Generating data with Faker

This package can be used to generate synthetic data, mostly categorical (we’ll look at alternatives for quantitative)

How to do it:

from faker import Faker

fake = Faker()

fake.name()

Faker has LOTS of categories to generate: check the documentation for more <https://faker.readthedocs.io/en/master/index.html>

You can even make your results local to your geography/language!

Once we’ve called Faker() we can add more data:

fake.country()

If you don’t want the data to continue to generate randomly, set the Faker seed:

from faker import Faker

fake = Faker()

fake.seed\_instance(1234)

fake.name()

To create a DataFrame with different categories and quantities, it’s worth defining a function first:

def generate\_data(num\_rows):

    data = {

        'Name': [fake.name() for \_ in range(num\_rows)],

        'Age': [fake.random\_int(min=18, max=80) for \_ in range(num\_rows)],

        'City': [fake.city() for \_ in range(num\_rows)],

        'Income': [fake.random\_int(min=30000, max=120000) for \_ in range(num\_rows)],

        'Category': [fake.random\_element(elements=('A', 'B', 'C')) for \_ in range(num\_rows)],

        'Score': [np.round(np.random.uniform(0, 100), 2) for \_ in range(num\_rows)]

    }

    return pd.DataFrame(data)

df = generate\_data(10)

EXERCISES

1. Generate 5 random email addresses (fake.email())
2. Generate 3 random city names (fake.city())
3. Modify your code to set a random seed
4. Create a DataFrame with 10 rows of fake
   1. Name (fake.name())
   2. Phone number (fake.phone\_number())
   3. Company (fake.company())
   4. Credit card number (fake.credit\_card\_number())
   5. Job (fake.job())

Monte Carlo – distribution of profits

Let’s build a user-driven model to simulate and plot distribution of profits given different assumptions