Product Dataset: Detailed Explanation & Insights

The Product Dataset contains information about various products, including categories, brands, manufacturers, and barcodes. The dataset was explored, cleaned, and analyzed using EDA (Exploratory Data Analysis) techniques. Below is a structured summary of the investigation.

Understanding the Data

The dataset includes the following key columns:

- **BARCODE**: Unique identifier for each product.
- **CATEGORY_1, CATEGORY_2, CATEGORY_3, CATEGORY_4**: Product classification hierarchy.
- **BRAND**: The brand associated with the product.
- **MANUFACTURER**: The company producing the product.

Initial Observations

- The dataset contains a large number of unique products, with multiple levels of categorization.
- Some records have missing values in key fields like brand, manufacturer, and product categories.
- Duplicate records were identified, which could impact data integrity.

Exploratory Data Analysis (EDA)

EDA was conducted to uncover patterns, distributions, and potential issues in the data. The following visualizations were created:

Missing Values Visualization

- **Bar Chart**:
    - Showed that some product categories and manufacturer details were missing.
    - Missing values were not randomly distributed but appeared to be concentrated in certain product types.

Duplicate Records Analysis

- **Bar Chart**:
  - Identified 215 duplicate product entries.
  - These duplicates could lead to overcounting of certain products in downstream analytics.

Category & Brand Distributions

- **Bar Chart (CATEGORY_1):**
  - Identifies the most frequent high-level product categories in the dataset.
  - Helps in understanding product diversity and which categories contribute the most.
- **Tree Map (CATEGORY_2)**:
  - Displayed product category distributions and their relative sizes.
  - Helps in category segmentation analysis for product grouping.
- **Box Plot (CATEGORY_3):**
  - Displays the spread and distribution of CATEGORY_3, showing frequency variation.
  - If the boxplot has long whiskers or outliers, it indicates that some CATEGORY_3 values have extremely high counts.
- **Word Cloud (Brands)**:
  - Highlighted the most frequent brands in the dataset.
  - Allowed quick identification of **leading product brands**.
- **Pie Chart (Top 5 Manufacturers)**:
  - Showed the **dominance of a few manufacturers** in the dataset.
  - Helps in identifying market concentration.

Assumptions Made

- **Missing Brands & Manufacturers**:
  - If missing, they were replaced with `"Unknown"`.
  - Assumed that missing values do not indicate non-existent brands but data entry issues.
- **Duplicate Records**:
  - Considered true duplicates if all key fields matched.
  - Assumed duplicates were errors and not intentional repeat listings.
- **Category Hierarchy**:
  - Assumed that missing category levels (`CATEGORY_2`, `CATEGORY_3`) could be inferred from higher-level categories where possible.

Handling Missing Values & Duplicates

Missing Values

- Filled missing values for:
  - `CATEGORY_1`, `CATEGORY_2`, `CATEGORY_3`, `CATEGORY_4` → `"Unknown"`
  - `BRAND`, `MANUFACTURER` → `"Unknown"`
  - `BARCODE` → `-1` (Placeholder for missing barcodes)

Duplicates

- Removed all duplicate rows based on exact matches in barcode, brand, and categories.

Key Takeaways

- The dataset contained missing product details, particularly in category and manufacturer fields.
- Duplicates were present and were successfully removed.
- Most products belong to a few dominant categories, as seen in the Tree Map & Word Cloud.
- Brand data was inconsistent, requiring imputation for missing values.