towards
data science

31 Datasets For Your Next Data Science Project

A compilation of task-based datasets that you can use for building your next data science project

Yash Prakash 2 days ago · 6 min read★

Photo by Nathan McBride on Unsplash

Browsing through the datasets on Kaggle, I always find it hard to settle on a dataset to try out a new machine learning concept that I recently learned. Fresh datasets are posted everyday on these popular websites and the effort to find the right one for a new project does tend to quickly become overwhelming.

With this mind, and because I could not find a single proper task-based compiled list anywhere on the internet, I decided to make another one of these lists myself — a master list actually, one that I'll always come back to when I need to choose a dataset to

practise a newly acquired skill. And hopefully, it will be beneficial for you as well!

If you want to read the older version of this list that I published a few months back, you can read it here.

What this article is ABOUT and NOT ABOUT

Before we get into the list, I want to be clear on just this one thing. *Let this article be a reference for you to consider trying out a new dataset for an algorithm that you learn and want to experiment with*

However, please do not consider it to be an exhaustive list of datasets for machine learning tasks — it is not possible for me or anyone to look over *every* single dataset available on every website out there.

Okay, now that we've gotten that over with, let's get on with the list. 😀

The list of datasets

I will try to cover as many machine learning/deep learning-based tasks and related datasets for their applications as I can think of, and I will also update this article from as well (this is the second iteration of the datasets article). So consider bookmarking it for safekeeping as well, if you want.

I'm including the datasets I've been using throughout my data science journey — personal favourites included — and also some that I am hoping to try in the future.

General Classification Problems

1. Campus Recruitment — Determine if a student gets placed in a company based on various features like their education, grades, and so on.
2. Australian Fatal Road Accident 1989–2021 — This is a fairly new dataset, and you need to classify the crash type from the various features available about the crash such as time and day of the crash, speed of the vehicle, etc.
3. Heart Disease UCI — To predict the presence of heart disease in the patient based on a set of 76 different physiological attributes of an individual.
4. CelebFaces Attributes (CelebA) Dataset — A popular one to use over 200k images of celebrities and use Computer vision concepts for implementing facial recognition.

Regression Datasets

1. Boston House Prices — A classic dataset for flexing your Regression muscles, also recommended in the part 1 of my dataset master list.
2. Tesla dataset — A stock price dataset for all the Tesla fans, and for those who enjoy dabbling into the intricacies of the financial industry.
3. WHO Life Expectancy — Another good one for experimenting with your EDA skills also.
4. Red Wine Quality — A dataset to predict the quality of wines using wine attributes such as fixed acidity, chlorides, citrus content and so on. This is a fun dataset I'd recommend experimenting with if you're already familiar with a little bit of regression and have practised on the dataset 1 above.

Recommender Systems

1. Popular Movies from IMDb— A classic crowd-sourced movie information database for starting out, in which you need to predict which movie to recommend.
2. Goodreads Books — Detailed information about books through numerous columns for building a book recommender engine. This is my personal favourite for getting a hang out of actually attempting the recommendation task.
3. Netflix Data — collection of movies and TV shows details until 2019, also a great one for some practical exposure to a real world application.
4. Subreddit Recommender — This is one of my recent favourites, and with this dataset, you need to take into account each user's comments in subreddits and then predict some new subreddits to recommend to them. If you're sick of all the repetitive movie datasets, I would say to try this one for sure!

Time Series Analysis

1. E-Commerse Sales — For predicting sales/transaction for a store. The classic time series forecasting job.
2. House Property Sales — One of the classics I'll definitely recommend if you're just starting out with time series analysis.
3. Minimum Daily Temperatures — This dataset describes the minimum daily temperatures over 10 years in the city Melbourne, Australia.
4. Microsoft Stock — Another stock dataset for you to experiment with, this one wants you to predict Microsoft's stock prices based on five-six years of historical data.
5. Household Electric Power Consumption — Has measurements of electric power consumption in one household with a one-minute sampling rate over a period of 4 years.

Text Summarization Datasets

1. CNN-DailyMail News — a wonderful dataset to start with text summarization. You will need to summarize 300k unique news articles written by CNN journalists all over. You can do both extractive and abstractive summarization for them.
2. WikiHow Dataset — Another good one for summarizing wikipedia how-to articles that you might frequently find online.
3. Arxiv Dataset — Collection of arxiv research papers for building text generation, abstractive summarization, and question answering systems.

Text Question-Answering Systems

1. Covid-19 Open Research Challenge — With a great number of Covid research articles with full text, this is a great dataset to start out with text summarisation, semantic search and question answering systems. Quite a good one to put on your CV if you manage to build a web app around it as well.
2. Stanford Question-Answering Dataset — This is a classic one, based on a reading comprehensions with of questions posed by crowd-workers on a set of Wikipedia articles.

Other Large Scale Text Datasets for NLP tasks

1. Amazon Reviews — A classic dataset for sentiment analysis task. Overused a little,

yes, but undeniably a great and a classic one if you're starting out.

2. StackOverflow Tags and Questions — It has 60k questions from StackOverflow to predict tags for questions, as well as classify a question based on scores as high or low quality.

3. News Headlines — This is another classic dataset and you can use it build a model to detect sarcasm in news headlines, a binary classification task.

4. The WikiBooks Dataset — This one consists of the complete set of contents of all the Wikibooks in 12 languages. You can use it to create an interactive knowledge base in a notebook.

The Multi-Purpose Datasets — For trying out any big or small algorithm

1. Kaggle Titanic Survival Prediction Competition — A dataset for trying out all kinds of basic + advanced ML algorithms for binary classification, and also try performing extensive *feature engineering*.

2. Fashion MNIST — A dataset for performing multi-class image classification tasks based on different categories such as apparels, shoes, handbags, etc.

3. Credit Card Approval — A binary classification task for good or bad credit scores, if the people can be a risk for defaulting credit card loans.

4. Rock Paper Scissors — Image classification for those three classes.

5. Loan Prediction Based on Customer Behaviour — Binary classification of risk flag in approving a loan for a customer based on numerous financial attributes available on the customer profile.

A few last words

I have tried to include as much versatility as possible in recommending these datasets. I have also tried to keep the Kaggle dataset usability factor in mind when making this list, so that you have a good time reading about them, and hopefully choosing to use them in your projects.

That is it for now. I will be coming back to update this list again as I come across and try out more datasets.

. . .

Join Medium if you want to read more from me. It helps support my writing and means the world to me!

Happy learning! :)

Another relevant content you might find interesting:

Python Data Science Programming Machine Learning Deep Learning