Work Task2 for the Ph.D. Position in GenAI @ BTH
Prepared By: Kanchana Weerasinghe

1. Objective

The objective of this project is to generate synthetic sample data that closely mimics the statistical properties and distributional characteristics of the original dataset. The generated data should align with the real data in both numerical and categorical aspects.

2. Approach Overview

The overall approach consists of **three main components**:

- **OriginalDataGenerator**: Simulates an original dataset with controlled randomness.
- **DistributionAnalyzer**: Analyzes the original dataset to understand and capture underlying statistical patterns.
- **SampleDataGenerator**: Generates new synthetic data that matches the distributional patterns of the original data.

Each component is modular, reproducible, and designed to ensure that the synthetic sample data remains **statistically and visually consistent** with the original data.

3. Detailed Methodology

3.1 Original Data Generation

Class Used: OriginalDataGenerator

- **Purpose**: To generate an initial dataset with:
  - Categorical values (following a specified probability distribution).
  - Numerical values (following specified normal distributions).
- **Key Design Choices**:
  - Fixed random seed for reproducibility.
  - Saved output to CSV for persistence and inspection.

3.2 Statistical Distribution Analysis

Class Used: DistributionAnalyzer

- **Purpose**: To analyze the original dataset and capture:
  - Type (Numerical or Categorical),
  - Statistical properties (mean, std, min, max),

    - Best-fit distribution (Normal, Log-normal, Exponential, or Gamma for numerical columns),
    - Category proportions for categorical columns.
- **Key Design Choices**:
    - Fit multiple distributions and select the best one using **Kolmogorov–Smirnov (KS) test** based on the highest p-value.
    - For categorical data, proportions are preserved exactly as observed.
    - Visualizations created to visualize the parameter distributions

3.3 Sample Data Generation

Class Used: SampleDataGenerator

- **Purpose**: To generate a synthetic sample dataset that:
    - Mimics the statistical patterns and distributions of the original dataset.
    - Ensures both global (mean, std, skewness, kurtosis) and local (distribution shape) alignment.
- **How it works**:
    - **Numerical Columns**:
        - Synthetic data is generated using the *best-fit distribution* found earlier.
        - Additional adjustment for skewness and kurtosis if discrepancies are observed.
    - **Categorical Columns**:
        - Sampling performed using the same category proportions as the original data.
- **Key Design Choices**:
    - Fix the random seed for reproducibility.
    - Adjust higher-order moments (skewness and kurtosis) to better mimic real-world asymmetries.

4. Validation Strategy

Validation of the synthetic data is done in two stages:

4.1 Statistical Validation

- **Kolmogorov–Smirnov (KS) test** for numerical columns:
    - Compares the cumulative distributions of original and sample data.
    - A p-value > 0.05 suggests that the distributions are statistically similar.
- **Chi-square test** for categorical columns:
    - Compares the observed frequencies of each category.
    - A p-value > 0.05 suggests that the proportions are statistically similar.

4.2 Visual Validation

- **Distribution plots** (histograms + density curves) for numerical data.
- **Bar plots** comparing category proportions for categorical data.

These plots provide an intuitive verification that the synthetic sample matches the original dataset.

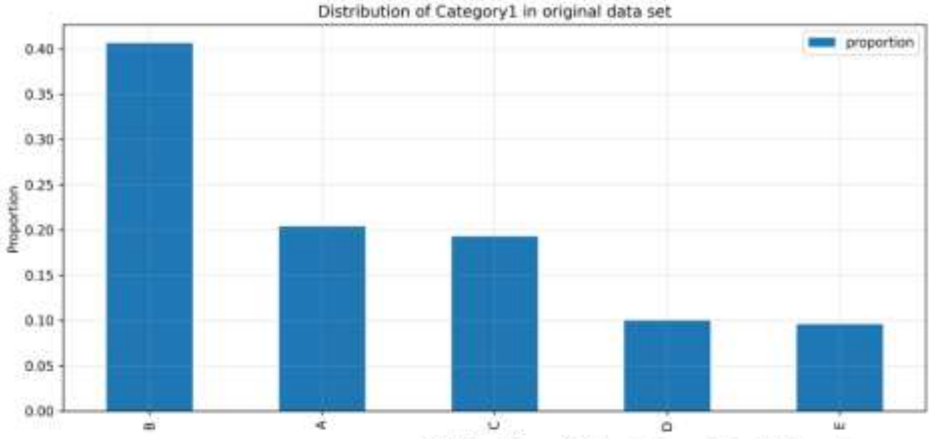
5. Justification for the Approach

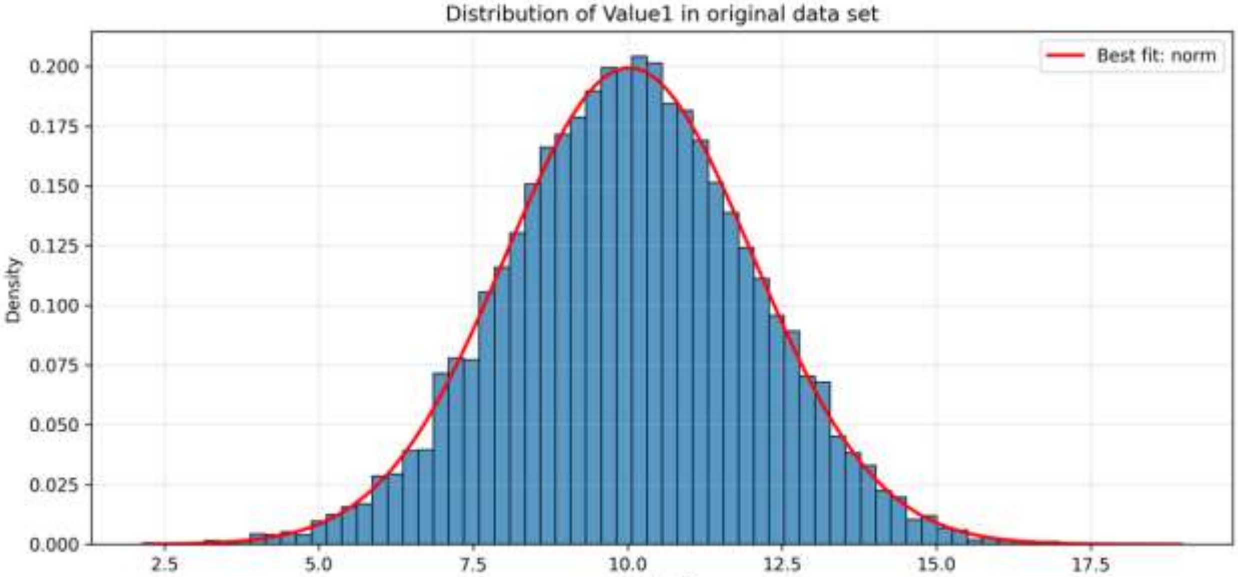
- **Comprehensive Distribution Modeling**: Captures both central tendency (mean, std) and shape (skewness, kurtosis).
- **Flexible Distribution Fitting**: Supports multiple distributions and selects the best based on KS test, rather than assuming normality.
- **Reproducibility**: Fixed random seeds ensure reproducible experiments.
- **Robust Validation**: Combines statistical tests and visual inspection.
- **Realism**: Adjusts for skewness and kurtosis to simulate realistic data properties, which simple mean-variance matching cannot achieve.
- **Extensibility**: New distributions or features can be easily added to the framework if the dataset evolves.

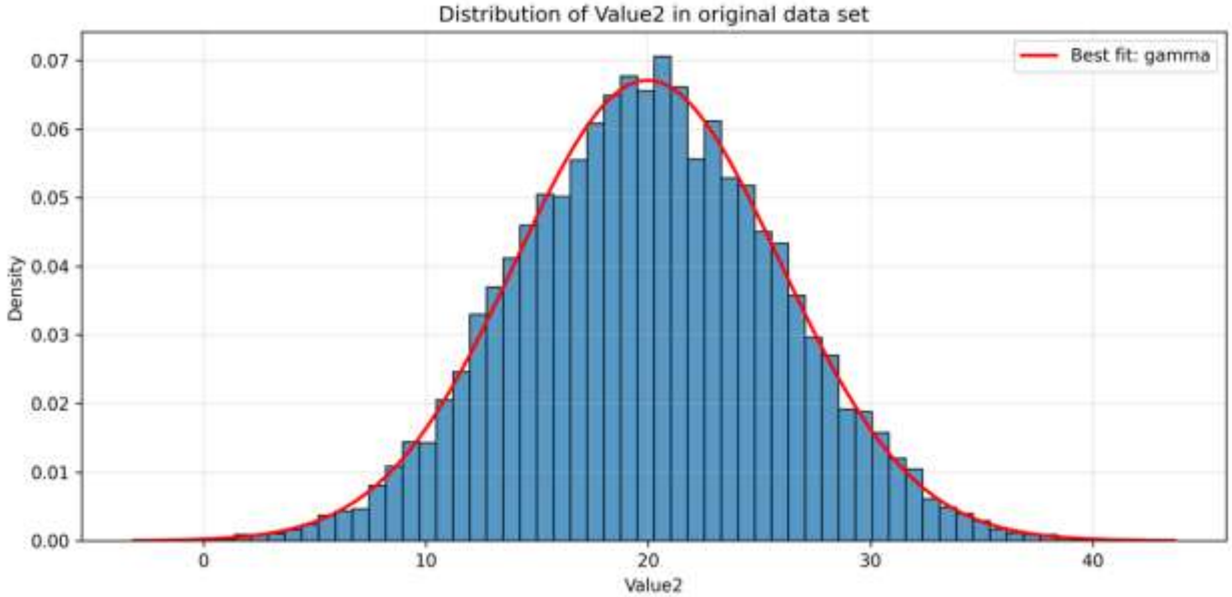
6. Results

6.1 Original Data Set

```
==================================================
Column: Category1
Type: Categorical
--------------------------------------------------
Number of Categories: 5
Most Common Category: B
==================================================

==================================================
Column: Value1
Type: Numerical
--------------------------------------------------
Fitted Distributions and KS-Test Results for original data set:
  - norm: KS Statistic = 0.0049, p-value = 0.9663
  - lognorm: KS Statistic = 0.0450, p-value = 0.0000
  - expon: KS Statistic = 0.4419, p-value = 0.0000
  - gamma: KS Statistic = 0.0053, p-value = 0.9366

Best Fit Distribution: norm
Reason: norm had the highest p-value (0.9663),
indicating the best agreement with the data according to the Kolmogorov-Smirnov test.
==================================================

==================================================
Column: Value2
Type: Numerical
--------------------------------------------------
Fitted Distributions and KS-Test Results for original data set:
  - norm: KS Statistic = 0.0066, p-value = 0.7710
  - lognorm: KS Statistic = 0.0711, p-value = 0.0000
  - expon: KS Statistic = 0.3646, p-value = 0.0000
  - gamma: KS Statistic = 0.0064, p-value = 0.8042

Best Fit Distribution: gamma
Reason: gamma had the highest p-value (0.8042),
indicating the best agreement with the data according to the Kolmogorov-Smirnov test.
```


Distribution of Category1 in original data set

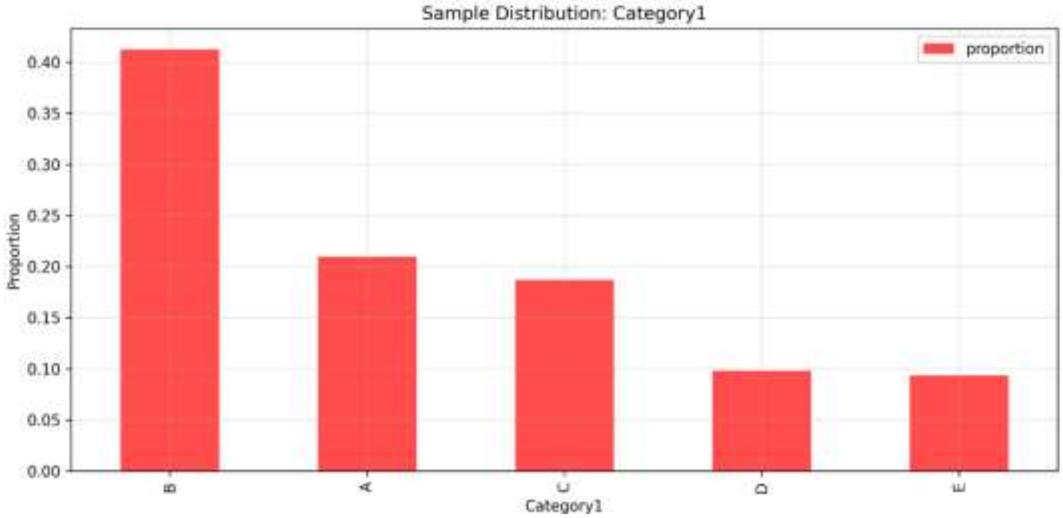

Distribution of Value1 in original data set


Distribution of Value2 in original data set

6.2 Sample Data Set

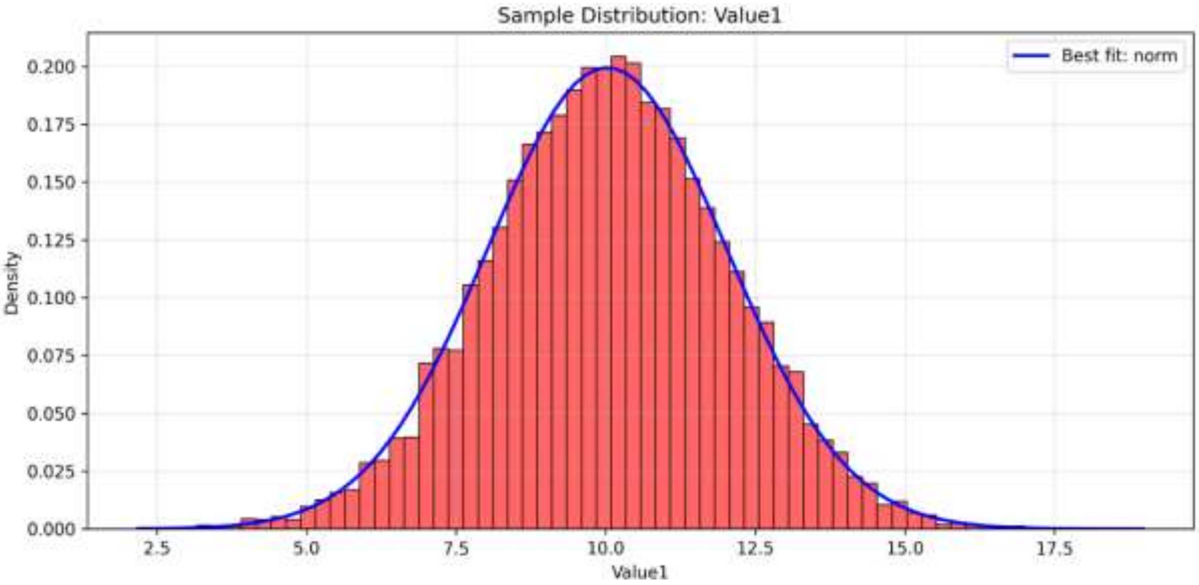
6.2 .1 Categorical Column: Category1

```
Sample proportions:
Category1
A    0.2095
B    0.4124
C    0.1868
D     0.0980
E    0.0933
```

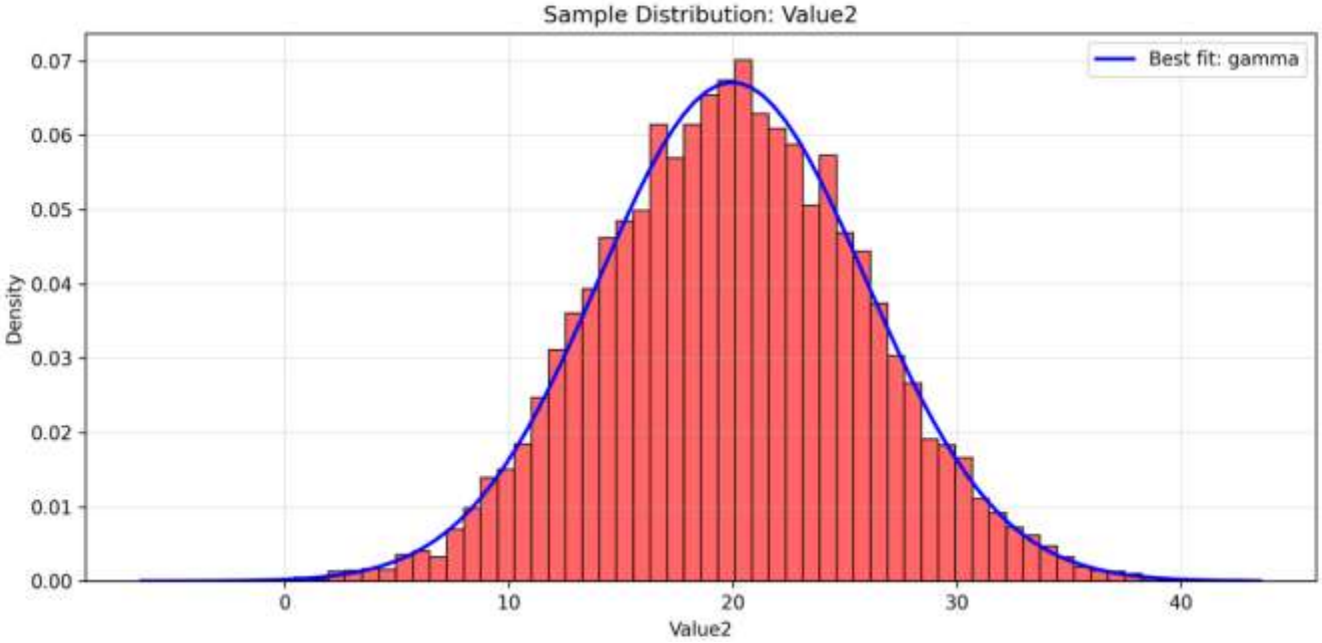


6.2 .2 Numerical Column: Value1

Distribution: Normal
- Sample mean: 10.05
- Sample std: 2.00
- Sample skew: -0.02
- Sample kurtosis: 0.06



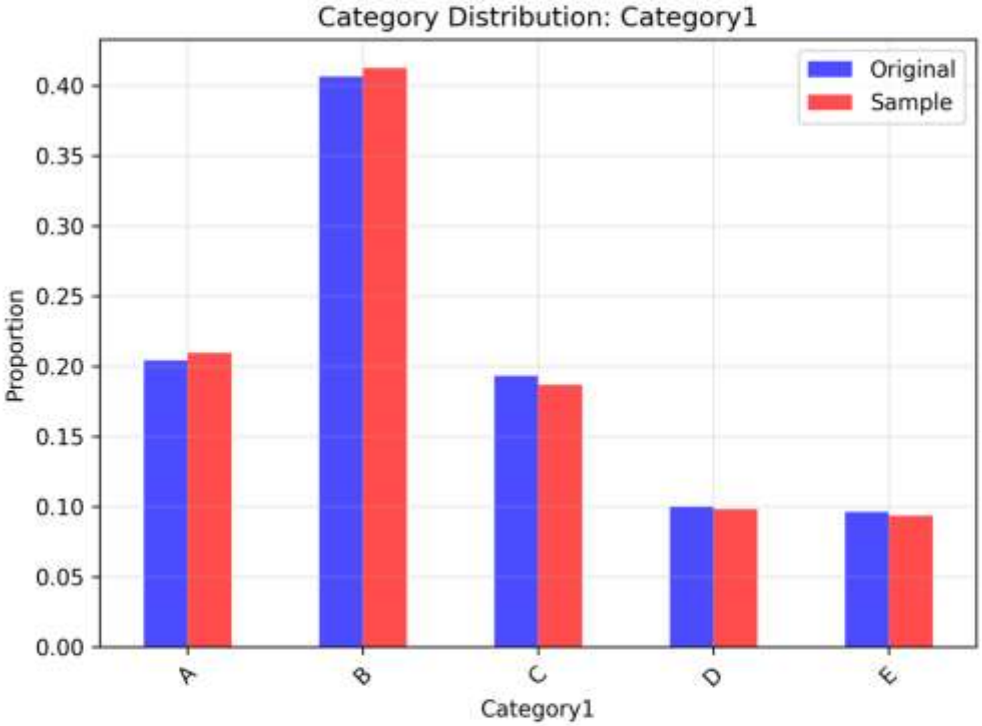
Distribution: Gamma
- Sample mean: 19.97
- Sample std: 5.97
- Sample skew: 0.02
- Sample kurtosis: 0.01



7. Comparison of Original vs Sample data distributions

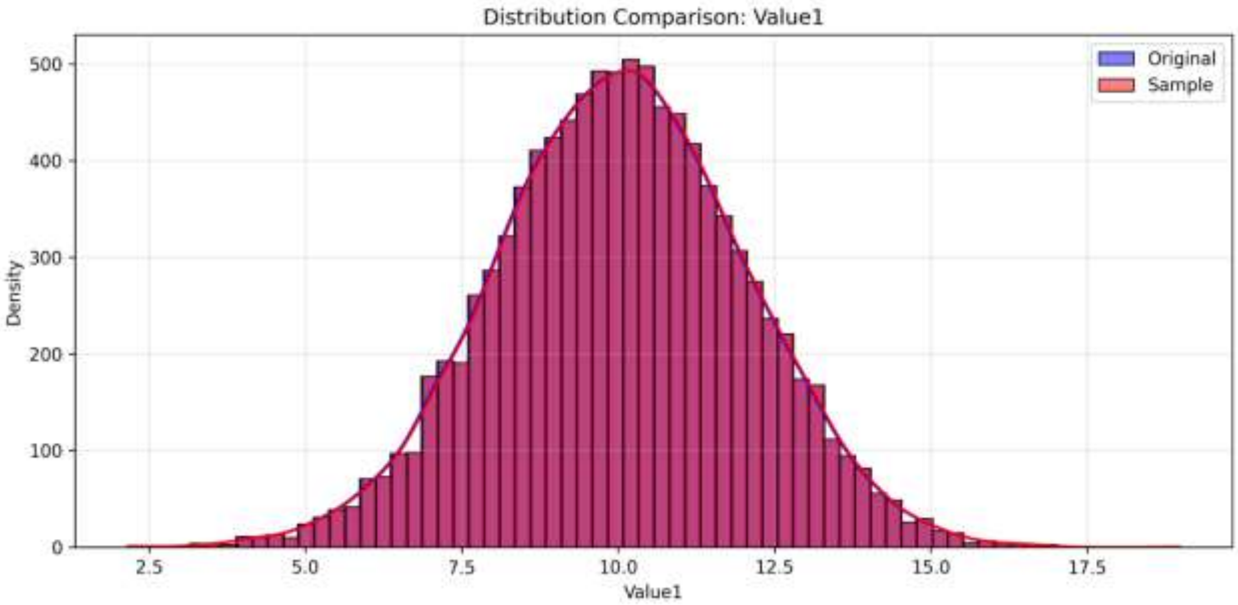
7.1 Categorical Variable

Original Category1 variable	Sample Category1 variable
A 0.2043	A 0.2095
B 0.4065	B 0.4124
C 0.1930	C 0.1868
D 0.1001	D 0.0980
E 0.0961	E 0.0933
Chi-Square Test: Statistic=5.4280, p-value=0.2461 (p > 0.05 indicates proportions match)	



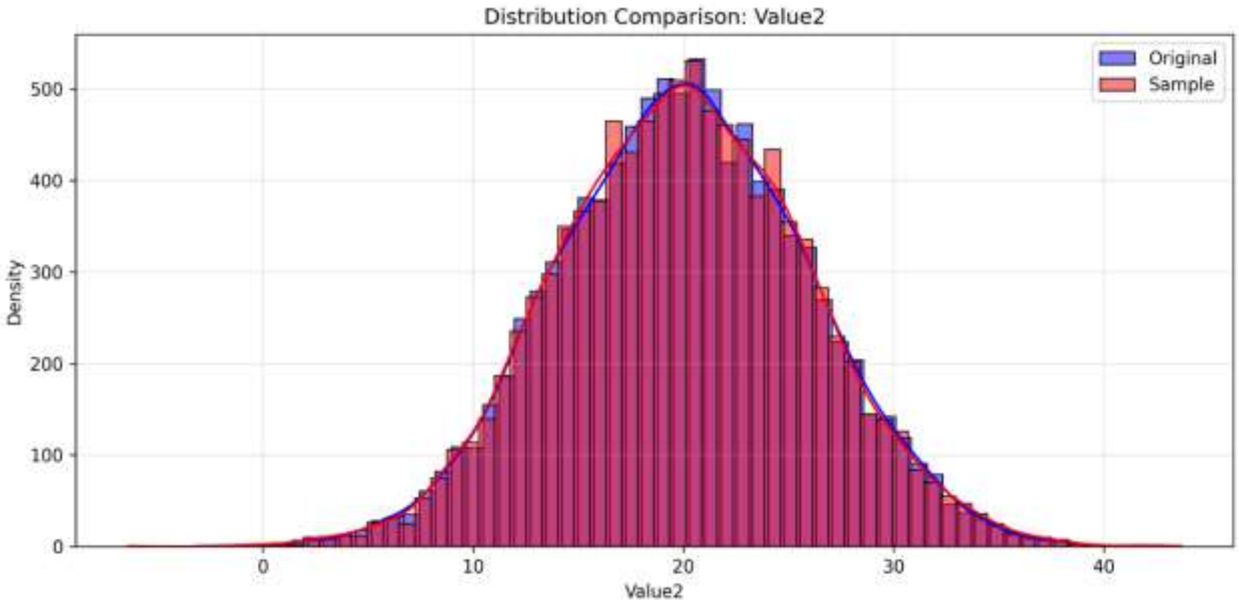
7.2 Numerical Variable – Variable 1

Original Value1 variable	Sample Value1 variable
mean: 10.02	mean: 10.05
std: 2.00	std: 2.00
skew: -0.02	skew: -0.02
kurtosis: 0.06	kurtosis: 0.06
KS Test: Statistic=0.0075, p-value=0.9412 (Values closer to 0 and p > 0.05 indicate good match)	



7.3 Numerical Variable – Variable 2

Original Value2 variable	Sample Value2 variable
mean: 19.99	mean: 19.97
std: 5.94	std: 5.97
skew: 0.01	skew: 0.02
kurtosis: -0.05	kurtosis: 0.01
KS Test: Statistic=0.0067, p-value=0.9783 **(Values closer to 0 and p > 0.05 indicate good match)**	



Alignment with the Instruction

“Use the Python code below to generate a dataset.”

- Used the OriginalDataGenerator class to generate the original dataset.

“Based on the characteristics of this dataset, your task is to generate new, similar samples.”

- SampleDataGenerator analyzes the original dataset's distribution characteristics using parameters such as best-fit distributions, skew, and kurtosis, then generates new samples that statistically and visually align with those characteristics.

“Do not use the provided sampling parameters when generating new samples.”

- Instead of reusing the original fixed parameters (e.g., loc=10, scale=2), this code dynamically use best-fit distributions derived from the data (batch_test_results) to generate new samples.

“It’s ok to generate a larger dataset, i.e. altering the num samples.”

- Designed a is flexible — num_samples is a configurable parameter in the constructor, and not hardcoding the size of the generated sample to match the original.