Hochschule Bochum
Bochum University
of Applied Sciences

Fachbereich Wirtschaft
Prof. Dr. Christian Bockermann

# Data Science 1

Sommersemester 2025

## Hausarbeit

Die Prüfungsleistung zum Modul *Data Science 1* findet als Hausarbeit statt. Die Aufgabenstellung zur Hausarbeit finden Sie in diesem Dokument.

Für die Bearbeitung der Aufgabenstellung und die Erstellung Ihrer Hausarbeit steht wieder der Jupyter-Notebook Server zu Verfügung. Die Abgabe der Hausarbeit erfolgt dann, indem Sie ihr erstelltes Jupyter-Notebook in der zugehörigen Aufgabe im Moodle Kurs hochladen. Die Abgabe ist bis zum 10.8.2025 um 23:59 Uhr möglich.
Andere Formen der Abgabe sind nicht vorgehen.

Die Hausarbeit kann in Gruppenarbeit von bis zu drei Personen bearbeitet werden. In diesem Falle genügt *eine* Abgabe pro Gruppe.

**In jedem Fall sind in der Hausarbeit am Anfang die Namen und Matrikelnummern aller daran beteiligten Personen zu vermerken (gilt auch für Einzelabgaben).**

Als Materialien können Sie sämtliche Unterlagen aus der Vorlesung und den Übungen mit benutzen, im Internet recherchieren oder weitere Bücher/Kurse mit verwenden. Geben Sie bitte bei Verwendung von umfangreicherem Programm-Code aus dem Netz (mehr als 3-4 Zeilen) die Quelle kurz mit an.

Die Verwendung von ChatGPT oder ähnlichen Hilfsmitteln ist nicht gestattet. Die Prüfungsordnung sieht für Verdachtsfälle die Möglichkeit mündlicher Nachprüfungen vor.

**Aufgabe 1** (Python Basics)

Solar Energie ist eine wichtige Energiequelle bei der Transformation hin zu einer Reduktion des $CO_2$ Ausstoßes. Dabei gab es in den letzten Jahren sowohl technische Entwicklungen, z.B. Effizienzsteigerungen bei den Solarpanelen, als auch politische Entwicklung, z.B. in Bezug auf die Förderung von Solarenergie.
Seit ca. 2017 gibt es die Möglichkeit, kleinere Solaranlagen (*Balkonkraftwerke*) ohne großen Aufwand aufzustellen und über einen Stecker an das Hausnetz anzuschließen. Diese Anlagen sind auf eine Leistung von 800 W beschränkt.
In Deutschland müssen diese Balkonkraftwerke allerdings im sogenannten *Marktstammdatenregister* angemeldet. Dieses Register ist öffentlich einsehbar[1] und dient als Datengrundlage für diese Hausarbeit.

## Vorbereitung

Laden Sie die Dateien **`Hausarbeit.ipynb`** und **`solar.py`** in Ihr Verzeichnis auf dem Jupyter-Server hoch. Die Datei **`solar.py`** ist ein kleines Modul, dass für diese Aufgabe benötigt wird. Beide Dateien müssen sich im gleichen Verzeichnis auf dem Jupyter-Server befinden.
Sobald Sie die Dateien hochgeladen haben, können Sie in Ihrem Jupyter-Notebook die ersten zwei Zellen ausführen, die das Modul **`solar`** importieren die Funktion für die Liste der Solaranlagen zeigt:

```
import solar
liste = solar.anlagen()
```

Die Funktion **`anlagen()`** liefert eine Liste der installierten Balkonkraftwerke in der Stadt Bochum. Jeder Eintrag der Liste ist ein Tupel in folgendem Format:

**`(nr, name, status, brutto, netto, datum, plz, ort, anzahl)`**

Dabei bezieht sich die Komponente **`name`** auf die Bezeichnung der Anlage, die vom Besitzer gewählt wurde. Die Werte in **`status`**, **`brutto`** und **`netto`** geben an, in welchem Status sich die Anlage aktuell befindet und wie hoch die Brutto- bzw. Nettoleistung der Anlage ist.
Das **`datum`** ist der Tag der Inbetriebnahme und **`anzahl`** die Anzahl der einzelnen Module, die zur Anlage gehören. **`plz`** und **`ort`** geben natürlich die Postleitzahl und den Ort an. Hier ist ein kleines Beispiel, wie die Daten zu benutzen sind:

```
import solar

liste = solar.anlagen()

anlage24 = liste[23]
#  (24, 'WAT-Garage-740Wp', 'Voruebergehend stillgelegt',
#    0.6, 0.6, '2022-03-01', 44866, 'Bochum', 2.0)
```

[1] Marktstammdatenregister: **`https://www.marktstammdatenregister.de/MaStR/Einheit/Einheiten/OeffentlicheEinheitenuebersicht`**

Die Anlage an der Position 23 hat die Nummer 24 und den Status *Vorübergehend stillgelegt*. Die Anlage hat 0.6 kWp, also 600 Watt maximale Leistung und wurde am 1.3.2022 in Betrieb genommen. Sie besteht aus 2 Modulen und ist irgendwo im PLZ Bereich 44866 in Bochum aufgestellt worden.
Der Name "WAT-Garage-740Wp" klingt ein wenig so, als wäre Sie auf einer Garage installiert worden. Dieser Namen kann allerdings vom Benutzer beliebig ausgefüllt werden.

Für eine derartige Liste sollen Sie die folgenden Aufgaben lösen:

1. Schreiben Sie eine Funktion `plzs(liste)`, die als Parameter die obige Liste bekommt und die Menge der Postleitzahlen zurückgibt, für die es Anlagen in der Liste gibt. Das bedeutet insbesondere, das jede PLZ nur einmal im Ergebnis vorkommt.

2. Schreiben Sie eine Funktion `anlagen_plz(liste, plz)`, die für die gegebene Liste und eine angegebene Postleitzahl die Liste der Anlagen mit dieser Postleitzahl zurückliefert.

3. Geben Sie eine Funktion `anzahl_plz(liste)` an, die die Liste von Anlagen bekommt und eine Liste mit Tupeln der folgenden Art berechnet:

   ```
   [ (plz, anzahlAnlagen), ...]
   ```

   D.h. jedes Tupel besteht aus der PLZ und der Anzahl der Anlagen, die in diesem PLZ-Gebiet aufgestellt wurden.

4. Schreiben Sie ein Funktion `nettoleistung_plz(liste, jahr)`, Summe der installierten Nettoleistung aller Anlagen für das angegebene Postleitzahlengebiet berechnet.

5. Schreiben Sie eine Funktion `nettoleistung_pro_plz(liste)`, die eine Liste von Tupeln berechnet, wobei jedes Tupel die Form

   ```
   (plz, summeNettoleistung)
   ```

   haben soll.

6. Ursprünglich waren Balkonkraftwerke mit einer Leistung von 600 Watt erlaubt. Dies wurde später auf 800 Watt erhöht. Schreiben Sie eine Funktion mit dem Namen `anlagen_groessen(liste)`, die ein Tupel mit 3 Werten zurückliefert:

   ```
   (anzahlBis300, anzahlBis600, anzahlUeber600)
   ```

   Der erste Wert soll die Anzahl der Anlagen mit maximal 300 Watt Nettoleistung enthalten, der zweite Wert, die Anzahl der Anlagen mit mehr als 300 Watt und maximal 600 Watt und der letzte Wert die Anzahl der Anlagen über 600 Watt Nettoleistung.

   Die Summe der Werte soll der Gesamtanzahl der Anlagen in der Liste entsprechen, d.h. eine Anlage mit 500 Watt wird nur im zweiten Wert mitgezählt.

   **Hinweis:** Die Netto- und Bruttoleistung in der Liste ist in Kilowatt angegeben!

**Aufgabe 2** (Pandas und Statistiken)

Über das Marktstammdatenregister sind noch ein Vielzahl weiterer Informationen erhältlich. Unter der URL `https://data.hsbo.de/solaranlagen.csv` finden Sie ein recht große Datei (ca. 250 MB) mit alle registrierten Solaranlagen und Batteriespeichern in NRW.
Die folgende Tabelle enthält einen gekürzten Ausschnitt der Daten (der Datensatz selbst enthält natürlich noch weiter Zeilen und Spalten). Dabei wurden die Spaltennamen aus Darstellungsgründen leicht geändert (Energietr = Energieträger, Solar = Solare Strahlungsenergie, Nettolstg = Nettoleistung, usw.)

| MaStR_Nr | Energietr | Nettolstg | Inbetriebnahme | PLZ | Ort | Ausrichtung | Kapazität |
|---|---|---|---|---|---|---|---|
| 517 | Speicher | 0.8 | 2025-05-17 | 46535 | Dinslaken | | 1.7 |
| 003 | Solar | 6.0 | | 53842 | Troisdorf | West | |
| 786 | Speicher | 0.8 | 2025-05-14 | 53173 | Bonn | | 1.6 |
| 369 | Solar | 8.8 | 2025-05-16 | 40593 | Düsseldorf | Ost-West | |
| 681 | Speicher | 0.8 | 2025-05-16 | 52531 | Übach-Palenberg | | 2.0 |
| 847 | Solar | 9.6 | 2025-04-03 | 42857 | Remscheid | Süd | |
| 125 | Speicher | 0.8 | 2025-05-18 | 47475 | Kamp-Lintfort | | 1.6 |
| 970 | Speicher | 0.8 | 2025-05-17 | 58332 | Schwelm | | 1.36 |
| 802 | Solar | 8.0 | 2024-11-05 | 53909 | Zülpich | Süd-West | |
| 740 | Solar | 10.0 | 2025-05-15 | 48431 | Rheine | Ost-West | |

Die Bedeutung der Spalten ist weitestgehend selbsterklärend. Mit *Ausrichtung* (im richtigen Datensatz heisst die Spalte *Hauptausrichtung_SolarModule*) ist z.B. die Himmelsrichtung gemeint, in die die Module hauptsächlich ausgerichtet sind.
Natürlich sind für einige Datensätze bestimmte Spalten leer, bzw. enthalten den Wert *nan*, was für einen leeren Wert steht. Dies liegt z.B. daran, dass ein Datensatz, der einen Energiespeicher (Batterie) darstellt, keine Himmelsrichtung hat, in die die Batterie ausgerichtet ist.

Wir sind in dieser Aufgabe auf der Suche nach Antworten auf Fragen wie z.B.:

- Wie hat sich die Gesamtleistung der Solaranlagen in NRW über die Jahre entwickelt?
- Welche Städte haben den meisten Zuwachs in Solar-Kapazitäten?
- Wie hat sich parallel dazu die Batterie-Kapazität entwickelt?

Hintergrund dieser Aufgabe ist es, dass Sie sich mit einem unbekannten Datensatz vertraut machen und mit Hilfe von Pandas untersuchen, welche Informationen aus den Daten herausgesucht werden können.

## Die Aufgaben:

1. Zunächst sollen ein paar generelle Informationen berechnet werden:
   - Wieviele verschiedene Anlagen gibt es in dem Datensatz? Über welchen Zeitraum sind überhaupt Daten verfügbar?
   - Gibt es ungewöhnliche Werte? Welche min/max Werte gibt es z.B. bei den numerischen Spalten?
     Sind irgendwelche Wert ggf nicht schlüssig?
   - Wieviele verschiedene Städte sind in dem Datensatz mit aufgeführt?
2. Berechnen Sie die Anzahl der in Betrieb genommenen Solaranlagen pro Jahr. Welchen Zeitraum würden Sie sinnvollerweise dazu betrachten? Lesen Sie dazu ggf. nochmal die Einleitung zu Aufgabe 1. Erstellen Sie einen Plot, der die Entwicklung der in Betrieb genommenen Anlagen zeigt.
3. Erstellen Sie auch einen Plot, der Summe der Nettoleistungen pro Jahr. Was fällt Ihnen dabei auf?
4. Berechnen Sie auch die Anzahl und die Gesamtkapazität der in Betrieb genommenen Batterie-Speicher. Die Preise für Batterie-Speicher sind in den letzten Jahren ja stetig gefallen. Erstellen Sie einen Plot, der die Summe der Speicherkapazitäten pro Jahr zeigt.

   Ist mit den gesunkenen Preisen auch die Kapazität der installierten Batteriespeicher stark gestiegen?
5. Betrachten Sie nochmal nur die Solaranlagen. Dort wird für jede Anlage die Nettoleistung und die Anzahl der Module angegeben. Man könnte nun als Kennzahl für die Effizienz von Modulen die Nettoleistung durch die Anzahl der Module teilen.

   Berechnen Sie basierend auf dieser Überlegung die durchschnittliche Effizienz der Module pro Jahr und erstellen Sie einen Plot dazu. Wie hat sich die Effizienz der Module über die Jahre entwickelt?

   Gibt es Probleme bei dieser Rechnung (vgl. die generellen Informationen aus 2.1)?
6. Interessant ist natürlich auch die unterschiedliche Entwicklung in den einzelnen Städten. Welche Städte haben die größte installierte Leistung (Nettoleistung) von Solarmodulen?
7. Nehmen Sie zwei Städte ihrer Wahl und vergleichen Sie die Entwicklung über die vergangenen zehn Jahre. Gibt es dort Unterschiede in der Zunahme der Solar-Leistung?
8. Betrachten Sie die Anlagen der Stadt Bochum. Wieviele Anlagen gibt es insgeamt in der Stadt? In welchem Stadtbezirk finden Sie die meisten Anlagen? Wo finden Sie die größten Anlagen (Nettoleistung)? Wie teilen Sie das auf?

## Hinweis zur Bearbeitung

Es geht bei der Bearbeitung dieser Aufgaben nicht nur um die reine Programmierung in Python. Ziel ist es, die Daten entlang der Teilaufgaben zu analysieren und die Ergebnisse in einem gewissen Rahmen zu interpretieren.
Dazu gehört zu jeder Teilaufgabe, dass Sie kurz skizzieren, wie Sie vorgehen wollen, welche Teil-DataFrames sie ggf. berechnen wollen und was Sie am Ergebnis ggf. kritisch betrachten (z.B. Datenqualität, etc.). Auch dafür haben Sie in Data Science und Kursen wie Wirtschaftsstatistik Methoden und Werkzeuge kennengelernt.
Überlegen Sie sich zudem, wie Sie die Ergebnisse Ihrer Analysen überprüfen können.

Für die Bearbeitung insbesondere die Verwendung von Zeitangaben, sei Ihnen hier nochmal die Foliensätze aus der Vorlesung *Wirtschaftsinformatik 2* nahegelegt, die einen guten Überblick über DataFrames und Datumsangaben enthalten:

- `winf2-06-pandas-dataframe.pdf`
- `winf2-08-pandas-dataframe-groupby.pdf`

Die Foliensätze finden Sie auf dem Server `https://datascience.hs-bochum.de` unter

Vorlesungen -> aktuelles Semester -> Wirtschaftsinformatik 2

Sie sind allerdings auch auf der Materialseite der DataScience 1 Vorlesung nochmal verlinkt.