OTH AMBERG-WEIDEN
Fakultät Elektrotechnik, Medien und Informatik
Prof. Dr. Christian Bergler
Wintersemester 2024/2025

Projektarbeit Data Analytics

Ausgabe: 16.12.2024 – 20:00:00 Deadline: 26.01.2025 – 23:59:59

Ziel der Projektarbeit

In dieser Projektarbeit soll die Stromerzeugung und die Entwicklung von Strompreisen in den vergangenen Jahren in Deutschland anhand historischer Daten analysiert werden. Betrachtungszeitraum sind die Jahre 2020-2024. Ziel ist es insbesondere, zu untersuchen, wie sich verschiedene Einflussfaktoren (z.B. der Ausbau und die Erzeugung von erneuerbaren Energien, das Wetter oder geopolitische Ereignisse) auf die Höhe der Börsenstrompreise auswirken. Weiterhin soll der Zusammenhang zwischen Börsenstrompreisen und Endkundenpreisen analysiert und ein Modell zur Vorhersage von Börsenstrompreisen abgeleitet werden.

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.
- 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 26.01.2025 um 23:59:59 Uhr** über Moodle zu erfolgen.
- Eine Präsentation der Ergebnisse in Form eines ca. 20-minütigen Kurzvortrags erfolgt voraussichtlich am 29.01.2025 und/oder 30.01.2025. Ein zeitlicher Ablaufplan wird auf der Basis der gebildeten Zweiergruppen erstellt und rechtzeitig vorher bekannt gegeben. Bitte füllen Sie bis spätestens 31.12.2024 eine Umfrage zur Gruppenbildung in Moodle aus. Im Fall einer Zweiergruppe genügt es, wenn nur ein Gruppenmitglied die Umfrage ausfüllt.
- 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. Alle verwendeten Hilfsmittel und Quellen sind anzugeben.
- Die editorielle Qualität des Dokuments fließt in die Bewertung ein.
- Die in den einzelnen Teilaufgaben gewonnenen Erkenntnisse sind ausführlich visuell und textuell dokumentiert und in den Anwendungskontext eingeordnet. Beschreiben Sie nicht nur, *was* Sie beobachten, sondern gehen Sie auch auf mögliche Ursachen und weiterführende Zusammenhänge ein, ggf. unter Einbeziehung externer Quellen.
- Alle Ausführungen sind klar formuliert, nachvollziehbar und durch die Daten belegbar.
- 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 zu Börsenstrompreisen und zur Energieeerzeugung
Ausgangspunkt für die Analysen sind Stromerzeugungs- und -preisdaten, 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 Stromerzeugung mit erneuerbaren und nicht erneuerbaren Energieträgern (ausgenommen Kernenergie) und zu den Börsenstrompreisen (Day Ahead Auktion an der Strombörse Leipzig) in Deutschland seit dem Jahr 2020, wobei je eine CSV-Datei pro Kalenderwoche bereitgestellt wird.

Daten zu Neukunden-Strompreisen
Weiterhin werden für das Jahr 2024 Daten zu tagesaktuellen Neukunden-Angebotspreisen, wie sie etwa in Preisvergleichsportalen abgerufen werden können, bereitgestellt, die Tarifangebote verschiedener Anbieter in 40 bayerischen Städten enthalten. Die angebotenen Tarife beziehen sich auf Endkunden mit einem jährlichen Stromverbrauch von 4.000 kWh.

Darüber hinaus sollen im Verlauf dieser Arbeit weitere Daten (z.B. Wetterdaten) aus externen

Quellen bezogen werden, vgl. nachfolgende Arbeitsaufträge.

Hinweis: die bereitgestellten Datensätze werden in der ersten Januarwoche aktualisiert, damit Sie mit den vollständigen Daten für das Jahr 2024 arbeiten können.

Aufgaben

Aufgabe 1 (Datenvorbereitung)

a) Lesen Sie die CSV-Dateien, die die Stromerzeugungsdaten und Börsenstrompreise enthalten ein und führen Sie sie in einem DataFrame namens `df_hourly` zusammen.

b) Passen Sie die dtypes der Spalten von `df_hourly` geeignet an. Überführen Sie insbesondere das Datum in ein DateTime-Format. Entfernen Sie alle Datensätze, die sich nicht auf den Betrachtungszeitraum 2020-2024 beziehen.

c) Beurteilen Sie die Datenqualität des Datensatzes und führen Sie, sofern aus Ihrer Sicht notwendig, geeignete Datenbereinigungsschritte durch.

d) Erzeugen Sie aus `df_hourly` einen weiteren DataFrame namens `df_daily`, der in jeder Zeile die erzeugte elektrische Energie mit erneuerbaren und nicht erneuerbaren Energieträgern sowie den an diesem Tag durchschnittlich gemessenen Börsenstrompreis enthält.

Aufgabe 2 (Explorative Analyse der Stromerzeugungs- und Preisdaten)

a) An welchen 10 Tagen im Betrachtungszeitrum wurde am meisten Strom aus erneuerbaren Energieträgern erzeugt?

b) An welchem Tag im Betrachtungszeitraum wurde der höchste Börsenstrompreis verzeichnet und wie hoch war er? An welchem Tag wurde der geringste Preis verzeichnet und wie hoch war er?

c) Wie viele Tage gab es im Betrachtungszeitraum 2020-2024, an denen ein negativer Börsenstrompreis aufgetreten ist?

d) Wie viel Strom wurde pro Jahr mit erneuerbaren und mit nicht erneuerbaren Energieträgern erzeugt?

Aufgabe 3 (Weiterführende Analyse der Stromerzeugungs- und Preisdaten)

a) Visualisieren Sie in geeigneten Diagrammen die Verteilung der Börsenstrompreise (insgesamt und pro Jahr).

b) Berechnen Sie bezogen auf die einzelnen Jahre des Betrachtungszeitraums verschiedene statistische Kenngrößen für den Börsenstrompreis.

c) Visualisieren Sie in einem Säulendiagramm die mittleren Börsenstrompreise pro Monat des Betrachtungszeitraums.

d) Visualisieren Sie die stündlichen Börsenstrompreise in einem interaktiven Liniendiagramm in Plotly. Versehen Sie dieses mit einem Range-Selektor und einem Range-Slider. Analysieren Sie auf der Basis der Ergebnisse der Teilaufgaben a)-d) die wesentlichen Entwicklungen und Trends der Börsenstrompreise im Betrachtungszeitraum 2020-2024.

e) Berechnen und visualisieren Sie die im Mittel mit erneuerbaren Energieträgern erzeugte elektrische Energie im Tagesverlauf, indem Sie sie auf die vollen Stunden eines Tages aggregieren.

f) Visualisieren Sie in einem geeigneten Diagramm die pro Tag mit erneuerbaren Energieträgern erzeugte elektrische Energie und analysieren Sie diese. Gehen Sie dabei sowohl auf einzelne auffällige Tage als auch auf übergeordnete Entwicklungen und Trends ein.

g) Visualisieren Sie auf geeignete Weise die Zusammensetzung des erzeugten Stroms (erneuerbar vs. nicht erneuerbar) im Zeitverlauf und analyisieren Sie diese.

Aufgabe 4 (Untersuchung von Einflussfaktoren auf den Strompreis)
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`.

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 anschließend an, um historische Wetterdaten der Jahre 2020-2024 für die Stadt Amberg zu beziehen. Reichern Sie den DataFrame `df_daily` um diese Wetterdaten an. Speichern Sie den resultierenden Datensatz in einer CSV-Datei namens `daily.csv` und laden Sie diese mit Ihrer Abgabe auf Moodle hoch.

b) Reichern Sie den DataFrame `df_daily` weiterhin um Börsenschlusskurse für Kohle und Erdgas an, die in den CSV-Dateien `gaspreise.csv` bzw. `kohlepreise.csv` gegeben sind. Visualisieren und untersuchen Sie auf geeignete Weise die Zusammenhänge (paarweise) zwischen Börsenstrompreis, Gaspreis, Kohlepreis und erneuerbarer Energieerzeugung.

c) Führen Sie eine Korrelationsanalyse für alle Variablen des DataFrames `df_daily` durch. Erzeugen Sie dazu eine interaktive Correlation HeatMap in Plotly.

Aufgabe 5 (Modellbildung)

a) Erstellen Sie unter Verwendung der Bibliothek `Scikit-learn` ein lineares Regressionsmodell zur Modellierung des mittleren Börsenstrompreises pro Tag in Abhängigkeit verschiedener Eingangsgrößen. Selektieren Sie basierend auf der vorherigen Aufgabe zunächst geeignete Merkmale als Eingangsgrößen. Teilen Sie die Daten anschließend in eine Trainings- und eine Testdatenmenge und erstellen Sie ein lineares Regressionsmodell auf dem Trainingsdatensatz. Wie lautet der ermittelte funktionale Zusammenhang zwischen den Input- und der Outputgrößen des Modells?

b) Beurteilen Sie die Güte des Modells, indem Sie den mittleren relativen Fehler (mean absolute percentage error, verfügbar in `sklearn.metrics`) berechnen und auswerten.

c) Gehen Sie auf mögliche Limitierungen Ihres Modells ein.

Aufgabe 6 (Analyse von Stromtarif-Angeboten für Endkunden)
In dieser Aufgabe sollen Neukunden-Angebotspreise für Endkunden in verschiedenen bayerischen Städten aus dem Jahr 2024 analysiert werden, wie sie beispielsweise auf Preisvergleichsportalen zu finden sind. Es ist pro Tag und Stadt eine JSON-Datei gegeben, in der bis zu 20 Tarifangebote aufgeführt sind, die an diesem Tag in dieser Stadt am günstigsten waren (Sortierung nach `Preis im 1. Jahr`) Die angebotenen Tarife beziehen sich jeweils auf einen jährlichen Gesamtverbrauch von 4000 kWh/Jahr.

a) Führen Sie die gegebenen Preisvergleichdaten in einem DataFrame namens `df_cust` zusammen. Exportieren Sie diesen als CSV-Datei namens `prices_customers.csv` und laden Sie diese mit Ihrer Abgabe auf Moodle hoch. Verwerfen Sie bitte zur Minimierung der Dateigröße alle Spalten, die im weiteren Verlauf nicht mehr verwendet werden. Kommentieren Sie nun den Code zur Datensatzgenerierung aus und lesen Sie die CSV-Datei in den DataFrame `df_cust` erneut aus dieser Datei ein.

b) Bereiten Sie die Daten auf die weitere Analyse vor, indem Sie geeignete Datentransformations- und -bereinigungsschritte durchführen.

c) Wie viele verschiedene Tarife wurden insgesamt angeboten? Zu wie vielen Tagen sind pro Stadt Daten vorhanden? Wie viele verschiedene Anbieter haben insgesamt Tarife angeboten?

d) Ermitteln Sie, welche unterschiedlichen Tarife in Amberg angeboten wurden und visualisieren Sie exemplarisch für die Stadt Amberg den Füllgrad der Daten. Erstellen Sie dazu eine HeatMap, aus der hervorgeht, an welchen Tagen es zu welchen der ermittelten Tarife Angebotsdaten gab.

e) Visualisieren Sie die durchschnittliche Preisentwicklung im Verlauf des Jahres 2024 über alle Tarife und Orte hinweg. Berücksichtigen Sie dabei nur Tarife, bei denen die Vertragslaufzeit mindestens 12 Monate beträgt. Verwenden Sie dazu den `Preis im 1. Jahr`, der den monatlichen Preis unter Berücksichtigung des Grundpreises, des Arbeitspreises und von Bonuszahlungen o.ä. enthält. Untersuchen Sie anschließend den Zusammenhang zum Börsenstrompreis, indem Sie geeignete Kenngrößen berechnen und weitere Diagramme erstellen.

f) Erstellen Sie ein interaktives Liniendiagramm in Plotly, um die zeitliche Entwicklung der Angebotspreise pro Stadt zu visualisieren. Das Diagramm soll eine Dropdown-Liste enthalten, über die man die Stadt auswählen kann. Das Diagramm soll die Preisentwicklung für diejenigen zehn Tarife zeigen, die in der jeweiligen Stadt im Verlauf des Jahres am häufigsten angeboten wurden (die sich also am häufigsten unter den günstigsten 20 Tarifen beim Preisvergleich befanden). Durch Klick auf den Namen des Tarifs in der Legende soll es möglich sein, dessen Kurve im Diagramm ein- und auszublenden.

g) Untersuchen Sie mit Hilfe des Diagramms die Preisentwicklung der verschiedenen Anbieter in Amberg. Welche Empfehlungen leiten Sie für den Abschluss eines neuen Vertrags ab?

h) Untersuchen Sie die durchschnittlichen Preisniveaus pro Stadt und visualisieren Sie diese auf einer Karte in Folium. Lassen sich bestimmte Trends und Einflussfaktoren erkennen?

i) Im Merkmal `Anbieter` befinden sich kurze Beschreibungen der Anbieter und der Tarife. Erstellen Sie mit Hilfe des Pakets `WordCloud` eine Wortwolke für die Anbieter-Beschreibungen und untersuchen Sie, welche Schlagworte besonders häufig auftreten.

Anlage zur Projektarbeit Data Analytics

Wintersemester 2024/2025
Prof. Dr. 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)