About Dataset

This is a simulated credit card transaction dataset containing legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants.

Source of Simulation

This was generated using Sparkov Data Generation | Github tool created by Brandon Harris. This simulation was run for the duration - 1 Jan 2019 to 31 Dec 2020. The files were combined and converted into a standard format.

Information about the Simulator

I do not own the simulator. I used the one used by Brandon Harris and just to understand how it works, I went through few portions of the code. This is what I understood from what I read:

The simulator has certain pre-defined list of merchants, customers and transaction categories. And then using a python library called "faker", and with the number of customers, merchants that you mention during simulation, an intermediate list is created.

After this, depending on the profile you choose for e.g. "adults 2550 female rural.json" (which means simulation properties of adult females in the age range of 25-50 who are from rural areas), the transactions are created. Say, for this profile, you could check "Sparkov | Github | adults_2550_female_rural.json", there are parameter value ranges defined in terms of min, max transactions per day, distribution of transactions across days of the week and normal distribution properties (mean, standard deviation) for amounts in various categories. Using these measures of distributions, the transactions are generated using faker.
What I did was generate transactions across all profiles and then merged them together to create a more realistic representation of simulated transactions.

Acknowledgements

- Brandon Harris for his amazing work in creating this easy-to-use simulation tool for creating fraud transaction datasets.
- Link : https://www.kaggle.com/code/matheusgou/fraud-prediction-and-threshold-analysis