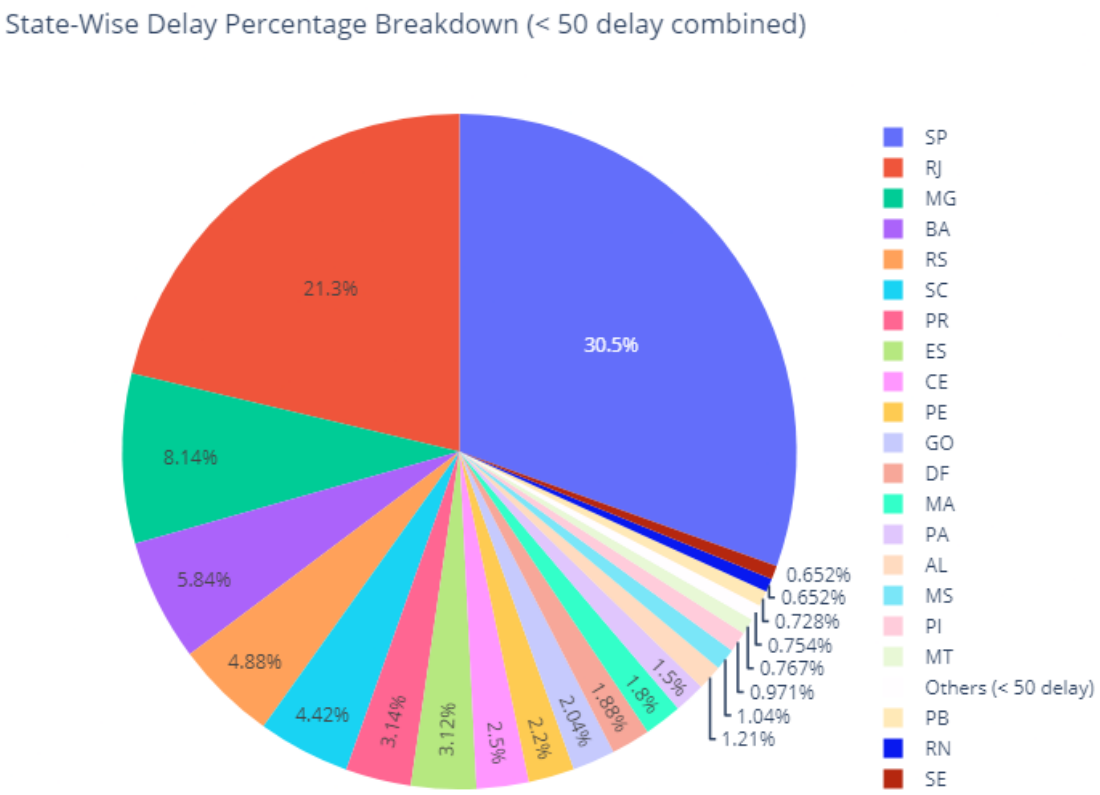


Fig 5.4.4 State-Wise Delay Percentage Breakdown

5.5 Customer Purchase Frequency Analysis

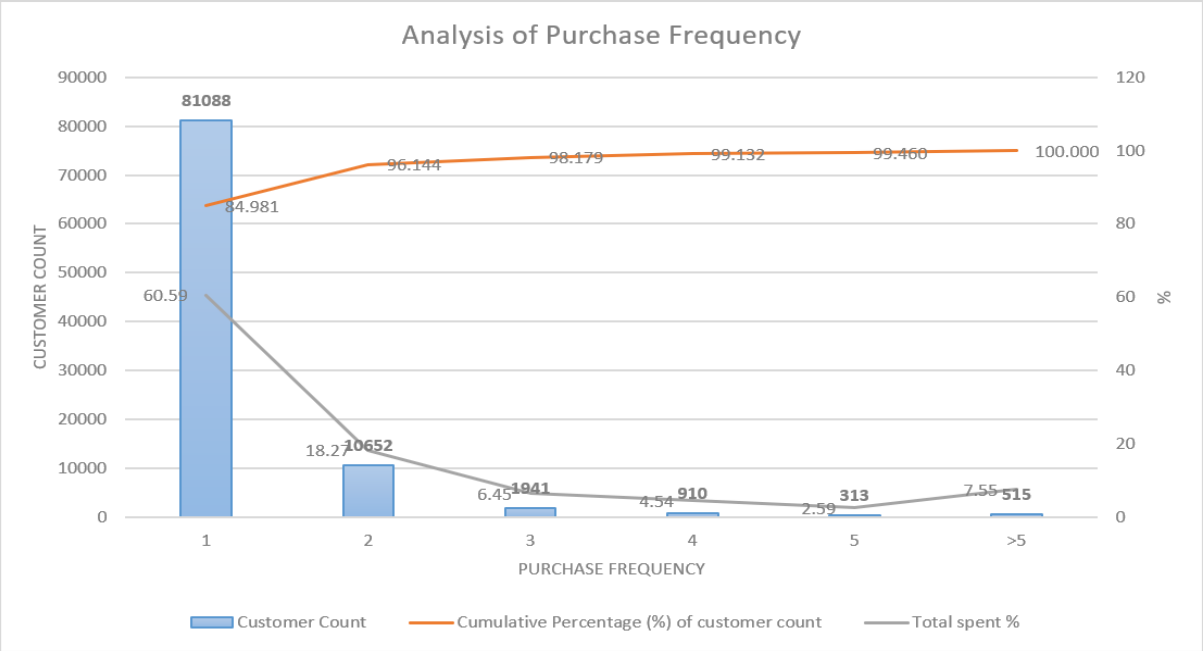


Fig 5.5.1 Customer Purchase frequency Analysis

- A significant portion of the customer base (**80%**) makes only **one purchase** while smaller percentage of customers (those with 2+ purchases) contributes to a disproportionately higher value (premium customers).
- Beyond the first purchase, customer retention drops significantly.

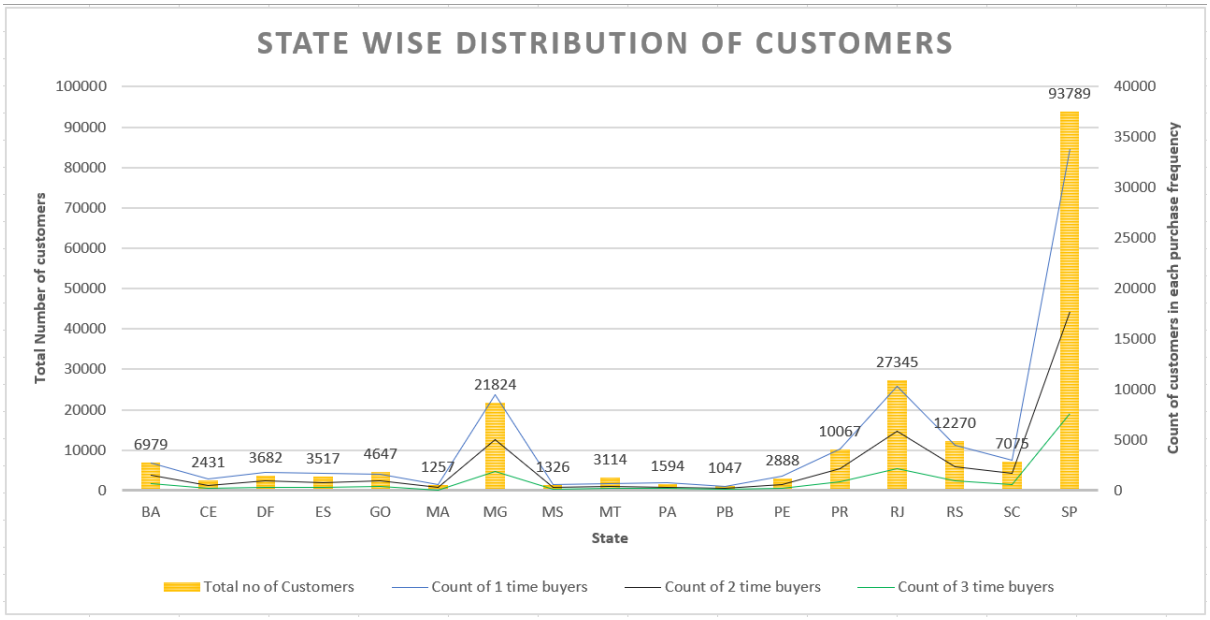


Fig 5.5.2 State-wise distribution of customers

- **SP** has the highest customer count at **93,789,** significantly higher than any other state while there is moderate concentration in **RJ** and **MG** states.
- States like **CE (2,431), DF (3,682), PA (1,594),** and **PB (1,047)** have relatively low customer counts and minimal repeat buyers.
- In every state, the majority of customers are 1-time buyers (blue line), with a very small percentage being repeat buyers (2-time and 3-time buyers).

5.6 Product Category Demand Distribution

The bar graph provides a detailed analysis of product categories based on order count, highlighting the most popular category in demand.

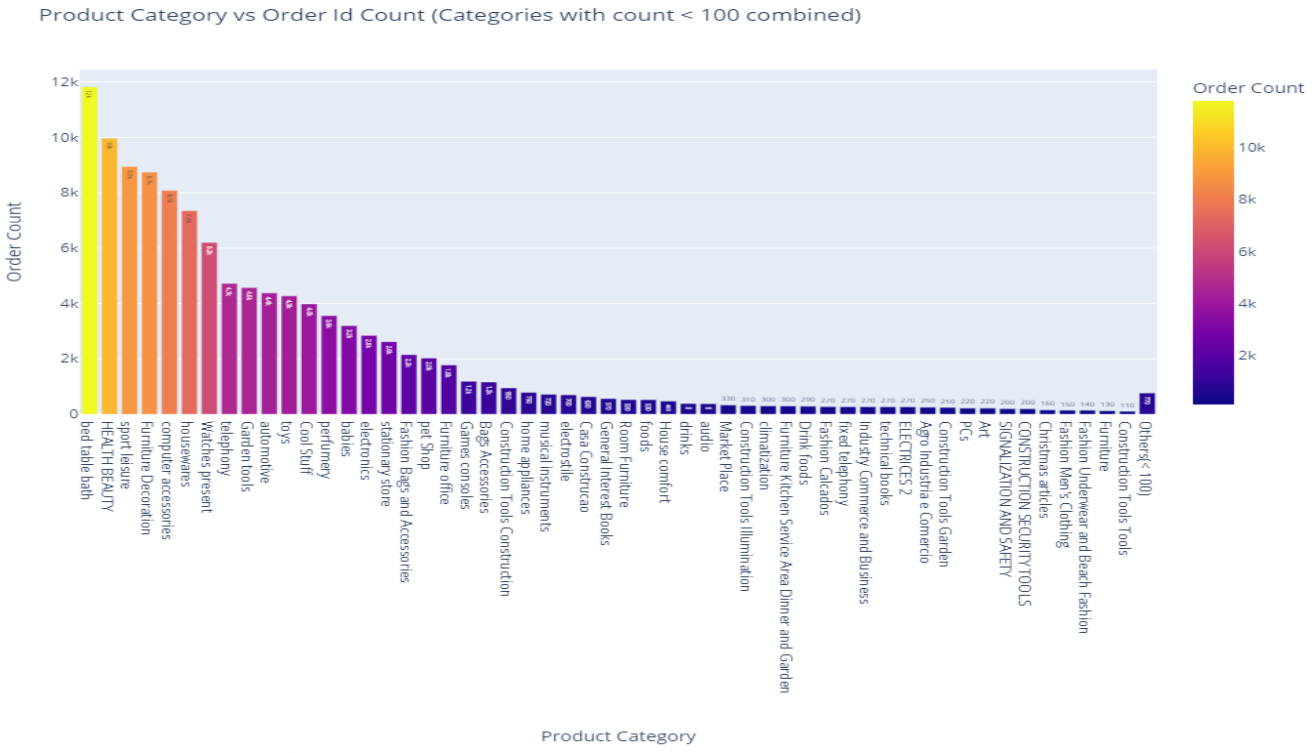


Fig 5.6.1 Product category by Order Count

- The **Bed Table & Bath** product category is the most preferred, with an order count of approx. **12k**, followed by **Health & Beauty** at around **10k**, and **Sports Leisure** at **9k**.
- The under-preferred categories include **Christmas Articles, Fashion Clothing, Furniture,** and **Construction Tools**, with order counts significantly less than 200.
- Many categories are grouped under "Others (<100 orders)", indicating a vast number of niche or underperforming product categories.

5.7 Product Affinity Analysis

The Product Affinity Analysis identified the following categories of products frequently purchased together.

Product Category 1	Product Category 2
House comfort	bed table bath
Cool Stuff	babies
babies	toys
home construction	Furniture Decoration
Construction Tools Illumination	Furniture Decoration

Fig 5.7.1 Product Category frequently Purchased Together

```python
import pandas as pd
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules

def encode(item_freq):
    res = 0
    if item_freq > 0:
        res = 1
    return res

orders_items = pd.merge(orders, order_items, on='order_id', how='inner')
orders_items['item_count_in_order'] = orders_items.groupby('order_id')['order_item_id'].transform('count')
orders_items_item_count_gt_2 = orders_items[orders_items['item_count_in_order'] > 1].sort_values(by='order_id').reset_index(drop=True)
orders_items_item_count_gt_2 = pd.merge(orders_items_item_count_gt_2, products, on='product_id', how='inner')

order_id_product_id_df = pd.crosstab(orders_items_item_count_gt_2['order_id'], orders_items_item_count_gt_2['product category'])

basket_input = order_id_product_id_df.map(encode)
frequent_itemsets = apriori(basket_input, min_support=0.001, use_colnames=True)

num_itemsets = frequent_itemsets['support'].count()
print(f'num_itemsets = {num_itemsets}')
rules_df = association_rules(frequent_itemsets, metric="lift", num_itemsets=num_itemsets)

rules_df = rules_df.sort_values(["support", "confidence","lift"],axis = 0, ascending = False).reset_index(drop=True)
rules_df.head()
```

Fig 5.7.2 Python code snippet for product affinity analysis

5.8 Customer Segmentation Analysis Using K-Means Clustering

The cluster segmentation analysis identified 4 clusters or 4 groups of customers as given below:

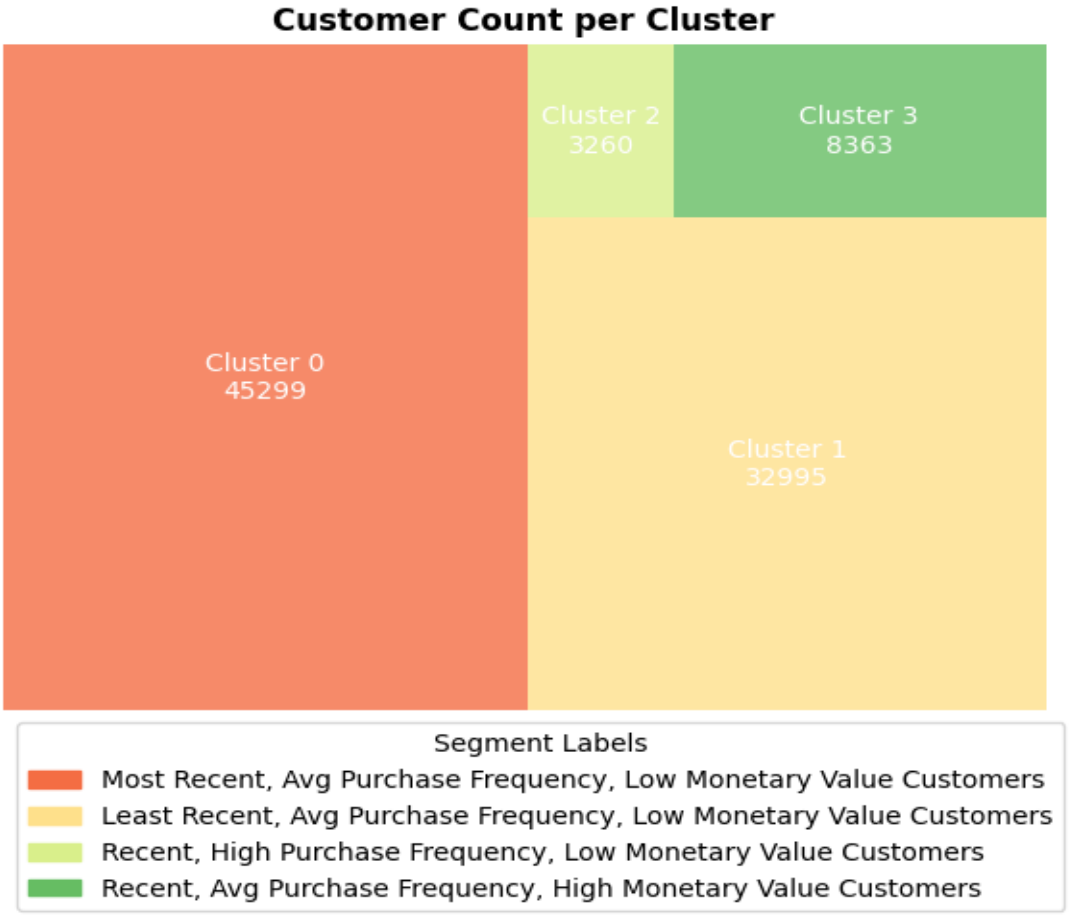


Fig 5.8.1 Customer Segments

- The **Cluster 0** the largest group with **45,299** customers. These customers recently made purchases but have low monetary value, suggesting they are active but low spenders.
- With **32,995** customers in **Cluster 1**, these customers have not made recent purchases, indicating they are at risk of churn and a small group of **3,260** customers in **cluster 2** are frequent purchasers but have low spending amounts.
- **Cluster 3** group has **8,363** customers and represents the most valuable segment. They have made recent purchases with high monetary value, making them key customers.

6. Interpretation of results and recommendation

- We found that **97.02%** of orders are delivered successfully. However, there is small percentage of orders in **shipped**, **invoiced** and **processing** status. Improving process management and enhancing software automation can help reduce the small proportion of orders that remain in the processing, shipped, or unavailable stages.

- The analysis of average actual (**12 days 13 hours**) vs estimated (**23 days 17 hours)** delivery time revealed that the estimated time is way more than actual. The business should update delivery estimation models with real-time data, reduce overly buffers, conduct regular audits to recalibrate estimates, and enhance transparency with realistic delivery timelines. Also, it is advised to establish clear communication through automated notifications will help manage customer expectations and satisfaction.

- To deal with increasing order with increasing number of sellers (**correlation 0.83**), the company should consider strategically partnering with sellers who are in high-demand regions (like states **SP** and **RJ**) or categories (**table bath**, **sports leisure** and **health beauty)** to optimize product availability, ensuring that products are delivered quickly and efficiently across various locations. Also, the logistic capacity should be enhanced by scaling delivery infrastructure (e.g., warehouses, delivery partners) proportionally with the growing seller base and order volume.

- To tackle the lateness in inter-regional deliveries (**7824/7826** late orders within different zip code) due to geographic distances and inefficient routing, it is advised to explore local fulfilment options or optimize logistics for long-distance deliveries using AI-driven routing systems. This can also reduce the high transient time which is also one of the reasons for delayed delivery.

- Almost **85%** of customers are one-time buyers. This can indicate lack of customer retention strategies, unsatisfactory customer experience and ineffective post-purchase engagement. To address this, it is suggested to implement retention strategies like loyalty programs, discounts on repeat purchases, or personalized follow-ups like personalized emails and product recommendations to bring customers back.

- The product categories **Christmas Articles**, **Fashion Clothing**, and **Construction Tools** are low-demand categories. It would be wise to introduce discounts, bundles, or strategic advertisements to drive more sales. The product portfolio should also be streamlined by reassessing inventory and resource allocation.

- The product affinity analysis revealed the products frequently purchased together like **House comfort and bed table bath**, **babies and toys** etc. Implementing an AI-driven recommendation system to suggest complementary products at checkout based on affinity and organizing these products closer together in both online listings and physical stores can boost overall sales and customer satisfaction.

- For the recent and high value customer segments (**Cluster 3**), the focus should be on retaining them through personalized services, loyalty programs, and premium experiences. The segment, which is frequent buyer and low spending (**Cluster 2**), can be offered exclusive rewards and discounts to incentivize higher spending. For At-risk of churn group (**Cluster 1**), re-engagement campaigns such as personalized offers, discounts, or reminders can be implemented to bring these customers back.

*** End Of The Report ***