# Lecture Datasets

Datasets of the MADS-Modules
MMS, ML, and DL

Stephan Doerfel

September 22, 2024

This document presents a list of datasets that are used in the MADS modules MMS, ML, and DL. Before you start a portfolio exam in any of the above modules, please carefully compare your selected dataset to the ones on this list.

*The dataset for your portfolio exam **must not** be*
*one of the datasets on this list*
*nor a derivative of such a dataset!*

## 1 The List of Datasets

### 1.1 Iris

Datasets of Iris plants. We are going to use two versions of the dataset:

1. A dataset containing only the measurements for clustering. This dataset will be uploaded to Moodle in the MMS module.
2. The same dataset additionally containing the actual species of each measured plant. This dataset can be found for example
   - at the UCI Machine Learning Repository,
   - in the python packages sklearn or seaborn.

### 1.2 Decathlon Athletes of Men's Decathlon at the 2019 World Athletics Championships

This is a very small dataset used to demonstrate PCA.

1. The data holds the results of each finishing participant of the men's decathlon as well as positions in the final ranking.
2. The data can be found on Wikipedia.

## 1.3 Wine Recognition

A dataset with chemical properties of different wines.

1. These wines are grown by three different cultivars.
2. We use the data in clustering experiments and try to determine different types of wines (not necessarily grouped by who cultivated them).
3. The data can be found in the UCI Machine Learning Repository, or
4. be downloaded directly in sklearn.

## 1.4 Yellow Trip Taxi Data

In the lecture, we investigate one subset of the overall available data.

1. NYC Taxi & Limousine Commission publishes monthly trenches of taxi trips.
2. Trenches can be downloaded from NYC, or
3. some of them – already as CSV – on Kaggle.

## 1.5 Stockprices

This is a dataset that holds daily stock market prices for several companies.

1. The dataset collects stock values for serveral days during 2020 and 2021.

## 1.6 Digits

A dataset for Optical Recognition of Handwritten Digits. The dataset can be found

- at the UCI Machine Learning Repository,
- in the python package sklearn.

## 1.7 Play Tennis

The play tennis dataset is a toy dataset containing weather conditions and the decision to play or not to play tennis. It can be found

- in the lecture slides of the ML module,
- in textbooks,
- at Kaggle

## 1.8 20 News Groups

A dataset containing posts to 20 different Usenet news groups.

1. A famous dataset with unstructured data. The goal is to determine the news group that a post belongs to, i.e. topic classification.
2. The data can be directly downloaded using scikit-learn
3. It is also described and offered at multiple other sources in different versions!

## 1.9 Glassdoor – Wages

Wages of employees of a hypothetical employer.

1. Glassdoor uses the data to demonstrate the analysis of gender paygaps[1] within a company.
2. The data is offered by Glassdoor
3. or available on Kaggle.

## 1.10 Mid-Atlantic Wage Data

Wages and employee features of employees in the mid atlantic region.

1. The data is part of the ISLP libraray
2. and has been extracted from a population survey.

## 1.11 Boston Housing Data

House prices based on various features.

1. The dataset was used to investigate the influence of air quality on housing prices.
2. The dataset can be found in the repository of Carnegie Mellon University
3. or on Kaggle

## 1.12 Health Insurance Charges

A simulated dataset based on real world data.

1. The dataset was created for the book "Machine Learning with R" by Brett lantz, published by Packt Publishing.
2. Available through GitHub.

[1] Warning: the demo analysis contains methodical errors (using least squares linear regression on logarithmized data)!

## 1.13 AirBnB

A dataset that was created through the Inside AirBnB initiative.

1. Available on Kaggle.

## 1.14 MNIST Dataset

MNIST is short for Modified National Institute of Standards and Technology.

1. Available as demo dataset in TorchVision.
2. Details can be found on Wikipedia.

## 1.15 Air Traffic

Data on traffic of planes in the US.

1. Published by the Bureau of Transportation Statistics of the United States Department of Transportation.
2. Found in yearly trenches under the key "T-100 Domestic Segment (All Carriers)".