Data cleaning and Exploratory data analysis to optimize the efficiency of inventory

Nikitha Kalali

1.Introduction

"Why is Supply Chain Management so important?" - Amid the complex net of the international trade chains, the management of supply networks becomes nuts and bolts to the success and sustainability of a business. From factory assembly belts when there is huge production to retail shelves in stores, Supply Chain Management orchestrates the flow of goods and services with precision and foresight.

Supply chain management covers the procurement of materials to distribution of final products to the end client through a wide range of processes. Ideally, this underlines SCM as the mechanism that aims at reducing the costs spent throughout the whole supply chain and improving efficiency. In other words, a successful SCM ensures that products reaching the right place at the right time and in the correct quantity are delivered. This also meets customer demand, thereby enhancing profitability to increase competitive advantage.

The current environment of the SCM resembles dynamic market needs, complex planning network, and an increasingly auditing consciousness that prioritizes sustainability and transparency. With the constant eruption of these complications, conventional SCM plans with their sequential computational strategy and reactionary decision-making paradigms are being rendered deficient in catering for the modern market demands. The impact of digital technologies on supply chain management is undeniable, as it creates an environment of data- led decisions and predictive analytics. This change is not a temporary phenomenon but a paradigm shift, in essence, in the structure and the model of a supply chain. The SCM department of companies in general face several challenges while operating in the day-to-day schedules, and data analytics turns out to be an effective solution that can seamlessly address any such challenges. Below mentioned are a few real time examples of challenges across SCM along with the data analytics solutions that have led to their effective resolution.

There are many challenges in SCM like supply chain planning, sourcing and procurement, logistics and distribution. We are going to solve some of the key challenges like demand forecasting, customer segmentation and fraud detection. Demand forecasting predicts customer demand for a product or item over a defined period. For an effective SCM, precise prediction of demand stands pivotal. Data analytics, especially machine learning models such as time series analysis and neural networks, is known to enhance the precision of demand forecasting. This digital revolutionary approach aids companies in analyzing past sales data, market trends, and consumer behavior that give them insights about expected demand in the future, giving them a broader picture of planning and allocating resources accordingly. Inventory management is a process of keeping track of stock levels and the logistics of goods. Effective inventory is the important crucial part that can make profit to the organization and makes supply chain efficient. Data analytics offers insights into stock levels, profit rates, and reorder stages thus assisting the organization in maintaining the required amount of stock and reducing the excess storage problems. A Fast fashion brand Zara demonstrates that SCM can be used in strategic fashion to swiftly react to trends in the market. Capitalizing powerful data analytics for demand prediction, Zara manages to lead in terms of time turnaround from design to store shelves and, as a result, decreases both overstock and stockouts. Therefore, the retailer enjoys the flexibility of their supply chain which enables them to update their inventory frequently so that their products are extremely close to consumer preferences and market demand. Starbucks uses advanced data analytics for demand forecasting to optimize inventory levels and minimize wastage. By examining transaction data, weather conditions, and nearby events, Starbucks can forecast changes in demand at specific store locations, guaranteeing that each store is supplied with the appropriate amount of goods. This precision in forecasting significantly improves operational efficiency and customer satisfaction by guaranteeing product availability without excessive wastage.

The Dataset used in this analysis is Mendeley dataset (Mendeley_dataset)

FIELDS	DESCRIPTION
Type	: Type of transaction made
Days for shipping (real)	: Actual shipping days of the purchased product
Days for shipment (scheduled)	: Days of scheduled delivery of the purchased product
Benefit per order	: Earnings per order placed
Sales per customer	: Total sales per customer made per customer
Delivery Status	: Delivery status of orders: Advance shipping, Late delivery, Shipping canceled, Shipping on time

Late_delivery_risk	: Categorical variable that indicates if sending is late (1), it is not late (0).
Category Id	: Product category code
Category Name	: Description of the product category
Customer City	: City where the customer made the purchase
Customer Country	: Country where the customer made the purchase
Customer Email	: Customer's email
Customer Fname	: Customer name
Customer Id	: Customer ID
Customer Lname	: Customer lastname
Customer Password	: Masked customer key
Customer Segment	: Types of Customers: Consumer, Corporate, Home Office
Customer State	: State to which the store where the purchase is registered belongs
Customer Street	: Street to which the store where the purchase is registered belongs
Customer Zipcode	: Customer Zipcode
Department Id	: Department code of store
Department Name	: Department name of store
Latitude	: Latitude corresponding to location of store
Longitude	: Longitude corresponding to location of store
Market	: Market to where the order is delivered: Africa, Europe, LATAM, Pacific Asia, USCA
Order City	: Destination city of the order
Order Country	: Destination country of the order
Order Customer Id	: Customer order code
order date (DateOrders)	: Date on which the order is made
Order Id	: Order code
Order Item Cardprod Id	: Product code generated through the RFID reader
Order Item Discount	: Order item discount value
Order Item Discount Rate	: Order item discount percentage
Order Item Id	: Order item code
Order Item Product Price	: Price of products without discount
Order Item Profit Ratio	: Order Item Profit Ratio
Order Item Quantity	: Number of products per order
Sales	: Value in sales
Order Item Total	: Total amount per order
Order Profit Per Order	: Order Profit Per Order
Order Region	: Region of the world where the order is delivered : Southeast Asia ,South Asia ,Oceania ,Eastern Asia, West Asia , West of USA , US Center , West Africa, Central Africa ,North Africa

	,Western Europe ,Northern , Caribbean , South America ,East Africa ,Southern Europe , East of USA ,Canada ,Southern Africa , Central Asia , Europe , Central America, Eastern Europe , South of USA
Order State	: State of the region where the order is delivered
Order Status	: Order Status: COMPLETE, PENDING, CLOSED, PENDING_PAYMENT, CANCELED , PROCESSING ,SUSPECTED_FRAUD ,ON_HOLD ,PAYMENT_REVIEW
Product Card Id	: Product code
Product Category Id	: Product category code
Product Description	: Product Description
Product Image	: Link of visit and purchase of the product
Product Name	: Product Name
Product Price	: Product Price
Product Status	: Status of the product stock: If it is 1 not available, 0 the product is available
Shipping date (DateOrders)	: Exact date and time of shipment
Shipping Mode	: The following shipping modes are presented: Standard Class, First Class, Second Class, Same Day

Data Analysis

3. Data Preprocessing

For the EDA, the dataset goes through the preprocessing stage to check for the data quality issues and their relevance for the applied models. A data frame summary is generated depicting critical attributes of the dataset to ensure dataset integrity and avoid any issues that may arise from the missing values to enable proper progress into deeper analysis undefined.

The data includes various data types, such as object, int64, and float64. Some columns contain a dense number of unique values, showing a lot of data (for example, 'Benefit per order' has 21,998 unique values). Several of the columns contain null values, and 'Order Zipcode' is among the greatest thus having 86.24% of missing data. There is no ' Product Description' at all (100% missing values). Certain table columns, such as 'Customer Email' and 'Customer Password', contain only one unique value which might indicate redundancy, meaning the columns might not have a significant contribution to the dataset. Columns such as 'Customer Lname' and 'Customer Zipcode' are uncoordinated and have missing values, even though these are essential customer information. This could eventually affect the dataset reliability and completeness and has effects on the model outcomes.

Using this summary, the variables likely to be affected by missing data will be identified, the columns that do not suffer much variation and contain high rates of null values will be dropped, and assurance of data type conversion will be ensured.

Dropping unnecessary columns

The dataset has been updated with the following modifications:
Columns Removed: We removed columns with 0 or 1 unique values, as well as specific customer-related columns like 'Customer Email', 'Customer Fname', 'Customer Lname', 'Customer Password', 'Customer Address', 'Customer Zipcode'. It also simplifies the dataset by getting rid of any irrelevant or trivial data. The deletion of these columns should not change the reliability of the rest of the data, hence allowing for accurate analysis or forecasting.
This dataset now has 46 columns, less than in the original one. First, this would make the dataset easier to handle and relevant for the subsequent analysis or modeling.
The main columns which include characteristics of shipping, sales, and products as well as order information have been retained, therefore, the important aspects of supply chain data are still available for analysis.

Data Precision Optimization

In the data preparation phase, all the floating-point variables within the data set, aside from latitude and longitude for geospatial accuracy, are rounded off to two decimal places. This step is the most important one in the process of achieving data uniformity and readability where such factors as financial figures which require high precision are dealt with. In the last step of post-processing, the dataset is presented to the users so that any changes made are confirmed. The dataset is now in standardized format to which further analysis can be performed.

Temporal Data Structuring for Analysis

The timestamp data preparation step requires the rearranging of temporal information. First, in order date and shipment "date" fields inconsistent with "/" replaced with "-" to provide uniform date format across the dataset. Toward the end, the columns get converted into datetime objects which allows to search the exact date and time components. This selected function assists department allocation of 'Order Date' and 'Order Time' including 'Shipping Date' and 'Shipping Time' that may be relevant in analyzing order processing and shipping efficiency. The data is downsized by removing datetime columns after the necessary data is extracted. The next step involves the first few rows of the dataframe being modified; presenting a confirmation of the accuracy and soundness of the transformation. This circumstance serves as a foundation for the time-sensitive supply chain analytics.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) plays a seminal role in examining data and serving as the very foundation for further modelling efforts. The purpose of EDA is to spot patterns, trends, and relationships among the data. Afterwards, the identified patterns, trends, and relationships aid the researcher to develop effective analytical strategies. The exploration in this initial step helps to understand the relationship between different dimensions of the dataset and make decisions regarding modeling later.

The graph in Fig.1 below provides a visible classification of sales performance in various categories, with the categories' performance showing a trend of concentration of specific customers' preferences fueled by trends, product quality, or marketing effectiveness. It also helps to understand the low sales of products and the factors contributing to them like the low awareness and lack of stock variety.

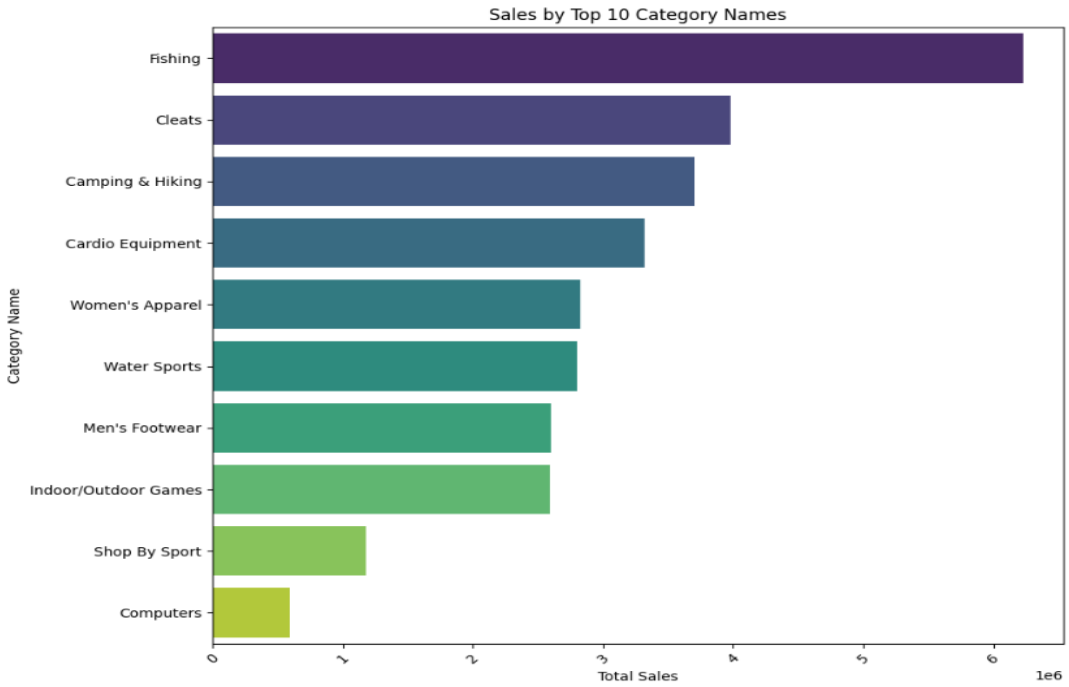

Fig. 1. Sales by Top 10 Category Names

The line graph in Fig.2 below exhibiting sales overtime furnishes the understanding about the sales volume pace across different periods since fluctuations attributable to seasonal changes, marketing campaigns, or other saw marketing factors are expected to occur during this period. For most of the time the sales did not drop below 0.8, indicating a constant trend of fluctuations. To realize the importance of trends in sales, one should understand the factors that affect the markets as well as sales strategy. Indexing peak and lull of sales makes it easier for business owners to examine the pattern of seasonality or events such as promotions which may have influenced consumer behavior.

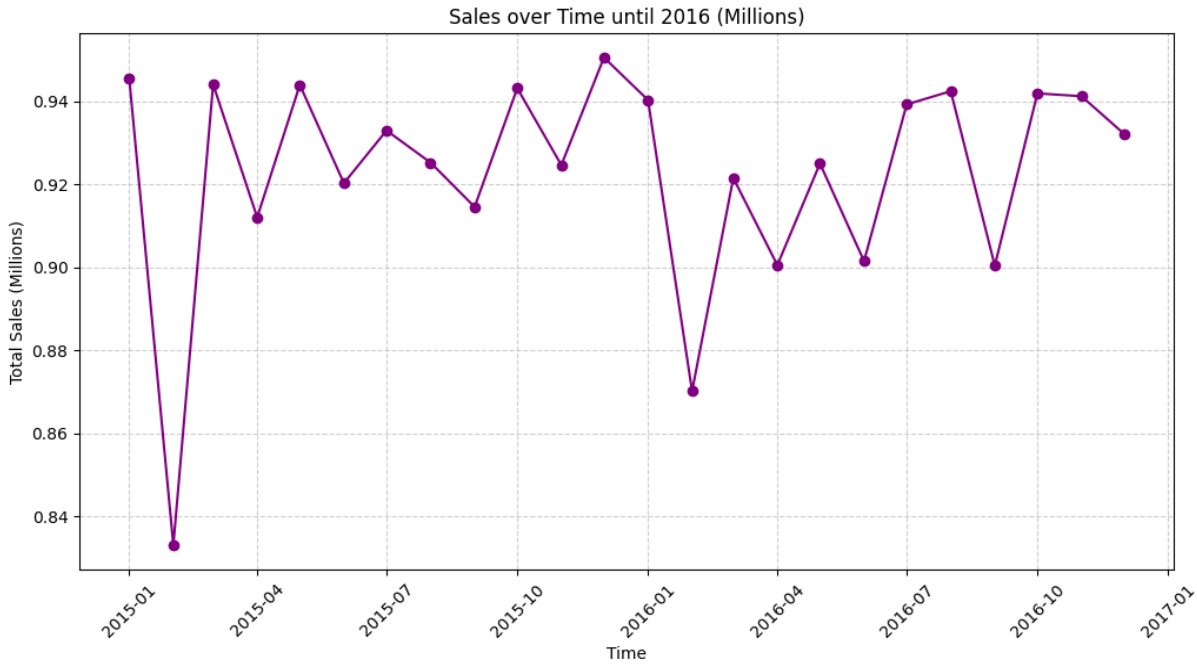

Fig.2. Line Graph of Sales over Time Insights and Analysis

The graph in Fig.3 below shows the sales by shipping mode. It is evident that the sales with standard class delivery account for 59.9% of all the shipping modes followed by second class delivery with 19.4% and the least is the same day delivery which accounts for 5.3%. This indicates the preferences of the customers shipping mode are based on price, speed, and convenience.

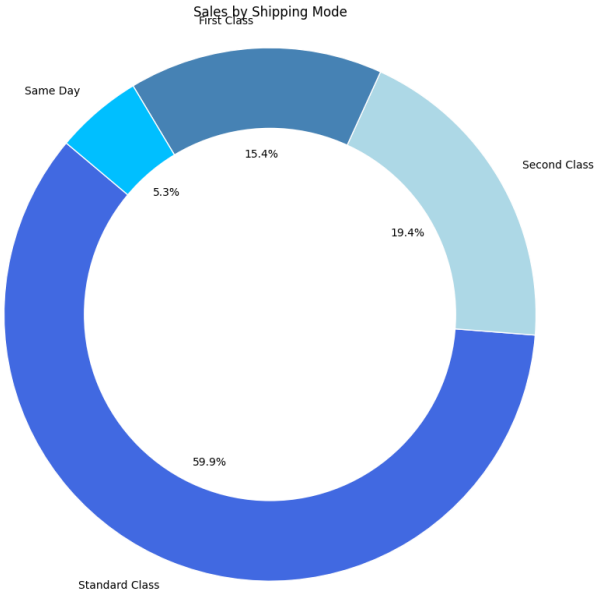

Fig.3. Sales by Shipping Mode

From the histogram in Fig.4 below it is evident that the average sales of computers are the highest of all accounts at 1300 followed by Garden. There is a marginal difference in the average sales per customer for the Top 1 product and the rest of the products.

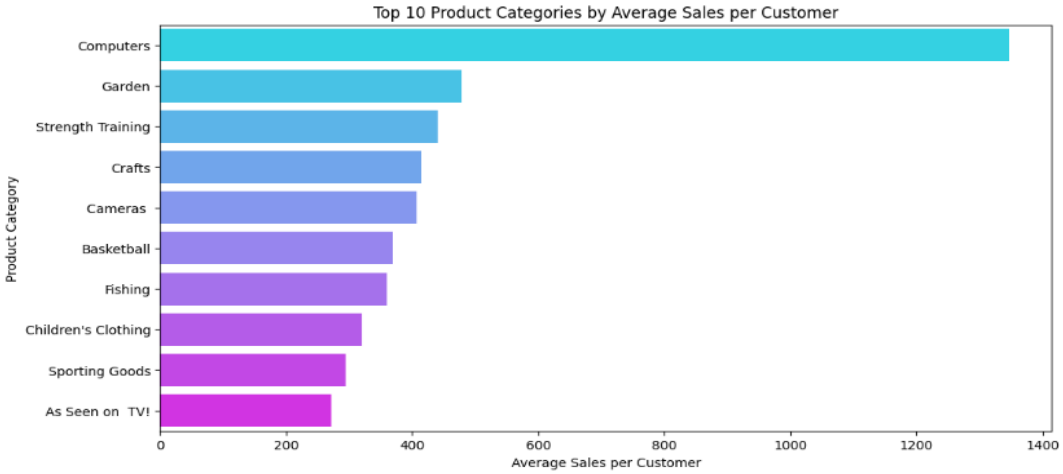

Fig.4. Top 10 Product Categories by Average Sales per Customer

The box plot in Fig.5 below describes the deviation "of" "Days for shipping (real)" being presented in "Shipping Modes" along with a Standard Class having the biggest interquartile range. The channel types of "Same Day" and "First Class" with narrower coverage suggest a generally optimal delivery time. Sporadic occurrence of outliers both in "Standard Class" and "Second Class" is an indication that occasionally some products were delivered with delays out of their usual delivery times. Implications that may arise include client dissatisfaction, the low-quality feeling of the economy class and the potency of the premium delivery options to carry higher fees and the need for improvement in the work processes which will be affected by the outliers.

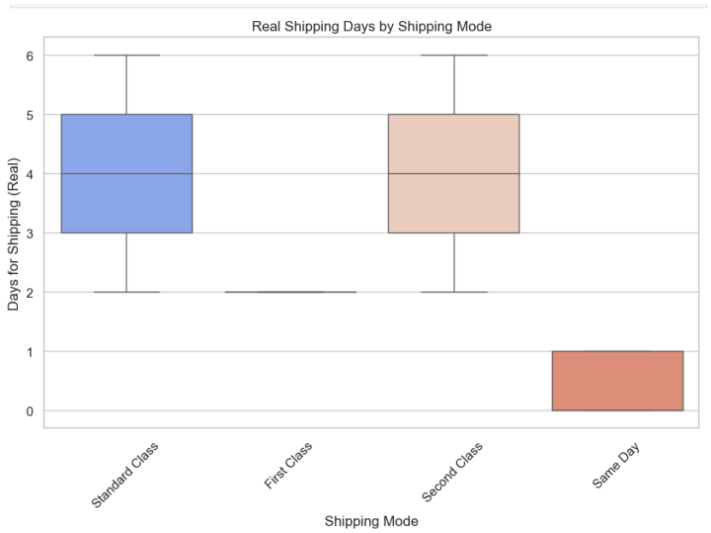

Fig.5. Real Shipping Days by Shipping Mode

The bar chart in Fig.6 below shows a descending order of the mean product price across various categories. The shop's organization into sales categories on the left can denote higher average prices, which might associate with premium products or production-cost concerned commodities. One of the business implications involves pricing that corresponds to the customer's perception of the market positioning, inventory management for categories that are often averagely priced, and then lastly sales and marketing strategies are devised to the model. Suggestion requires reviewing pricing strategies to ensure they accurately reflect target markets, sophisticated marketing campaigns for more expensive segments should be developed, and market analysis will be conducted since the main objective is to maintain competitiveness while protecting the margins (especially in high average price categories).

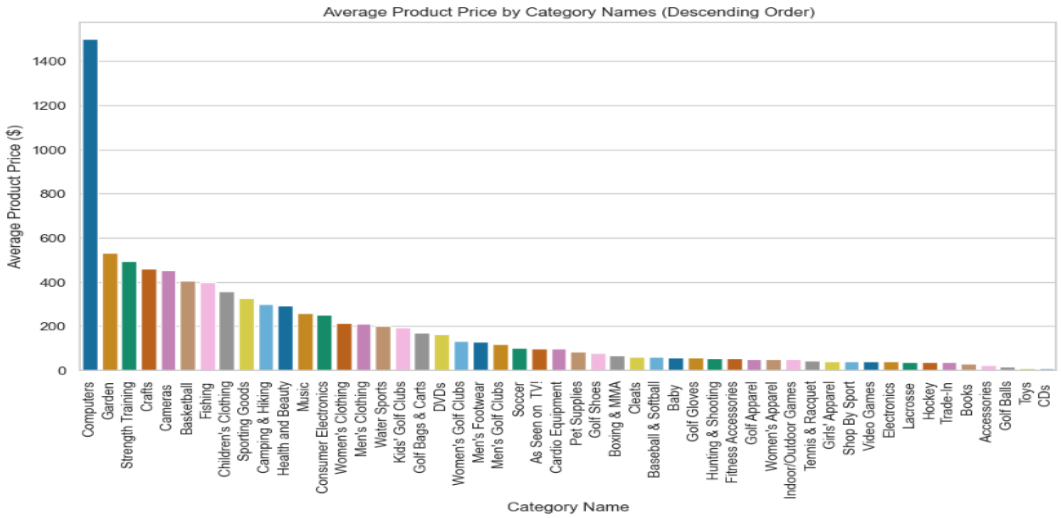

Fig.6. Graph Average Product Price by Category Names (Descending Order)

The graph in Fig.7 below reflects the shipping volumes' change through the days of the calendar month, showing the highest activity on days. We can see that the number of shipments is above 6000 on specific days of the calendar month which include 4,11,24. The shipments count has been reduced drastically on the month end with only above of 3000 shipments. Apart from the month end, all the days in the month have an average of above 5000 shipments per day.

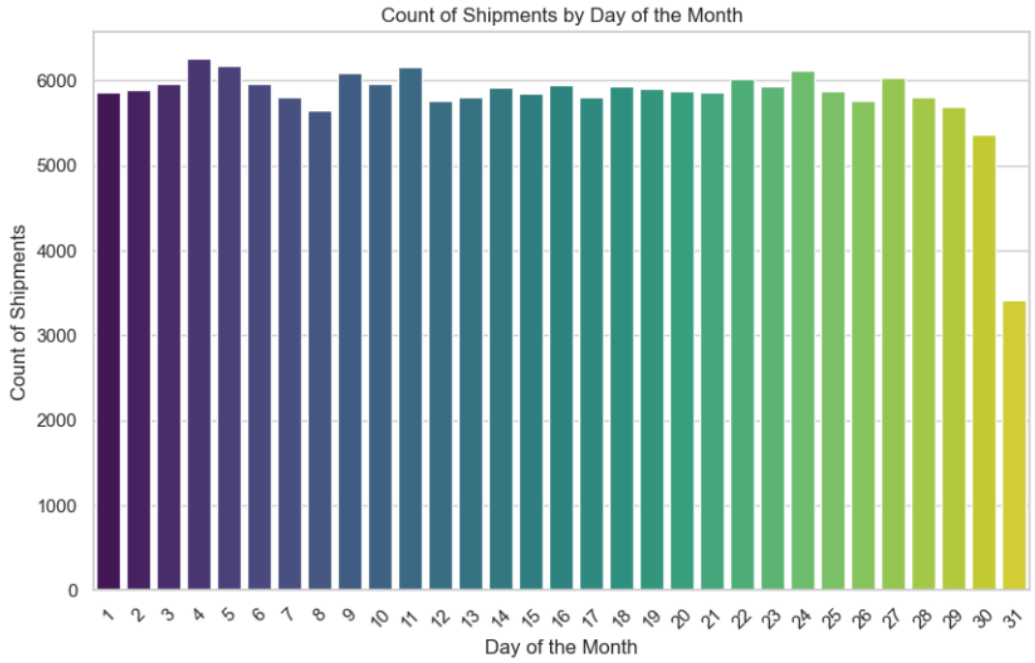

Fig.7. Count of Shipments by Day of the Month

Fig.8 below indicates the shipments by month in a year. Firstly, the total shipments in the first month are the highest, accounting for about 17500 shipments. There is no specific trend from the data as we can see fluctuations in the shipments based on customers' demand, indicating seasonality. However, by the end of the year, i.e.,10 month, the

shipments started to decline below 15000 shipments and in the 11-month shipments were below 12500, which is the least throughout the year**.**

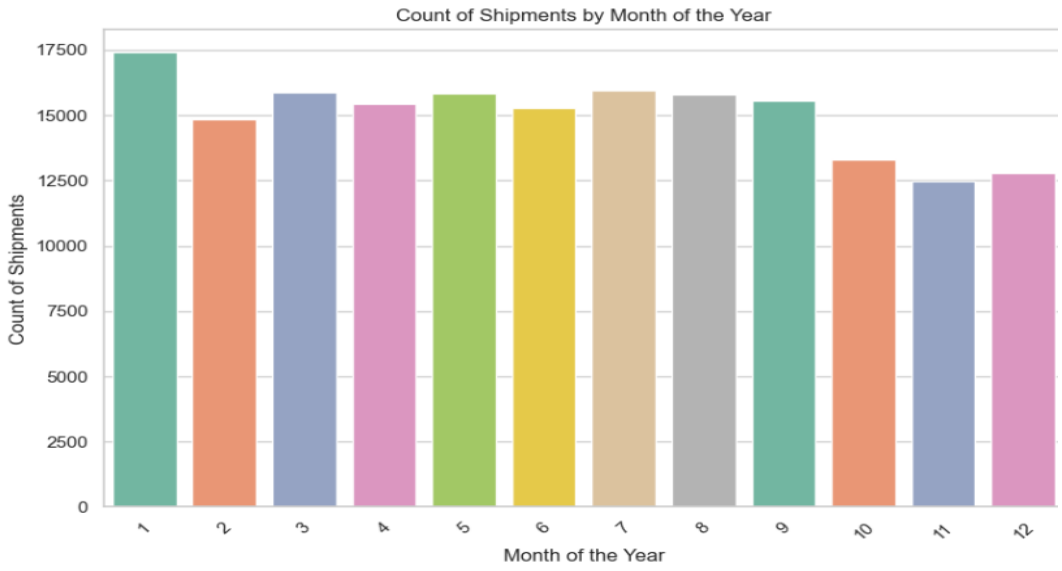

Fig.8. Count of Shipments by Month of the Year

The analysis in Fig.9 below emphasizes the late delivery occurrences in all market segments. Late delivery has the risk of customer dissatisfaction and brand reputation. It also indicates operational inefficiency. The value 1 indicates the risk of late delivery being highest and 0 indicates the less risk of late delivery. The graph shows the frequence of occurrence of 1 for 0 in late delivery column from the data frame as count. The highest misses were in the 'Consumer' group in both on-time and late deliveries. But it should be mentioned that the 'Home Office' segment aggravates the timely delivery problem impacting others yet.

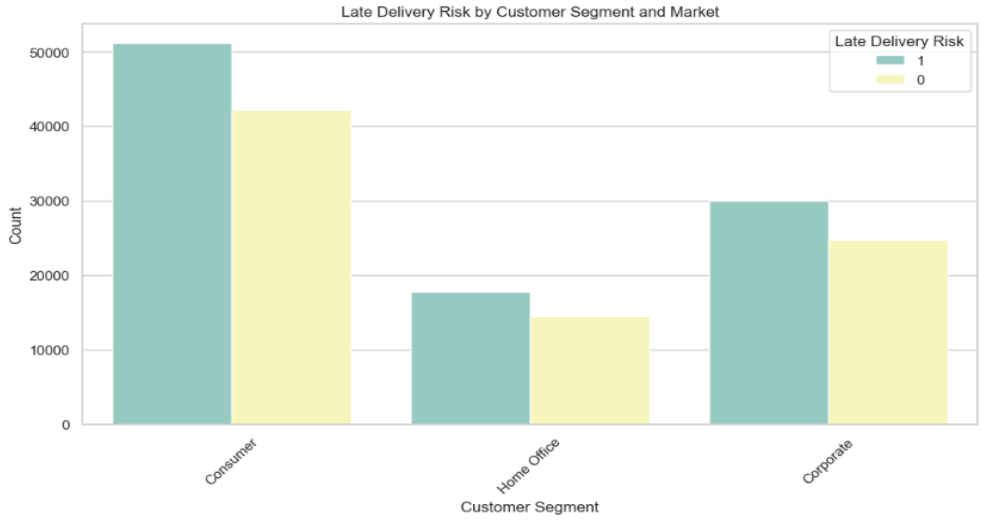

Fig.9. Late Delivery Risk by Customer Segment and Market

In Fig.10 below the study reports that Standard Class shipping is dominant against other market segments, what shows that people value low cost more than speed or meet those delivery times. While the differential is not striking in terms of shipment mode proportion by market, Pacific Asia, which broadly carries LATAM markets, clearly shows the

inclination toward Standard Class. Among all three selected markets, same day shipping is the least used one. It can be altered by some other factors such as the more expensive costs or the limited accessibility of this mode. Preferences of customers combined with different requirements for speed of delivery, which depend on the quality of infrastructure in a particular area and its customer base, are indicators of the market. The economics of Standard Class drive the preferences. On the contrary, dummy Same Day shipping service seems to open the window of chance to market the premium services better. Advocated options include a possibility to be developed: slower shipping options, elaborate market-specific strategies; also, a much-needed cost-benefit analysis must be carried out before introducing or expanding expedited shipping.

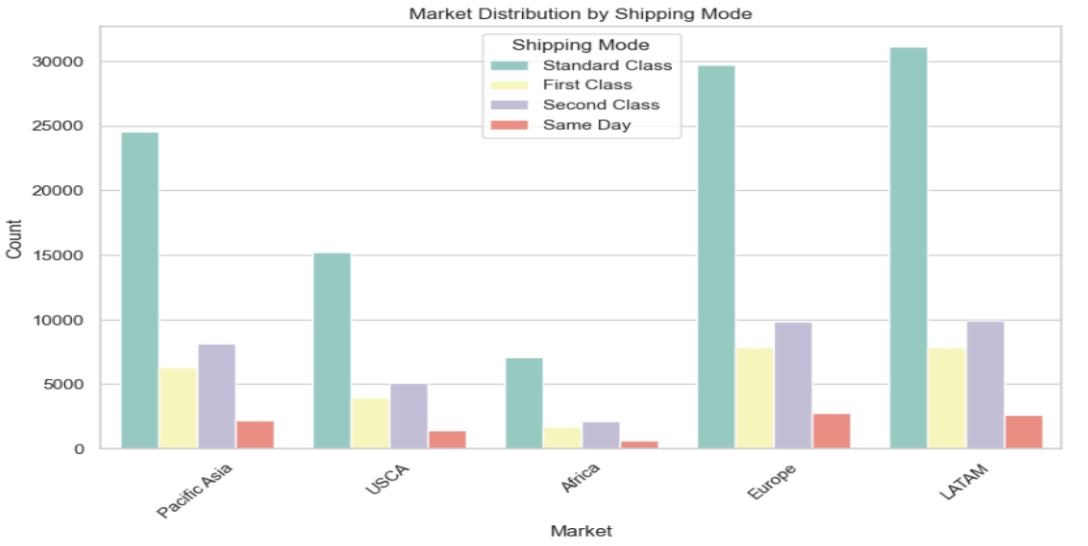

Fig.10.Market Distribution by Shipping Mode

In the Fig.11 below findings show that the US accounted for 61.6% consumers, while the customer base in PR made up 38.4% of the total impactor ratio. Hence, a noticeable customer base exists in both locations. The US is a bigger market in terms of revenues and to successfully sell there one needs not only to have grander logistics and customer service systems but also to be noticeably larger than the one in the US. The importance of the customer base in Puerto Rico and the demand for localized strategy that meets local requirements are testified by the fact that the local market plays a decisive role here. A diversified customer base involving markets brings market-specific risks which mitigate the risk of a particular market. Marketing is tailored to each market, there's optimization done specific to logistics in every country, and products are on offer which suit the tastes of customers in both areas.

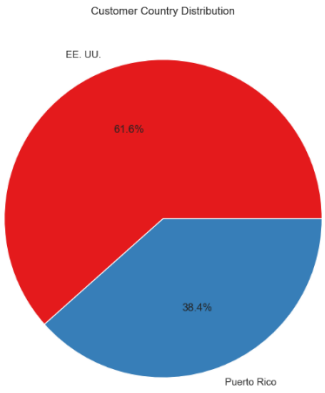

Fig.11.Customer Country Distribution

The Fig.12 below analysis reveals (GIS) difference in the risk of late delivery across different states and the frequence of occurrence of 1 for 0 in late delivery column from the data frame as count the 0 in the graph indicated lower and no delivery risk whereas the 1 indicates the considerable risk of late delivery based on many different criteria. The count value is the frequency of occurrence of specific state in the data frame Taking for instance PR that has (to a great extent) more punctual deliveries due to several issues such as (transport and external factors) Delivery delays cover almost any state, meaning there is almost a systematic problem for the Federal Healthcare Disc. Certain states display an elevated percentage of delayed deliveries, which is ambivalent that they might be associated with high order quantities. Firms should direct their efforts to improve late delivery risk areas by examining the local logistics, backing off the performance from carriers and distribution centers to improve brand loyalty and customer satisfaction. Recommendations such as reviewing logistics partners, improving customer communication, and using data-driven automatic tracking of high-risk states for timely delivery might be beneficial to the firm.

Fig.12.Customer State Distribution by Late Delivery Risk

Furthermore, the analysis in Fig.13 below shows frequency of the top 10 cities in three different customer segments. The count indicates frequency of each city from the customer city column in the dataframe corresponding to the three different customer segments in the dataset. The top customer cities are just a few of them; it is thus Caguas that has larger numbers in this field. Office/Housing clients are usually the first term in terms of the number of customers, but Corporate/ Home Office segment is remarkable in some areas. Town and city centers look to have average, high, and extremely high concentration levels of the customers across all segments. Targeted marketing campaigns should aim at the high customers' neighborhoods where transportation trucks should also be highly optimized for the efficiency via traffic restrictions amongst the cities. Strategies specific to a segment should be devised to meet the specific requirements in urban areas, and further, local marketing efforts, municipal development, and customer-oriented programs are recommended to enlarge brand recognition and trust.

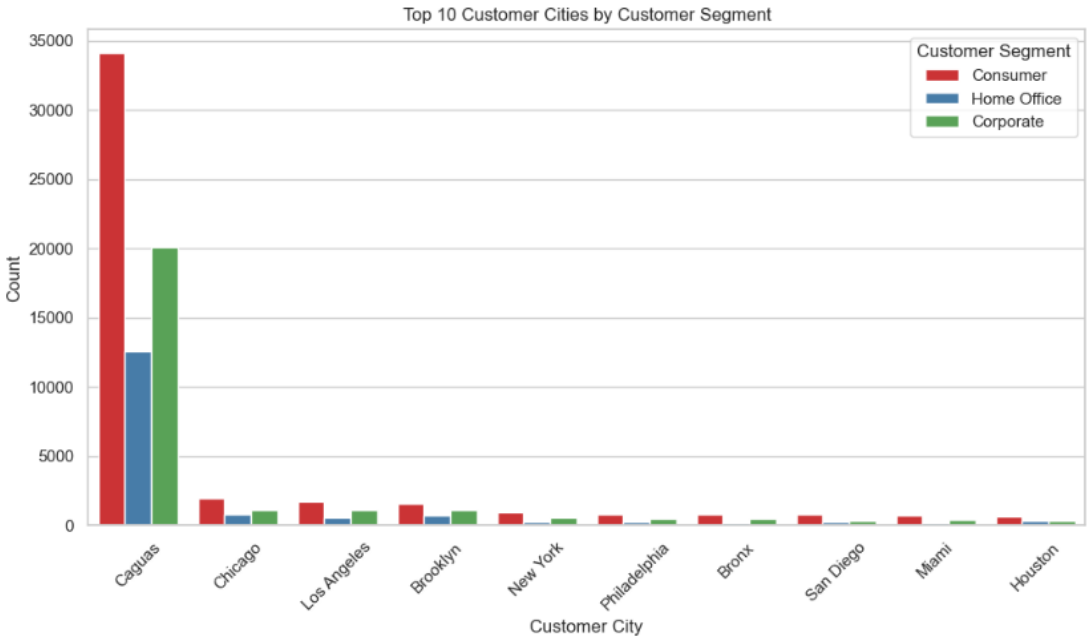

Fig.13.Top 10 Customer Cities by Customer Segment

The graph in Fig .14 below shows sales per customer by market. The X-axis indicates the different markets (taken from the dataset Market column) where the order needs to be delivered and Y axis is the sales per customer. The real outcome is the Median Balance of Payment per Customer value in every market is like the unexpected measurements in Asia-Pacific which may imply that the sales can be improved. Segment customers based on spend and trend personalize the offerings. While markets with high sales fluctuations might result in market volatility, this risk can be curbed by employing policies on price stabilization. The suggestions comprise of development of high-value customer programs; adaptations of sales strategies to suit the market-specific sales patterns; also, deeper market analysis should be done on the customer spending; this can reveal the underlying drivers of the spending/consumption patterns, especially the pockets with significant differences in outlook.

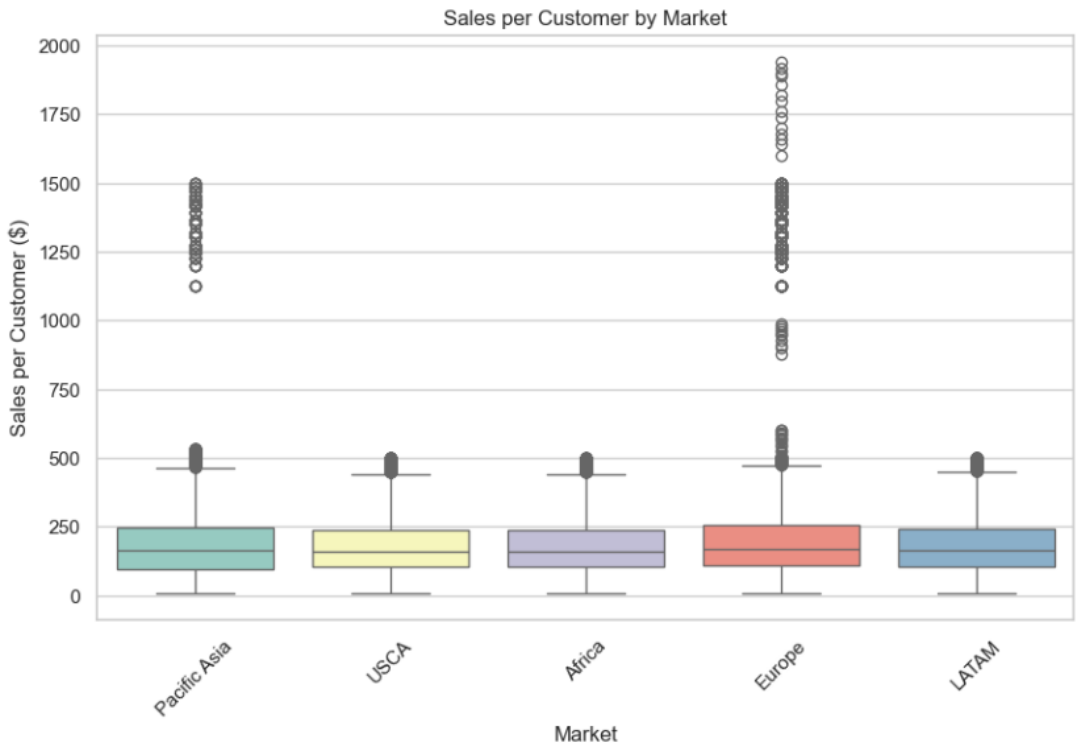

Fig.14.Sales per Customer by Market

The analysis in Fig.15 below indicates the substantial effect of the top expenditures on complete sales, showing high client concentration and the chances of revenue dependence relying on these customers. As we tried to use the customer names the dataset, we have is based on customer first name last name. Using the customer's first name sums up the sales of all the customers with the same first name even with different last names. Hence, we used the customer ID as it is unique value for each customer. However, we dropped the customer Fname and Customer Lname columns to avoid any irregularities. To be able to avoid the risks and diversify business, you need customer retention, establishing strong relationships, and providing personalized benefits for key customers through hiring of a dedicated manager. In addition to this, the process of establishing a niche customer base which is not dependent on only a few customers should be initiated to ensure continued normal performance and development of a business.