OTH AMBERG-WEIDEN
Fakultät Elektrotechnik, Medien und Informatik
Prof. Dr. F. Brunner
Sommersemester 2020

Ostbayerische Technische Hochschule
Amberg-Weiden

Studienarbeit Data Analytics

Ausgabe: 08.06.2019

Ziel der Studienarbeit

In dieser Studienarbeit soll die Qualität der Luft in Bayern anhand der öffentlich verfügbaren Daten aus dem Messnetz des Umweltbundesamts untersucht werden. An den einzelnen Messstationen werden Daten zur Belastung der Luft durch verschiedene Schadstoffe wie Feinstaub, Stickoxide und Ozon erhoben. Betrachtungszeitraum sind die Jahre 2016 bis 2019.

Modalitäten

- Die Bearbeitung der Studienarbeit 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 und erläuternden Text) enthält. 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 angegeben, ebenfalls mit abzugeben.
- Die Bearbeitung der Studienarbeit 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 05.07.2020 um 23:59:59 Uhr über Moodle zu erfolgen.
- Eine Vorstellung der Ergebnisse in Form eines ca. 15-minütigen Kurzvortrags erfolgt am 06.07.2020 und 07.07.2020 zu den im Stundenplan vorgesehenen Vorlesungs- und Übungszeiten. Die Vorträge finden online über BigBlueButton statt. Eine Ablaufplan und eine Einteilung wird auf der Basis der gebildeten Zweiergruppen erstellt und bekannt gegeben, sobald die Frist zur Prüfungsanmeldung abgelaufen ist (15.06.2020) .
- 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.
- Die gewonnenen Erkenntnisse sind ausführlich textuell und ggf. visuell dokumentiert. Die Ausführungen sind klar formuliert, nachvollziehbar und durch die Daten belegt. 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 worden.
- Die gewonnenen 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

Ausgangspunkt für die Analysen sind die durch das Umweltbundesamt über verschiedene APIs bereitgestellten Daten zur Luftqualität, die hier zu finden und dokumentiert sind:

`https://www.umweltbundesamt.de/daten/luft/luftdaten/doc`

Die Akquise und Aufbereitung der Daten ist Teil der Aufgabenstellung. Über die APIs können Metadaten zu den Stationen und Messungen sowie die Messdaten selbst bezogen werden. Es liegen Messdaten zu folgenden Luft-Schadstoffen vor:

ID	Abkürzung	Bezeichnung	Einheit
1	PM10	Feinstaub	$\mu\mathrm{g}/\mathrm{m}^3$
2	CO	Kohlenstoffmonoxid	mg/m^3
3	O3	Ozon	$\mu\mathrm{g}/\mathrm{m}^3$
4	SO2	Schwefeldioxid	$\mu\mathrm{g}/\mathrm{m}^3$
5	NO2	Stickstoffdioxid	$\mu\mathrm{g}/\mathrm{m}^3$

Abhängig vom gewählten Luftschadstoff stehen verschiedene Arten der Auswertung („scopes“) zur Verfügung, z.B. Tagesmittelwerte, Stundenmittelwerte etc. Je nach Lage werden drei Arten von Messstationen unterschieden: verkehrsnahe Stationen („traffic“), industrienahe Stationen („industry“) und Stationen mit Hintergrundbelastung („background“). Weitere Informationen können der Dokumentation der APIs entnommen werden. Weiterhin werden auf Moodle Wetterdaten zum Download zur Verfügung gestellt.

Aufgaben

Aufgabe 1 (Messstationen, Datenakquise, Semistrukturierte Daten, Geovisualisierung)

a) Beziehen Sie über die Metadaten-API des Umweltbundesamts die Daten zu den Messstationen zum Stand 01.01.2020, indem Sie, z.B. unter Verwendung der Bibliothek `requests`, einen geeigneten HTTP-Request absetzen. Überführen Sie die erhaltenen (semistrukturierten) Daten in einen DataFrame namens `stations`, der für jede Station eine Zeile mit den verfügbaren Informationen enthält (z.B. Name, Adresse, Geokoordinaten, Bundesland etc.). Speichern Sie den DataFrame in eine CSV-Datei namens `stations_2020.csv` und laden Sie diese mit Ihrer Einreichung auf Moodle hoch.

b) Wie viele Messstationen sind derzeit bundesweit in Betrieb?

c) Visualisieren Sie mit Hilfe eines Kreisdiagramms, wie sich die Stationen hinsichtlich ihres Typs zusammensetzen.

d) Erstellen Sie mit `folium` eine interaktive Karte, auf der die einzelnen Messstationen als Kreise eingezeichnet sind. Industrienahe Stationen sollen gelb, verkehrsnahe rot und die Stationen mit Hintergrundbelastung grün eingezeichnet werden. Beim Klick auf die Kreise sollen die Namen der Stationen angezeigt werden.

e) Erzeugen Sie durch Filterung des DataFrames `stations` einen DataFrame `stations_BY`, der die Informationen zu allen Messstationen in Bayern enthält.

Aufgabe 2 (NO_2-Daten, Datenvorbereitung, Datenqualität)

a) Laden Sie über die Measurements-API für alle bayerischen Stationen (wie oben ermittelt) die Ein-Stunden-Mittelwerte für die NO_2-Konzentrationen für den Zeitraum 01.01.2016 bis 31.12.2019 herunter und überführen Sie diese in einen DataFrame namens `data_no2`. Dieser soll die Spalten `STATION_ID`, `DT` und `NO2` besitzen, die die Stations-ID, das Messdatum mit Uhrzeit sowie die gemessene NO_2-Konzentration enthalten.

b) Setzen Sie den `dtype` der Spalte `NO2` auf `float` und wandeln Sie die Spalte `DT` in ein DateTime-Format um.

c) Entfernen Sie alle Zeilen, bei denen der Wert in der Spalte `NO2` fehlt. Geben Sie an, wie viele Zeilen dadurch entfernt wurden.

d) Entfernen Sie die Daten zu allen Stationen, die nicht für mindestens 95% der Messzeitpunkte im Auswertezeitraum einen gültigen Messwert enthalten.

e) Für wie viele Stationen enthält der DataFrame `data_no2` nun noch Daten?

f) Zu welchen der bayerischen Stationen enthält er keine Daten (mehr)? Geben Sie deren IDs und Namen aus.

Aufgabe 3 (Explorative Datenanalyse)

a) Welches ist der in den Jahren 2016-2019 höchste gemessene Ein-Stunden-Mittelwert für NO_2? Wann und an welcher Station wurde er gemessen?

b) An welchem Tag im Auswertezeitraum war die durchschnittliche NO_2-Konzentration über alle bayerischen Stationen am höchsten und welchen Wert hatte sie?

c) Ermitteln Sie die 10 höchsten Messwerte und die zugehörigen Messzeitpunkte für die Station in der Nikolaistraße in Weiden.

d) Berechnen Sie die Mittelwerte der gemessenen NO_2-Konzentrationen über die einzelnen Jahre. Wie haben sich diese zeitlich entwickelt? Unterscheiden Sie dabei auch nach dem Stations-Typ.

Aufgabe 4 (Verletzung der zulässigen NO_2-Grenzwerte)

a) Ermitteln Sie, an welchen bayerischen Stationen jeweils in den Jahren 2016-2019 Überschreitungen des Stundengrenzwerts (d.h. der Ein-Stunden-Mittelwert überschreitet $200\mu g/m^3$) gemessen wurden, und geben Sie an wie viele Überschreitungen es jeweils waren. Dieser darf innerhalb eines Jahres höchstens 18-Mal pro Station überschritten werden. Welche Stationen haben dieses Kriterium verletzt?

b) Ermitteln Sie, an welchen bayerischen Stationen und in welchen Jahren der Jahresmittelwert der Ein-Stunden-Mittelwerte die Grenze von $40\mu g/m^3$ überschritten hat und geben Sie die zugehörigen Jahresmittelwerte an.

Aufgabe 5 (Visualisierung)

a) Erstellen Sie ein Histogramm über alle gemessenen NO_2-Konzentrationen im Auswertungszeitraum 2016-2019.

b) Stellen Sie den jahreszeitlichen Verlauf der gemessenen NO_2-Konzentrationen in einem geeigneten Diagramm dar. Was ist zu beobachten und wie kann dies erklärt werden?

c) Visualisieren Sie in einem geeigneten Diagramm den Zeitverlauf der Tagesmittel der gemessenen NO_2-Konzentrationen im Beobachtungszeitraum. Lassen sich Trends erkennen?

Aufgabe 6 (Interaktives Diagramm)

a) Erzeugen Sie ein interaktives Säulendiagramm in `Plotly`, in welchem die Mittelwerte der NO_2-Konzentrationen im Tagesverlauf über die (vollen) Stunden aufgetragen werden. Verwenden Sie als Datengrundlage die Messwerte der bayerischen Stationen aus dem DataFrame `no2_data`. Das Diagramm soll zwei Radio-Buttons enthalten. Über den ersten Radio-Button kann der Stations-Typ gefiltert werden (Auswahlmöglichkeiten `all`, `background` und `traffic`), über den zweiten Radio-Button kann der Wochentag eingeschränkt werden (Auswahlmöglichkeiten `All`, `Monday`, ..., `Sunday`).

b) Analysieren Sie anhand des erstellten Diagramms den tageszeitlichen Verlauf der NO_2-Konzentration in Abhängigkeit des Stations-Typs und des Wochentags. Beschreiben und interpretieren Sie die beobachteten Zusammenhänge.

Aufgabe 7 (Abhängigkeit zwischen der Ozonkonzentration und der Temperatur)

a) Laden Sie sich die Wetterdaten aus der Datei `wetterdaten.csv` in einen DataFrame namens `df_weather`. Dieser enthält historische Wetterdaten für die Oberpfalz für die Jahre 2016-2019. Für unsere Analyse ist die Spalte `temperatureMax` relevant, die die Tageshöchsttemperaturen in Grad Fahrenheit beinhaltet.

b) Wandeln Sie die Temperaturwerte in Grad Celsius um.

c) Laden Sie über die Measurements-API für die Station in der Nikolaistraße in Weiden die Ein-Stunden-Mittelwerte für die Ozon-Konzentrationen für den Zeitraum 01.01.2016 bis 31.12.2019 herunter und überführen Sie diese in einen DataFrame namens `data_o3`. Dieser soll die Spalten `DT` und `O3` besitzen, die das Messdatum mit Uhrzeit (Beginn der Stunde, über die gemittelt wird) und die gemessene O_3-Konzentration enthalten.

d) Aggregieren Sie die Messwerte, indem Sie die O_3-Maximalkonzentrationen pro Tag ermitteln und diese in einen DataFrame namens `o3_data_max` speichern.

e) Stellen Sie den zeitlichen Verlauf der aggregierten Tageswerte für die Jahre 2016-2019 grafisch dar und beschreiben Sie die beobachteten Trends.

f) Erstellen Sie ein Streudiagramm, in dem die Maximalkonzentration (y-Achse) und die Maximaltemperatur (x-Achse) der einzelnen Tage für die Jahre 2016-2019 gegeneinander aufgetragen sind. Beschreiben und erklären Sie den beobachteten Zusammenhang.

g) Ermitteln Sie anhand der Tageswerte den empirischen Korrelationskoeffizienten zwischen der Maximalkonzentration und der Maximaltemperatur.

h) Erstellen Sie ein lineares Modell zur Modellierung der Beziehung zwischen den beiden Größen. Erzeugen Sie das Streudiagramm erneut und zeichnen Sie die berechnete Gerade mit ein.

Aufgabe 8 (Einfluss der Coronavirus-Pandemie auf die Luftqualität)
Untersuchen Sie anhand geeigneter Analysen den Einfluss des Lockdowns im Zusammenhang mit der Coronavirus-Pandemie auf die Luftqualität in Bayern. Stellen Sie Hypothesen auf und überprüfen Sie diese, indem Sie geeignete Luft-Messdaten akquirieren und auswerten. Stellen Sie Ihre Ergebnisse textuell und visuell dar.