כריית נתונים ב- R // תרגיל בית מספר 1

בתרגיל זה נשתמש במסד ה- IMDB, המכיל את שני קבצי הנתונים שהוצגו בכיתה:

IMDB_movies.csv

IMDB_players.csv

1. המירו את עמודת budget (ב imdb.movies) לעמודה נומרית (בדומה לקוד הנלמד).
2. (2 נק') צרו עמודה חדשה בשם is.over.1m, המכילה TRUE במידה והתקציב (budget) גדול או שווה למיליון, אחרת FALSE. הדפיסו לפלט את שורת הפקודה.
מצאו באמצעות פקודת table כמה רשומות גדולות/ קטנות ממיליון. הוסיפו לפונקציית table את הפרמטר useNA עם ערך "ifany". ציינו לכמה רשומות לא מופיע תקציב (budget)? הדפס לפלט.
3. (3 נק') חשבו באמצעות פקודת table מספר בעלי תפקידים מכל סוג (actor, director וכו')

לפי סרט. הכניסו את החישוב למשתנה מסוג data.frame בשם n.actors.

הדפיסו את 6 השורות הראשונות של n.actors.

דוגמא לשורת פלט:

```
> head(n.actors)
                Actor Cinematographer Composer Director Producer Writer
10000bc.htm         0               0        0        1        2      2
```

4. (2 נק') חשבו קורלציה בין budget ל- total gross, עבור סרטים להם התקציב וההכנסות מדווחים. השתמשו בפונקציית cor. הדפיסו את הקורלציה לפלט.
5. (3 נק') חשבו את ההכנסות (total.gross) הממוצעות פר דירוג (MPAA.rating) וז'אנר רק לז'אנרים מסוג "Crime Comedy". השתמשו בפונקציית aggregate. הדפיסו לפלט עד שש שורות ראשונות.

כעת בצעו זאת באמצעות ChatGPT. ציינו בקובץ הפלט מה ביקשתם ומה קיבלתם. האם דרך הפעולה והתוצאות זהות? ציינו באיזו גירסה של הצ'אט השתמשתם.

<u>אופן הגשה:</u>

- ✓ הגשה דרך אתר למידה
- ✓ הגשה בזוגות או ביחידים (רק אחד מבני הזוג צריך להגיש באתר. על הקובץ יופיעו השמות של שני המגישים)
- ✓ יש להגיש קובץ R (אחד) **מתועד**, וקובץ **פלט** בפורמט word/ pdf המכיל את התשובות לשורות המסומנות בצהוב