DATA PREPROCESSING SUMMARY REPORT

1. Data Loading, Cleaning and Augmentation

By: Marion Mwangi

The initial dataset, ***customer_transactions.csv***, was loaded, containing the following columns:

- customer_id_legacy – Legacy customer identifier
- transaction_id – Unique transaction identifier
- purchase_amount – Amount spent in the transaction
- purchase_date – Date of the transaction
- product_category – Category of the purchased product
- customer_rating – Customer's rating for the transaction

Data Cleaning Process:

- Identified missing values for later imputation.
- Checked for duplicate records (none were found).
- Verified data types to ensure consistency.
- Analyzed outliers using box plots and decided whether to retain or adjust extreme values.

Data Augmentation:

The following augmentation techniques were applied:

- Handling Missing Values: Used linear interpolation to fill missing values in the *customer_rating* column.
- Synthetic Data Generation: Introduced small variations (Gaussian noise) in *customer_rating* and *purchase_amount* to create a slightly altered version of the original dataset.
- Data Expansion: Generated synthetic transactions based on numerical features.

The cleaned and augmented dataset was saved as customer_transaction_augmented.csv

2. Dataset Merging and Feature Transformation

By: Sifa Mwachoni

The following datasets were merged:

- Transaction Data – Purchase amounts, ratings, and legacy customer IDs
- ID Mapping – Links legacy to new customer IDs
- Social Profiles – Engagement data, platform usage, and sentiment analysis

Key Challenges & Solutions:

1. Multiple transactions per customer**:** Grouped data by customer ID, aggregated purchases (sum) and ratings (average).
2. Multiple social profiles per customer**:** Merged platform data per customer, joined text fields, and averaged numeric values.
3. Missing data**:** Used a left join to retain all customers, filling gaps with default placeholders.
4. Pandas warnings: Updated outdated operations with modern, recommended practices.

The merged dataset was saved as merged_dataset.csv

Feature Engineering:

The following features were created:

- Binary: One-hot encoded social platforms and review sentiments
- Aggregated: Platform count, sentiment diversity
- Interaction: Purchase-engagement and purchase-sentiment interactions
- Ratios: Purchase per engagement, interest-to-purchase
- Variability: Purchase amount and rating fluctuations
- Composite Scores: Customer value and engagement complexity
- Normalized Metrics: MinMax-scaled key metrics, percentile rankings
- Customer Segments: Value-based, engagement-based, and combined categories

The following insights were gained during the feature engineering process**:**

- Left joins preserved all transaction records, highlighting data gaps.
- Summing purchase amounts and averaging ratings maintained interpretability.
- Text consolidation removed duplicate values while preserving readability.
- One-hot encoding expanded feature space, revealing cross-domain relationships.
- MinMax scaling enabled valid composite scores.
- Quantile-based segmentation ensured balanced customer groups.

The final dataset, incorporating merged data and engineered features, was saved as

final_customer_data_group3.csv