# Dataset

* **UCI's Machine Learning Repository** - UC Irvine maintains a fantastic collection of datasets for machine learning, tagged by machine learning task (among other things). This is a great place to start.
* **Kaggle** - Online collection of datasets.  Kaggle also has competitions for data mining and information about jobs in data science.
* **KDnuggets Dataset** - Another collection of datasets - most of the datasets are free. Within KDnuggets, there are links for government data,  Data APIs,  and Data Mining Competitions.
* **US Government Data**, **UK Government Data**, **Canada's Open Data Exchange**, **World Health Organization**, and the **World Bank**.

**Semester Project**

**PROJECT SPECIFICATIONS:**

* Pick a data set that you and your group find interesting. (Example source: [UC Irvine Machine Learning Repository](http://archive.ics.uci.edu/ml/index.php). Feel free to select your data from any other source as appropriate.)
* Perform data pre-processing, data cleaning, outlier removal, and so on to sanitize your data, if necessary.
* Save your data in a .csv file (or other format as appropriate for your data set and project scenario).
* Read in data to your program from the .csv file.
* (*Optional* - do as appropriate) Process the data or perform any calculations or statistics on it before storing the data into a data frame (see next step).
* Save the data into one or more data frames (or other structures as appropriate) (*review lecture notes/examples*).
* Once you have stored your data, query your data to reveal interesting/useful information based on your project scenario.
  + Query your data using at least 4 different queries.
* Capture the results of the queries appropriately (either write results to a file, or store into another data structure, or do something else with the results as appropriate based on your project scenario).
* (*Do as appropriate*) Process the results, or submit additional queries to the obtained results (if results were saved to a file or another data structure).
* Display final results in a presentable way (use tables and/or visualizations). You will present this information during the Module 12 Live Session.
* Perform adequate testing (TDD and/or unit testing). Submit this in a separate .py file.
* Submit a write-up along with your project (see details in the next section below), in your write-up be sure to display the results and describe what you have learned, as well as how these results can be used by others (mention the relevance/significance of the results you've obtained).
* **Group size:** Groups of 3 or 4. No individuals or pairs.