Introduction

An overview of data analysis with pandas.

In the **Data Processing** section, you will be using pandas to analyze Major League Baseball (MLB) data. The data comes courtesy of Sean Lahman, and contains statistics for every player, manager, and team in MLB history. The full database can be found and downloaded here.

A. Data analysis

Before doing any task with a dataset, it is a good idea to perform preliminary data analysis. Data analysis allows us to understand the dataset, find potential outlier values, and figure out which features of the dataset are most important to our application.

B. pandas

Since most machine learning frameworks (e.g. TensorFlow) are built on Python, it is beneficial to use a Python-based data analysis toolkit like pandas. pandas (all lowercase) is an excellent tool for processing and analyzing real world data, with utilities ranging from parsing multiple file formats to converting an entire data table into a NumPy matrix array.

In the following chapters we'll dive into the main data analysis functionalities of pandas. For a complete overview of the pandas toolkit, you can visit the official pandas website.

C. Matplotlib and pyplot

An essential part of data analysis is creating charts and plots to visualize the data. Similar to the saying, "a picture is worth a thousand words", data visualization can convey key data trends and correlations through a single figure.

The library we will use for data visualization in Python is Matplotlib.

Specifically, we'll be using the pyplot API of Matplotlib, which provides a variety of plotting tools from simple line plots to advanced visuals like heatmaps and 3-D plots. While we will only touch on the basic necessities for our data analysis (e.g. line plots, boxplots, etc.), a full overview of Matplotlib can be found at the official website.