Data Preprocessing Project

Introduction:

This project focuses on preprocessing, augmenting, and merging two real-world datasets to predict customer spending behavior using machine learning. The project is divided into three main tasks: Data Augmentation on CSV Files , merging data with transitive properties and Data Consistency and quality Checks.
Additionally, we tackled the Bonus Challenge which involves training a machine learning model to predict customer spending.

Datasets

The following datasets were used in this project:

- Customer_transactions.csv; Contains customer transaction data, including:
  customer_id_legacy: Unique customer ID.
  transaction_id: Unique transaction ID.
  purchase_amount: Amount spent in the transaction.
  purchase_date: Date of the transaction.
  product_category: Category of the product purchased.
  customer_rating: Customer rating for the transaction.

- Customer_social_profiles.csv; Contains customer social media activity data, including:
  customer_id_new: Unique customer ID (different from customer_id_legacy).
  social_media_activity: Social media activity score.
  purchase_interest_score: Customer's interest in purchasing.

- id_mapping.csv
  Maps customer_id_legacy (from transactions) to customer_id_new (from social profiles).

Methodology:

Data Augmentation on CSV's:

Before starting the whole task, The dataset has categorical features which was the product_category column. It was converted to numerical features using Label Encoder to prevent running into errors.

- **Handling missing values;**The customer_rating column had 10 missing values. These were handled using mean imputation, where the missing values were replaced with the mean value of the column. After imputation,

the dataset was checked again, and no missing values remained. Just for curiosity, Bernice also tried to implement the predictive modelling way of handling missing values using the random forest regressor model and it also worked.

- **Data Augmentation strategies;**
  To introduce variability, random noise was added to the purchase_amount column. The noise was generated using a uniform distribution between *-0.1 and 0.1*, and the purchase amounts were adjusted accordingly.
  SMOTE was applied to balance the dataset, particularly for the product_category column. This technique generates synthetic samples for the minority classes to address class imbalance. After applying SMOTE, the dataset was resampled, and the product_category column was inversely transformed back to its original categorical labels.

  Log transformation was used to stabilize variance and make the distribution more normal. By applying *log1p (logarithm of 1 + x)*, we handle zero values effectively.
  Generating Synthetic Data: To create realistic synthetic transactions, we employed K-Means clustering to uncover patterns in customer behavior. The dataset was segmented into three distinct clusters based on two key features: purchase_amount and customer_rating. Each cluster's centroid represents the average characteristics of customer transactions within that group. By generating new transactions around these centroids, we were able to produce synthetic data points that closely mimic real customer behaviors. This approach not only compresses large values but also enhances model performance by providing a more representative and balanced dataset.

  After that, this Augmented data was saved to be used in Task 2.

Merging Data with Transitive Properties:

- **Data Merging:** The dataset underwent a structured merging process to ensure consistency and integrity:
  - The *transactions_df* was merged with *id_mapping_df* using *customer_id_legacy* to establish accurate customer identification.
  - This merged dataset was further combined with *social_profiles_df* on `customer_id_new` to integrate social profile information.
- **Conflict Resolution** to address inconsistencies and ensure reliability. The most frequently occurring customer_id_new was selected for each *customer_id_legacy* to resolve duplicate mappings. Missing values were handled appropriately as Numerical attributes were filled with the median value. Categorical attributes were

assigned default values to maintain data uniformity.

- **Feature Engineering:** In enhancing the dataset's analytical value, several new features were engineered: Customer Engagement Score was computed based on multiple factors:
    - A 7-day Moving Average was calculated to track purchase trends over time.
    - Monthly Spending was aggregated to understand customer purchasing behavior.
    - Sentiment Analysis was applied, mapping review_sentiment to numerical values for better interpretation.
    - TF-IDF (Term Frequency-Inverse Document Frequency) was used for text-based analysis to extract key insights from textual data.

      Afterwards, the final cleaned and processed data was saved to final_customer_data_[Group 8].csv to be used in the third task.

Data Consistency and Quality Checks

- **Data Integrity Check:** To ensure the reliability and accuracy of the dataset, the following integrity checks were performed: Duplicate entries and missing values were identified and addressed. All customer transactions were validated against a valid social profile to ensure data authenticity.
- **Statistical Summarization:** A comprehensive statistical analysis was conducted to gain insights into the dataset: Summary statistics were generated for numerical columns to understand central tendencies and distributions. The distribution of *purchase_amount* was visualized to detect potential anomalies and assess spending patterns.
- **Feature Selection for Machine Learning:** Feature selection was conducted to optimize model performance, Highly correlated features were identified using a correlation heatmap to avoid multicollinearity. The top 10 most important features were selected based on feature importance analysis.

  The data was then saved in final_dataset_ready_[group 8].csv.

Bonus Challenge: Predict Customer Spending

A Random Forest Regression model was trained to predict purchase_amount based on the selected features.The model achieved a R^2 Score of **0.85**, indicating a high level of predictive accuracy.The Mean Squared Error (MSE) was **0.195**, demonstrating the model's performance in estimating customer spending. The trained model was saved as

customer_spending_model.pkl for future use in customer spending predictions.

.