Data Science

Homework in preparation of

Lecture 4 – Missing Values

Prof. Dr. Markus Breunig

Photo by Pascal van de Vendel on Unsplash

Pandas (Part 1)

In the upcoming lecture we will be discussing the problem of *Missing Values*. In pretty much all real-world datasets contain some values are missing for various reasons, which can have a negative impact on the data science project if not handled carefully.

We will be using the pandas package to analyse and handle missing values. In this homework, you will start familiarizing yourself with pandas (the following homework will cover more advanced pandas subjects).

Pandas is a package built on top of NumPy, and provides an efficient implementation of a *DataFrame*. DataFrames are 2-dimensional tables with row and column labels, and often with heterogeneous types and/or missing data, very similar to tables in relational databases. As well as offering a convenient storage interface for labelled data, Pandas implements a number of powerful data operations familiar to users of both relational databases and spreadsheet programs. And it also has some basic functionality to handle missing values.

We will be using the "Python Data Science Handbook" by Jake VanderPlas to learn about pandas. The full text of this book (written as Jupyter Notebooks) is available online at [https://jakevdp.github.io/PythonDataScienceHandbook/] (note that we are using the first edition, which is quite old, a second edition has been announced but is not available yet).

This homework assignment consists of the following parts and expected time:

- Read Chapter 3 (Data Manipulation with Pandas) up to and including *Hierarchical Indexing* – (expected time: **60min**)
- Complete the notebook ”h04a...” - (expected time: **90min**)
  - Note that this notebook may look short, but will take some time – take a break between each of the 4 sections!