Here are some resources for finding data:

- Kaggle: `https://www.kaggle.com/datasets`
  See `https://github.com/Kaggle/kaggle-api` for downloading data.
- Machine learning datasets: `https://www.datasetlist.com`
- World Bank Data Catalog `https://datacatalog.worldbank.org`
- U.S. Government's Open Data `http://data.gov`
- HealthData.gov `https://healthdata.gov`
- Open Data Network `https://www.opendatanetwork.com`
- United States Census `https://www.census.gov/data.html`
- Project Gutenberg (thousands of free eBooks) `http://www.gutenberg.org/wiki/Gutenberg:The_CD_and_DVD_Project`
- American National Election Studies `http://electionstudies.org/data-center`
- Purdue datasets list `http://llc.stat.purdue.edu/2018/29000/datasets.html`
- Wikipedia `https://en.wikipedia.org/wiki/Wikipedia:Database_download`
- Forbes list of free data sources `https://www.forbes.com/sites/bernardmarr/2018/02/26/big-data-and-ai-30-amazing-and-free-public-data-sources-for-2018/#4936a3975f8a`

Here are many more links to data sets from Sebastian Raschka's course:

- Datasetlist.com – Datasets by domain
  `https://www.datasetlist.com`
  - Lets you sort datasts by category (image, NLP, audio, etc.)
  - Sortable by license and year
  - Links to original papers (if applicable) are provided
- Awesome Public Datasets – Large GitHub README list by application domain
  `https://github.com/awesomedata/awesome-public-datasets`
  - Links to approximately 650 datasets
  - Organized by application domain (e.g., Agriculture, Biology, etc.)
- Papers with Code – Datasets with benchmarks
  `https://www.paperswithcode.com/datasets`

  - 3,095 machine learning datasets and links to original paper if applicable
  - Contains number of papers that used the dataset
  - Compiles benchmark information and links to the benchmark sources

- Google Dataset Search – A search engine for datasets
  `https://datasetsearch.research.google.com`
  - Let's you search datasets by name or description
  - Returns summary results and links to various source where a dataset can be obtained

- r/datasets – A subreddit for sharing and discussing datasets
  `https://www.reddit.com/r/datasets/`
  - A site where you can discover datasets and/or engage in discussion around a dataset

- Data is Plural – A dataset newsletter
  `https://tinyletter.com/data-is-plural`
  - A weekly newsletter that compiles a list of interesting new datasets each week

- Jupyter Tutorial Data – List of dataset repositories
  `https://jupyter-tutorial.readthedocs.io/en/latest/data/index.html`
  - A list linking the most common dataset repositories and search engines

- Huggingface Datasets – A Python library for loading NLP datasets
  `https://github.com/huggingface/datasets`
  - A tool that makes NLP datasets directly available in Python

- Roboflow Public Datasets – Datasets for computer vision
  `https://public.roboflow.com`
  - A list of publicly available computer vision datasets
  - Categories include classification and object detection

- Public APIs – A list of public dataset API
  `https://github.com/public-apis/public-apis`
  - A list of approximately 650 dataset APIs

- VisualDataDiscovery
  `https://www.visualdata.io/discovery`

  - A collection of more than 500 computer vision datasets
  - Can be filtered by license and code/model availability

- Kaggle Datasets `https://www.kaggle.com/datasets`
  - A search engine for datasets available through Kaggle
  - Datasets can be discovered by search terms, category tags, and file types

- Bob Fisher's Compilation of Computer Vision Datasets:

  `http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm`
  - An impressive collection of more than 1000 datasets for computer vision sorted by category (agriculture, general images, etc.)

- Penn Machine Learning Benchmarks

  `https://github.com/EpistasisLab/pmlb/tree/master/datasets`
  - A collection of preprocessed datasets in tabular form
  - More appropriate for traditional machine learning rather than deep learning

- UCI Machine Learning Repository

  `https://archive.ics.uci.edu/ml/index.php`
  - The classic repository for machine learning datasets taht can be searched by task (classification, regression etc.), application area, data type, and size
  - Most datasets in this data base are more suitable for traditional machine learning rather than deep learning