DSE 230 - Datasets for Project

You can use one of the datasets listed below for your project. You can also find a dataset on your own to use. If you will be using your own dataset, choose one that is (1) publicly available and (2) has at least 100K samples.

Housing

- Available on Canvas
- This dataset contains housing information. Data includes number of bedrooms, square footage, location. Data for 2016 and 2017 are provided in separate files.
- Each file has 2,985,217 samples

e-Commerce

- Available on Canvas
- This dataset contains e-commerce orders for a company in Brazil. Data includes information on customers, orders, products, and sellers.
- 100,000 samples

NOAA World Ocean

- https://registry.opendata.aws/noaa-wod/
- This dataset provides historical subsurface ocean data, including temperature, salinity, oxygen, nutrients, and others.

US Climate Data

- This dataset contains three-decade averages of climatological variables, including temperature and precipitation from 1981 - 2010.
- https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/climate-normals/1981-2010-normals-data

New York Taxi Data

- https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- This dataset contains NYC taxi data, including pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

Smartphone and Smartwatch Activity and Biometrics Data

- https://archive.ics.uci.edu/ml/datasets/WISDM+Smartphone+and+Smartwatch+Activity+and+Biometrics+Dataset+
- This dataset contains accelerometer and gyroscope time-series sensor data collected from a smartphone and smartwatch as 51 test subjects perform 18 activities for 3 minutes each.
- 15,630,426 samples

Individual household electric power consumption data

- https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption
- This dataset contains measurements of electric power consumption in one household with a one-minute sampling rate over a period of almost 4 years. Different electrical quantities and some sub-metering values are available.
- 2,075,259 samples

Additional resources with datasets:

- UCSD Library datasets
  - https://ucsd.libguides.com/data-statistics/finddata
- KDnuggets: Public datasets on github
  - https://www.kdnuggets.com/2015/04/awesome-public-datasets-github.html
- Data is Plural
  - https://www.data-is-plural.com/
- Amazon reviews
  - http://jmcauley.ucsd.edu/data/amazon/
- Awesome Public Datasets - Machine Learning
  - https://github.com/awesomedata/awesome-public-datasets#machinelearning
  - ○ Yahoo! Music User Ratings (423 MB)
  - ○ Restaurants Health Score data (54k rows)
- Registry of Open Data on AWS
  - https://registry.opendata.aws/
  - NEXRAD
  - IRS 990 filings
  - NOAA water column sonar data archive
- NOAA Climate Data
  - https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets