# **Dataset Exploration – Univariate Analysis**

# **Name – Aradhana Hirapara**

# **LinkedIn - https://www.linkedin.com/in/aradhana-hirapara/**

# \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# **Dataset Description & Dictionary:**

The dataset used for this project includes various attributes related to order transactions, including information on order dates, shipping details, item types, sales channels, and financial metrics. It includes both categorical and numerical variables.

This dataset aims to understand sales trends, inventory efficiency, and profitability across different regions, item types, and sales channels. This analysis can give insights to optimize inventory management, pricing strategies, and shipping processes in a retail business. By examining variables like total profit, order priority, and region, the dataset will provide valuable information for decision-making and strategic planning.

**Data Cleaning:**performeddata check for missing, duplicate and incorrect value in excel but could not find such errors. Hence proceeding with current data**.**

**Data Dictionary:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| **Order ID** | Unique identifier, used for tracking and differentiation. | Categorical |
| **Order Month** | The month when the order was placed | Categorical |
| **Order Weekday** | The day of the week when the order was placed | Categorical |
| **Order Date** | The exact date when the order was placed | Date |
| **Order\_Ship\_Days** | Number of days between the order placement and shipping | Numerical |
| **Ship Date** | Date when the order was shipped | Date |
| **Region** | The geographic region where the order was placed | Categorical |
| **Country** | The country where the order was placed | Categorical |
| **Item Type** | Category of item (e.g., cosmetics, office supplies, snacks) | Categorical |
| **Item Type Encoded** | Encoded value representing the item type | Numerical |
| **Order Priority** | Priority level of the order (e.g., high, medium, low) | Categorical |
| **Order Priority Encoded** | Encoded value representing the order priority | Numerical |
| **Sales Channel** | Channel through which the order was placed (e.g., online, offline) | Categorical |
| **Sales Channel Encoded** | Encoded value representing the sales channel | Numerical |
| **Units Sold** | Number of units sold in the order | Numerical |
| **Unit Price** | Price per unit of the product | Numerical |
| **Unit Cost** | Cost per unit of the product | Numerical |
| **Total Revenue** | Total revenue generated by the order | Numerical |
| **Total Cost** | Total cost incurred for the order | Numerical |
| **Total Profit** | Profit earned from the order | Numerical |
| **Unit Margin** | Margin per unit (Unit Price - Unit Cost) | Numerical |

**Data Encoding & Rearrangements:**

The dataset is arranged logically to follow the flow of an order lifecycle, starting from order identification, placement, and shipping details, followed by regional and categorical attributes, and ending with financial metrics. This structure ensures a clear and intuitive analysis, facilitating both timeline-based and attribute-based insights.

In this dataset, I have added encoded columns for **Item Type**, **Order Priority**, and **Sales Channel** to facilitate more efficient analysis. These encoded columns represent the categorical values numerically, allowing for easier comparison and correlation analysis in the later stages of the project. The encoded values are included in both the dataset and the data dictionary for reference.

# **General Assumptions & Extra Data:**

# **General Assumptions:**

# The dataset represents a sample of transactions rather than the entire population, given its limited scope and synthetic nature. This assumption aligns with the dataset's description and origin.

# The data is assumed to be complete, accurate, and consistent over a defined period, with no significant missing or duplicate entries.

# Categorical variables, such as Order Priority and Sales Channel, are correctly labeled and encoded, ensuring their numerical representations accurately align with their original categorical meanings.

# Financial data (e.g., Total Revenue, Total Cost, and Total Profit) are assumed to be consistent across all orders and measured in the same currency.

# Dates for orders and shipping are assumed to follow the same time zone or region to accurately calculate shipping durations.

# **Extra Data Requirements:**

# **Customer Surveys:** Adding customer survey data (e.g., how customers learned about the store) could provide insights into marketing effectiveness and customer acquisition channels.

# **Demographics:** Including customer demographic information (e.g., age, gender, income level) would enable a more comprehensive understanding of purchasing behaviors and preferences.

# **Seasonality:** Data on seasonal trends or promotional events (e.g., holidays, sales campaigns) could help identify patterns in sales performance.

# **Marketing Spend:** Information on marketing and advertising expenditures by region or item type would facilitate analysis of their correlation with sales trends and profitability.

# **Customer Loyalty Data:** Data on repeat customers and loyalty programs would allow analysis of customer retention and lifetime value.

# **Univariate Statistical Analysis:**

# We will perform Univariate statistical analysis on Units Sold, Unit Price, Unit Cost, Total Revenue, Total Cost, Total Profit and Unit Margin.

# **Unit Sold:**

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 5053.988 |
| Standard Error | 91.749543 |
| Median | 5184 |
| Mode | 6283 |
| Standard Deviation | 2901.3753 |
| Sample Variance | 8417978.7 |
| Kurtosis | -1.2213013 |
| Skewness | -0.0512172 |
| Range | 9985 |
| Minimum | 13 |
| Maximum | 9998 |
| Sum | 5053988 |
| Count | 1000 |
| Largest(1) | 9998 |
| Smallest(1) | 13 |
| Confidence Level(95%) | 180.04393 |

# To create a histogram we have calculated the bins range using some of the statistical values

# Bin Range = (Max Value - Min Value) / Number of Bins

# Hence we got Bin 1: 0 - 1000 Bin 2: 1000 - 2000 Bin 3: 2000 - 3000 ... Bin 10: 9000 – 10000 etc. Here is the histogram of the column.

# **Histogram Analysis for Units Sold**

# **Distribution:**

# The histogram shows an approximately uniform distribution, with frequencies relatively consistent across most bins.

# No single bin appears to dominate the dataset, indicating no significant concentration of orders in a particular range.

# **Identifying Outliers:**

# The "More" bin represents extreme values above the highest defined bin (9000+). These could indicate outliers, requiring further analysis to determine if they are valid data points or anomalies.

# **Suggested Business Strategy:**

# The uniform distribution suggests diverse order quantities, which could indicate consistent customer demand across different product categories or regions.

# For inventory management, businesses may need to maintain balanced stock levels across ranges to accommodate varying order sizes.

# **Comparing Sales Segments:**

# Comparing the frequencies of bins such as 1000–2000 and 8000–9000 can provide insights into the common order sizes.

# Marketing campaigns can be tailored based on common order quantities,

# e.g., offering discounts for bulk orders in higher-frequency bins.

# **Unit Price:**

|  |  |
| --- | --- |
| **Statistics** | **Values** |
| Mean | 262.10684 |
| Standard Error | 6.8311857 |
| Median | 154.06 |
| Mode | 47.45 |
| Standard Deviation | 216.02106 |
| Sample Variance | 46665.099 |
| Kurtosis | -0.7349003 |
| Skewness | 0.7920352 |
| Range | 658.94 |
| Minimum | 9.33 |
| Maximum | 668.27 |
| Sum | 262106.84 |
| Count | 1000 |
| Largest(1) | 668.27 |
| Smallest(1) | 9.33 |
| Confidence Level(95%) | 13.405119 |

As per the given formula given for Units sold here for Unit Price, we have calculated the bin range from 0,30 to 470.

* 1. **Distribution:**
* The histogram indicates a positively skewed distribution, with the highest frequency observed in the "More" bin, representing unit prices above 270.
* A smaller concentration of unit prices is evident in lower ranges, such as 0–90 and 90–180, reflecting a gradual increase in frequency as the range progresses.
  1. **Identifying Outliers:**
* The "More" bin suggests a concentration of outliers or exceptionally high unit prices beyond 270. These prices may indicate premium or luxury items and should be examined to verify their significance or correctness.
  1. **Suggested Business Strategy:**
     1. **Low-Range Prices (0–90)**: Represents low-cost items that may drive high-volume sales. Marketing strategies like bulk discounts or promotions could boost sales in this category.
     2. **Mid-Range Prices (90–210)**: These ranges exhibit moderate frequency and could represent mid-tier products. These items may target a balanced audience and could benefit from focused marketing campaigns.
     3. **High-Range Prices (270+ and "More")**: Represents high-value products that, despite lower sales volume, could significantly impact revenue. Exclusive marketing and targeting affluent customers can maximize profitability in this segment.
  2. **Comparing Revenue Segments:**
* Products priced in the higher ranges contribute more per unit but may have lower sales frequency. Optimizing inventory for such products could reduce storage costs while maintaining profitability.
* Low- and mid-range products should be managed for volume sales, ensuring consistent availability to meet demand.

1. **Unit Cost:**

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 184.96511 |
| Standard Error | 5.5431347 |
| Median | 97.44 |
| Mode | 31.79 |
| Standard Deviation | 175.28931 |
| Sample Variance | 30726.343 |
| Kurtosis | -0.5996022 |
| Skewness | 0.9480311 |
| Range | 518.04 |
| Minimum | 6.92 |
| Maximum | 524.96 |
| Sum | 184965.11 |
| Count | 1000 |
| Largest(1) | 524.96 |
| Smallest(1) | 6.92 |
| Confidence Level(95%) | 10.877523 |

Here, we have selected bin ranges from 30,60 to 540.

* 1. **Distribution:**
* The histogram of unit cost reveals a right-skewed distribution, as evidenced by the positive skewness (0.95). This means a majority of unit costs are concentrated in the lower ranges (e.g., below 200), with fewer instances of very high costs.
* Peaks around 60–120 and 240–270 suggest common pricing ranges, while the frequency drops significantly for higher ranges.
  1. **Identifying Outliers:**
* Outliers are apparent due to the wide range (6.92–524.96). Values close to the maximum (e.g., 524.96) are much higher than the median (97.44) and may represent premium products or data entry errors.
* By analyzing these outliers, you can investigate if they are justified based on the product type, demand, or errors.
  1. **Guiding Business Strategy:**
* **Pricing Strategy:** With most unit costs clustering around specific ranges (e.g., 60–120), businesses can focus on optimizing production and sales within these cost-effective ranges.
* **Cost Optimization:** High-cost outliers might indicate inefficiencies or opportunities to reduce costs by negotiating with suppliers or streamlining production processes.
* **Product Differentiation:** The presence of both low-cost and high-cost units can guide segmentation strategies, ensuring pricing and marketing align with customer expectations.
  1. **Comparing Revenue Segments:**
* The variability in unit costs suggests segmentation opportunities. For example, lower-cost products (below 100) might cater to budget-conscious customers, while higher-cost products (above 300) could target premium markets.
* Analyzing revenue contribution from these segments could reveal whether higher-cost items justify their impact on profitability or if the business should focus more on high-frequency cost ranges like 60–120.

1. **Total Revenue:**

|  |  |
| --- | --- |
| **Statistics** | **value** |
| Mean | 1327321.8 |
| Standard Error | 47007.718 |
| Median | 754939.18 |
| Mode | 90211.77 |
| Standard Deviation | 1486514.6 |
| Sample Variance | 2.21E+12 |
| Kurtosis | 2.0682412 |
| Skewness | 1.6315375 |
| Range | 6615166.3 |
| Minimum | 2043.25 |
| Maximum | 6617209.5 |
| Sum | 1.327E+09 |
| Count | 1000 |
| Largest(1) | 6617209.5 |
| Smallest(1) | 2043.25 |
| Confidence Level(95%) | 92245.192 |

Here, as per the formula we have kept a bin width of 500,000 and the last value is 6500000.

**Histogram Analysis:**

* 1. **Distribution**
     + - * The histogram shows that most of the revenue data falls in the lower bins, with the largest frequency (373) in the revenue range of up to **500,000**.
         * The frequency decreases as the revenue ranges increase, indicating a skewed distribution where lower revenue transactions are more in numbers in the dataset.
         * Hence, we can say that the majority of orders are of smaller monetary value.
  2. **Identify Outliers**
* The "More" bin, representing revenues exceeding **5,000,000**, contains **42** instances. These transactions could be considered outliers, as they deviate significantly from the majority of the data.
* Outliers need closer examination to understand their nature—whether they represent high-value clients, special orders, or data errors.
  1. **Suggested Business Strategy**
* The heavy concentration in lower revenue ranges indicates an opportunity to analyze why most transactions fall under this category. Strategies could focus on increasing the average transaction size or converting lower-tier customers into mid or high-tier buyers.
* High-value transactions in the upper bins may represent premium customers. Understanding these customers' behaviors and preferences could guide loyalty and retention strategies.
  1. **Compare Revenue Segments**
* Breaking down the histogram into revenue segments allows businesses to identify which range contributes the most to total revenue.
* For example, the segment between **500,000 and 1,500,000** accounts for a combined **58.2%** of cumulative revenue-related transactions. These segments can be targeted for upselling and cross-selling strategies.

**Actionable Insights:**

* Focus marketing efforts on customers generating revenue in the **1,000,000 - 2,000,000** range to move them to higher segments.
* Review outlier transactions for patterns that could be replicated across other customer groups or orders.
* Explore strategies to engage small-scale customers contributing to the majority of transactions below **500,000**.

1. **Total Cost:**

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 936119.2288 |
| Standard Error | 36763.71519 |
| Median | 464726.065 |
| Mode | 66909.48 |
| Standard Deviation | 1162570.753 |
| Sample Variance | 1.35157E+12 |
| Kurtosis | 2.558918675 |
| Skewness | 1.798451421 |
| Range | 5203561.65 |
| Minimum | 1416.75 |
| Maximum | 5204978.4 |
| Sum | 936119228.8 |
| Count | 1000 |
| Largest(1) | 5204978.4 |
| Smallest(1) | 1416.75 |
| Confidence Level(95%) | 72142.96094 |

**Histogram Analysis:**

* 1. **Distribution:**
* The histogram displays a positively skewed distribution with a majority of values concentrated in the lower ranges (0–500,000). This indicates that most of the data points lie within smaller revenue figures, while very few achieve higher revenues.
* The bars significantly decrease as the bin range increases, with a sharp decline beyond 1,000,000.
  1. **Identifying Outliers:**
* Outliers may exist in the upper range, particularly in the bins beyond 3,000,000, where the frequencies are very low. These could indicate exceptional transactions or data entry anomalies.
* The maximum revenue values (e.g., above 5,000,000) could represent rare cases and warrant further investigation to confirm their validity.
  1. **Suggesting Business Strategy:**
* The concentration in lower bins suggests that most transactions or customers generate lower revenues. Efforts to upsell or cross-sell can target customers in these lower bins to increase revenue.
* For the higher ranges, it may be beneficial to analyze these exceptional cases to identify the factors contributing to significant revenue spikes, such as specific items, regions, or promotional strategies.
  1. **Comparing Revenue Segments:**
* The histogram divides the data into clear segments, allowing comparison between low, medium, and high-revenue-generating transactions.
* Segments above 2,000,000 might represent premium customer bases or unique business operations, whereas bins below 1,000,000 represent regular operations.
* By focusing on these segments, the business can tailor marketing strategies and prioritize investments in profitable areas.

1. **Total Profit :**

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 391202.61 |
| Standard Error | 12131.768 |
| Median | 277225.98 |
| Mode | 23302.29 |
| Standard Deviation | 383640.19 |
| Sample Variance | 1.472E+11 |
| Kurtosis | 1.604344 |
| Skewness | 1.4053775 |
| Range | 1725648.8 |
| Minimum | 532.61 |
| Maximum | 1726181.4 |
| Sum | 391202612 |
| Count | 1000 |
| Largest(1) | 1726181.4 |
| Smallest(1) | 532.61 |
| Confidence Level(95%) | 23806.671 |

**Histogram Analysis:**

**1. Distribution:**

* The data is right-skewed, with most of the profit values concentrated in the lower ranges (0–400,000).
* Higher profit values (beyond 1,200,000) are relatively rare.

**2. Identifying Outliers:**

* There may be potential outliers in the higher profit ranges (above 1,600,000). These could represent unusually profitable orders and merit further investigation.

**3. Guiding Business Strategy:**

* The high frequency of lower-profit orders suggests focusing on optimizing costs or increasing revenue in this segment.
* High-profit orders, though less frequent, might represent premium products or bulk orders and could be targeted for retention or upselling strategies.

**4. Comparing Profit Segments:**

* Comparing the low-profit and high-profit segments reveals opportunities for business growth. For example, understanding what drives high-profit orders can help replicate these factors across lower-profit transactions.

1. **Region:**

|  |  |  |
| --- | --- | --- |
| **Regions** | **Count of Region** | **Frequency** |
| Asia | 136 | 13.6 |
| Australia and Oceania | 79 | 7.9 |
| Central America and the Caribbean | 99 | 9.9 |
| Europe | 267 | 26.7 |
| Middle East and North Africa | 138 | 13.8 |
| North America | 19 | 1.9 |
| Sub-Saharan Africa | 262 | 26.2 |
| **Grand Total** | **1000** |  |

# 

# 

# The analysis of Regions reveals that Europe and Sub-Saharan Africa are the largest contributors to the dataset, accounting for 26.7% and 26.2% of orders, respectively. This distribution suggests a stronger presence in **Europe** **and Sub-Saharan Africa**, which may align with **strategic market focuses** or **regional sales.**

# **Order Priority:**

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Count of Order Priority** | **Frequency** |
| C | 262 | 26.2 |
| H | 228 | 22.8 |
| L | 268 | 26.8 |
| M | 242 | 24.2 |
| **Grand Total** | **1000** |  |

# The analysis of Order Priority shows a relatively even distribution among the four priority levels: Critical (C), High (H), Low (L), and Medium. This balanced distribution indicates that orders are spread across various levels of urgency, with no extreme skew toward any single priority level.

# **Suggested Outliers and Cleaning:**

# 

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

**Based on the above box plots:**

1. **Unit Sold, Unit Price, Unit Cost:**

* These plots suggest relatively symmetric distributions without extreme outliers, as the whiskers extend reasonably on both ends. However, any points beyond the whiskers (none visible here) would indicate potential outliers. These data show consistent variability and central tendency.

1. **Total Revenue, Total Cost, Total Profit:**

* These plots show significant outliers on the higher end, as visible from the points above the upper whisker. These values suggest that some transactions or operations resulted in exceptionally high revenue, cost, or profit compared to the majority. These outliers are likely valid entries reflecting rare but impactful business events.

1. **Suggested Improvements:**

* For clearer identification, statistical tests (e.g., Z-scores or IQR thresholds) can supplement the box plot to confirm whether these points significantly deviate from the central data.

# **Coding and/or Categorization**

* **Order Priority** was encoded as a numerical variable (High = 3, Medium = 2, Low = 1) to help in future analysis. This encoding can help correlate priority with profitability or sales trends in the dataset.
* **Item Type** was encoded as a numerical variable to facilitate easier analysis of different product categories. Each item type was assigned a unique number (e.g., Cosmetics = 1, Vegetables = 2, Baby Food = 3, etc.). This encoding allows for comparisons between different item categories and their impact on metrics like total revenue and total profit.
* **Sales Channel** was also encoded as a numerical variable (Online = 1, Offline = 2) to simplify the analysis of sales performance across different sales channels. This encoding will help correlate sales channel type with other variables such as units sold, total profit, and region-specific sales performance.

# **FINER Research Questions**

* **Broad Question**: How do the top-performing product categories contribute to overall revenue, and how do their sales trends fluctuate throughout the year?
* **Supporting Question**: Is there a correlation between the number of units sold and the profit margin?
* **Supporting Question**: How does the revenue contribution of each top product category compare to others over monthly and quarterly periods?
* **Supporting Question**: How does the sales performance of different item types vary across different regions?
* **Supporting Question**: What factors drive peaks in sales for the top-performing product categories across different seasons?

# **Tracking**

1. **Steps Taken:**

* Performed univariate analysis on key numerical variables (**Units Sold, Unit Price, Unit Cost, Total Revenue, Total Cost, Total Profit, Unit Margin**).
* Used box plots and histograms to identify potential outliers and understand data distribution.
* Cleaned data to remove inconsistencies, ensuring accuracy in analysis.
* Encoded categorical variables (**Item Type, Order Priority, Sales Channel**) for ease of correlation and visualization.
* Visualized categorical and numerical distributions with **pie charts, bar graphs, and histograms**.

1. **Assumptions:**

* Data represents a synthetic sample, not a complete population.
* Financial metrics are consistently recorded in the same currency.
* Encoded variables accurately reflect their original meaning for comparison.

1. **Challenges and Resolutions:**

* **Incorrect ranges in visualizations:** Recalculated bin ranges and reformatted numerical data.
* **Outlier identification:** Combined box plots and statistical methods like Z-scores and IQR thresholds to flag anomalies.
* **Balancing categorical distribution:** Used pivot tables for accurate frequency analysis and visualization.

1. **Improvements for Future Analysis:**

* Incorporate additional variables (e.g., customer demographics) for deeper insights.
* Use advanced statistical methods (e.g., regression, clustering) for enhanced outlier detection.
* Automate data preparation and visualization steps for efficiency.