Transactions Dataset: Detailed Explanation & Insights

The Transactions Dataset contains records of purchases made by users, including transaction dates, store details, purchased items, and sales amounts. This dataset is critical for analyzing spending behavior, identifying sales trends, and understanding user engagement. Below is a structured summary covering EDA, assumptions, data cleaning, and key insights.

Understanding the Data

The dataset includes the following key columns:

- **RECEIPT_ID** → Unique identifier for each purchase transaction.
- **SCAN_DATE** → Date when the receipt was scanned.
- **STORE_NAME** → Name of the store where the purchase was made.
- **USER_ID** → Links the transaction to a specific user.
- **BARCODE** → Identifies the purchased product.
- **FINAL_QUANTITY** → Number of items purchased in a transaction.
- **FINAL_SALE** → The total amount spent in the transaction.
- **PURCHASE_DATE** → The actual date of the transaction (if available).

Initial Observations

- The dataset contains transaction-level data, linking users to their purchases.
- Missing values were found in key fields like BARCODE, FINAL_QUANTITY, FINAL_SALE, and PURCHASE_DATE, which could impact sales analysis.
- Duplicate transactions were identified, possibly indicating double-entry errors or receipt resubmissions.
- Data type inconsistencies were found in `FINAL_QUANTITY` and `FINAL_SALE`, which needed conversion for accurate numerical analysis.

Exploratory Data Analysis (EDA)

EDA was conducted to uncover **patterns, distributions, and potential issues** in the dataset. The following **visualizations and transformations** were applied:

Handling Data Type Issues

- Converted `FINAL_QUANTITY` & `FINAL_SALE` to Numeric

    - Some records contained non-numeric values (e.g., errors, blanks).
    - Used `pd.to_numeric(errors='coerce')` to ensure proper numerical formatting.
    - Ensures correct aggregation of transaction amounts and product counts.

Missing Values Visualization

- **Bar Chart:**
    - Highlighted missing values in barcodes, sales amounts, and purchase dates.
    - Ensured that missing values were handled appropriately before analysis.

Sales Distribution Analysis

- **Histogram: Distribution of `FINAL_SALE`**
    - Identified skewness in sales data, indicating whether most transactions are low or high-value.
    - Helped detect outliers, such as unusually high-value purchases.

Feature Correlation Analysis

- **Correlation Heatmap for Numeric Columns**
    - Showed relationships between key numeric variables, such as `FINAL_QUANTITY` and `FINAL_SALE`.
    - Helped identify redundant or highly correlated features that may impact model building.

Quantity vs. Sales Analysis

- **Scatter Plot: `FINAL_QUANTITY` vs. `FINAL_SALE`**
    - Confirmed whether higher quantities lead to higher sales.
    - Helped detect anomalies, such as high sales with low quantities (potential discounts or errors).

Pairwise Feature Relationships

- **Pairplot for Numeric Columns**
    - Provided pairwise comparisons of all numeric features to detect patterns.
    - Helped identify clusters and potential outliers in the data.

Time Series Analysis: Transactions Over Time

- **Line Chart: Monthly Transactions**
    - Grouped transactions by month to identify sales trends over time.
    - Helped detect seasonal patterns, showing peak shopping periods.

Assumptions Made

- **Missing Barcodes**
  - If `BARCODE` was missing, assumed that the product was manually entered or barcode scanning failed.
  - Missing barcodes were replaced with `"-1"` as a placeholder.
- **Final Quantity & Final Sale**
  - If `FINAL_QUANTITY` was missing, assumed to be "0" (indicating no recorded quantity).
  - If `FINAL_SALE` was missing, assumed to be "0"` (indicating an unknown sales amount).
  - These assumptions helped prevent skewed sales analysis.
- **Duplicate Transactions**
  - Duplicate `RECEIPT_ID` entries were assumed to be errors from users submitting the same receipt multiple times.
- **Purchase Date Validation**
  - Ensured `PURCHASE_DATE` exists before performing date-based analysis.
  - Converted `PURCHASE_DATE` to datetime for proper time-series analysis.

Handling Missing Values & Duplicates

Missing Values

- Filled missing values for:
  - `BARCODE` → Replaced with `"-1"` (placeholder for unknown product codes).
  - `FINAL_QUANTITY` → Filled with `"0"`.
  - `FINAL_SALE` → Filled with `"0"` to avoid incorrect sales calculations.
  - `PURCHASE_DATE` → Converted to datetime with invalid values replaced as `NaT`.

Duplicates

- Removed all duplicate transactions where `RECEIPT_ID` and `USER_ID` matched.
- Ensured that each transaction is counted only once.

Key Takeaways

- The dataset contained missing product details, particularly in barcode information.
- Duplicates were present and needed to be removed for accurate transaction analysis.
- Sales distribution showed significant variation, requiring further breakdown by category and user segments.
- Some transactions had zero recorded sales or quantities, which could indicate data entry errors or refunds.
- Sales trends over time revealed seasonal spending patterns, useful for business forecasting and marketing strategies.