Dataset Description

Dataset Description (Updated)

A. DensityReports.xlsx

Purpose: This dataset serves as the central repository for evaluating and optimizing the packaging processes at FashionWorld Retail. It contains 500,000 records, where each entry represents a recommendation on how a product should be packaged and a final operational assessment of its packaging quality.

Key Variables:

- **ReportID:** Unique identifier for each packaging report.
- **SupplierName:** Name of the supplier involved (including possible minor taxonomical variations).
- **DateOfReport:** Date when the report was generated, ranging from January 1, 2023, to June 30, 2024.
- **GarmentType:** Type of garment (e.g., Shirt, Pants, Jacket, Dress, Skirt, Suit, Coat, Sweater).
- **Material:** Material of the product (Cotton, Polyester, Wool).
- **ProductReference:** Unique product code used to link with the ProductAttributes dataset.
- **ProposedUnitsPerCarton:** Recommended number of units per carton to optimize efficiency.
- **ProposedFoldingMethod:** Recommended folding method (“Method1”, “Method2”, or “Method3”).
- **ProposedLayout:** Recommended box layout (LayoutA, LayoutB, LayoutC, LayoutD, LayoutE).
- **PackagingQuality:** Operational label indicating the packaging quality ("Good" or "Bad"), based on predefined criteria.

Important Note: PackagingQuality represents an operational assessment and does not necessarily determine the definitive ground truth. A "Good" label does not guarantee that no issues will arise, and a "Bad" label does not always imply critical failure. Packaging incidents and anomalies may still occur independently, as captured in the HistoricalIncidents dataset.

B. ProductAttributes.xlsx

Purpose: This dataset provides detailed characteristics for approximately 10,000 unique products, allowing analysis of how intrinsic product attributes impact packaging outcomes.

Key Variables:

- **ProductReference:** Product code (linked to DensityReports).
- **GarmentType:** Type of garment.
- **Material:** Material of the product.
- **ProductName:** Product’s descriptive name.
- **Size:** Size variant of the product.
- **Collection:** Product collection (e.g., Summer, Winter, Spring, Autumn).

- **Weight:** Weight of the product.

C. SupplierScorecard.xlsx

Purpose: This dataset captures monthly supplier performance metrics over 18 months, providing an aggregated view of their operational consistency and quality.

Key Variables:

- **SupplierName**: Supplier's name (maintaining any taxonomy errors from operational records).
- **Month:** Year and month of the evaluation (from 2023-01 to 2024-06).
- **PackagesHandled: Total packages managed that month.**
- **BadPackagingRate (%):** Percentage of packages classified as "Bad."
- **TotalIncidents:** Total number of incidents associated with the supplier that month.
- **AverageCostPerIncident (€):** Average cost impact per incident.
- **OnTimeDeliveryRate (%):** Rate of on-time deliveries.
- **AnomaliesDetected:** Number of anomalies detected by quality inspections.

D. HistoricalIncidents.xlsx

Purpose: This dataset documents past packaging-related incidents, providing critical insights into the real-world consequences of operational and packaging decisions.

Key Variables:

- **ProductReference:** Product involved in the incident.
- **SupplierName:** Supplier responsible (aligned with supplier names across datasets).
- **DateOfIncident:** Date when the incident occurred.
- **IssueDescription:** Nature of the issue (e.g., Packaging Damage, Labeling Error).
- **ResolutionStatus:** Status of the issue resolution (Resolved, In Progress, Not Resolved).
- **CostImpact (€):** Financial impact associated with the incident.

Global Relational Schema

Key Relationships:

- **ProductReference:**
  This key is used to link DensityReports and ProductAttributes, establishing a many-to-one relationship (i.e., many packaging reports correspond to one unique product). This linkage is crucial for integrating product-specific attributes into the analysis of packaging quality.
- **SupplierName:**
  This field connects DensityReports, SupplierScorecard, and HistoricalIncidents, enabling the analysis of how supplier performance and historical incidents affect packaging outcomes.

Data Integration Overview:

- **DensityReports** serves as the central dataset containing the operational packaging evaluations and quality labels.
- **ProductAttributes** enriches these reports with the inherent characteristics of each product.

- SupplierScorecard provides performance metrics that can be correlated with packaging quality.
- HistoricalIncidents offers background on past packaging issues, adding context to the quality assessments.s

Step-by-Step Roadmap for Solving the Exercise

Step 1: Understand the Case and Familiarize Yourself with the Data

- Objective:
  Develop a predictive model that classifies packaging quality ("Good" or "Bad") based on multiple data sources.
- Actions:
  - Read through the dataset documentation to understand the purpose and meaning of each variable.
  - Identify the key fields (ProductReference and SupplierName) that link the different datasets.

Step 2: Explore the Data

- Actions:
  - Perform a descriptive analysis of each dataset, noting the structure, distribution, and any potential issues (e.g., missing or inconsistent values).
  - Analyze the distribution of the packaging quality labels and examine variations across product attributes and dates.

Step 3: Preprocess and Clean the Data

- Actions:
  - Identify and correct inconsistencies such as typographical errors in supplier names.
  - Handle missing or inconsistent entries in critical fields, ensuring that ProductReference values are consistent across datasets.
  - Standardize date formats and properly encode categorical variables to prepare the data for further analysis.

Step 4: Integrate the Datasets

- Actions:
  - Merge DensityReports with ProductAttributes using the ProductReference key to enrich each report with detailed product attributes.
  - Consider incorporating aggregated information from SupplierScorecard (e.g., average adherence score) and HistoricalIncidents (e.g., total number of incidents or cost impact) based on SupplierName or ProductReference.
- Validation:
  Confirm that the merging process preserves the integrity of the data and that the key relationships (ProductReference and SupplierName) are maintained.

Step 5: Conduct Exploratory Data Analysis (EDA)

- Actions:

    - Generate visualizations and summary statistics to understand the distribution of PackagingQuality and other key variables.
    - Explore relationships and correlations between product attributes (such as GarmentType and Material) and packaging quality.
    - Investigate temporal patterns and assess whether supplier performance or historical incidents have a discernible impact on quality outcomes.

Step 6: Engineer Features

- Actions:
    - Create derived variables that capture additional information, such as extracting month, quarter, or year from DateOfReport.
    - Develop aggregated features from SupplierScorecard and HistoricalIncidents, such as average performance metrics or incident counts and cost impact per product or supplier.
    - Encode and, if necessary, normalize categorical and numerical variables to ensure compatibility with the chosen modeling techniques.

Step 7: Develop the Predictive Model

- Actions:
    - Select one or more supervised classification algorithms (for example, Random Forest, XGBoost, or Logistic Regression) to predict PackagingQuality.
    - Split the integrated dataset into training and testing subsets (or use cross-validation) to ensure a robust evaluation of model performance.
    - Train the model with the selected features and evaluate its performance using appropriate classification metrics (accuracy, precision, recall, F1-score, AUC-ROC, etc.).
    - Analyze feature importance to identify which variables have the greatest impact on the quality predictions.

Step 8: Interpret the Results and Document Findings

- Actions:
    - Interpret model outcomes and discuss how product attributes, supplier performance, and historical incidents are associated with packaging quality.
    - Prepare a comprehensive report that outlines the full workflow—from data exploration and cleaning through integration, feature engineering, model development, and evaluation.
    - Include visualizations and clear explanations of how the integrated data supports decision-making in the packaging process.

Step 9: Present and Discuss Your Findings

- Actions:
    - Develop a presentation that summarizes your methodology, key findings, and the potential operational benefits (such as reduced costs and improved efficiency) identified through the analysis.
    - Engage in a critical discussion highlighting the challenges encountered during data integration and model development and propose strategies for further refinement.