Data Intake Report

Name: Customer Segmentation for XYZ Bank
Report date: 18th September 2024
Internship Batch: LISUM36
Data intake by: KRISHNAKANT SONJI
Data storage location: GitHub - https://github.com/Thir13een/DG-PROJECT.git

Tabular data details:

Total number of observations	1 million observations
Total number of files	1 file
Total number of features	48 columns
Base format of the file	.csv format
Size of the data	161.7 Mb

Proposed Approach:
Deduplication Validation:

- **Check for duplicate rows using customer ID, age, and join date.**
- **Count distinct customer IDs to spot duplicate records.**
- **Use profiling tools to find duplicates automatically.**

Assumptions:

- **Missing income values will be filled using the median.**
- **The dataset is a good sample for training customer segmentation models.**
- **Outliers in age or seniority may be errors and will be handled during feature engineering.**