This is a set of data sources curated by the instructional staff. Feel free to suggest new data sources in the forums. The initial list was provided by Kevyn Collins-Thomson from the University of Michigan School of Information.

* [Long general-purpose list of datasets](https://vincentarelbundock.github.io/Rdatasets/datasets.html)
* [The Academic Torrents site](http://academictorrents.com/browse.php?cat=6) has a growing number of datasets, including a few text collections that might be of interest (Wikipedia, email, twitter, academic, etc.) for current or future projects.
* [Google Books n-gram corpus](http://books.google.com/ngrams)
  + [N-grams: COCA corpus (1 billion words)](https://www.ngrams.info/)
  + [Google Books Ngram Datasets](http://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html)
* [Common Crawl: Currently 6 billion Web documents (81 Tb)](https://commoncrawl.atlassian.net/wiki/display/CRWL/About+the+Data+Set)
  + [Amazon S3 Public Data Set](http://aws.amazon.com/datasets/41740)
  + [About the Common Crawl Data Set](https://commoncrawl.atlassian.net/wiki/display/CRWL/About+the+Data+Set)
  + [Award project using Common Crawl](http://norvigaward.github.io/entries.html)
  + [Python example](http://www.freelancer.com/projects/Python-Data-Processing/Python-script-for-CommonCrawl.html)
* [Internet Archive](http://archive.org/help/json.php) (huge, ever-growing archive of the Web going back to 1990s)
* [WikiData](https://www.wikidata.org/wiki/Wikidata:Main_Page)
* [World Food Facts](http://world.openfoodfacts.org/data)
* [Data USA](https://datausa.io/) - a variety of census data
* [Center for Disease Control](https://data.cdc.gov/browse) - variety of data sets related to COVID
* [U.S. Government open data](https://data.gov/) - datasets from 75 agencies and subagencies
* [NASA data portal](https://data.nas.nasa.gov/) - space and earth science