שאלה מספר 7

סעיף 1

שלב 1 – בחירת 10 סרטים ישראליים

בחרתי 10 סרטים ישראליים ידועים, כך שיש סיכוי גבוה שחלקם יהיו ב־DB (אבל ייתכן שלא כולם קיימים):

1. Waltz with BashirWaltz with Bashir (ואלס עם באשיר), 2008
2. Beaufort (בופור), 2007
3. The Band's Visit (ביקור התזמורת), 2007
4. Aviva My Love (אביבה אהובתי), 2006
5. Nina's Tragedies (האסונות של נינה), 2003
6. Yossi & Jagger (יוסי וג׳אגר), 2002
7. Late Marriage (חתונה מאוחרת), 2001
8. Campfire (מדורת השבט), 2004
9. Walk on Water (ללכת על המים), 2004
10. Ushpizin (אושפיזין), 2004

שלב 2 – בדיקת קיום במערכת

חיפוש ישיר לפי שם הסרט (movies.name)

```
SELECT id, name, year
FROM movies
WHERE lower(name) IN (
  'waltz with bashir','beaufort','the band''s visit',
  'aviva my love','nina''s tragedies','yossi & jagger',
  'late marriage','campfire','walk on water','ushpizin'
);
```

תוצאה:
נמצאו רק 2 אופציות מתוך 10.

id	name	year
358586	Walk On Water	2004
372024	Yossi & Jagger	2002
NULL	NULL	NULL

חיפוש לפי שם + שם חלופי של במאים ישראליים (אם יש טעויות בשם הסרט)

- כאן נצטרך להכניס את שמות הבמאים האמיתיים של הסרטים, כדי לאתר התאמות גם אם שם הסרט רשום אחרת.

```
SELECT m.id, m.name, m.year, d.first_name, d.last_name
FROM movies m
JOIN movies_directors md ON md.movie_id = m.id
JOIN directors d ON d.id = md.director_id
WHERE lower(concat(d.first_name,' ',d.last_name)) IN (
  'ari folman','joseph cedar','eran kolirin','savi gabizon',
  'savi gabizon','eytan fox','dover kosashvili',
  'joseph cedar','eytan fox','gidon raf','gidi dar'
);
```

תוצאה: נמצאו 2

id	name	year	first_name	last_name
144207	Hesder, Ha-	2000	Joseph	Cedar
209521	Medurat Hashevet	2004	Joseph	Cedar
65170	Clara Hakedosha	1996	Ari	Folman
199505	Made in Israel	2001	Ari	Folman
7794	After	1990	Eytan	Fox
25912	Ba'al Ba'al Lev	1997	Eytan	Fox
299241	Shirat Ha'Sirena	1994	Eytan	Fox
358586	Walk On Water	2004	Eytan	Fox
372024	Yossi & Jagger	2002	Eytan	Fox
388943	"Florentine 1995"	1997	Eytan	Fox
193959	Long Journey, The	2004	Eran	Kolirin

שלב 3 – איך להתייחס לסרטים שלא קיימים?

1. **ניסיון לאיתור עקיף**
לעיתים סרט לא מופיע בשם, אבל כן דרך במאי (directors) או שחקנים (actors + roles).
נבדוק גם את הטבלאות האלה כדי לוודא אם הסרט באמת חסר או רק לא מתויג נכון.

2. **לא להמציא רשומות**
אם סרט באמת חסר, לא מוסיפים אותו מלאכותית - אלא מתעדים את החוסר ומבצעים ניתוח על בסיס מה שקיים.
לכל היותר ניתן למזג (merge) מידע ממקורות חיצוניים או ליצור mapping, אבל לא להוסיף "דאטה מומצא".

3. **בדוחות - חישוב שיעור הכיסוי (coverage)**
מציינים במפורש: מתוך רשימת 10 סרטים ישראליים שנבחרו, כמה נמצאו וכמה חסרים.
לדוגמה: "נמצאו 2 מתוך 10 → כיסוי 20%".
זה מבהיר למשתמשים שההמלצות מבוססות על דאטה חלקי, ונותן הקשר לתוצאות.

<u>לסיכום:</u>
כאשר סרטים חסרים בבסיס הנתונים, אין זה אומר שיש טעות, אלא שמדובר במגבלת ייצוג. כלומר, יש לגיטימציה לייצור בסיס נתונים מסרטים אמריקאיים בלבד בכדי לשמור על טוהר המידע ואמינות המידע.

לכן אנו מתייגים את הסרטים החסרים בטבלה נפרדת, מנסים לאתרם גם בעקיפין דרך במאים או שחקנים, ולא מוסיפים רשומות מומצאות.

בנוסף, אנו מחשבים את שיעור הכיסוי (coverage) ומציינים אותו בדוחות - לדוגמה, 2 מתוך 10 סרטים ישראליים נמצאו - כדי שהמשתמש יבין שההמלצות מבוססות על מידע חלקי ולא יוסקו מסקנות מוטעות.

סעיף 2

במערכת נתונים כמו שלנו, העובדה שיש רשומה ב־movies לא בהכרח מבטיחה שמדובר בסרט אמיתי. יכולות להיות טעויות כתיב, כפילויות, סרטים מומצאים, או רשומות שגויות שהוזנו בטעות.

בנוסף, ייתכן שחלק מהרשומות ב־movies מייצגות סרטים שבפועל אין להם שום קשרים לנתונים אחרים במערכת — למשל, אין להם במאי (movies_directors), אין שחקנים (roles), ואין ז'אנר (movies_genres). זה סימן חזק לכך שהסרט "לא קיים" בפועל או שהנתונים עליו חסרים בצורה קיצונית.

<u>איך לאתר סרטים כאלה?</u>

רק נציין שזהו מצב בעייתי, ניתן לדגום כמה סרטים ולמצוא לקיומם אישוש חיצוני. זה לא יאתר את כל הסרטים שלא קיימים אבל כן יאפשר לנו לקבל יותר מידע על ההסתברות לקיומם.

1. **איתור סרטים "מנותקים"**
נבדוק אילו סרטים בטבלת movies לא מופיעים באף טבלת קשר אחרת (movies_directors, roles, movies_genres). סרט כזה מנותק לגמרי ממידע נוסף - וזה מעלה חשד לאמינותו.

2. **איתור כפילויות עם שמות דומים מאוד**
סרטים עם שם זהה או כמעט זהה אבל עם מזהי id שונים, במיוחד אם אין הבדלים משמעותיים בשנה או בפרטים, יכולים להיות כפילות של אותה רשומה.

סעיף 3

```
SELECT m1.name, m1.year AS year1, m2.year AS year2

FROM movies m1

JOIN movies m2

  ON m1.name = m2.name

 AND m1.id < m2.id

WHERE ABS(m1.year - m2.year) = 1;
```

תוצאה:

2759 row(s) returned

סיבה אפשרית לכך: ייתכן שמדובר בהמשכים מהירים, רימייקים, או טעויות הקלדה בשנה או שפשוט הסרט הוזן פעם שנייה בטעות.
כדי לקבוע אם זו כפילות או יצירה אחרת, משווים בין הבמאים, השחקנים והז'אנרים.
קיום רשומות כאלה יכול לבלבל את מערכת ההמלצות ולגרום לה לחשוב שמדובר באותו סרט או להמליץ פעמיים על משהו דומה מאוד.

סעיף 4 – סרטים פוליטיים

החלטה האם להחשיב נאום כסרט תשפיע ישירות על אופן הצגת הנתונים והניתוח שלהם. ההחלטה היא של מקים בסיס הנתונים, ויש החלטות לגיטימיות רבות. לאור השפעת ההחלטה, כדאי שהיא תוצג ותהיה מבוססת על כללים ברורים.
אם נאום עומד במרכז והופך ליצירה קולנועית שלמה, כמו במקרה של תיעוד נאומי תעמולה, הרי שהכללתו כסרט תשקף את חשיבותו התרבותית וההיסטורית ותתרום לסטטיסטיקות של הפקות קולנועיות.
לעומת זאת, כאשר מדובר בנאום שהוא רק חלק מסצנה בתוך יצירה רחבה יותר, כמו נאום של מרטין לות'ר קינג המשולב בסרט על תקופה מסוימת, הכללתו כסרט עצמאי עלולה לעוות את התמונה הכוללת וליצור רושם מוטעה לגבי כמות וסוג היצירות. לכן, ההחלטה אם לכלול נאומים מוסרטים כסרטים או לא קובעת אם ניתוח הנתונים יתמקד בהיסטוריה קולנועית נקייה או יכלול גם תיעוד מצולם בעל אופי אחר.

סעיף 5

התשובה מאוד תלויה בצורך , לכן ככלל אצבע : "לא אומרים לא לדאטה".

לפעמים אנחנו פוגשים הרבה גרסאות שנראות שונות מבחינה טכנית (קובץ שונה, hash שונה) אבל בפועל מתנהגות בדיוק אותו הדבר. זה כמו להוציא את אותו הסרט עם עטיפה אחרת - אין באמת שינוי בתוכן.

במצב כזה עדיף לשמור "ישות־אב" אחת שמייצגת את הפונקציונליות, ולחבר אליה את כל הווריאנטים הדומים. נפריד לגרסאות שונות רק אם באמת השתנה משהו מהותי - למשל נוספו פיצ'רים, השתנה API, או יש הבדל משמעותי בביצועים.

כך אנחנו שומרים על הנתונים נקיים, חוסכים מקום, ומקבלים המלצות וניתוחים הרבה יותר מדויקים - בלי רעש מיותר.