Vamos listar 5 fontes de dados que permitem você fazer o download e usar gratuitamente diversos datasets públicos, que podem ser usados como ponto de partida em seus projetos de Data Science.

A propósito: lembre-se sempre de referenciar suas fontes de dados. Não use dados privados e nem use dados sem a devida autorização. Estamos na era da LGPD (Lei Geral de Proteção de Dados) além de outras leis que regem o uso de dados ao redor do mundo.

Os datasets disponíveis aqui são públicos e podem ser usados livremente, mas certifique-se de checar os termos de uso.

1- Scikit-Learn

O Scikit-Learn é o principal framework Python para construção de modelos de aprendizado de máquina e contém várias APIs para diversos conjuntos de dados, desde dados simples, passando por dados reais, até a geração de dados para um propósito específico. Aqui estão os links para você:

Toy Datasets

Real World datasets

Generated Datasets

Other Datasets

2- NLTK

NLTK é um pacote Python específico para o trabalho de Processamento de Linguagem Natural. O NLTK também fornece conjuntos de dados de texto que você pode usar para seus projetos.

Existem dezenas de conjuntos de dados de texto do NLTK disponíveis para uso. Consulte a lista completa aqui: **NLTK Corpora**

3- Statsmodels

Statsmodels é um pacote Python para modelagem estatística, mas o pacote também fornece vários conjuntos de dados que podem ser usados em seus projetos. Aqui a lista completa: **Statsmodels Datasets**

4- Pydataset

Pydataset é um pacote Python que fornece vários conjuntos de dados de código aberto. Os datasets são básicos, mas podem ser um bom ponto de partida para um projeto ou para um experimento com uma nova biblioteca de Machine Learning. Confira o pacote aqui: **Pydataset**

5- Datasets

Datasets é um pacote Python da HuggingFace criado especificamente para acessar e compartilhar conjuntos de dados.

O que é ótimo no pacote datasets é que, não importa o tamanho do conjunto de dados, você pode processar o conjunto de dados com leituras de cópia zero sem nenhuma restrição de memória, pois o pacote datasets usa o Apache Arrow em segundo plano.

Você pode examinar o hub HuggingFace do pacote datasets para obter a lista completa com milhares de conjuntos de dados: **Datasets**