פרוייקט חלק א' – DESIGN

מגישים: איתי בכר 206948218

אלעד וינר 209639855

נתחיל בכך שניתן לחלק ולשמור את הdata בכל אתר בהמון דרכים שונים ומגוונים ולעשות אנליזות שונות (כל דרך והיתרונות והחסרונות שלו).
לדוגמה: הדרך שבו אנחנו בחרנו לעשות את חלק 2 (data analysis) הוא עפ"י עיר מגורים של המשתמשים, באותו אופן ניתן לעשות עפ"י עיר חיפוש בשאילתות של המשתמשים.
הסיבה שאנחנו בחרנו להשתמש בעיר המגורים היא מכיוון שלפי תובנה 1 (#insight1), ראינו שיש קשר הדוק מאוד בין עיר המגורים לעיר החיפוש (משתמשים מקיבוץ גשר בקורלציה עם טבריה).
לדוגמה יותר מ-85% מהמשתמשים שגרים בתל אביב מחפשים סרטים שזמינים בתל אביב -
$\frac{21151}{21151+3688} * 100 = 85.152$

כמו בחלק 1 נניח שסוגי השאילתות המופיעים בטבלת queries הם הסוגים היחידים הקיימים ועל סמך ההנחה הנ"ל נוכל לבצע אנליזות על הdata ולחלק לפרגמנטציות על פי הרלציות: movies & credits.

בתור התחלה ביצענו כמה אנליזות מקדימות להבנת ההתפלגויות של הdata בנוגע למשתמשים, שאילתות והזמנת כרטיסים (ראינו כי ההתפלגויות הן זהות: 30.5%~ ירושלים, 5.5%~ קיבוץ גשר, 25% תל אביב, 21.5% חיפה, 17.5% אילת).

לאחר מכן ביצעתי בדיקות עבור כל עיר מגורים על אחוז החיפושים לפי שדות מסויימים, מה שעוזר לנו לחלק לפי חלוקה אנכית.

<u>תובנה מספר 1 (#insight1):</u>
בדקנו עבור כל המשתמשים מעיר מגורים מסוים לפי איזה ערים המשתמשים מחפשים וראינו כי הרוב המכריע מחפש סרטים לפי עיר מגורם (משתמשים מקיבוץ גשר מחפשים באתר של טבריה), כלומר נבצע חלוקה אופקית ונשמור בכל אתר את הסרטים המוקרנים באותה עיר. בנוסף, ראינו כי משתמשים הגרים באילת תמיד מחפשים לפי העיר אילת ולכן נוכל למחוק באתר של אילת את העמודה cities המציינת באיזה ערים הסרט זמין (הסרט יהיה זמין באילת עבור חיפוש באתר של אילת). בנוסף, נשים לב שמשתמשים מירושלים מחפשים 60% מהשאילתות שלהם לפי ירושלים ותל אביב ביחד ורק 40% ירושלים, לכן נשמור גם את הסרטים המשודרים בתל אביב באתר של ירושלים.

<u>תובנה מספר 2 (#insight2):</u>
בדקנו עבור כל המשתמשים מאילת את אחוז השאילתות לפי שחקנים וראינו כי אף אחד לא חיפש לפי שחקנים בכלל ולכן נוכל להוריד את העמודה cast (actors) מהאתר של אילת.

<u>תובנה מספר 3 (#insight3):</u>
בדקנו עבור כל המשתמשים משאר הערים (בלי אילת) את אחוז השאילתות לפי שחקנים וראינו כי המשתמשים שגרים בירושלים משתמשים בשדה באופן תדיר (יותר מ95% מהשאילתות) בחיפוש לפי שחקנים, ומשתמשים שגרים בתל אביב, חיפה וקיבוץ גשר מחפשים באופן פחות תדיר (50%~) אך עדיין משמעותי, ולכן נחליט לשמור את השדה cast (actors) בכל ארבעת האתרים (בלי אילת).

<u>תובנה מספר 4 (#insight4):</u>
בדקנו עבור כל המשתמשים לפי עיר מגורים את אחוז השאילתות לפי במאיים וראינו כי המשתמשים שגרים בחיפה משתמשים בשדה באופן תדיר (91.91% מהשאילתות) בחיפוש לפי במאיים, ואילו משתמשים

משאר הערים משתמשים בשדה באופן משמעותי פחות (3%~ מהשאילתות עבור כל עיר). לכן נשמור את השדה crew (directors) באתר של חיפה ושאר האתרים יגשו אליו.

תובנה מספר 5 (#insight5):
בדקנו עבור כל המשתמשים לפי עיר מגורים את אחוז השאילתות לפי מדינות שהסרטים הופקו בהן וראינו כי המשתמשים שגרים באילת, קיבוץ גשר וירושלים משתמשים בשדה הנ"ל תמיד (100% מהשאילתות), ואילו משתמשים מחיפה ומתל אביב משתמשים בשדה באופן משמעותי פחות (3%~ מהשאילתות). לכן נרצה לשמור את השדה production_countries בכל שלושת האתרים – אילת טבריה וירושלים (כלומר נשכפל את השדה), וחיפה תיגש לאתר של טבריה, תל אביב לאתר של ירושלים.

עד כאן התובנות לחלוקות אנכיות, כעת נעבור לתובנות לחלוקות אופקיות.

תובנה מספר 6 (#insight6):
בדקנו את השפות שהכי הרבה חיפשו על פיהן בכל עיר, ראינו שהשפה השנייה שהכי הרבה חיפשו לפיה (התעלמנו מאנגלית כי כולם חיפשו לפי אנגלית) היא השפה העברית ע"י משתמשים מירושלים (מהשפה השלישית ואילך אין משהו משמעותי). לאחר מכן ראינו כי שאר המשתמשים מחפשים הרבה פחות בשפה העברית (98.8% מחיפושי הסרטים בשפה העברית נעשו ע"י המשתמשים מירושלים) ולכן נוכל לבצע חלוקה אופקית עבור סרטים בשפה העברית שמשודרים בירושלים (הסבר על החלוקה לפרגמטים יתוארו בהמשך).

תובנה מספר 7 (#insight7):
בדקנו את המדינות שהסרטים הופקו בהן שהכי הרבה חיפשו על פיהן בכל עיר, ראינו שהמדינה שהכי הרבה מחפשים לפיה היא ישראל ע"י משתמשים מירושלים. לאחר מכן ראינו כי שאר המשתמשים מחפשים הרבה פחות לפי מדינת ישראל ולכן נוכל לבצע חלוקה אופקית עבור סרטים שהופקו בישראל שמשודרים בירושלים (הסבר על החלוקה לפרגמטים יתוארו בהמשך).

תובנה מספר 8 (#insight8):
בדקנו את השנת הוצאה לאור הכי "ישנה" שחיפשו לפיה בכל עיר, ראינו שהשנה הכי ישנה בערים אילת, ירושלים וקיבוץ גשר היא 1990 ולכן ניתן לבצע חלוקה אופקית עבור סרטים שיצאו לאור מ-1990 ומשודרים באילת, ירושלים וטבריה בהתאמה. בנוסף, ראינו שהשנה הכי ישנה בערים חיפה ותל אביב היא 2010 ולכן ניתן לבצע חלוקה אופקית עבור סרטים שיצאו לאור מ-2010 ומשודרים בחיפה ובתל אביב בהתאמה.

תובנה מספר 9 (#insight9):
בדקנו את החברות הפקה שהכי הרבה חיפשו לפיה בכל עיר, ראינו שהמשתמשים מאילת מחפשים המון לפי חברות הפקה ובמיוחד לפי שלושה ספציפיים: Pixar Animation Studios, Walt Disney Pictures, Warner Bros ולכן נוכל לבצע חלוקה אופקית עבור סרטים שאחד משלושת החברות האלה הפיקו אותם ומשודרים באילת.

תובנה מספר 10 (#insight10):
בדקנו את הז'אנרים שהכי הרבה חיפשו לפיה בכל עיר, ראינו שהמשתמשים מתל אביב מחפשים המון לפי הז'אנר Action, המשתמשים מחיפה מחפשים המון לפי Drama והמשתמשים מקיבוץ גשר מחפשים המון לפי Family ולפי Documentary ולכן נוכל לבצע חלוקה אופקית עבור סרטים מז'אנר Action ומשודרים בתל אביב, עבור סרטים מז'אנר Drama ומשודרים בחיפה ועבור סרטים מז'אנר Family ומז'אנר Documentary ומשודרים בטבריה.

כעת נסכם ונראה את החלוקות לפי האתרים עצמם.

אתר מספר 1 – אילת:

- לפי תובנה 1, נשמור רק את הסרטים שמשודרים בעיר אילת, בנוסף נוכל לבצע חלוקה אנכית ונוריד את עמודת cities מmovies שכן תמיד יהיה בו אילת.
- לפי תובנה 2, נבצע חלוקה אנכית ונוריד את עמודת cast (actors) מcredits.
- לפי תובנה 4, נבצע חלוקה אנכית ונוריד את עמודת crew (director) מcredits – תיגש לחיפה.
  - מכיוון שהורדנו גם את עמודת cast וגם את עמודת crew נוכל להתעלם לגמרי מטבלת credits.
- לפי תובנה 5, נרצה לשמור את עמודת production_countries – שכן יש 100% שימוש.
- לפי תובנה 8 ולפי תובנה 9, נרצה לבצע חלוקה אופקית באופן הבא:
  - פרגמנט 1 – מכיל את כל הסרטים שחברות ההפקה שלהם הם: Pixar Animation Studios, Walt Disney Pictures, Warner Bros שמשודרים באילת ויצאו לאור משנת 1990 ואילך.
  - פרגמנט 2 – מכיל את שאר הסרטים שמשודרים באילת ויצאו לאור משנת 1990 ואילך.
  - נוריד את כל הסרטים שמשודרים באילת ויצאו לאור לפני 1990.
- בסוף לא צריך לעשות join.

אתר מספר 2 – חיפה:

- לפי תובנה 1, נשמור באתר רק את הסרטים שמשודרים בעיר חיפה.
- לפי תובנה 3 ולפי תובנה 4, נשמור באתר את עמודת cast (actors) ועמודת crew (director) ועמודת movie_id מcredits.
- לפי תובנה 5, נבצע חלוקה אנכית ונוריד את עמודת production_countries.
- לפי תובנה 8 ולפי תובנה 10, נרצה לבצע חלוקה אופקית באופן הבא:
  - פרגמנט 1 – מכיל את כל הסרטים מז'אנר "דרמה" שמשודרים בחיפה ויצאו לאור משנת 2010 ואילך.
  - פרגמנט 2 – מכיל את שאר הסרטים שמשודרים בחיפה ויצאו לאור משנת 2010 ואילך.
  - נוריד את כל הסרטים שמשודרים בחיפה ויצאו לאור לפני 2010.
- בסוף נעשה join בין העמודות מcredits לבין הפרגמנטים שיצרנו מmovies עפ"י movie_id.

אתר מספר 3 – ירושלים:

- לפי תובנה 1, נשמור באתר את הסרטים שמשודרים בעיר ירושלים ובעיר תל אביב.
- לפי תובנה 3, נשמור באתר את עמודת cast (actors) ועמודת movie_id מcredits.
- לפי תובנה 4, נבצע חלוקה אנכית ונוריד את עמודת crew (director) – תיגש לחיפה.
- לפי תובנה 5, נרצה לשמור את עמודת production_countries – שכן יש 100% שימוש.
- לפי תובנה 6 ולפי תובנה 7 ולפי תובנה 8, נרצה לבצע חלוקה אופקית באופן הבא:
  - פרגמנט 1 – מכיל את כל הסרטים שלפחות אחת מהמדינות שבהן הופקו הסרט היא ישראל ולפחות אחת מהשפות הזמינות היא השפה העברית שמשודרים בירושלים ויצאו לאור משנת 1990 ואילך.
  - פרגמנט 2 – מכיל את שאר הסרטים שמשודרים בירושלים ויצאו לאור משנת 1990 ואילך.
  - נוריד את כל הסרטים שמשודרים בירושלים ויצאו לאור לפני 1990.
- בסוף נעשה join בין העמודות מcredits לבין הפרגמנטים שיצרנו מmovies עפ"י movie_id.

אתר מספר 4 – תל אביב:

- לפי תובנה 1, נשמור באתר רק את הסרטים שמשודרים בעיר תל אביב.
- לפי תובנה 3, נשמור באתר את עמודת cast (actors) ועמודת movie_id מcreditsמ.
- לפי תובנה 4, נבצע חלוקה אנכית ונוריד את עמודת crew (director) – תיגש לחיפה.
- לפי תובנה 5, נבצע חלוקה אנכית ונוריד את עמודת production_countries.
- לפי תובנה 8 ולפי תובנה 10, נרצה לבצע חלוקה אופקית באופן הבא:
  - פרגמנט 1 – מכיל את כל הסרטים מז'אנר "אקשן" שמשודרים בתל אביב ויצאו לאור משנת 2010 ואילך.
  - פרגמנט 2 – מכיל את שאר הסרטים שמשודרים בתל אביב ויצאו לאור משנת 2010 ואילך.
  - נוריד את כל הסרטים שמשודרים בתל אביב ויצאו לאור לפני 2010.
- בסוף נעשה join בין העמודות מcreditsמ לבין הפרגמנטים שיצרנו מmoviesמ עפ"י movie_id.

אתר מספר 5 – טבריה:

- לפי תובנה 1, נשמור באתר רק את הסרטים שמשודרים בעיר טבריה.
- לפי תובנה 3, נשמור באתר את עמודת cast (actors) ועמודת movie_id מcreditsמ.
- לפי תובנה 4, נבצע חלוקה אנכית ונוריד את עמודת crew (director) – תיגש לחיפה.
- לפי תובנה 5, נרצה לשמור את עמודת production_countries – שכן יש 100% שימוש.
- לפי תובנה 8 ולפי תובנה 10, נרצה לבצע חלוקה אופקית באופן הבא:
  - פרגמנט 1 – מכיל את כל הסרטים מז'אנר "Family" ומז'אנר "Documentary" שמשודרים בטבריה ויצאו לאור משנת 1990 ואילך.
  - פרגמנט 2 – מכיל את שאר הסרטים שמשודרים בטבריה ויצאו לאור משנת 1990 ואילך.
  - נוריד את כל הסרטים שמשודרים בטבריה ויצאו לאור לפני 1990.
- בסוף נעשה join בין העמודות מcreditsמ לבין הפרגמנטים שיצרנו מmoviesמ עפ"י movie_id.

- **בגדר רעיון בלבד:** חשבנו על כך שניתן היה לוותר על הפרגמנטים בתוך האתרים עצמם (לדוגמה 2 הפרגמנטים באתר של חיפה) ולייעל בעזרת האופטימיזציה שחישבנו בחלק 2, בעזרת מיון אופטימלי לפי עמודות לפי ממוצע מספר הכניסות בכל עמודה בשאילתות (כמה שפחות כניסות/ערכים ככה יותר טוב אך נוודא שלא קצת מדי כי אז זה יחלק לקבוצה גדולה מדי):
  - Haifa : sorted by (actors, release date, language, director)
  - Tel Aviv : sorted by (actors, release date, language, genres)
  - Jerusalem : sorted by:(release date , production company, genre , actor , language)
  - Tiberius : sorted by (actor, release date, production company, language, genre)
  - Eilat: sorted by (release year, language, genres ,country, production company)