JOURNEY OF ANALYTICS

Deep dive into data analysis tools, theory and projects

50+ free Datasets for Data Science Projects

/ ANU - JOURNEY OF ANALYTICS TEAM / COMMENTS OFF

[Updated as on Jan 31, 2020]

50+ free-datasets for your DataScience project portfolio

There is no doubt that having a project portfolio is one of the best ways to master Data Science whether you aspire to be a data analyst, machine learning expert or data visualization ninja! In fact, students and job seekers who showcase their skills with a unique portfolio find it easier to land lucrative jobs faster than their peers! (For project ideas, check this post, for job search advice look here.)

To create a custom portfolio, you need good data. So this post presents a list of Top 50 websites to gather datasets to use for your projects in R, Python, SAS, Tableau or other software. Best part, these datasets are all free, free, free! (Some might need you to create a login)

The datasets are divided into 5 broad categories as below:

1. Government & UN/ Global Organizations
2. Academic Websites
3. Kaggle & Data Science Websites
4. Curated Lists
5. Miscellaneous

Government and UN/World Bank websites:

- [1] US government database with 190k+ datasets – link . These include county-level data on demographics, education/schools and economic indicators; list of museums & recreational areas across the country, agriculture/ weather and soil data and so much more!
- [2] UK government database with 25k+ datasets . Similar to the US site, but from the UK government.
- [3] Canada government database. Data for Canada.

- [4] Center for Disease Control – link
- [5] Bureau of Labor Statistics – link
- [6] NASA datasets – link
- [7] World Bank Data – link
- [8] World Economic Forum. I like the white paper style reports on this site too. It teaches you on how to think what Qs to answer using different datasets as well as how to present results in a meaningful way! This is an important skill for senior data scientists, academics and analytics consultants, so take a look.
- [9] UN database with 34 sets and 60 million records – link . Data by country and region.
- [10] EU commission open data – link
- [11] NIST – link
- [12] U.S. National Epidemiological Survey on Alcohol and Related Conditions (NESARC) – dataset from survey to determine magnitude of alcohol use and psychiatric disorders in the U.S. population. The dataset and descriptive codebook are available here.
- [13] Plants Checklist from US Department of Agriculture – link .

Academic websites:

- [14] Yelp academic data – link
- [15] Univ of California, Irvine Machine Learning Repository – link

- [16] Harvard Univ: link
- [17] Harvard Dataverse database: link
- [18] MIT – link1 and link2
- [19] Univ of North Carolina, adolescent health – link
- [20] Mars Crater Study, a global database that includes over 300,000 Mars craters 1 km or larger. Link to Descriptive guide and dataset.
- [31] Click Dataset from Indiana University (~2.5TB dataset) – link .
- [32] Pew Research Data – Pew Research is an organization focused on research on topics of public interest. Their studies gauge trends in multiple areas such as internet, technology trends, global attitudes, religion and social/ demographic trends. Astonishingly, they not only publish these reports but also make all their datasets publicly available for download!
- [33] Million Song Dataset from Columbia University , including data related to the song tracks and their artist/ composers.

Kaggle & Datascience resources:

Few of my favorite datasets from Kaggle Website are listed here. Please note that Kaggle recently announced an Open Data platform, so you may see many new datasets there in the coming months.

- [34] Walmart recruiting at stores – link
- [35] Airbnb new user booking predictions – link
- [36] US dept of education scorecard – link
- [37] Titanic Survival Analysis – link
- [38] Edx – link
- [39] Enron email information and data – link
- [40]Quandl – an excellent source for stock data. This site has both FREE

and paid datasets.

- [41] Gapminder – link

Curated Lists:

curated-datasets

- [42] KDnuggets provides a great list of datasets from almost every field imaginable – space, music, books, etc. May repeat some datasets from the list above. link
- [43] Reddit datasets – Users have posted an eclectic mix of datasets about gun ownership, NYPD crime rates, college student study habits and caffeine concentrations in popular beverages.
- [44] Data Science Central has also curated many datasets for free – link
- [45] List of open datasets from DataFloq – link
- [46] Sammy Chen (@transwarpio) curated list of datasets. This list is categorized by topic, so definitely take a look.
- [47] DataWorld – This site also has a list of paid and FREE datasets. I

have not used the site, but heard good reviews regarding the community.

Others:

- [48] MRI brain scan images and data – link
- [49] Internet Usage Data from the Center for Applied Internet Data Analysis –link .
- [50] Google repository of digitized books and ngram viewer – link.
- [51] Database with geographical information – link
- [52] Yahoo offers some interesting datasets, the caveat being that you need to be affiliated with an accredited educational organization. (student or professor) – you can view the datasets here.
- [53] Google Public Data – Google has a search engine specifically for searching publicly available data. This is a good place to start as you can search a large amount of datasets in one place. Of course, there is a NEWER link that went live a couple days ago! 🙂
- [54] Public datasets from Amazon – see link.

Make sure you do attribute the datasets to the appropriate origin sites. Happy vizzing and coding! 🙂

Learning Resources, R programming

50 FREE DATASETS 50 PUBLIC DATASETS FOR ANALYTICS PROJECTS

DATASETS FOR DATASCIENCE PROJECTS

MACHINE LEARNING AND PREDICTIVE ANALYTICS DATA

PREVIOUS POST
Yelp College Search – Shiny Based App

NEXT POST
Twitter Sentiment Analysis

Comments are closed.

RECENT POSTS

How to Become a Data Scientist in 2020

50+ Free DataSets for DataScience Projects