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. המר את עמודת budget (ב imdb.movies) לעמודה נומרית (בדומה לקוד הנלמד).
2. הוסף לקובץ הקוד (בחלקו העליון) את השורה הבאה - options(scipen = 999) (לצורך תצוגה קריאה יותר של המספרים) והרץ אותה.
3. (2 נק') צור עמודה חדשה בשם is.over.1m, המכילה TRUE במידה והתקציב (budget) גדול או שווה למיליון, אחרת FALSE. הדפס לפלט את שורת הפקודה.
4. (1 נק') כמה רשומות בעלות תקציב גדול ממיליון? כמה רשומות בעלות תקציב קטן ממיליון? מצא באמצעות פקודת table. הוסף לפונקציית table את הפרמטר useNA עם ערך "ifany". לכמה רשומות לא מופיע תקציב (budget)? הדפס לפלט.
5. (2 נק') חשב באמצעות פקודת table מספר בעלי תפקידים מכל סוג (actor, director וכו')

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

הדפס 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
```

6. (3 נק') חשב קורלציה בין budget ל- total gross, עבור סרטים להם התקציב וההכנסות מדווחים. השתמש בפונקציית cor. הדפס את הקורלציה לפלט.
7. (2 נק') חשב את התקציב (budget) הממוצע לכל ז'אנר רק לז'אנרים מסוג "Action" או "Comedy". השתמש בפונקציית aggregate (יש לחפש ברשת כיצד רושמים את האופרטור or). הדפס לפלט את התוצאה.

<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