Hochschule Bochum
Bochum University
of Applied Sciences

Fachbereich Wirtschaft
Prof. Dr. Christian Bockermann

Data Science

Wintersemester 2020/2021

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 `Vorlesung/data/`. Wenn Sie ihr Notebook im Hauptverzeichnis anlegen, müssen Sie den Dateinamen mit Pfad angeben, also: `Vorlesung/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 Planzenarten 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 Plfanzenarten *Iris Setosa* und *Iris Versicolor* enthält.

Aufgabe 3 * (Daten Auswählen / Transformieren / Gruppieren)

Das RKI veröffentlicht jeden Tag die aktuellen Corona-Fallzahlen über seine Homepage. Die Daten sind z.B. als CSV Datei verfügbar. Im Verzeichnis **`Vorlesung/data/`** finden Sie die Datei **`rki-covid19-2020-10-22.csv`**, die die veröffentlichten Daten vom 22. Oktober enthalten.
Die darin enthaltenen Daten umfassen unter anderem die Spalten

`Bundesland`
`Landkreis`
`Meldedatum`
`Altersgruppe`
`Geschlecht`
`AnzahlFall`
`AnzahlTodesfall`
`AnzahlGenesen`
...

Bei Interesse finden Sie Hintergrundinformationen zu dem Datensatz und den darin enthaltenen Informationen auf der Seite:

https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0

Dort finden Sie unter anderem eine genauere Beschreibung der Bedeutung der Attribute (Merkmale) des Datensatzes. Derartige Beschreibungen zu Lesen und zu Verstehen ist natürlich Bestandteil des *Data Understanding* (vgl. CRISP-DM Modell).

1. Laden Sie den Datensatz in ein DataFrame Objekt. Welchen Datentyp haben die jeweiligen Attribute?

2. Berechnen Sie eine Liste der Attribute, die mit **`datum`** enden. Mit der Funktion **`pd.to_datetime(s)`** können Sie aus einer Series **`s`** mit z.B. einem **`str`** Type eine Series mit einem Datumstyp machen.

   Benutzen Sie die Funktion **`pd.to_datetime(..)`** um alle Spalten, die mit **`datum`** enden, zu Datumsspalten zu konvertieren.

3. Ermitteln Sie die Menge und Anzahl der Landkreise, die in dem Datensatz enthalten sind. Passt auch die Anzahl der Bundesländer zu der von Ihnen erwarteten Anzahl?

4. Mit der Methode **`groupby(spalte)`** eines DataFrames werden die Zeilen nach den Werten der angegeben Spalte gruppiert. Auf dem Resultat kann dann mit der Funktion **`sum()`** die Summe gebildet werden.

   Beispiel:

   ```
   gruppiert = df.groupby("Bundesland").sum()
   ```

   Welchen Typ hat das Ergebnis **`gruppiert`**? Welche Form (**`shape`**)?

5. Gruppieren Sie die Daten nach Altersgruppe. Welche Altersgruppe verzeichnet aktuell die meisten Fälle?