Hochschule Bochum
Bochum University
of Applied Sciences

Fachbereich Wirtschaft
Prof. Dr. Christian Bockermann

# Data Science

Sommersemester 2022

## Übungsblatt 3

Dieses Übungsblatt beschäftigt sich mit dem Einlesen, dem Filtern und der Exploration von Daten. Die Daten liegen als CSV-Dateien in Ihrem Verzeichnis auf dem Notebook-Server

`https://datascience.hs-bochum.de/`

Die Dateien finden Sie im Verzeichnis `Kurse/DataScience1/data/`. Wenn Sie ihr Notebook im Hauptverzeichnis anlegen, müssen Sie den Dateinamen mit Pfad angeben, also: `Kurse/DataScience1/data/iris.csv`

### **Aufgabe 1** (Daten Einlesen)

Erstellen Sie ein neues Notebook und lesen Sie die Datei `iris.csv` ein. Denken Sie daran, dass Sie zunächst das Pandas Modul importieren müssen!

1. Welche Größe hat der Datensatz (Zeilen/Spalten?)
2. Geben Sie die Liste der Spaltennamen aus!
3. Welchen Datentyp haben die einzelnen Spalten?
4. Welchen Mittelwert/Standardabweichung hat die Spalte `petal_length`?

### **Aufgabe 2** (Daten Filtern)

Mit dem Iris-Datensatz aus der Aufgabe 1 geht es jetzt hier weiter. Der Datensatz enthält die Bezeichnung der Pflanze in der Spalte `species`.

1. Extrahieren Sie aus dem Datensatz die Menge der unterschiedlichen Planzenarten!

   **Hinweis**: Aus einer Liste können Sie mit `set(..)` eine *Menge* machen, die dann jedes Element der Liste nur einmal enthält.
   Schauen Sie sich dazu auch das `.values` Attribute von Series an (siehe Folie 14).

2. Wählen Sie aus dem Datensatz die ersten 50 Zeilen aus. Welche Pflanzenarten sind in diesen ersten 50 Zeilen enthalten?
   Wiederholen Sie das mit den ersten 100 Zeilen.

3. Welchen Mittelwert/Standardabweichung haben die Merkmale `petal_length` und `petal_width` für die Klasse *Iris Versicolor*?

4. Welche Pflanzenarten haben eine Kelchblattlänge (`sepal_length`) größer als 6?

5. Erstellen Sie einen Datensatz `zweiArten`, der nur die Daten für die Pflanzenarten *Iris Setosa* und *Iris Versicolor* enthält.