OTH AMBERG-WEIDEN
Fakultät Elektrotechnik, Medien und Informatik
Prof. Dr.-Ing. Christian Bergler
Wintersemester 2023/2024

Ostbayerische Technische Hochschule
Amberg-Weiden

# Projektarbeit Data Analytics

Ausgabe: 18.12.2023 – 18:00:00 Uhr Deadline: 28.01.2024 – 23:59:59

## Ziel der Projektarbeit

In dieser Projektarbeit soll die Stromerzeugung der vergangenen Jahre in Deutschland anhand historischer Daten analysiert werden. Betrachtungszeitraum sind die Jahre 2018-2023. Ziel ist es insbesondere, Erkenntnisse zum Strom-Mix in Deutschland zu gewinnen und den Ausbau der erneuerbaren Energien zu untersuchen, wobei der Schwerpunkt auf dem Ausbau der Photovoltaik-Stromerzeugung liegt.

## Modalitäten

- Die Bearbeitung der Projektarbeit erfolgt in der Programmiersprache Python.
- Als Ergebnis ist ein Jupyter-Notebook namens `nachname_vorname.ipynb` zu erstellen und elektronisch via Moodle abzugeben, das den vollständigen Programmcode und alle Analyseergebnisse (eingebettete Grafiken, Outputs der Zellen, erläuternder Text) enthält. Eine separate schriftliche Ausarbeitung ist nicht erforderlich. Die Ergebnisse zu den einzelnen Aufgaben und die abgeleiteten Erkenntnisse sollen jedoch ebenso wie das methodische Vorgehen in Form von Fließtext innerhalb des Notebooks ausführlich textuell dokumentiert werden. Darüber hinaus ist für den Vortrag ein Foliensatz zu erstellen, der als Teil der Dokumentation mit abzugeben ist. Datensätze, die als Ergebnisse einzelner Aufgaben resultieren, sind, sofern explizit gefordert, ebenfalls mit abzugeben.
- Der eingereichte Programmcode soll im EDV-Labor MBUT224 mit der dort installierten Version von Python lauffähig sein. Geben Sie im Kopf des Dokuments (Jupyter-Notebook) an, ob und ggf. welche zusätzlichen Pakete installiert werden müssen.
- Die Bearbeitung der Projektarbeit ist in Gruppen von maximal zwei Personen zulässig. Im Fall einer Zweierabgabe genügt es, wenn ein Gruppenmitglied die Arbeit elektronisch einreicht. Im Kopf des Dokuments sind alle Gruppenmitglieder zu benennen.
- Die Abgabe der Dokumente hat **bis spätestens 28.01.2024 um 23:59:59 Uhr** über Moodle zu erfolgen.
- Eine Präsentation der Ergebnisse in Form eines ca. 15-minütigen Kurzvortrags erfolgt voraussichtlich am 05.02.2024 und/oder 07.02.2024. Ein zeitlicher Ablaufplan wird auf der Basis der gebildeten Zweiergruppen erstellt und rechtzeitig vorher bekannt gegeben. Bitte füllen Sie bis spätestens 31.12.2023 die Abstimmung zur Terminauswahl in Moodle aus und tragen Sie sich in einen der vorgeschlagenen Time-Slots ein. Im Fall einer Zweiergruppe ist es zwingend notwendig das beide Gruppenmitglieder denselben Time-Slot auswählen.
- Die unten beigefügte schriftliche Erklärung (s. Anhang) ist von allen Gruppenmitgliedern auszufüllen, zu scannen und mit den eingereichten Dokumenten hochzuladen.

# Anforderungen und Bewertungsgrundlagen

- Der Code ist lauffähig und erfüllt die in den Aufgaben gestellten Anforderungen.
- Der Code ist klar strukturiert, gut lesbar, nachvollziehbar und ausreichend kommentiert.
- Der Code ist elegant, effizient und verwendet, sofern verfügbar, bereits vorhandene Python-Funktionen zur Bearbeitung der gestellten Analyseaufgaben.
- Das eingereichte Jupyter-Notebook ist ansprechend und übersichtlich gestaltet. Verwenden Sie dazu die Strukturierungsmöglichkeiten, die die Markdown-Sprache bietet. Das Dokument soll mit einer Gliederung mit Verlinkung zu den Lösungen der einzelnen Aufgaben versehen werden und eine abschließende Zusammenfassung der Analyseergebnisse und der gewonnenen Erkenntnisse enthalten. Sofern Sie externe Quellen verwenden, geben Sie diese in einem Quellenverzeichnis an.
- Die in den einzelnen Teilaufgaben gewonnenen Erkenntnisse sind ausführlich visuell und textuell dokumentiert und in den Anwendungskontext eingeordnet. Die Ausführungen sind klar formuliert, nachvollziehbar und durch die Daten belegbar. Die editorielle Qualität des Dokuments fließt in die Bewertung ein.
- Die erstellten Diagramme sind ansprechend und übersichtlich gestaltet und transportieren eine klare Botschaft. Sie sind insbesondere ausreichend beschriftet (z.B. Titel, Achsenbeschriftungen, Einheiten etc.).
- Die bei der Datenvorbereitung und -analyse durchgeführten Schritte sind fachlich und methodisch korrekt ausgeführt und hinreichend motiviert und dokumentiert worden. Beschreiben Sie nicht nur, *wie* Sie vorgehen, sondern auch *warum*.
- Die Ergebnisse werden im Rahmen eines Vortrags ansprechend und überzeugend präsentiert. Die Qualität der Folien, die zur Dokumentation gehören, fließt in die Bewertung ein.

# Gegebene Daten

**Daten zur Nettostromerzeugung**
Ausgangspunkt für die Analysen sind Energiedaten, die von Wissenschaftlern des Fraunhofer-Instituts für Solare Energiesysteme ISE aus verschiedenen Quellen stündlich oder täglich abgerufen und unter `https://www.energy-charts.info` veröffentlicht werden. Für die Analysen im Rahmen dieser Arbeit stehen Auszüge dieser Datensätze auf Moodle zum Download zur Verfügung. Diese enthalten Daten zur Nettostromerzeugung in Deutschland seit dem Jahr 2018, wobei je eine CSV-Datei pro Kalenderjahr bereitgestellt wird.

**Daten zu Photovoltaikanlagen aus dem Marktstammdatenregister**
Weiterhin werden Daten zu Stromerzeugungseinheiten aus dem Martkstammdatenregister analysiert. Dabei handelt es sich um ein von der Bundesnetzagentur geführtes Register, in dem u.a. die Stammdaten aller Strom- und Gaserzeugungsanlagen geführt werden. Das Martkstammdatenregister ist unter `https://www.marktstammdatenregister.de` abrufbar. Auf Moodle finden Sie einen Auszug des Registers mit Angaben zu allen Photovoltaikanlagen in Deutschland.

Darüber hinaus sollen im Verlauf dieser Arbeit weitere Daten (z.B. Wetterdaten) aus externen Quellen bezogen werden, vgl. nachfolgende Arbeitsaufträge.

Hinweis: die bereitgesellten Datensätze werden in der ersten Januarwoche aktualisiert, damit Sie mit den vollständigen Daten für das Jahr 2023 arbeiten können.

# Aufgaben

**Aufgabe 1 (Datenvorbereitung)**

a) Lesen Sie die CSV-Dateien zur Nettostromerzeugung in den einzelnen Kalenderjahren ein und führen Sie sie in einem DataFrame namens `df_el` zusammen.

b) Passen Sie die dtypes der Spalten von `df_el` geeignet an.

c) Überprüfen Sie den Datensatz auf fehlende Werte und wenden Sie ggf. eine geeignete Strategie an, um mit diesen zu verfahren.

d) Beurteilen Sie die Datenqualität des Datensatzes.

**Aufgabe 2 (Explorative Datenanalyse)**
In dieser Aufgabe sollen anhand verschiedener Diagramme und Statistiken erste Erkenntnisse über die Stromerzeugung in den vergangenen Jahren gewonnen werden.

a) An welchen 10 Tagen wurde insgesamt am meisten Strom erzeugt?

b) An welchem Tag wurde bisher am meisten Windstrom erzeugt?

c) Ermitteln Sie, wie viel Strom mit jedem der Energieträger im Betrachtungszeitraum 2018-2023 pro Jahr (absolut) erzeugt wurde. Visualisieren Sie diese Ergebnisse in einem geeigneten Diagramm.

d) An wie vielen Tagen wurde mehr als 30% des Stroms durch Sonnenenergie erzeugt?

e) Visualisieren und untersuchen Sie anhand von Histogrammen die Verteilung des täglich erzeugten Windstroms. Erstellen Sie je ein Histogramm für Onshore und Offshore und vergleichen Sie die beobachteten Verteilungen.

f) Berechnen Sie mit Hilfe von `describe` verschiedene statistische Kenngrößen für den täglich erzeugten Solarstrom bezogen auf die einzelnen Jahre des Betrachtungszeitraums. Visualisieren Sie mit Hilfe von Box-Whisker-Plots die Verteilungen der Werte für den täglich generierten Solarstrom. Erstellen Sie für jedes Jahr ein separates Diagramm. Beschreiben Sie die Entwicklungen vor dem Hintergrund der historischen jährlichen Sonnenscheindauer in Deutschland und dem PV-Ausbau (vgl. Aufgabe 5).

**Aufgabe 3 (Weiterführende Analyse der Stromerzeugung)**

a) Visualisieren Sie in einem Säulendiagramm, wie viel Strom pro Monat im Betrachtungszeitraum insgesamt erzeugt wurde. Beschreiben Sie den Verlauf.

b) Berechnen und visualisieren Sie die mittlere erzeugte Energie pro Wochentag und beschreiben Sie mögliche Trends.

c) Visualisieren Sie die Zusammensetzung des Strommixes im Zeitverlauf. Differenzieren Sie dabei nur nach erneuerbaren und nicht erneuerbaren Energieträgern.

d) Visualisieren und untersuchen Sie den Anteil der Kernenergie an der Nettostromerzeugung im Zeitverlauf.

e) Erstellen Sie ein interaktives Diagramm in Plotly, in dem die zeitlichen Verläufe der Energieerzeugung mit den einzelnen Energieträgern als Liniendiagramm dargestellt werden. Wenden Sie unter Verwendung der Methode `DataFrame.rolling` einen gleitenden Mittelwert an, um die Zeitreihen zu glätten. Durch Klick auf die Legende sollen die Kurven zu den einzelnen Energieträgern aus- und eingeblendet werden können.

f) Untersuchen Sie mit Hilfe des interaktiven Diagramms, wie sich die Stromerzeugung ausgewählter Energieträger in Deutschland seit 2018 entwickelt hat. Beschreiben Sie jeweils die beobachteten Trends und erklären Sie diese. Untersuchen Sie insbesondere, wie sich die Energieerzeugung im zeitlichen Zusammenhang mit der Abschaltung der letzten Atomkraftwerke am 15.04.2023 geändert hat.

**Aufgabe 4 (Erzeugter Solarstrom in Abhängigkeit vom Wetter)**
In dieser Aufgabe soll die Solarstromerzeugung in Beziehung zum Wetter gesetzt werden. Verschaffen Sie sich zunächst anhand der Online-Dokumentation unter

`https://open-meteo.com/en/docs/historical-weather-api`

einen Überblick über die `historical-weather-api` von `open-meteo`. Über diese API sollen im Folgenden historische Wetterdaten bezogen werden.

a) Implementieren Sie eine Funktion namens

`get_weather_data(lat, lon, start_date, end_date),`

die die (tagesbezogenen) Wetterdaten für die durch `(lon,lat)` gegebene Geo-Position im Zeitraum zwischen `start_date` und `end_date` von der `open-meteo-API` bezieht und das Ergebnis als DataFrame zurückgibt. Verwenden Sie für den Zugriff auf die API das Paket `requests`. Jede Zeile des DataFrames soll die Wetterdaten zu einem Tag enthalten. Wenden Sie diese Funktion nun an, um historische Wetterdaten der Jahre 2018-2023 für die Stadt Amberg zu beziehen. Speichern Sie diese in einen DataFrame namens `df_weather` und verwenden Sie diesen als Datengrundlage für alle weiteren Teilaufgaben dieser Aufgabe.

b) Untersuchen Sie die verschiedenen Variablen des Wetterdatensatzes auf Korrelationen, indem Sie eine interaktive CorrelationHeatmap in Plotly erzeugen.

c) Ermitteln Sie für jedes Jahr des Betrachtungszeitraums die Anzahl der Sonnenscheinstunden und die jährliche Sonneneinstrahlung.

d) Erzeugen Sie für jedes Jahr des Betrachtungszeitraums ein Säulendiagramm, in dem die aggregierte Sonneneinstrahlung pro Monat auftragen ist. Ordnen Sie die sechs Diagramme mit Hilfe von `pyplot.subplots` in einem Raster der Dimension $2 \times 3$ an.

e) Visualisieren Sie in geeigneten Diagrammen die maximale Tagestemperatur, die Sonnenscheindauer und die Sonneneinstrahlung pro Tag im Betrachtungszeitraum.

f) Untersuchen Sie anhand der Ergebnisse der vorherigen drei Teilaufgaben die wesentlichen Entwicklungen und Trends für die Sonneinscheindauer, Sonneneinstrahlung und Temperaturen im Betrachtungszeitraum.

g) Untersuchen Sie nun die Zusammenhänge zwischen den sonnenbezogenen Wettervariablen (exemplarisch betrachtet für die Stadt Amberg) und dem in Deutschland pro Tag erzeugten Solarstrom. Visualisieren und quantifizieren Sie diese geeignet.

**Aufgabe 5 (Photovoltaik-Ausbau in Deutschland)**
Der absolut erzeugte Solarstrom hängt natürlicherweise mit der in Deutschland installierten Leistung für Photovoltaik zusammen. Diese soll mit Hilfe von Daten aus dem Marktstammdatenregister analysiert werden.

a) Auf Moodle finden Sie einen Auszug des Marktstammdatenregisters mit Angaben zu allen Photovoltaikanlagen in Deutschland. Laden Sie sich die CSV-Dateien herunter, lesen Sie sie ein und führen Sie sie in einen DataFrame namens `df_pv` zusammen.

b) Untersuchen Sie die Qualität des Datensatzes und bereiten Sie diesen für die weiteren Analysen geeignet vor.

c) Wie viele Photovoltaikanlagen sind derzeit insgesamt und pro Bundesland in Betrieb? Verwenden Sie für die weiteren Analysen nur diese Anlagen.

d) Berechnen Sie den Mittelwert und den Median der Bruttoleistungen und erläutern Sie Ihre Beobachtung.

e) Geben Sie die zehn Anlagen mit der höchsten Bruttonennleistung aus.

f) Ermitteln und visualisieren Sie den monatlichen Zubau der Bruttonennleistung pro Monat seit dem Jahr 2015. Beschreiben Sie die Entwicklung und ordnen Sie diese in den Kontext politischer und wirtschaftlicher Rahmenbedingungen ein.

g) Ermitteln Sie die in Deutschland installierte Bruttoleistung im Zeitverlauf seit dem Jahr 2000 und visualisieren Sie diese in einem geeigneten Diagramm.

h) Berechnen und visualisieren Sie die derzeit installierte Bruttoleistung pro Bundesland. Erstellen Sie dazu unter Verwendung von `folium` eine Choroplethen-Karte von Deutschland, in der die Bruttoleistung der einzelnen Bundesländer farblich dargestellt wird. Ein Geo-Json-File steht in Moodle zum Download bereit. Beschreiben Sie anhand der Karte die regionale Verteilung der PV-Leistung.

i) Der Anlagenname kann durch den Betreiber selbst gewählt werden. Untersuchen Sie typische Namen für die PV-Anlagen, indem Sie mit Hilfe des Pakets `WordCloud` eine Wortwolke für die Anlagennamen erstellen.

j) Untersuchen Sie nun den Zusammenhang zwischen der an einem Tag installierten Leistung und den an diesem Tag erzeugten Solarstrom. Führen Sie dazu die Stromerzeugungsdaten und die Daten zur installierten Bruttoleistung geeignet zusammen. Laden Sie den resultierenden Datensatz als CSV-Datei namens `erzeugung_leistung.csv` zusammen mit Ihrer Abgabe auf Moodle hoch.

**Aufgabe 6 (Modellbildung)**
In den vorangegangenen Aufgaben wurde bereits festgestellt, dass der täglich erzeugte Solarstrom einerseits von der installierten Photovoltaik-Gesamtleistung und andererseits vom Wetter abhängt. In dieser Aufgabe soll nun der Zusammenhang zwischen diesen Größen modelliert werden.

a) Erstellen Sie anhand der gegebenen historischen Daten für die Jahre 2018-2023 ein Lineares Regressionsmodell, das den an einem Tag in Deutschland erzeugten Solarstrom (vgl. Aufgabe 1) in Abhängigkeit der zu diesem Zeitpunkt installierten Bruttoleistung für Photovoltaik (vgl. Aufgabe 5) sowie der Tageslicht-Dauer, der Sonnenschein-Dauer und der Sonneneinstrahlung über Amberg (vgl. Aufgabe 4) prognostiziert. Führen Sie dazu zunächst die Daten geeignet zusammen. Laden Sie den resultierenden Datensatz als CSV-Datei namens `dataset_model.csv` mit Ihrer Abgabe auf Moodle hoch. Teilen Sie die Daten anschließend in eine Trainings- und eine Testdatenmenge auf und erstellen Sie unter Verwendung der Bibliothek `Scikit-learn` auf dem Trainingsdatensatz das Modell. Beurteilen Sie die Güte des resultierenden Modells, indem Sie den mittleren relativen Fehler (mean absolute percentage error) auf dem Trainings- und auf dem Testdatensatz auswerten.

b) Wenden Sie Ihr Modell nun geeignet an, um den in Deutschland produzierten Solarstrom für die einzelnen Tage vom 29.01.2024-04.02.2024 zu prognostizieren. Bereiten Sie für Ihre Präsentation eine Auswertung vor, wie gut die Prognose im Nachgang war.

c) Gehen Sie abschließend auf mögliche Limitierungen des Modells ein und erläutern Sie, durch welche Maßnahmen es verbessert werden könnte.

# Anlage zur Projektarbeit Data Analytics

Wintersemester 2023/2024
Prof. Dr.-Ing. Christian Bergler

Füllen Sie die nachfolgende Erklärung entweder gemeinsam oder pro Gruppenmitglied aus und laden Sie eine gescannte Version mit Ihrer Einreichung auf Moodle hoch.

Name, Vorname – Gruppenmitglied 1:

Matrikelnummer – Gruppenmitglied 1:

Name, Vorname – Gruppenmitglied 2:

Matrikelnummer – Gruppenmitglied 2:

## Erklärung

Hiermit wird erklärt, dass die eingereichte Projektarbeit ausschließlich von den o.g. Personen erstellt wurde. Alle verwendeten Hilfsmittel und Quellen sind in der Arbeit angegeben worden.

Ort, Datum

Unterschrift(en)