**Exploratory Data Analysis (EDA) for Retail Sales Dashboard**

Submitted by**: Ashfaq Javed**

**1. Introduction**

The dataset used in this project is an **Online Retail dataset**, comprising transactions between a UK-based online retailer and international customers between 2010 and 2011. The primary objective of this Exploratory Data Analysis (EDA) is to uncover insights into sales patterns, customer behavior, and regional performance. These insights will play a key role in supporting future AI/ML tasks such as customer segmentation, sales prediction, and demand forecasting.

Our focus is on extracting meaningful insights by visualizing key metrics such as **total sales, quantity sold, unit price, and invoice date**, broken down by **countries and customer IDs**. The resulting dashboard will be interactive and informative, providing users with the ability to explore sales trends across different dimensions.

**2. Visualization Process**

**Bar Chart:**

* **Purpose:** The bar chart helps in comparing **total sales** by country, providing a quick visual comparison of the performance of each region.
* **Steps:**
  + **Country** was placed on the X-axis, and **Total Sale** was mapped to the Y-axis.
  + **Quantity** was used as a **color scale** to provide additional information about the volume of products sold in each region.
  + **Total Sale** was used as the **label** to display exact sales values.
  + **Unit Price** was set for the **tooltip** to provide information when hovering over the bar.
  + **Rationale:** The bar chart is ideal for comparing discrete categories (in this case, countries). The use of color (quantity sold) adds another layer of insight, while tooltips offer an interactive element, giving more details as needed.

**Line Chart:**

* **Purpose:** The line chart helps analyze **sales trends over time**, broken down by country.
* **Steps:**
  + **Invoice Date** was placed on the X-axis to represent time, and **Total Sale** on the Y-axis.
  + **Country** was used to color the lines, representing different countries.
  + **Customer ID** was added as a **tooltip** to highlight customer-specific insights when interacting with the chart.
  + **Rationale:** A line chart is well-suited for showing trends over time, and using color for different countries allows for easy comparison of sales performance across regions. The tooltip adds depth by revealing customer-specific information.

**Box Plot:**

* **Purpose:** The box plot provides a **summary of sales distributions** across different countries, revealing outliers and the spread of sales data.
* **Steps:**
  + **Total Sale** was used on the Y-axis to represent the distribution of sales.
  + **Country** was used for color-coding the different groups.
  + The **median total sale** was displayed as a **label** to emphasize central tendencies in each country.
  + **Customer ID** was again utilized as a **tooltip** to allow further exploration of individual sales contributions.
  + **Rationale:** Box plots are useful for summarizing distributions and identifying potential outliers. Using countries as categories allows for quick comparison of sales patterns across regions.

**Scatter Plot:**

* **Purpose:** The scatter plot helps explore the relationship between **quantity sold and unit price**, giving insights into pricing strategy and customer behavior.
* **Steps:**
  + **Quantity** was mapped to the X-axis and **Unit Price** to the Y-axis.
  + **Description** was used for color-coding, which enabled a quick distinction between product categories.
  + **Total Sale** was set as the **size** of the dots, with larger dots indicating higher total sales.
  + **Customer ID** was added as a **label**, and **Country** was used for the **tooltip**.
  + **Rationale:** Scatter plots are perfect for showing relationships between two numerical variables. The use of dot size to indicate total sales and color to distinguish between products enhances the plot's utility for exploring multi-dimensional relationships.

**Dashboard Integration:**

* The charts were combined into a dashboard, providing an interactive, cohesive environment for exploration. Each chart complements the others, with different perspectives on the same data, creating a holistic view of sales performance.

**3. Decision-Making Justification**

* **Chart Type Choices:**
  + **Bar Chart** was chosen for its ability to provide a clear comparison across countries.
  + **Line Chart** effectively visualizes temporal trends, which is crucial for sales analysis over time.
  + **Box Plot** reveals distribution patterns, essential for identifying outliers or skewness in sales data.
  + **Scatter Plot** is optimal for discovering relationships between unit price and quantity sold, with an added layer of total sales and product categories.
* **Color and Aesthetics:**
  + Each visualization uses color purposefully. In the bar chart, different hues indicate quantity sold, while in the scatter and line charts, colors are used to differentiate between countries and product descriptions.
  + Tooltips enhance interactivity, offering further details without overwhelming the visual design.
* **Clarity and Interactivity:**
  + Labels, colors, and tooltips are employed to improve clarity. By embedding interactive elements like tooltips and hover effects, users can gain detailed information without cluttering the dashboard.
  + The choice of **median** labels in the box plot ensures that central tendencies are emphasized without obscuring the distribution.

**4. Challenges and Solutions**

* **Data Quality:** Some records in the dataset had missing or erroneous values, which required cleaning. This was handled through **filtering** and **data imputation** techniques in Tableau.
* **Visual Clarity:** Maintaining a balance between displaying enough data without overwhelming the viewer was challenging. I achieved clarity by using tooltips for additional information, labels for critical insights, and appropriate color schemes.
* **Chart Integration:** Combining different chart types into a single dashboard presented a challenge in terms of maintaining interactivity without compromising on performance. This was solved by optimizing filter settings and ensuring that charts were responsive to interactions.

**5. Conclusion**

The EDA revealed key insights:

* **Country-wise Sales Performance:** Certain countries consistently outperform others, indicating high-priority regions for sales focus.
* **Temporal Sales Trends:** Sales follow predictable seasonal trends, with certain spikes in the data indicating promotional periods or holiday effects.
* **Distribution of Sales:** The box plot exposed several countries with significant outliers, pointing to potential high-value customers that could be targeted for loyalty programs.
* **Product-Customer Interaction:** The scatter plot shows that some products with high unit prices are sold in small quantities, possibly indicating a niche market that requires targeted marketing strategies.

These findings provide a solid foundation for future AI/ML tasks such as **customer segmentation**, **demand forecasting**, and **pricing strategy optimization**. The dashboard serves as a valuable tool for interactive data exploration and decision-making.

