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

**Nikitha Kalali**

**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.

A graph of sales by top 10 categories

Description automatically generated

**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.

A graph with purple lines and numbers

Description automatically generated

**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.

A pie chart with numbers and text

Description automatically generated

**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.

A graph of sales

Description automatically generated with medium confidence

**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.

A diagram of a shipping chart

Description automatically generated with medium confidence

**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).

A graph showing the average product price

Description automatically generated

**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.

A graph of different colored bars

Description automatically generated

**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**.**

A graph of different colored bars

Description automatically generated

**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.

A graph of a customer segment

Description automatically generated with medium confidence

**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.

A graph of different colored bars

Description automatically generated.

**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.

A red and blue circle with a number of percentages

Description automatically generated

**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.

A graph of a customer state distribution

Description automatically generated

**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.

A graph with different colored bars

Description automatically generated with medium confidence

**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.

A graph of sales per customer

Description automatically generated

**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.

A graph of different colored bars

Description automatically generated