MyFaker - Fake data generator

Introduction

Data is king and for almost all the ML models need a large volume of quality data. With increase in privacy and security concerns, companies are restricting customer data usage so it is becoming hard to get good quality data for development and testing. So developers need to generate their own fake data to reflect real world scenarios.

Objective

The objective of this project is to build a package to generate random sample data set defined by user

- Flexibility/Configurability: Simple option to provide configuration:, so users need flexibility to generate data based on their conditions.
- Extensibility: Support custom data input: Each field needs different data, for example, medical data is different from sales data. So it is helpful if the package supports bringing their own data.
- Manage relation between data columns – ex: Product & Product Category, should go together to provide meaningful information
- Generate pandas data frame: Simple to output or re structure in the required format

Targeted users

This package is helpful for Students, programmers/developers, data scientists and analysts/researchers. They can use this library to generate data for their development and testing. Most data science work is of no use without data, similarly need data for testing as well as to understand any existing package. Researchers often need different data sets to analyze various scenarios.

Use cases

- David, a student, would like to create a model using eCommerce sales data, but it is hard to find real credit card data for his experiment. He can use this package to

generate his own data. He can specify the format of the card and other properties and generate thousands of records in no time.

- Mary, a data scientist, wants to analyze a model's performance and needs quality data. She can use this package to generate necessary data.