Einführung in R für SozialwissenschaftlerInnen

Übungsblatt Woche 4

Vorgehensweise

Füge, wenn möglich, nur die Befehle in ihre Abgabe ein, die gefordert sind. Alle anderen Befehle können Dein Ergebnis verfälschen. Achte unbedingt darauf, keinen Plot-Befehl einzufügen!

Hintergrund

Ziel: Umgang mit Dataframes üben
In diesem Arbeitsblatt wirst du mit einem Auszug aus den Daten der Internet Movie Database arbeiten. Die IMDb ist eine von Amazon betriebene Datenbank mit Einträgen über mehr als 6 Millionen Filmproduktionen. Registrierte Nutzer können auf dieser Website Filme auf einer Skala von 1 (sehr schlecht) bis 10 (sehr gut) bewerten.

Der Originaldatensatz wurde erstellt, indem ein Programm am 01. Januar 2020 automatisch alle Einträge zu Filmen, zu denen mehr als 100 Bewertungen vorlagen, ausgelesen hat. Den originaldatensatz, der insgesamt 85.855 Filme umfasst, findest du unter diesem Link: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

Um Speicherplatz zu sparen gibt es in dieser Aufgabe nur einen kleinen Auszug aus dem Datensatz, der 328 Filme umfasst. Lade den Datensatz "Filmauswahl.csv" herunter und speichere ihn in deinem Ordnern.

Vorbereitung

Im Abgabesystem ist der Datensatz in einem Ordner mit den Namen "resources" gespeichert. Damit das Abgabesystem die Datei öffnen kann, musst Du folgenden Befehl an den Anfang deines Skripts setzen:

```
film<-read.csv("resources/Filmauswahl.csv", encoding="UTF-8")
```

Wenn du die Hausaufgaben in R-Studio machst musst du evtl den Pfad ändern und noch zuvor dein Working-Directory setzen. **Bevor du das Skript ins Abgabesystem hochlädst musst du diese Befehle wieder löschen. Im Skript sollen sich nur der Befehl, den du hier siehst und die Befehle für die Hausaufgaben, befinden.**

Der Datensatz enthält Daten zu einer internen ID, einer Datenbank-ID, Titel und Original-Titel, Jahr und Datum der Veröffentlichung, Genre, Filmlänge, Herkunftsland, Sprache, Regisseure, Drehbuchautoren, Produktionsfirma, Schauspieler, eine Beschreibung des Films, die durchschnittliche Bewertung, die Anzahl der Bewertungen sowie Budget und Einnahmen der Filme. Zusätzlich beinhalten der Datensatz den Metascore zwischen 0 und 100 (100=ausgezeichnete Qualität) der Website metacritic für diesen Film, die Anzahl der Nutzer, die diesen Film bewertet haben und die Anzahl externer Kritiken, die über diesen Film verfasst wurden.

a) Maximale Punktzahl: 1 Punkt

Nicht alle Daten sind für diese Aufgabe von Interesse. Speichere deshalb in der Variable film_a den Inhalt der Variable film, aber nur mit den folgenden Spalten: imdb_title_id, original_title, year, genre, duration, budget, avg_vote, metascore, reviews_from_users, reviews_from_critics. Achte darauf, dass die Spalten genau in dieser Reihenfolge aufgeführt sind.

b) Maximale Punktzahl: 1 Punkt

Prüfe, in wie vielen Spalten aus film_a fehlende Werte vorkommen und schreibe das Ergebnis in die Variable b.

c) Maximale Punktzahl: 1 Punkt

Berechne den Mittelwert aller Metascores aus film_a und speichere das Ergebnis in der Variable c. Fehlende Werte sollen nicht in die Berechnung miteinfließen.

d) Maximale Punktzahl: 2 Punkte

Entferne alle Zeilen aus film_a mit fehlenden Werte in der Spalte reviews_from_users und speichere das Ergebnis in der Variable film_d1. Entferne alle Zeilen aus film_a mit fehlenden Werte in der Spalte reviews_from_critics und speichere das Ergebnis in der Variable film_d2. Achte darauf, dass dabei keine Zeilen entfernt werden, die nur fehlende Werte in anderen Spalten haben. Verändere nicht die Variable film_a.

e) Maximale Punktzahl: 1 Punkt

Entferne alle Zeilen aus film_a, die in irgendeiner Spalte fehlende Werte haben und speichere das Ergebnis in der Variable film_e. Verändere nicht die Variable film_a.

f) Maximale Punktzahl: 2 Punkte

Im Jahr 2001 ist der 10-minütige Kurz-Film “Staplerfahrer Klaus” erschienen. Der Film hat auf IMDb ein durchschnittliches Voting von 7,8, sowie 20 User reviews und 13 externe Kritiken. Die Werte für den Metascore und das Budget fehlen. Kopiere den Datensatz film_a in die Variable film_f. Füge am Ende des Datensatzes film_f eine neue Zeile für den Staplerfahrer Klaus hinzu. Nutze hierbei eine imdb-title-id von “tt8820590” und das Genre “Lehrvideo”. Achte darauf, dass die Zeilen anschließend noch die gleichen Datentypen haben, wobei eine Veränderung von int zu num jedoch in Ordnung ist.

g) Maximale Punktzahl: 2 Punkte

Kopiere den Datensatz film_a in die Variable film_g. Entferne anschließend die dreihundertste Zeile aus film_g. Achte darauf, dass am Ende die Zeilenpositionen mit den Zeilennamen wieder übereinstimmen.

h) Maximale Punktzahl: 2 Punkte

Kopiere den Datensatz film_a in die Variable film_h. Füge eine neue Spalte am Ende des Datensatzes film_h hinzu. Die Spalte soll duration_hours heißen und die Länge des Filmes in Stunden (gerundet auf eine Nachkommastelle) enthalten. Entferne anschließend die Spalte duration aus dem Datensatz film_h.

i) Maximale Punktzahl: 2 Punkte

Kopiere den Datensatz film_a in die Variable film_i. Wandle die Spalte avg_vote in film_i in den Datentyp numeric um. Achte dabei darauf, die Kommata durch Punkte zu ersetzen.

j) Maximale Punktzahl: 2 Punkte

Speichere den Originial-Titel des am besten bewertetsten Filmes aus film_i in der Variable best. Speichere den Originial-Titel des am schlechtesten bewertetsten Filmes aus film_i in der Variable worst. Die Bewertung soll anhand der Spalte avg_vote getroffen werden. Achte darauf, dass die Variablen jeweils eine Zeichenkette enthalten sollen und keinen Datensatz.

k) Maximale Punktzahl: 1 Punkt

Kopiere den Datensatz film_a in die Variable film_k. Wandle die Spalte genre in film_k in den Datentyp factor um. Du kannst davon ausgehen, dass im Datensatz alle Kategorien vertreten sind.

l) Maximale Punktzahl: 3 Punkte

Arbeite mit dem Datensatz film_k:

- Finde den Originaltitel und das Durchschnitts-Voting aller Filme mit dem Genre "Adventure" und speichere das Ergebnis in der Variable l_adv.
- Finde den Originaltitel aller Animationsfilme, die länger dauern als 90 Minuten und speichere das Ergebnis in der Variable l_ani.
- Finde den Originaltitel aller Actionfilme, die im Jahr 2006 oder 2016 erschienen sind und speichere das Ergebnis in der Variable l_act.

Achte darauf, dass deine Variablen Dataframes enthalten.