Generating Synthetic Transaction Data with Australian Locale Using the Faker Library

Introduction

This document describes the process used to generate a synthetic dataset that mirrors the structure and contents of an existing retail transactions dataset, incorporating realistic Australian locale data such as state and suburb. The Python Faker library is utilized to ensure the synthetic data is convincingly real, supporting various testing and development needs without compromising privacy.

Objectives

The primary objective is to produce a robust synthetic dataset for use in applications where real transaction data is sensitive. This dataset includes realistic demographic and transactional data, conforming to Australian geographic specifics such as state and suburb, enhancing the dataset's utility for regional analysis.

Methodology

- Tools Used

- Python: The primary programming language for scripting.
- Pandas Library: For handling data structures and operations.
- Faker Library: For generating fake data, customized to Australian locales.

- Data Structure

The synthetic dataset incorporates several fields, each designed to replicate the complexity and variety found in real-world data:

- Transaction_ID
- Date
- Customer_Name
- Product_Name
- Total_Items
- Total_Cost
- Payment_Method
- Suburb
- Postal_Code
- State
- Sub_Category
- Product_Group
- Unit_Price
- Unit_Price_Unit
- SKU

- Steps for Synthetic Data Generation

1. Setup and Configuration:

- Install Python and necessary libraries (Faker, Pandas).
- Configure Faker to use the Australian locale ('en_AU') to ensure geographic accuracy for names, suburbs, and postal codes.

2. Loading Original Data:
   - Load "Australia_Grocery_2022Sep.csv" to use real SKUs, product names, and sub-categories, maintaining product reference integrity.

3. Generating Data Entries:
   - Develop a function to generate each synthetic data entry, ensuring each field is populated with appropriate fake data or data derived from the original dataset.
   - For each entry:
     - Transaction_ID: Generate a unique identifier.
     - Customer_Name, Date, Total_Items, Total_Cost, Payment_Method: Use Faker to generate realistic transaction details.
     - Suburb, State, Postal_Code: Use Australian localized settings in Faker to generate authentic Australian addresses.
     - SKU, Product_Name, Sub_Category: Randomly select from the original dataset to ensure consistency.
     - Product_Group, Unit_Price, Unit_Price_Unit: Generate using Faker with set categories and price ranges to reflect typical retail scenarios.

4. Compiling and Saving Data:
   - Compile the generated entries into a pandas DataFrame.
   - Perform any necessary data cleaning or transformation.
   - Save the DataFrame to a CSV file for ease of use in subsequent applications.

Conclusion

This methodology ensures the creation of high-quality, realistic synthetic data tailored to Australian locales, suitable for a variety of applications including software testing, data analysis training, and development environments where the use of real data is impractical or restricted.