**Dataset Description**

**A. DensityReports.xlsx**

**Purpose:**  
This dataset is the central repository used to evaluate and optimize the packaging process at FashionWorld Retail. It contains 500,000 records of packaging reports, with each record representing an operational recommendation for how a product should be packaged and a final evaluation of the packaging quality.

**Key Variables:**

* **ReportID:** A unique identifier for each report.
* **SupplierName:** The name of the supplier involved (three suppliers are considered: SupplierA, SupplierB, and SupplierC).
* **DateOfReport:** The date on which the report was generated, ranging between January 1, 2023, and June 30, 2024.
* **GarmentType:** The type of garment, categorized into 8 defined groups (e.g., Shirt, Pants, Jacket, Dress, Skirt, Suit, Coat, Sweater).
* **Material:** The material of the product, selected from 3 options (Cotton, Polyester, Wool).
* **ProductReference:** A unique product code constructed based on the GarmentType and Material. This field serves as a key to link the report with detailed product attributes.
* **ProposedUnitsPerCarton:** The recommended number of products per carton designed to optimize packaging efficiency. This value reflects operational considerations and may include some variability.
* **ProposedFoldingMethod:** The recommended method for folding the product, classified into three categories (“Method1”, “Method2”, and “Method3”).
* **ProposedLayout:** The proposed type of box or layout, now categorized into five distinct options (LayoutA, LayoutB, LayoutC, LayoutD, LayoutE) representing diverse logistical solutions.
* **PackagingQuality:** The final label indicating the quality of the packaging, recorded as either “Good” or “Bad”, based on operational criteria linked to product attributes and logistics conditions.
* **DataLabeled:** An indicator that confirms every record includes a quality label, which is essential for training and evaluating predictive models.

**B. ProductAttributes.xlsx**

**Purpose:**  
This dataset provides detailed information for approximately 5,000 unique products. It is essential for understanding how specific product attributes impact packaging quality.

**Key Variables:**

* **ProductReference:** A unique identifier for each product, which aligns with the ProductReference in DensityReports.
* **GarmentType:** The type of garment.
* **Material:** The material of the product.
* **ProductName:** A descriptive name for the product.
* **Size:** The size of the product.
* **Collection:** The collection to which the product belongs (e.g., Summer or Winter).
* **Weight:** The weight of the product, an attribute that may influence packaging decisions.

**C. SupplierScorecard.xlsx**

**Purpose:**  
This dataset captures the performance metrics of the suppliers. It enables the evaluation of supplier efficiency and consistency in meeting packaging standards, which may influence the overall quality of packaging.

**Key Variables:**

* **SupplierName:** The supplier’s name, which is used to link supplier performance with the packaging reports.
* **ReportDate:** The date when the supplier performance was evaluated.
* **AdherenceScore:** A score reflecting the supplier’s adherence to the established packaging standards.
* **NumberOfReminders:** The number of reminders sent to the supplier regarding deviations from packaging guidelines.
* **OnTimeDelivery:** An indicator of whether the supplier delivered on time, typically recorded as “Yes” or “No”.
* **CostSavings:** The cost savings achieved through effective packaging practices.

**D. HistoricalIncidents.xlsx**

**Purpose:**  
This dataset documents historical incidents related to packaging issues. The recorded information provides insight into past problems, which may help in understanding and predicting which packaging operations fail to meet standards.

**Key Variables:**

* **IncidentID:** A unique identifier for each incident.
* **DateOfIncident:** The date when the incident occurred.
* **SupplierName:** The supplier involved in the incident, linking to the relevant performance data.
* **ProductReference:** The product code associated with the incident, used to connect with ProductAttributes and DensityReports.
* **IssueDescription:** A description of the encountered problem (e.g., packaging error, damaged product).
* **ResolutionStatus:** The current status regarding the resolution of the incident.
* **CostImpact:** The economic 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.