Public Datasets

This is a list of datasets and data collections available online. Resources in this list are publicly available and do not require memberships or subscriptions to gain access. For this reason, Kaggle datasets were kept off the list.

- The first section points to generalized dataset resources
- The second section lists organizational websites and collections of data.
- The third lists specific datasets you can download for various purposes.
- The fourth section contains datasets used specifically within this course.

STARTING POINTS

Google Public Data

https://www.google.com/publicdata/directory

Quora Public Data

https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public

FlowingData: Resources to Find the Data You Need, 2016 Edition

https://flowingdata.com/2016/11/10/find-the-data-you-need-2016-edition/
It can be frustrating to sleuth for the data you need, so here are some tips on finding it (the openly available variety) and some topic-specific resources to begin your travels.

Duke University Statistics Lab

http://stat.duke.edu/resources/datasets and http://www2.stat.duke.edu/courses/Fall03/sta290/datasets.html

DATA COLLECTIONS

Government Repositories:

National Centers for Environmental Information (NCEI)

https://www.ncei.noaa.gov/access
https://www.ncdc.noaa.gov/data-access
Formerly the National Climatic Data Center (NCDC), NOAA's National Centers for Environmental Information (NCEI) hosts and provides public access to one of the most significant archives for environmental data on Earth. Through the Center for Weather and Climate and the Center for Coasts, Oceans, and Geophysics, we provide over 25 petabytes of comprehensive atmospheric, coastal, oceanic, and geophysical data.

Educational Repositories:

UC Irvine Machine Learning Repository

http://mlr.cs.umass.edu/ml/
We currently maintain 22 data sets as a service to the machine learning community.

Data Journalism Sites:

FiveThirtyEight

http://fivethirtyeight.com/
News outlet created by Nate Silver, datasets available from github

FlowingData

http://flowingdata.com/category/statistics/data-sources/
Have fun and play with some numbers.

Application Built-ins:

Plotly Datasets

https://github.com/plotly/datasets

The R Datasets Package

http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html

Seaborn Datasets

https://github.com/mwaskom/seaborn-data
Data repository for seaborn examples. *This is not a general-purpose data archive*.
This repository exists only to provide a convenient target for the seaborn.load_dataset function to download sample datasets from. Its existence makes it easy to document seaborn without confusing things by spending time loading and munging data. However, the datasets may change or be removed at any time if they are no longer useful for the seaborn documentation.

SPECIFIC DATASETS

Chicago Data Portal - Motor Vehicle Theft

https://data.cityofchicago.org/Public-Safety/motor-vehicle-theft/7ac4-d9tk
This dataset reflects reported incidents of motor vehicle theft that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified.
9 fields, 307644 records (and counting)

CO2 PPM - Trends in Atmospheric Carbon Dioxide

http://datahub.io/core/co2-ppm
Data are sourced from the US Government's Earth System Research Laboratory, Global Monitoring Division. Two main series are provided: the Mauna Loa series (which has the longest continuous series since 1958) and a Global Average series (a global average over marine surface sites).

Federal Housing Finance Agency - FHFA House Price Indexes (HPIs)

https://catalog.data.gov/dataset/fhfa-house-price-indexes-hpis
The FHFA House Price Index (HPI) is a broad measure of the movement of single-family house prices. The HPI is a weighted, repeat-sales index, meaning that it measures average price changes in repeat sales or refinancings on the same properties. This information is obtained by reviewing repeat mortgage transactions on single-family properties whose mortgages have been purchased or securitized by Fannie Mae or Freddie Mac since January 1975.

The Reddit Button

A large (if frivolous) set of time series data derived from a Reddit social experiment in 2015. On April 1, Reddit posted a simple button with a 60-second timer that counted down to zero. Every time the button was pressed by a unique Reddit user, the timer reset to 60 seconds. After more than two months and 1 million presses, the timer finally made it to zero seconds without a press.
Discussion: https://redditblog.com/2015/06/08/the-button-has-ended/
Data: https://github.com/reddit/thebutton-data/blob/master/thebutton_presses.csv
4 fields, 1008316 records, 44MB

DATASETS USED IN THIS COURSE

U.S. Census Bureau

https://www.census.gov/data/datasets/2017/demo/popest/nation-total.html#ds
https://www2.census.gov/programs-surveys/popest/datasets/2010-2017/national/totals/nst-est2017-alldata.csv
National, State, and Puerto Rico Commonwealth Totals Datasets: Population, population change, and estimated components of population change: April 1, 2010 to July 1, 2017
121 fields, 57 records

2018 Winter Olympics in PyeongChang, South Korea

http://time.com/5143796/winter-olympic-medals-by-country-2018/
https://www.pyeongchang2018.com/en/game-time/results/OWG2018/en/general/medal-standings.htm
Medal standings by competing country in the 2018 Winter Olympics
6 fields, 30 records

mpg.csv

https://gist.github.com/omarish/5687264
Miles per gallon and other statistics for automobiles manufactured from 1970-1982.
9 fields, 399 records

Mark Twain and the Quintus Curtius Snodgrass Letters

Adapted from https://www.math.utah.edu/~treiberg/M3074TwainEg.pdf citing data from Brinegar, C. S. (1963) Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship. *Journal of the American Statistical Association*. 58, 85-96
2 fields, 18 records

Fremont Bridge Bicycle Traffic

https://data.seattle.gov/Transportation/Grouped-by-Hour/7mre-hcut
The Fremont Bridge Bicycle Counter records the number of bikes that cross the bridge using the pedestrian/bicycle pathways. Inductive loops on the east and west pathways count the passing of bicycles regardless of travel direction. The data consists of a date/time field: Date, east pathway count field: Fremont Bridge NB, and west pathway count field: Fremont Bridge SB. The count fields represent the total bicycles detected during the specified one hour period. Direction of travel is not specified, but in general most traffic in the Fremont Bridge NB field is travelling northbound and most traffic in the Fremont Bridge SB field is travelling southbound.
3 fields, 47400 records (and counting)

National Oceanographic and Atmospheric Administration (NOAA), U.S. Climate Reference Network (USCRN)

Site: https://www.ncdc.noaa.gov/crn/qcdatasets.html
Datasets: https://www1.ncdc.noaa.gov/pub/data/uscrn/products/hourly02/2010/
The USCRN is a systematic and sustained network of climate monitoring stations with sites across the conterminous U.S., Alaska, and Hawaii. These stations use high-quality instruments to measure temperature, precipitation, wind speed, soil conditions, and more.

Iris Dataset

https://archive.ics.uci.edu/ml/machine-learning-databases/iris/
The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by Sir Ronald Fisher in the 1936 as an example of discriminant analysis. The set consists of 50 samples from each of three species of Iris (*Iris setosa*, *Iris virginica* and *Iris versicolor*), for a total of 150 samples. Four features were measured from each sample: the length and the width of the sepals and petals, in cm.
5 fields, 150 records

Abalone Dataset

http://mlr.cs.umass.edu/ml/datasets/Abalone
Used in machine learning, the Abalone dataset provides 8 features that can be used to predict the age of abalone. The age is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age.
We've added the following column names:
'sex', 'length', 'diameter', 'height', 'whole_weight', 'shucked_weight', 'viscera_weight', 'shell_weight', 'rings'
9 fields, 4177 records

Old Faithful Geyser Eruptions

Old Faithful is a cone-type geyser. Since 2000 its intervals have varied from 44 to 125 minutes, with an average of about 90-92 minutes, its duration is 1 1/2 to 5 minutes and its height is 90 to 184 feet.
It is not possible to predict more than one eruption in advance. Old Faithful is currently bimodal. It has two eruption durations, either a long (over 4 minutes) or more rarely a short (about 2-1/2 minutes). Short eruptions lead to an interval of just over an hour and long eruptions lead to an interval of about 1-1/2 hours.
Raw data (electronic recordings of eruptions, and tabulated visitor logs of eruption durations) can be found at http://www.geyserstudy.org/geyser.aspx?pGeyserNo=OLDFAITHFUL and
http://www.geyserstudy.org/ofvclogs.aspx

The dataset used for this exercise comes from the Duke University Statistics Lab at
http://www2.stat.duke.edu/courses/Fall03/sta290/datasets.html
It's comprised of 3 fields (D,Y,X) where
D = date of recordings in month (in August),
X = duration of the current eruption in minutes (to nearest 0.1 minute),
Y = waiting time until the next eruption in minutes (to nearest minute).
The first 107 records were recorded August 1-8, 1978 and the next 115 records were recorded a year later, August 16-23, 1979.

Seaborn Flights Data

Options include downloading the set from https://github.com/mwaskom/seaborn-data or importing directly from the Seaborn module with

```
import seaborn as sns
df = sns.load_dataset("flights")
```

3 fields, 144 records

Dash gapminderDataFiveYear

Site: https://github.com/plotly/datasets
Dataset: https://raw.githubusercontent.com/plotly/datasets/master/gapminderDataFiveYear.csv
This dataset is used in Dash online tutorials.
It's comprised of 6 fields:

country	
year	
pop	- Population
continent	- [Asia, Europe, Africa, Americas, Oceania]
lifeExp	- Life Expectancy
gdpPercap	- Gross Domestic Product Per Capita

and contains 1704 records.

NASDAQ Companies and Stock Symbols

Dataset:
http://www.nasdaq.com/screening/companies-by-industry.aspx?exchange=NASDAQ&render=download
Selection of companies listed on the NASDAQ stock exchange. 3299 original records whittled down to a handpicked 256 companies based on market capitalization.
Symbol, Name, LastSale, MarketCap, IPOyear, Sector, Industry (data as of 4/6/2018)
7 fields, 256 records

Arrhythmia

Source:
https://archive.ics.uci.edu/ml/datasets/arrhythmia
420 records culled from the arrhythmia dataset. For our histogram example we only took columns for Age, Sex (0=Male, 1=Female) and Height(in centimeters). We omit anyone younger than 20 and who weighs less than 30kg.
3 fields, 420 records

OPTIONAL OTHERS:

Dash - Chris Parmer's indicators.csv

Dataset:
https://gist.githubusercontent.com/chriddyp/cb5392c35661370d95f300086accea51/raw/8e0768211f6b747c0db42a9ce9a0937dafcbd8b2/indicators.csv
5 fields [(unnamed index), Country Name, Indicator Name, Year, Value], 85,536 records, 36,616 values