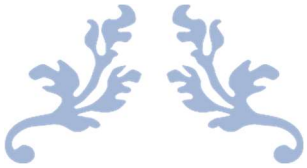


FAKE DATA GENERATOR

Hari Kishor Chintada

1. Objective

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.

2. Goals

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

- Generate pandas data frame: Simple to output or re structure in the required format

3. Use Cases

3.1. 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.

3.2. 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.

4. Design/Implementation

Create a package to generate a dataset using configuration. Provide flexibility to users to bring their own data or define rules to generate data. My primary objective is to create a project using some of the concepts I learned from Data 515 class.

4.1. Design concepts

The following design concepts were considered in this project implementation:

- Object oriented design: Created Data Generator class and defined various methods in it.
- General purpose deep modules: Defined helper class as general purpose deep module. Depending on user input, the data generator performs various tasks and returns requested data.
- Separation of concern: Created modules or functions to perform each separate/independent tasks, for example defined independent functions to process data from data frames and process regular expressions
- Information hiding: Most of the complex processing logic is hidden from the user and defined in helper class.

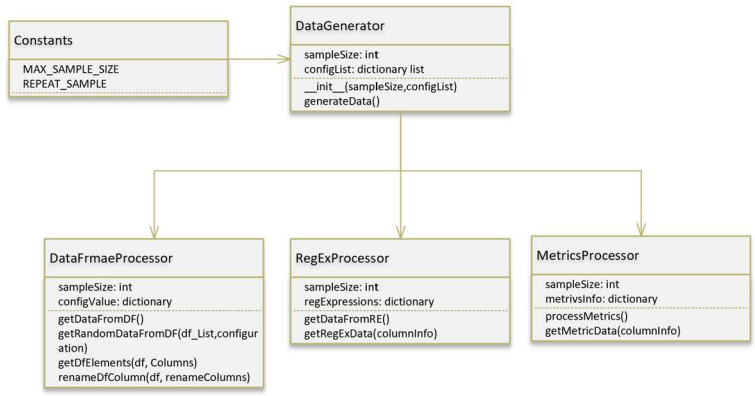
- Error handling: Included error handling using try and except and raised error to end user if there is any failures
- Testing: Included a few unit test cases and examples to understand how to prepare the configuration.

4.2. Libraries used:

The following libraries/modules are used in this project.

- pandas: To use pandas data frame and operations on it
- random: To generate random value or sample
- rsts: To generate random string using simple regular expression

5. Class diagram



6. How to use

Follow below steps to use the MyFaker package.

- Get latest code from Repo

  git clone https://github.com/HarikcUW/myfaker.git

- Check and update constants in constants.py file.
  - You can define Max number of rows and want to repeat any values or not
- Install MyFaker package using setup

  python setup.py install
- In your script file, Import myfaker package
- Define data schema and prepare configuration dictionary list
- Create an object and call generateData() function with parameters (number of rows to generate, configuration dictionary list)
- Capture return data frame

7. Limitations

- rstr module doesn't support all regular expressions, any unsupported complex expression required own implementation
- Not using distributes design, so result data set size & volume depends on user system capacity

8. Future Extensions

- Simplify how user can pass configuration, something like user pass (categorical, categorical, categorical, int, int, float)
- Extend to generate how many distinct values should generate for each categorical feature
- For Metrics – right now we generate random number, extend it to support any specific distribution

9. Summary

Data is required for almost all projects and often developers need to generate data themselves. The myfaker package can help to generate fake data using either from another data frame and regular expression.

10. References

10.1. https://github.com/leapfrogonline/rstr

10.2. https://github.com/joke2k/faker