How to Source Real-World Datasets for Data Analytics & Machine Learning Projects

Introduction

Working with diverse, real-world datasets during this 8-week course will broaden your experience, prevent overfitting, and foster creativity. This guide provides tips and links to help you source high-quality, unique datasets for your projects.

***** Students must not use datasets that are provided or utilized during the course. Instead, they are required to find new, unique datasets for their projects.*****

How to Source Datasets

1. Identify Your Problem Domain
    - Consider the area of interest (e.g., finance, healthcare, social media) and define the problem you want to solve.
    - Ask questions: What insights are you trying to uncover? What predictions do you want to make?

2. Explore Open Data Platforms

There are numerous platforms where you can find free and open datasets for analysis.

Below are some recommended sources:

Platform	Description	Link
Kaggle	A vast collection of datasets across various domains, ideal for machine learning and analytics.	https://www.kaggle.com/datasets
UCI Machine Learning Repository	A popular repository with diverse datasets, often used in academic research and machine learning projects.	https://archive.ics.uci.edu/
Google Dataset Search	A search engine specifically for datasets, pulling from various open data sources.	https://datasetsearch.research.google.com/
Data.gov	A comprehensive resource for U.S. government datasets on a wide range of topics.	https://data.gov/
Data .Gov Ireland	A comprehensive resource for Ireland public sector data	https://data.gov.ie/
FiveThirtyEight	A collection of data-driven journalism datasets, often accompanied by context and analysis.	https://data.fivethirtyeight.com/
World Bank Open Data	Global development data on various indicators, useful for socio-economic research.	https://data.worldbank.org/

3. Search Using Keywords
   Use these keyword searches based on your chosen domain:
   a. Finance: Stock prices, cryptocurrency trends, financial statements.
   b. Healthcare: Patient health records, disease outbreak data, drug efficacy studies.
   c. Social Media: Twitter sentiment analysis, Facebook ad performance, Reddit comment analysis.
   d. Environment: Climate change data, air quality indexes, renewable energy statistics.

4. Evaluate Dataset Quality
   a. Completeness: Assess missing data and determine if it's manageable.
   b. Size: Ensure the dataset's size is feasible within the course's timeframe.
   c. Ethics: Confirm the data is legally and ethically sourced, with no sensitive personal information unless anonymized. If you are using any organization/company's data, please get appropriate permissions.

5. Dataset Overview and Characteristics
   a. Description: Provide a brief description of the selected dataset, including its purpose and scope.
   b. Volume and Variety: Ensure the dataset has sufficient data volume and variety to support meaningful analysis. (Minimum : 12 variables, 2000 rows)*
   c. Size and Structure: Detail the size and structure of the dataset, including the number of records and variables.
   d. Sources and Collection Methods: Identify the sources of the data and the methods used for its collection: Online platforms, Databases, APIs.
   e. Data Types: Specify the types of data included such as numerical, categorical, or textual.
   f. Limitations: Note any known limitations or challenges associated with the dataset that may impact your analysis.