

Class Resources

Vicki Hertzberg

January 22, 2017

Data Sources

CDC Wonder (<https://wonder.cdc.gov>)

NYC Open Data (<https://data.cityofnewyork.us/browse>)

KD Nuggest Datasets (<http://www.kdnuggets.com/datasets/index.html>) (Note, this site has many links to many other lists of data.)

Yelp Datasets (https://www.yelp.com/dataset_challenge)

World Health Organization data (<http://www.who.int/gho/en/>)

Reddit datasets (<https://www.reddit.com/r/datasets/>)

BigDataMadeSimple datasets (<http://bigdata-madesimple.com/70-websites-to-get-large-data-repositories-for-free/>)

StackExchange datasets (<http://opendata.stackexchange.com>)

Kevin Chai's list of datasets (<http://kevinchai.net/datasets>)

US Government datasets (<https://catalog.data.gov/dataset>)

US Census datasets (<http://www.census.gov/data.html>)

Open fMRI datasets (<https://openfmri.org/dataset/>)

World Bank data (<http://data.worldbank.org>)

Pew Research Center's Internet Project datasets (<http://www.pewinternet.org/datasets/>)

Robert Seaton's blog post about 100+ interesting datasets, with hot links (<http://rs.io/100-interesting-data-sets-for-statistics/>)

UF's Virtual Library of datasets (<http://www.stat.ufl.edu/vlib/statistics.html>)

Chance magazine datasets (https://www.dartmouth.edu/~chance/teaching_aids/data.html)

From: NASA Socioeconomic Data and Applications Center
(<http://sedac.ciesin.columbia.edu/data/sets/browse>)

From International Panel on Climate Change (<http://www.ipcc-data.org/guidelines/pages/approvedDatasetLinks.html>)

Data from Gapminder (<https://www.gapminder.org/data/>)

Environmental Organization WebDirectory (<http://www.webdirectory.com/Databases/>)

Analyzing Baseball Data with R (https://github.com/maxtoki/baseball_R)

Yahoo! datasets (<https://webscope.sandbox.yahoo.com>) (Note: requires and agreement with Yahoo!)

Last.fm Million Song dataset (<http://labrosa.ee.columbia.edu/millionsong/lastfm>)

Grouplens (<https://grouplens.org>) (See dataset pulldown)

mldata.org's machine learning datasets (<http://mldata.org>)

Kaggle datasets (<https://www.kaggle.com/datasets>)

HetRec datasets (<http://ir.ii.uam.es/hetrec2011//datasets.html>)

University of California at Irvine Machine Learning Repository (<http://archive.ics.uci.edu/ml/>) (Click on "View All Datasets")

Xiaming Chen's list of Awesome Public Datasets (<https://github.com/caesar0301/awesome-public-datasets>)

bigml's list of public datasets (<https://blog.bigml.com/list-of-public-data-sources-fit-for-machine-learning/>)

Airbnb data (<http://insideairbnb.com/get-the-data.html>)

NYC taxi ride data (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml)

R open gov (<http://ropengov.github.io>)

Open population datasets (<https://code.facebook.com/posts/596471193873876/open-population-datasets-and-open-challenges/>)

R Resources

Swirl (<http://swirlstats.com>) Learn R, in R!

Introductory R tutorial (<http://www.cyclismo.org/tutorial/R/>)

DataCamp's Intro to R course (<https://www.datacamp.com/courses/free-introduction-to-r>) **FREE!**

Try R from Code School (<http://tryr.codeschool.com/levels/1/challenges/2>) **Also free!**

R Tutorial (<http://www.r-tutor.com/r-introduction>)

R Tutorial from tutorialspoint (<https://www.tutorialspoint.com/r/>)

Note, you can also access R tutorials on lynda.com (<https://www.lynda.com>), which you can access for free from anywhere on the Emory campus.

Data journalism with R (<https://rddj.info>)

RStudio cheat sheets (<https://www.rstudio.com/resources/cheatsheets/>)

R data import/export (<https://cran.r-project.org/doc/manuals/r-release/R-data.html>)

R for Data Science (<http://r4ds.had.co.nz>)