Group 4: Formative 2 Report

Steps taken in preprocessing

1. *Data augmentation on CSV files (Nina)*

After loading the data, a search for missing values was done and it was found that the **customer_rating** had a total of 10 missing values. The column was then visualized using a histogram and KDE to check its distribution. The distribution was uneven, therefore we chose to impute the missing values using the median. For data augmentation we applied random noise from a Gaussian distribution.
We proceeded to visualise the skewness of the data and sampled 10% of the dataframe to create a new dataframe, **new_transactions**, that contains synthetic transactions. The new dataframe is concatenated with the original dataframe which extends the original dataframe with synthetic transactions.

2. *Merging datasets with transitive properties (Madol)*

In this task, two datasets were merged, customer_transactions_augmented.csv and customer_social_profiles.csv, which use different customer ID systems (customer_id_legacy and customer_id_new). To link these datasets, a third mapping file, id_mapping.csv was also loaded. After merging, the dataset was enhanced through feature engineering and the final dataset was saved as final_customer_data_group4.csv.

3. *Data consistency and quality checks (AlHassan)*

This section focuses on understanding the dataset before applying any preprocessing or modeling. Proper data exploration helps us identify patterns, detect anomalies, and make informed decisions about feature engineering and model selection. The key steps taken to validate this data are: 1. I checked for duplicate records, 2. Histogram plots help understand variable distributions, and the purchase_amount column was modified by adding random noise to simulate real-world variations, and 3. a correlation heatmap to identify the top 10 features for the model training.

4. *Bonus (Florent Hirwa)*

The bonus section extends the data pipeline by incorporating a machine learning model to predict customer spending. In this part, the final preprocessed dataset is loaded and split into features and a target variable (purchase amount). Missing values in both the predictors and the target are handled appropriately before the dataset is divided into training and testing sets. A simple Linear Regression model is then trained on the training data, evaluated using metrics like Mean Squared Error and R^2, and finally saved to disk using joblib for future use. This section demonstrates how to convert your cleaned and enriched data into actionable insights through predictive modeling.

Summary of Key Insights Found During Preprocessing

1. During aggregation we mainly focused on the numeric categories and didn't include the categorical columns, we later realised we needed to aggregate them as well in order to feature engineer the new **customer_engagement_score** column.
2. Combining datasets from different sources can surely provide valuable insights, but requires considerable handling, for example, when we merged customer_transaction_augmented.csv and customer_social_profiles.csv, revealed connections between transaction history and social media activity.

Challenges faced and how they were resolved.

1. Conflicting schedules within the team. We resolved this through keeping each other informed on our assigned tasks' progress on Google Chats.

2. The two datasets used different Custom ID systems making it difficult for a direct merge. To provide a solution for this, we used the'' Id_mapping.csv'' file to establish a link between the two ID systems.

Video link

https://drive.google.com/file/d/1oF4HcNGi9wsEqoo4_jSNngioaCFyKKc3/view?usp=share_link