**Contribution Report**

**Synthetic Data Generation**

**Objective**

The goal was to generate realistic synthetic datasets for loan applications and KYC/watchlist checks. These datasets simulate real-world scenarios for testing the loan review agent logic in a controlled environment, without using sensitive personal data.

**Implementation Approach**

* Used **Python** with pandas, random, and Faker libraries to generate datasets.
* Ensured realistic data by applying rules and constraints on numerical and categorical values.

Introduced **some missing or randomized values** to mimic incomplete or real-world applications.

**Tasks Completed**

|  |  |
| --- | --- |
| **Task** | **Description** |
| **NIC Generator** | * Function to create valid Sri Lankan NICs in both old (90######V) and new (1990#########) formats. NICs stored as strings to prevent scientific notation issues in CSV files. |
| **Loan Application Dataset** | * Generated 30 synthetic loan applications with fields: *application\_id, full\_name, nic, age, employment\_status, monthly\_income\_lkr, existing\_debt\_lkr, requested\_amount\_lkr, loan\_purpose, credit\_score, kyc\_status, aml\_flag, delinquency\_12m, dti.* * Randomized data maintained logical constraints (e.g., loan amount vs income, debt-to-income ratios). |
| **KYC Watchlist Dataset** | * Created watchlist data for testing KYC/AML logic. Some NICs matched loan applications, others were new. Fields: *nic, flagged\_reason, on\_watch/risk\_level.* |
| **Data Integration and Quality Control** | * Ensured some NICs matched between datasets while others did not, simulating real-world incomplete coverage. Checked data integrity and maintained consistent formats across CSVs. |

**Challenges Faced**

* Prevented NIC numeric values from being saved in scientific notation.
* Balanced randomness and realism in loan amounts, income, and employment types.
* Intentionally included missing or partial data to test agent robustness.

**Outcomes**

* Generated two high-quality synthetic datasets: Loan\_applications.csv and Kyc\_watchlist.csv.
* Some NICs matched to simulate realistic KYC/AML scenarios.
* Datasets now serve as input for testing and validating the loan review agent prototype.