**פרויקט - בינה עסקית**

**מגישות :**

**דין חחיאשווילי 311558183**

**אלמוג דרור 209139682**

**גל שטיינברג 312407604**



## שאלה עסקית + KPIs

* המטרה הפרויקט היא לקבוע את הערך והתעריף שהמשתמשים מעניקים לסרטים המקוריים ביחס לסרטים הנרכשים, ולהבין אם ישנה הבדל משמעותי בדירוגים ובהערכה של המשתמשים בין השניים.

בנוסף , אנו נרצה לבדוק את התפלגות הדירוגים וההתרשמות של המשתמשים ולהשוות בין הדירוגים של סרטים מקוריים לסרטים שנרכשו, על מנת להבין איזה סוג של תוכן מקבל יותר הערכה מהמשתמשים ולעזור בהחלטות עסקיות נוספות בעתיד בנוגע לרכישת תוכן נוסף או ייצור תוכן עצמאי.

* השאלה העסקית : האם סרטים מקוריים של נטפליקס שיצאו בשנת 2021 מדורגים גבוה יותר ע"י משתמשים מאשר סרטים שנרכשו ע"י נקפליקס בשנת 2021.

KPI :

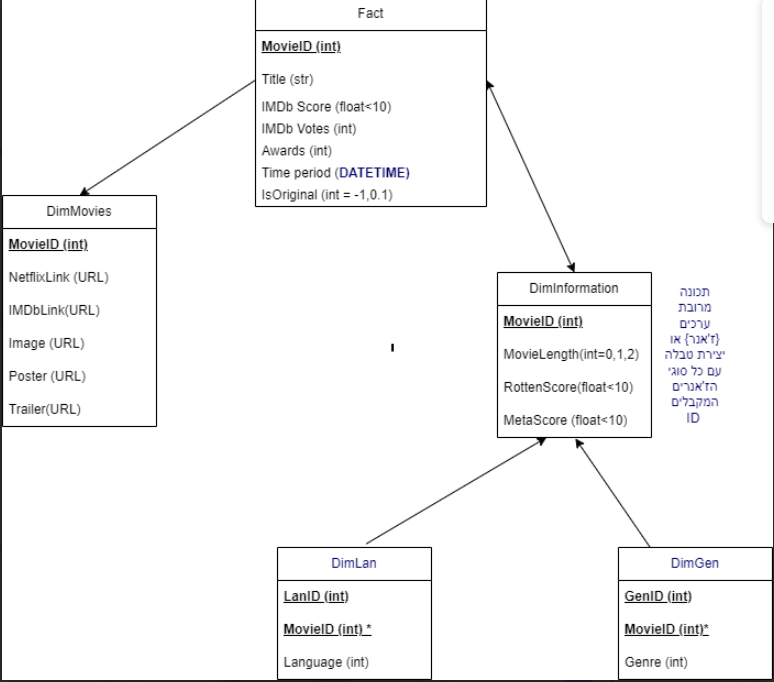
1. דירוג ממוצע של סרטים מקוריים של Netflix שהופקו ב-2021 לעומת סרטים שנרכשו ב-2021.
2. אחוז הסרטים המקוריים של נטפליקס שהופקו ב-2021 שקיבלו דירוג מעל רף מסוים בהשוואה לאחוז הסרטים שנרכשו ב-2021 שקיבלו דירוג מעל אותו סף.

KPI זה הוא ספציפי מכיוון שהוא מתמקד בהיבט ברור ומוגדר היטב של השאלה העסקית. זה ניתן למדידה מכיוון שניתן לכמת אותו באמצעות מערכי הנתונים הזמינים. זה בר השגה כי מערכי הנתונים מספקים דירוגים הן לסרטים מקוריים והן לסרטים שנרכשו. זה רלוונטי מכיוון שהוא קשור ישירות לשאלה העסקית, והוא מוגבל בזמן מכיוון שהוא מתמקד בשנת 2021.

## הגדרת Data Warehouse

1. בחרנו בסכמת Snowflake עבור אוסף הנתונים שלנו משום שהיא מספקת מבנה פשוט ומאורגן לנתונים, ידועה באופטימיזציה של ביצועי שאילתות וניתנת להרחבה ולהתאמה לצרכים עסקיים משתנים.

הבחירה שלנו במודל נתונים snowflake היא יעילה ועוזרת בקבלת החלטות למשתמש- כאשר יצרנו חלוקה לתתי טבלאות- בצורה היררכית, טבלת ז'אנרים, וטבלת שפות, שהיא תחת טבלת על מנת להקל על משתמש לבצע שליפה מהירה יותר של הסרטים באמצעות סינון ושליפה מהירה לפי פרמטרים אלו.

2. ERD

3.הציגו שני use-case המתארים מדוע הבחירה שלכם היא יעילה ועוזרת בקבלת החלטות למשתמש. למשל: הפחתת joins, שליפה מהירה יותר, וכו׳.

(לא מצאתי)

Use-case 1: המלצות על סרטים מבוססי ז'אנר-

המבנה ההיררכי של סכימת Snowflake, עם טבלה נפרדת לז'אנרים, מאפשר המלצות סרטים יעילות וממוקדות על סמך העדפות המשתמש. על ידי הימצאות אינפורמציה על הז'אנר בטבלה נפרדת, ניתן לשלוף סרטים השייכים לז'אנרים ספציפיים ללא צורך בפעולות Join מורכבות או חיפוש בדאטה כולו

Use-case 2: נגישות לסרטים מבוססי שפה

בעזרת טבלה נפרדת של שפות בתוך סכימת Snowflake מתאפשר סינון ושליפת נתונים יעיל ומהיר של סרטים על סמך העדפות שפה. משתמשים המעדיפים סרטים בשפות ספציפיות יכולים לגשת במהירות ולקבל החלטות לגבי הסרטים הזמינים בשפות המועדפות עליהם.

## תהליך הETL

טבלת Fact

|  |  |  |
| --- | --- | --- |
| **Source** | **Target** | **Mapping** |
| Title | Movie Title | * **הסרת רווחים מיותרים strip** * **המרה לאותיות קטנות lower** * **הפעלת פונקציית title** |
| IMDb Score | IMDb Score | * להעביר כמו שהוא * לשים -1 במקום. ואז לשלוף ערכים חיוביים. |
| IMDB Votes | IMDB Votes | * להעביר כמו שהוא |
| Awards Received | Awards | * להעביר כמו שהוא * לשים -1 במקום. ואז לשלוף ערכים חיוביים. |
| Production House | isOriginal | **ערך 1-** לתאים ריקים (לא ידוע)  **ערך 0** לתאים בהם נטפליקס לא שותפה   * **ערך 1** לתאים בהם נטפליקס שותפה |
| Release Date | TimePeriod | * נמיר את התאריך לפורמט:   **DD/MM/YYYY**   * נפצל לפי ״/״ וניקח את הערך האחרון (YYYY) * המרה לbins |

טבלת DimMovies

|  |  |  |
| --- | --- | --- |
| **Source** | **Target** | **Mapping** |
| Netflix Link | Netflix Link | * להשאיר |
| IMDb Link | IMDb Link | * **להשאיר כמו שהוא** |
| Image | Image | * **להשאיר כמו שהוא** |
| Poster | Poster | * **להשאיר כמו שהוא** |
| IMDb Trailer | Trailer | * **להשאיר כמו שהוא** |

DimInformation טבלת

|  |  |  |
| --- | --- | --- |
| **Source** | **Target** | **Mapping** |
| Genre | Genre | * שליפת כל סוגי הז'אנרים הקיימים בדאטה פעם אחת * הסרת רווחים מיותרים strip * המרה לאותיות קטנות lower * הפעלת פונקציית title * יצירת טבלה ונתינת אינדקס לכל ז'אנר |
| Language | Language | * שליפת כל סוגי השפות הקיימים בדאטה פעם אחת * הסרת רווחים מיותרים strip * המרה לאותיות קטנות lower * הפעלת פונקציית title * יצירת טבלה ונתינת אינדקס לכל שפה |
| RunTime | MovieLength | * הערך 0 לסרטים בין 30-60 דקות, הערך 1 לסרטים בין שעה לשעתיים, ערך 2 לסרטים מעל שעתיים |
| RottenScore | Rotten Tomatoes | * לחלק את העמודות ב10 |
| MetaScore | Metacritic Score | * לחלק את העמודות ב10 |

## ניתוח הData Warehouse

10 שאילתות SQL :

1. שאילתה הבודקת דירוגים ממוצעים לסרטי נטפליקס מקוריים שיצאו עד 2021:

SELECT AVG(`IMDb Score`) OVER () AS avg\_original\_score

FROM fact

WHERE isOriginal = 1 AND `Time Period` <= '2021';

2. שאילתה השולפת כמה סרטים מקוריים של נטפליקס זכו בפרסים לעומת סרטים שנרכשו?

SELECT isOriginal,

COUNT(\*) OVER (PARTITION BY isOriginal) AS Total\_Count,

COUNT(\*) OVER (PARTITION BY isOriginal, Awards\_Received > 0) AS Award\_Winning\_Count

FROM

fact

WHERE

isOriginal = '1' OR Production\_House = '0'

3. שאילתה הבודקת דירוגים ממוצעים לסרטי נטפליקס לא מקוריים שיצאו עד 2021:

SELECT AVG(`IMDb Score`) OVER () AS avg\_original\_score

FROM fact

WHERE isOriginal =0 AND `Time Period` <= '2021

4. שאילתה הבודקת את ממוצע כמות ההצבעות לסרטים מקוריים

Select AVG(IMDb Votes) OVER (PARTITION BY IsOriginal)

From fact

Where IsOriginal='1'

5. שאילתה הבודקת את ממוצע כמות ההצבעות לסרטים לא מקוריים

Select AVG(IMDb Votes) OVER (PARTITION BY IsOriginal)

From fact

Where IsOriginal='0'

6. שאילתה המחלקת לאחוזונים לפי הדירוג, ובודקת באחוזון העליון כמה מהסרטים הם מקוריים.

NTILE

SELECT `Movie Title`, `IMDb Score`,

NTILE(4) OVER (ORDER BY `IMDb Score`) AS percentile\_bucket

FROM fact

Where isOriginal = '1'

7. 10 הסרטים המקוריים של נטפליקס המובילים בדירוג IMDb

SELECT `Movie Title`, `IMDb Score`

FROM(

SELECT `Movie Title`, `IMDb Score`, ROW\_NUMBER()

OVER (ORDER BY `IMDb Score` DESC) AS rn

FROM fact

WHERE isOriginal = '1' AND `Time Period` <= '2021'

) subquery

WHERE rn <= 10;

8. שאילתה השולפת את שני סרטים (מקורי ולא מקורי) באותה השפה שיצאו באותו שנה המדורגים הכי גבוה.

SELECT `Movie Title`, `Release Year`, `IMDb Score`,

LAG(`IMDb Score`) OVER (PARTITION BY isOriginal ORDER BY `Release Year`) AS previous\_year\_score

FROM your\_table\_name

WHERE isOriginal = 'Yes'

ORDER BY `Release Yea

9. שאילתה הבודקת את מגמת הדירוג של סרטים מקוריים לפי שנים.

SELECT `Movie Title`, `Release Year`, `IMDb Score`,

LAG(`IMDb Score`) OVER (PARTITION BY isOriginal ORDER BY `Release Year`) AS previous\_year\_score

FROM your\_table\_name

WHERE isOriginal = 'Yes'

ORDER BY `Release Year`;

10. שאילתה הבודקת האם סרטים מקוריים שזכו בפרס הם בעלי דירוג גבוה.

SELECT

`Movie Title`,

`IMDb Score`,

CASE WHEN `IMDb Score` >= AVG(`IMDb Score`) OVER (PARTITION BY isOriginal, hasAward) THEN 'High Rating'

ELSE 'Low Rating'

END AS rating\_category

FROM your\_table\_name

WHERE isOriginal = 'Yes'

AND hasAward = 'Yes'

## מסקנות

1. ממוצע הדירוגים לסרטים מקוריים של נטפליקס שיצאו עד 2021 גבוה יותר מממוצע הדירוגים לסרטים שנרכשו. הממוצע המדורג של סרטים מקוריים עשוי לשמש כמדד לכימיה חיובית של המשתמשים והערכה גבוהה יותר שלהם כלפי סרטים מקוריים.

2. כמות הסרטים המקוריים שזכו בפרסים יכולה לשמש כקרדיט לאיכותם והצלחתם של סרטים מקוריים של נטפליקס. אם כמות הסרטים המקוריים שזכו בפרסים גבוהה מסך הכול, זה מראה על הצלחה וערך המשקיעים שהסרטים מקבלים מהתעשייה והקהל.

3. ממוצע הדירוגים לסרטים לא מקוריים של נטפליקס עשוי להיות נמוך יותר מהדירוגים לסרטים מקוריים. ההבדל בדירוגים עשוי להעיד על איכות והערכה נמוכה יותר של הסרטים שנרכשו במולדת ולהעדיף סרטים מקוריים.