**Homework 1 Group D – Business Intelligence**

מגישים:

דור דלמדיגו – 205973175

עדן גבריאל – 318868700

חנן חן - 301647814

**חלק 1**

בחרנו אוסף נתונים מאתר Kaggle. הקובץ מכיל מידע בנוגע לסרטים ע"פ אתר IMDB. ביצענו נרמול לטבלה הבודדת והפכנו אותה ל-2 טבלאות.

<https://www.kaggle.com/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows>

**חלק 2 – שאלות מחקר**

**שאלה מונחית:** האם סרטי דרמה מעל 120 דקות מרוויחים יותר מסרטי פשע מתחת ל 120 דקות אשר התפרסמו בין השנים 2019- 2016.

**KPI:** סך הרווחים מסרטי דרמה שאורכם מעל 120 דקות חלקי סך הרווחים מסרטי פשע מתחת   
ל- 120 דקות אשר התפרסמו בין השנים 2019- 2016.  
אם המדד יוצא גדול מ-1, הרווח מסרטי דרמה שאורכם מעל 120 דקות גבוה יותר מאשר סרטי פשע מתחת ל 120 דקות בהתייחס לשנים המצוינות.  
אם המדד יוצא קטן מ-1, הרווח מסרטי פשע מתחת ל 120 דקות גדול יותר מאשר הרווח מסרטי דרמה מעל 120 דקות בהתייחס לשנים המצוינות לעיל.   
אם המדד שווה ל-1, הרווח מהסרטים המצוינים לעיל זהה.

המדד הנ״ל עומד בדרישות מודל **SMART**:  
  
S - המדד ספציפי ומכוון ישירות לסך הרווחים המופקים מהסרטים בחתך של סוג הסרט )לפי עמודת Genre), אורך הסרט (לפי עמודת Runtime) ושנת פרסומו (לפי עמודת (Released\_Year.  
  
M- הוא מדיד מכיוון שניתן לקבל את ערך המדד לפי עמודת ה-Gross, הרווח הישיר מהסרטים, תוך הסתכלות על סוג הסרטים- דרמה ופשע, אורך זמן הסרט- חתך של מעל 120 דקות ומתחת ל- 120 דקות בהתאמה וטווח השנים שבהם התפרסמו הסרטים:   
 2019- 2016.

A – המדד ניתן ליישום על ידי פתרונות שונים כגון התמקדות בפרסום סוגי סרטים ספציפיים אשר הרווח המופק מהם גדול יותר, צמצום/הארכת משך זמני הסרטים לפי הרווח, פרסום סרטים ישנים/מודרניים יותר על מנת להגדיל את הרווח מהסרטים.

R – המדד רלוונטי כיוון שמודד את סך הרווחים מסרטי דרמה שאורכם מעל 120 דקות חלקי סך הרווחים מסרטי פשע מתחת ל- 120 דקות אשר התפרסמו בין השנים 2019- 2016.

T – המדד קצוב בזמן ומתפרש על פני טווח השנים 2019- 2016. שבהם הסרטים התפרסמו.

**שאלה לא מונחית:** האם לסרטים שהדירוג הממוצע שלהם ב-IMDB גדול מ-8.5 והתפרסמו בין השנים 1991-1999 קיימים מאפיינים משותפים?

**KPI:** ממוצע אורך הסרטים שהדירוג הממוצע שלהם ב-IMDB גדול מ-8.5 והתפרסמו בין השנים 1991-1999

המדד הנ״ל עומד בדרישות מודל **SMART**:  
  
S - המדד ספציפי ומכוון ישירות לממוצע אורך הסרטים (עמודת Runtime) שהדירוג הממוצע שלהם ב-IMDB גדול מ-8.5 (לפי עמודת Meta\_score) והתפרסמו בין השנים 1991-1999 (לפי עמודת (Released\_Year.  
  
M- הוא מדיד מכיוון שניתן לקבל את ערכו על ידי חישוב סך אורך הסרטים שהדירוג הממוצע שלהם ב-IMDB גדול מ-8.5והתפרסמו בין השנים 1991-1999 חלקי סך אורכי הסרטים הכולל.

A – המדד ניתן ליישום על ידי פתרונות שונים שיביאו להעלאת הדירוג הממוצע של הסרטים   
ב-IMDB, כמו הארכת/קיצור אורך הסרט או פרסום סרטים ישנים/מודרניים יותר.

R – המדד רלוונטי כיוון שמודד אורך הסרטים שהדירוג הממוצע שלהם ב-IMDB גדול מ-8.5 והתפרסמו בין השנים 1991-1999

ישנם כ-16 סרטים שהדירוג הממוצע שלהם ב-IMDB גדול מ-8.5 והתפרסמו בין השנים   
 1991-1999

T – המדד קצוב בזמן ומתפרש על פני טווח השנים 1991-1999 שבהם הסרטים התפרסמו.

**חלק 3 – הבנת הנתונים**

1. **מדדי פיזור**

Table

Description automatically generatedטבלה 1:

תמונה שמכילה שולחן

התיאור נוצר באופן אוטומטי

טבלה 2:

1. **קורלציה:**

**מתאם פירסון:**

טבלה 1: Table

Description automatically generated

טבלה 2:

Graphical user interface, text, application, chat or text message

Description automatically generated

**Correlation Map:**

טבלה 1:

Calendar

Description automatically generated with medium confidence

טבלה 2:

A picture containing graphical user interface

Description automatically generated

בדיקת קשר בין 2 עמודות – רגרסיה:

Runtime vs IMDB\_Rating

Chart, scatter chart

Description automatically generated

Released Year vs Runtime

Chart, scatter chart

Description automatically generated

Num of votes vs Gross

Chart, scatter chart

Description automatically generated

**Chart

Description automatically generated**

1. **מדדים סטטיסטיים נוספים**

אפשרויות של סוגי סרטים וכמויות מכל סוג:

**Table

Description automatically generated**

Chart, histogram

Description automatically generated

 

**פיזור הנתונים לשאלת המחקר המונחית:**

סרטי דרמה מעל 120 דקות:

Text

Description automatically generated



סרטי פשע מתחת ל-120 דקות:

Graphical user interface, text

Description automatically generated



**CDF**

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Table

Description automatically generatedפיזור נתונים של השאלה הלא מונחית: סרטים אשר הציון שלהם גבוה מ8.5:

Table

Description automatically generated

1. **אנטרופיה**

Text, letter

Description automatically generated

1. **מדדי GINI ו-Information Gain**

**Gini:**

**A picture containing graphical user interface

Description automatically generated**

**Information gain:**

Genre: 0.08

Released Year: 0.5

1. **הסקת נתונים**

ניתן ללמוד מחקר הנתונים הראשוני כי סט הנתונים מכיל 1000 שורות (חלקן Null). ממוצע שנת יציאת הסרטים הינו 1991 עם ממוצע רווחים של 68 מיליון דולר. בנוסף, ישנו קשר חיובי בינוני בין התכונה Runtime לציון IMDB וקשר שלילי בינוני בין עמודת Runtime לציון הסרט. גילינו גם, ע"פ מדד הinfo-gain, כי תכונת שנת הסרט יכולה להוות אינדיקציה לעמודת המטרה שלנו. יתר על כן, ניתן לראות מהרגרסיה שביצענו שהסרטים שקיבלו את הציון הכי גבוה נמצאים בטווח משוער בין 120 ל-150 דקות זמן סרט.