Hochschule Bochum
Bochum University
of Applied Sciences

Fachbereich Wirtschaft
Prof. Dr. Christian Bockermann

Data Science 1

Wintersemester 2024 / 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 als PDF-Export Ihres Jupyter-Notebooks. Das PDF Ihres Notebooks muss bis spätestens 23:59 Uhr am 9.2.2025 in der zugehörigen Aufgabe im Moodle Kurs hochgeladen werden.
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.

In jedem Fall sind in der Hausarbeit am Anfang des Notebooks 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)

Streaming Dienste haben das lineare Fernsehen längst abgelöst. Dabei übernehmen große Streaming Anbieter teilweise auch die Rolle des Produzenten und entwickeln und produzieren eigene Shows und Filme.
Auf dem DataScience Notebook Server finden Sie das Modul **`netflix`**, das eine Funktion **`shows()`** bereitstellt, die eine Liste von Shows des bekannten Anbieters *Netflix* zurückliefert.
Jedes Element dieser Liste ist ein Tupel mit Kompontenten wie der *Show ID*, dem *Titel* und *Direktor*, sowie dem *Erscheiungsjahr* und dem *Cast* (Liste der Schauspieler). Die Struktur der Tupel hat also folgendes Schema:

```
(showId, art, titel, direktor, cast, land,
  erscheinungsjahr, bewertung, kategorien )
```

Es sei darauf hingewiesen, dass die Komponenten mit dem *cast* und den Kategorien (*kategorien*) selbst wieder Listen mit ggf. mehreren Werten sind (vgl. nachfolgendes Beispiel).

Hier ist ein kleines Beispiel, wie die Daten zu benutzen sind:

```
import netflix

sendungen = netflix.shows()

show = sendungen[1555]
# ( 's1556', 'TV Show', 'Grizzy et les Lemmings', '',
#   ['Pierre-Alain de Garrigues', 'Josselin Charier'],
#   'France', '2018', 'TV-Y', ["Kids' TV", 'TV Comedies'] )
```

Wie in dem Python Code zu sehen ist, ist die Sendung am Index 1555 eine TV Show mit dem Titel *Grizzy et le Lemmings* und der ID **`s1556`**. Der Name des Direktors ist nicht angegeben (leerer String), zwei Schauspieler tauchen im Cast auf. Die Sendung wurde in Frankreich produziert und erschien zuerst im Jahr 2018.
Das Rating für den Jugendschutz ist *TV-Y* (Y für *young*) und die Sendung befindet sich auf Netflix in den Kategorien *Kids' TV* und *TV Comedies*.

Für eine derartige Liste sollen Sie die folgenden Aufgaben lösen:

1. Schreiben Sie eine Funktion **`ratings(liste)`**, die als Parameter die obige Liste bekommt und die Menge der Ratings zurückgibt, für die es Filme/Shows in der Liste gibt. Dabei soll jedes Rating nur einmal in der Ergebnisliste vorkommen.

2. Schreiben Sie eine Funktion **`schauspieler(liste)`**, die als Paramter die obige Liste bekommt und die Menge (**`set`**) der Schauspieler zurückliefert, die in all diesen Filmen und Shows auftreten.

3. Schreiben Sie eine Funktion **`hat_schauspieler(show,name)`**, die für einen Eintrag (**`show`**) aus der Liste und einen Schauspielernamen überprüft, ob dieser Schauspieler in der Liste der Schauspieler enthalten ist. Welchen Rückgabetyp sollte diese Funktion sinnvollerweise haben?

4. Schreiben Sie eine Funktion `shows_mit(liste, schauspieler)`, die für die gegebene Liste und einen gegebenen Schauspielernamen alle Listeneinträge zurückliefert, in denen der Schauspieler mitgespielt hat.

5. Schreiben Sie eine Funktion `kategorien_mit(liste, schauspieler)`, die für die Liste und einen Schauspielernamen die Menge (`set`) der Kategorien der Filme/Shows zurückgibt, in denen dieser Schauspieler mitgewirkt hat.

6. Geben Sie eine Funktion `anzahl_nach_schauspieler(liste)` an, die eine Liste von Sendungen im obigen Format bekommt, und eine Liste mit Tupeln der folgenden Art als Ergebnis zurückliefert:

   ```
   [ (schauspieler, anzahlShows), ...]
   ```

   D.h. im Ergebnis steht jeweils ein Tupel aus dem Namen eines Schauspielers und der Anzahl an Shows, in denen er mitgewirkt hat.

7. Nutzen Sie die Funktion `anzahl_nach_schauspieler(..)` und berechnen Sie den/die Schauspieler/in, die in den meistens Shows mitgewirkt hat. (Das Ergebnis kann auch eine Liste mehrerer Schauspieler/innen sein, sofern es mehrere gibt, die in gleich vielen Shows maximal mitgespielt haben).

Aufgabe 2 (Pandas und Statistiken)

Neben Shows auf Streaming Portalen sind natürlich Kinos und Blockbuster ein wichtiger Teil der Unterhaltungsindustrie. Die *Internet Movie Database* (IMDB) ist eine Datenbank, die versucht, möglichst alle Filme inklusive der Regisseure, Schauspieler, usw. zu speichern und durchsuchbar zu machen.
Ein Teil der Datenbank ist öffentlich und dient als Grundlage für diese Aufgabe. Unter der URL

`https://data.hsbo.de/imdb/`

finden Sie eine Reihe von CSV-Dateien, die die Filme, Schauspieler und Regisseure der IMDB seit dem Jahr 2000 enthalten. Die Grundlage bildet die Datei `filme.csv`, die die folgende Struktur hat:

ID	Titel	Jahr	Länge (min)	Genre
215727	Downward Angel	2001	97.0	Thriller
232098	BattleQueen 2020	2001	95.0	Action,Sci-Fi
243585	Stuart Little 2	2002	77.0	Adventure,Comedy,Family
250698	Das Ritual - Im Bann des Bösen	2002	99.0	Horror
254279	Glissement de terrain	2001	nan	nan

Table 1: Die Datei `filme.csv`

Die Spalten *ID*, *Titel* und *Jahr* enthalten die ID des Films, den Titel und das Erscheinungsjahr. Die Spalte *Laenge* ist die Länge des Films in Minuten und *Genre* enthält eine Liste von Genres, denen der Film zugeordnet wurde. Wie man in der letzten Zeile sehen kann, gibt es Filme, für die z.B. die Länge oder das Genre nicht vorhanden sind.

Regisseur- und Produzenten-Daten

Wichtige Merkmale für Filme sind natürlich die Regiesseure und Schauspieler. Die Datei `regie.csv` enthält zu jedem Film einen oder mehrere verantwortliche Regisseure, die Datei `produzent.csv` die jeweiligen Produzenten:

FilmID	Regisseur
211946	Hannes Stöhr
426073	Zsombor Dyga
445692	Theresa Jessouroun
466231	Birta Frodadottir
815246	Brandon Schmid

FilmID	Produzent
983193	Steven Spielberg
997188	Markus Selin
10136634	Tyler Savino
10275534	Aaron Ryder
11090818	Arlington Gordon

Table 2: Die Tabellenstruktur der Dateien `regie.csv` und `produzent.csv`.

Die Spalten sind eigentlich selbsterklärend: die FilmID ist die ID, die auch in der Film-Tabelle vorkommt und angibt, zu welchem Film der Regisseur (Spalte *Regisseur*) gehört und die Spalte *Produzent* entsprechend den Namen des Produzenten zum Film mit der zugehörigen *FilmID*.

User-Bewertungen

Entscheidend für den Erfolg eines Films sind u.a. die Schauspieler und die Zufriedenheit der Zuschauer. Die Datei `schauspieler.csv` enthält die Schauspieler, die in einem Film mitgewirkt haben und zusätzlich noch das Geburtsjahr der Schauspieler (nicht für alle Schauspieler ist das Geburtsjahr bekannt).
Um die Zufriedenheit der Zuschauener zu messen gibt es die Möglichkeit Filme auf IMDB zu bewerten. Die Datei `bewertungen.csv` enthält für eine Vielzahl der Filme (nicht alle!) die durchschnittliche Punktzahl und die Anzahl der Bewertungen. Die Bewertungsskala reicht von 0 bis 10 Punkten:

FilmID	Name	Geburtsjahr
306561	Michelle Paradise	1972.0
307479	George Clooney	1961.0
313166	Caroline Hay	nan
314213	Valeria Bruni Tedeschi	1964.0
315963	Kara Zediker	1969.0

FilmID	Punktzahl	Bewertungen
192722	6.9	41
215727	4.5	141
250698	4.8	2161
261992	5.0	1509
274861	4.2	528

Table 3: Die Struktur der Tabellen `schauspieler.csv` und `bewertungen.csv`.

Sie sollen sich im Folgenden mit diesen Daten beschäftigen. Wieviel Filme gibt es in den Daten? In welchen Jahren sind die meisten Filme erschienen? Gab es während Corona oder kurz danach weniger Neuerscheinungen?
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 Filme und Schauspieler sind in den Daten enthalten?
   - Wieviele Bewertungen hat ein Film im Schnitt? Wieviele maximal?
   - Was sind die am besten bewerteten Filme?
   - Bei wie vielen Filmen war George Clooney als Produzent tätig?
2. Betrachten Sie das Erscheinungsjahr der Filme. Für welchen Zeitraum enthalten die Daten erschienene Filme? Erstellen Sie einen Plot, die für jedes Jahr die Anzahl der erschienenen Filme zeigt.
3. Es gab immer mal wieder externe Auswirkungen auf die Filmindustrie. Sicherlich zum einen Corona, zum anderen aber auch der Streik der Drehbuchautoren in Hollywood. Recherchieren Sie einen der beiden Zeiträume und schauen Sie ob vorher/während/nachher deutlich mehr/weniger Filme veröffentlicht wurden.

4. Komplexere Fragen lassen sich untersuchen, wenn die Tabellen miteinander verbunden werden. Dazu finden Sie weiter unten ein paar Details zur weitergehenden Recherche.

   Suchen Sie in dieser Aufgabe zunächst nach einem Produzenten, der mindestens 5 Filme produziert hat und berechnen Sie die durchschnittliche Bewertung, die die Filme dieses Produzenten erreicht haben.

5. Suchen Sie in den Daten nach einem Film ihrer Wahl, der mindestens eine Bewertung von 6,0 erreicht hat und in dem mehr als 5 Schauspieler mitwirken. Berechnen Sie für diesen Film das Durchschnittsalter der Schauspieler zum Zeitpunkt der Filmveröffentlichung.

6. Betrachten wir als nächstes den Schauspieler George Clooney. In wie vielen Filmen hat er insgesamt mitgewirkt? Wie viele Filme hat Herr Clooney im Zeitraum der Daten pro Jahr als Schauspieler mitgemacht?

Hinweis zur Bearbeitung

Die Vorlesung hat einige der Grundlagen zu Pandas vermittelt. Natürlich ist Pandas deutlich umfangreicher, als man es innerhalb einer 1-semestrigen Vorlesung vermitteln kann. Für die Lösung einiger dieser Aufgaben ist es daher erforderlich, sich weiter mit Pandas zu beschäftigen. Dazu gehört u.a. die Verbindung mehrerer Tabellen (JOIN), was Sie beispielsweise aus dem Bereich der Datenbanken in Wirtschaftsinformatik 2 bereits kennen sollten.
Die Dokumentation für den JOIN findet sich z.B. unter

`https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.join.html`

Es sei hier noch angemerkt, dass es hilfreich ist, wenn der *index* des DataFrames, den man an einen bestehenden DataFrame heften möchte, für den JOIN relevant ist. So sollte z.B. der DataFrame für die Produzenten als Index am besten die FilmID enthalten, bevor dieser an die Filme ge-joint wird.

Alternativ sei an dieser Stelle auf den in der Vorlesung erwähnten Foliensatz aus Wirtschaftsinformatik 2 hingewiesen, der das Zusammenführen von verschiedenen DataFrames behandelt. Der Foliensatz ist auf der DataScience 1 Vorlesungsseite zusammen mit der Hausarbeit verlinkt.

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.