Common Dataset Sources

- UCI Machine Learning Repository (http://mlr.cs.umass.edu/ml/)
  - One of the oldest data source
- Kaggle (https://www.kaggle.com/)
  - One of the site that every data scientist must visit
- New York Open Data (https://opendata.cityofnewyork.us/)
  - Real dataset, good source for geographical related data
- Amazon AWS Open Data (https://aws.amazon.com/opendata/)
- Open Data for Nonprofit Research (https://lecy.github.io/Open-Data-for-Nonprofit-Research/)

Common Dataset

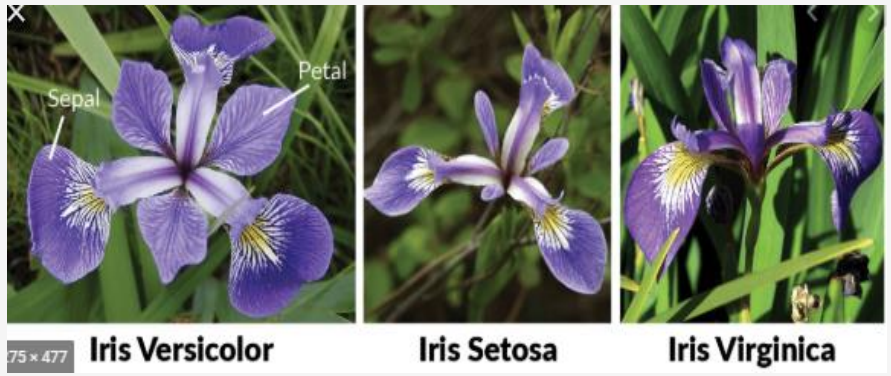
Datasets throughout the course

- Iris (https://archive.ics.uci.edu/ml/datasets/iris)
- US Housing data (https://www.kaggle.com/aariyan101/usa-housingcsv/version/1)
- Boston Housing data (https://archive.ics.uci.edu/ml/machine-learning-databases/housing/) or from scikit-learn
- https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
- https://www.kaggle.com/c/boston-housing
- Titanic (https://www.kaggle.com/c/titanic/data)
- Adult Income Data from Census https://archive.ics.uci.edu/ml/datasets/adult
- UCI Air Quality (https://archive.ics.uci.edu/ml/datasets/Air+quality)

Common Dataset

Many common dataset can be loaded directly from the Seaborn library package

- Import seaborn as sns
- Iris = sns.load_datasets('iris')
- mpg
- Tips
- Titanic



Common Dataset

Learning by doing