Customer Transactions Preprocessing Summary Report

Group Members and roles

Member	Role (task)
Glen Miracle	Part 1: Data Augmentation on CSV files
Nguepi Jordan	Part 2: Merging Datasets with Transitive properties
Peter Johnson	Part 3: Data consistency and Quality checks

Preprocessing Steps

The preprocessing of the `customer_transactions.csv` dataset involved several steps to clean, augment, and prepare the data for predicting customer_rating:

1. Data Cleaning and Augmentation (Glen Miracle)

- Loaded the **customer transactions dataset**.
- Handled **missing values** using mean, median, mode imputation, and predictive modeling.
- Applied **synthetic data generation** techniques, including:
  - Adding random noise to numerical values.
  - Using **SMOTE** for data balancing.
  - Applying log transformation for skewed data.
- Saved the cleaned dataset as `customer_transactions_augmented.csv`.

2. Merging Datasets (Nguepi Jordan)

- Merged **customer transactions** with **social profile data**, linking them using an **ID mapping dataset**.
- Standardized different customer ID formats to maintain consistency.

- Handled cases where a single customer ID had multiple records.
-

3. Feature Selection (Peter Johnson)

- Created a **Customer Engagement Score** using transaction history and social media activity.
- Engineered new features like:
  - Moving averages of transactions.
  - Time-based purchase aggregation.
  - Text-based features using **TF-IDF**.
- Identified **highly correlated features** using a **correlation heatmap**.
- Selected the **top 10 most important features** using **SelectKBest**.
- Saved the merged dataset as `final_customer_data_[Databases-Peer-2].csv`.

OUTPUT

- Saved the augmented, transformed dataset (300 rows) as `final_dataset_ready.csv.`

Key Insights

- **Missing Data**: customer_rating had 6.67% missing values (10/150), addressed via mean imputation to retain all rows.
- **Spending Patterns**: purchase_amount ranged from 51 to 495 (mean ≈280.78), with synthetic data maintaining this distribution.
- **Feature Relevance**: Temporal (e.g., days_since_purchase) and behavioral (e.g., `avg_monthly_purchases`) features strongly correlated with customer_rating, suggesting predictive power.
- **Data Expansion**: Doubling the dataset improved robustness for modeling without overfitting risks.

Challenges and Solutions

- **Challenge: Limited Data Size**
  - *Solution*: Augmented with synthetic data, adding controlled noise and customer-specific category sampling to mimic real transactions.
- **Challenge: Missing Values in customer_rating**
  - *Solution*: Imputed with the mean to avoid dropping rows, given the low missing rate and numeric nature.
- **Challenge: Skewed purchase_amount**
  - *Solution*: Applied normalization (purchase_amount_normalized) to reduce skewness, aiding clustering and selection.
- **Challenge: Temporal Feature Accuracy**
  - *Solution*: Used the current date (March 16, 2025) for days_since_purchase, assuming recent data relevance; future adjustments could use a fixed reference date.
- **Challenge: Feature Overload**
  - *Solution*: Employed `SelectKBest` to focus on the top 10 predictive features, balancing complexity and utility.

Conclusion

The preprocessing transformed a small, partially incomplete dataset into a clean, augmented, and feature-rich version suitable for machine learning. The steps addressed data quality, quantity, and relevance, setting the stage for accurate customer_rating predictions.