

שם הפרויקט: דירוגי סדרות וסרטים

מגיש:

אביב בסלו 322294315

משפט המתאר את הפרויקט: סדרות\סרטים ודירוגם ע"י מספר אתרי ביקורת (IMDb, Rotten Tomatoes).

שלב 3:

קובץ Input1_df מכיל מידע כללי על סרטים וסדרות כגון: שנת הפקה, משך זמן הסרט\סדרה ודירוגים מכל הפלטפורמות כגון דיסני, פריים וידיאו וכו'.

מהאינפורמציה של הנתונים ניתן לראות שיש 5380 אינדקסים, 14 עמודות, מכילות שם עמודה, כמה ערכים לא חסרים יש בכל עמודה ואיזה סוג עמודה כל אחת.

למשל, אחת מהעמודות היא country שהיא בעצם הסוג שקיים, בה ניתן לראות שישנם בסה"כ 4004 ערכים לא חסרים ושהיא מסוג object. עמודה נוספת release year שמופיעה בקובץ, בה ניתן לראות שישנם בסה"כ 5380 ערכים לא חסרים ושהיא מסוג int64.

בסטטיסטיקה התאורית של הנתונים ניתן לראות את המדדים השונים לכל עמודה כמו חציון, ממוצע, מינימום ומקסימום. לדוגמה, בעמודה של release year הממוצע הוא 2012.642193.

שלב 5:

pv1_df		
type	Movie	TV Show
country		
Argentina	7.500000	6.928571
Australia	NaN	7.210811
Austria	NaN	6.500000
Belarus	NaN	6.800000
Belgium	NaN	7.262500
Brazil	4.700000	6.666001
Canada	7.200000	6.940000

טבלה 1:

בטבלה הראשונה הגדרתי 'country' כ-index, עמודות כ-'type' (Movie, TV Show) וערכי IMDb כדירוג והחלתי על טבלת הציר פונקציית צבירה של ממוצע.

בטבלה זו ניתן לראות לכל מדינה את הממוצע דירוג IMDb שלה לפי סרט וסדרה. בחלק מהמדינות יש דירוג גם לסרט וגם לסדרה וחלק יש רק אחד מהם.

טבלה 2:

pv2_df					
Age	13+	16+	18+	7+	all
type					
Movie	0	84	90	82	67
TV Show	79	96	100	93	83

בטבלה השנייה הגדרתי 'type' כ-index, עמודות כגילאים וערכי Rotten Tomatoes כדירוג החלתי על טבלת הציר פונקציית צבירה של מקסימום וכל ערכי ה-NaN הוצגו כ-0.

בטבלה זו ניתן לראות לכל קבוצת גיל את המקסימום דירוג של Rotten Tomatoes לפי סרט וסדרה. בקטגוריה של +13 בסרטים ניתן לראות שאין דירוג כלל.

pv3_df		
type	Movie	TV Show
release year		
1904	0	1
1931	0	1
1932	0	1
1934	0	1
1943	0	1
...
2017	18	587
2018	12	550

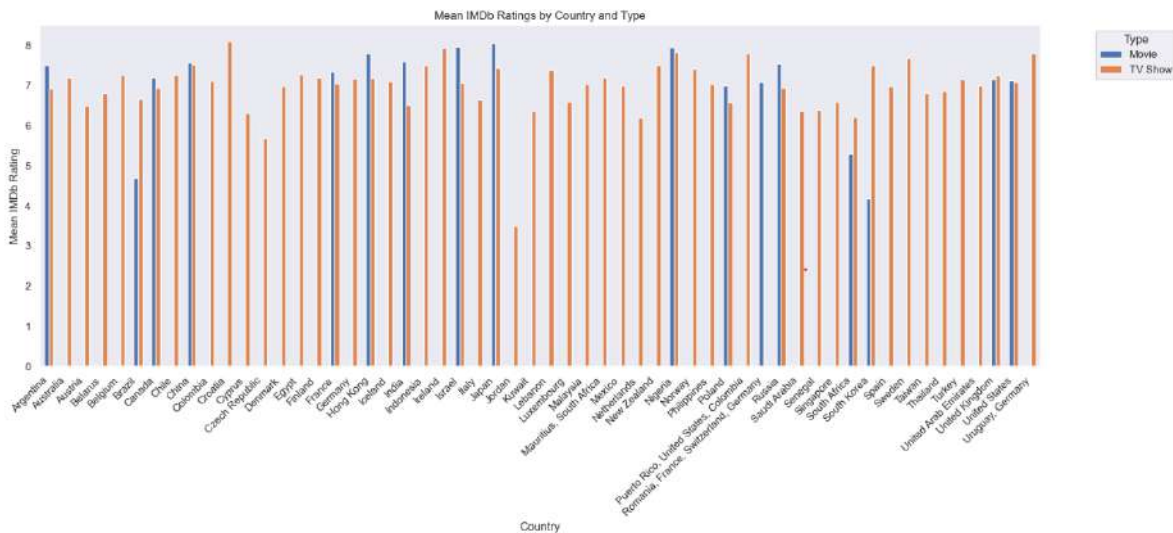
טבלה 3:

בטבלה השלישית הגדרתי 'release year' כ-index, עמודות כ-'type' וערכי Rotten Tomatoes כדירוג והחלתי על טבלת הציר פונקציית צבירה של ספירה וכל ערכי ה-NaN הוצגו כ-0.

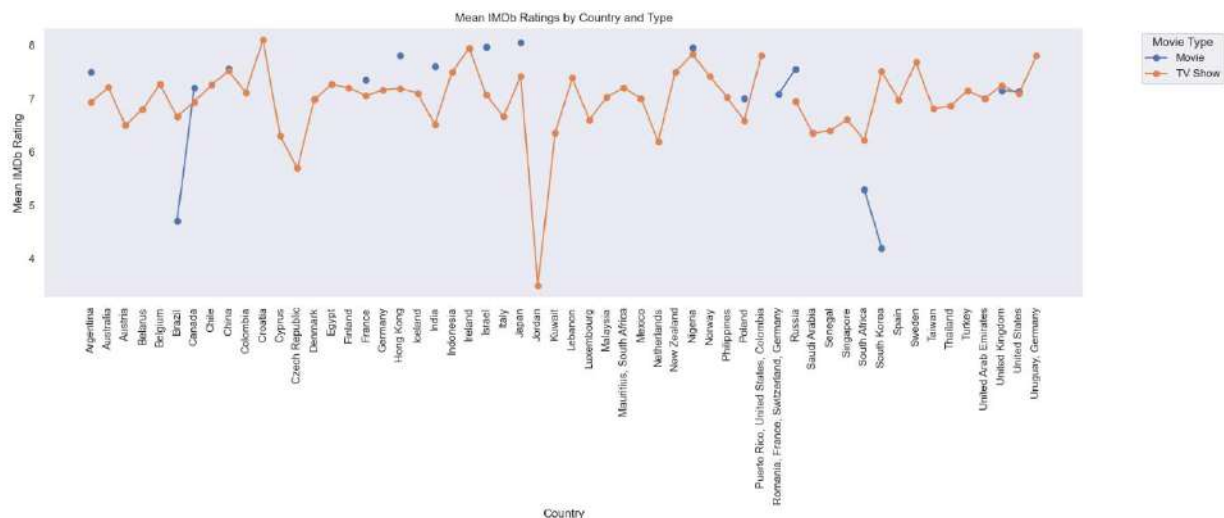
בטבלה זאת ניתן לראות כמות של סרטים וסדרות שקיימת לכל שנה.

יש שנים שיש להם רק סרטים או רק סדרות ויש שנים שיש גם וגם.

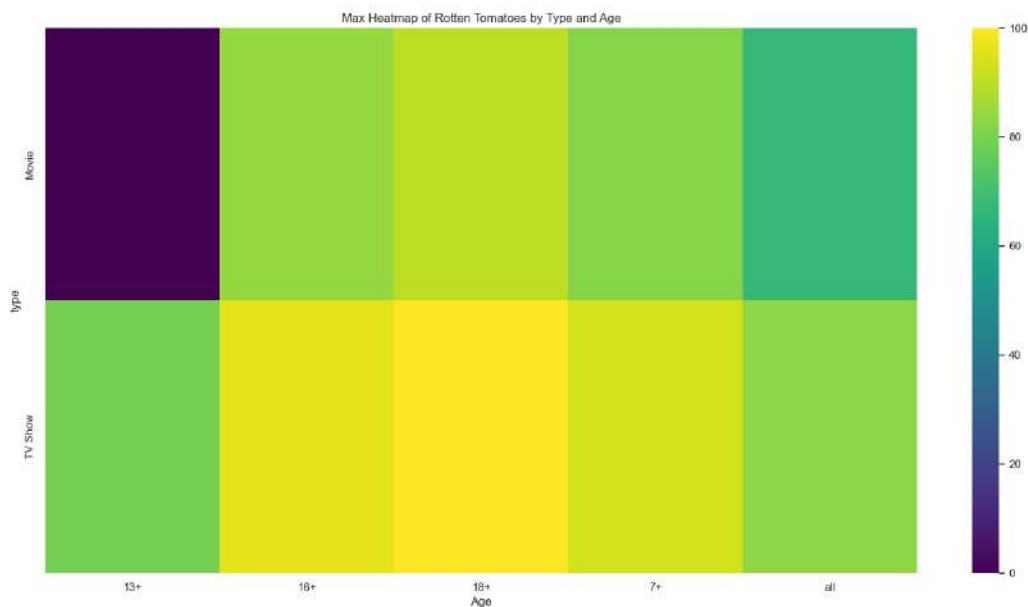
שלב 6:



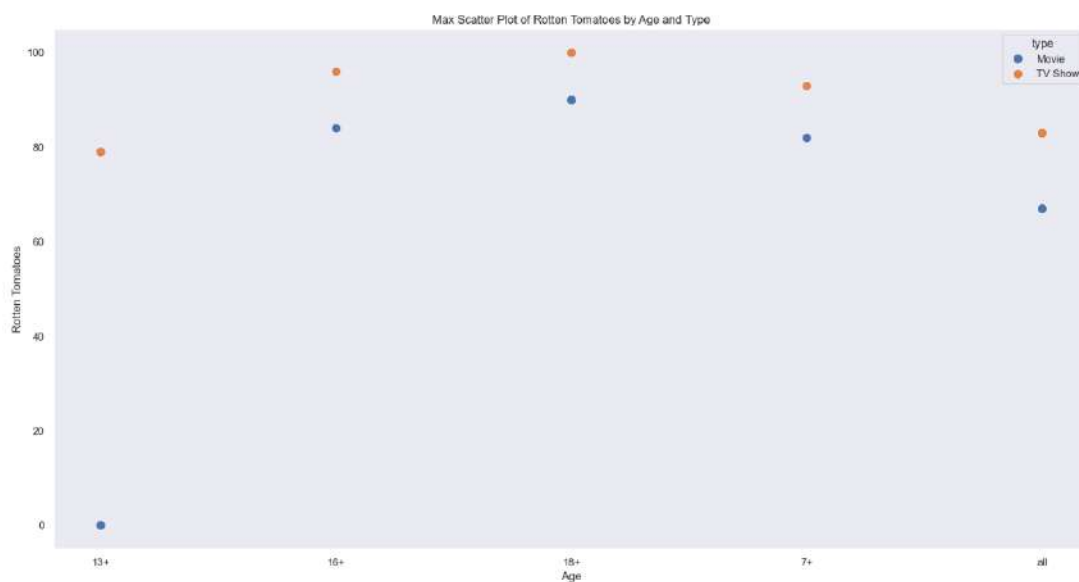
גרף 1: מייצג את טבלת הציר הראשונה, מייצג את ממוצע הדירוגים לכל מדינה לפי סרט(בצבע כחול) וסדרה(בצבע כתום). בגרף הזה ניתן לראות שממוצע הדירוג הכי נמוך (כ-3.5) נמצא בירדן בקטגוריית סדרות. בשאר המדינות ניתן לראות שטווח ממוצע הדירוגים הוא באזור ה-5 ל-7.5.



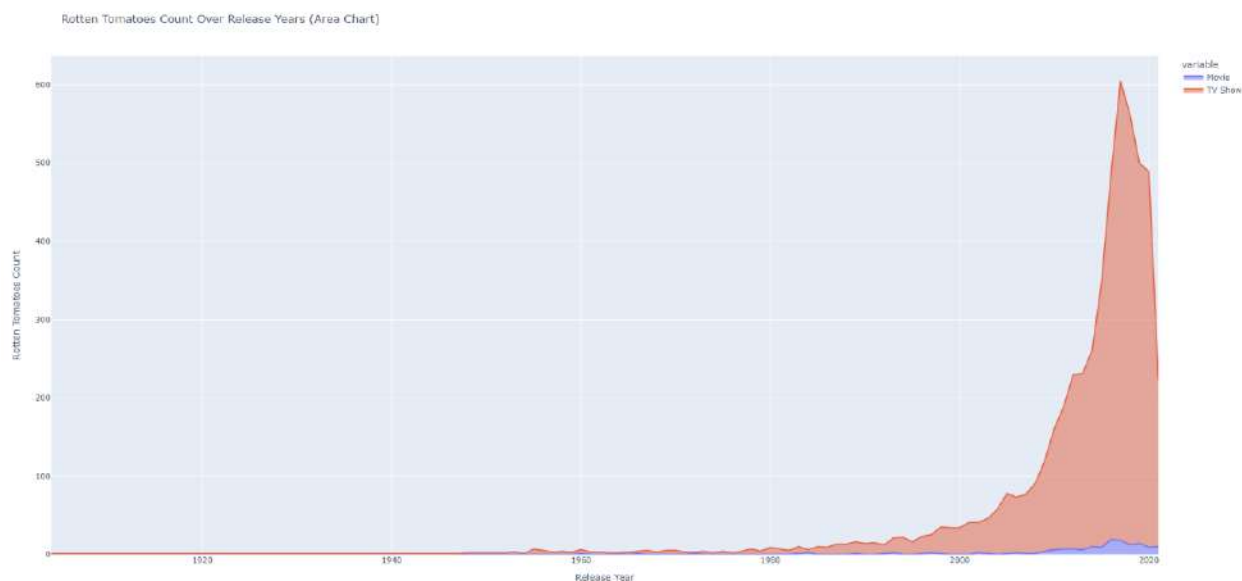
גרף 2: מייצג את טבלת הציר הראשונה, מייצג את ממוצע הדירוגים לכל מדינה לפי סרט(בצבע כחול) וסדרה(בצבע כתום). ממוצע הדירוג בקטגוריית הסדרות נמצא לרוב באותו הטווח בין 6 ל-8 לעומת קטגוריית הסרטים יש יותר פיזור כאשר רוב ממוצע דירוג זה נמצא בחלק העליון ובודדים בחלק התחתון.



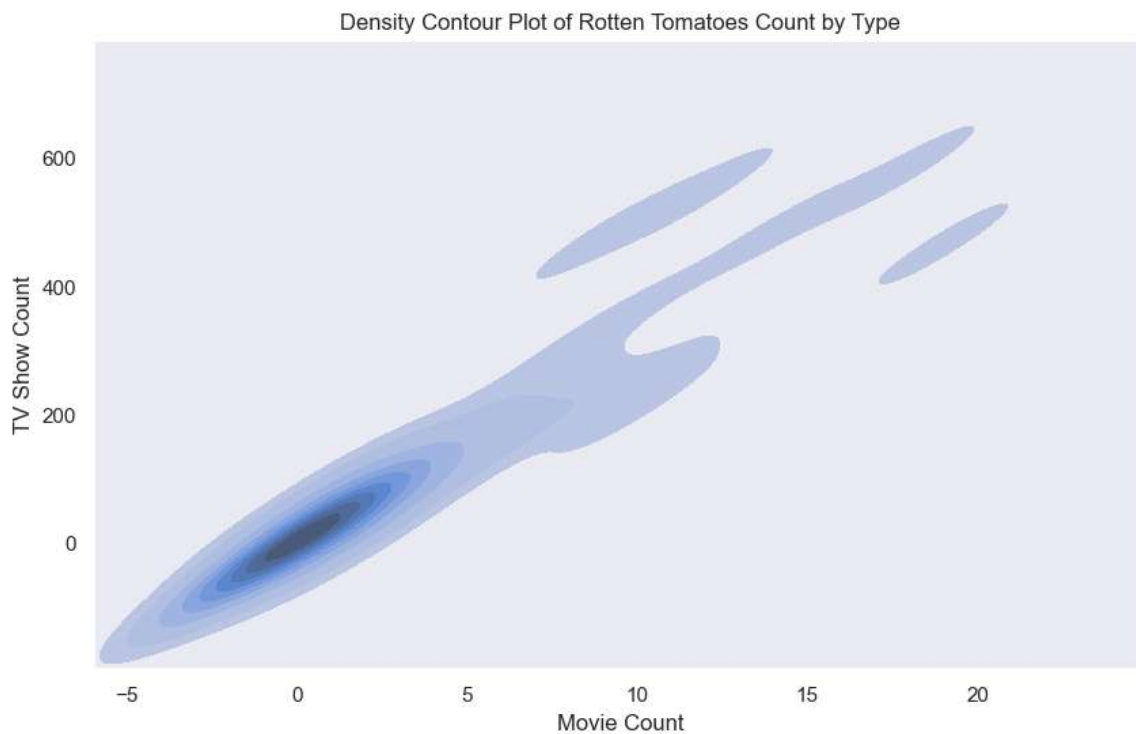
גרף 3: מייצג את טבלת הציר השנייה, מייצג את מקסימום הדירוג לכל קבוצת גיל לפי סרט וסדרה. ניתן לראות בגרף שהציון המקסימלי הכי נמוך נמצא בקבוצת הגיל +13 בסרטים, הציון המקסימלי הכי גבוה נמצא בקבוצת הגיל +18 בסדרות. בסדרות ניתן לראות שרובם עם דירוג גבוה מ-75 ובסרטים זה יחסית מפוזר יש דירוגים גבוהים ויש גם נמוכים.



גרף 4: מייצג את טבלת הציור השנייה, מייצג את מקסימום הדירוג לכל קבוצת גיל לפי סרט וסדרה. ניתן לראות בגרף שהציון המקסימלי הכי נמוך נמצא בקבוצת הגיל 13+ בסרטים. ניתן לראות שבין סדרות לסרטים באותם קבוצות גיל הדירוגים בסדרות יותר גבוהים מהדירוגים בסרטים.

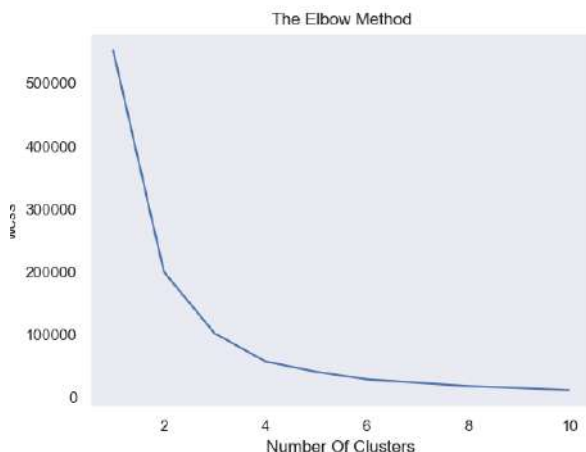


גרף 5: מייצג את טבלת הציור השלישית, מייצג את כמות הדירוג הקיימת לכל שנת יציאה לאור לפי סרט (בצבע כחול) וסדרה (בצבע אדום). ניתן לראות בגרף עלייה בכמות דירוגים בסדרות לאורך השנים בערך משנת 2000 כשכמות הדירוג הייתה 34 וב2017 כמות הדירוג הייתה 584 ולאחר מכן ישנה ירידה. לעומת זאת בסרטים רואים שיש עלייה של כמות הדירוג באזור שנת 2010 אבל בכללי אין עלייה דרסטית לעומת כמות דירוג הסדרות.



גרף 6: מייצג את טבלת הציר השלישית, מייצג את כמות הדירוג הקיימת לפי סרט וסדרה. בגרף זה ניתן לראות שיש קורולציה בכמות דירוג בין הסדרות לסרטים כאשר כמויות הדירוג קטנות. לאחר מכן ניתן לראות שהקשר בכמות דירוג בין הסדרות לסרטים פוחת כאשר כמות הדירוג של הסדרות גדלה בצורה הרבה יותר גדולה במהלך השנים לעומת כמות דירוג הסרטים ובכך ניתן לראות שבחלק מהגרף קשר זה נעלם.

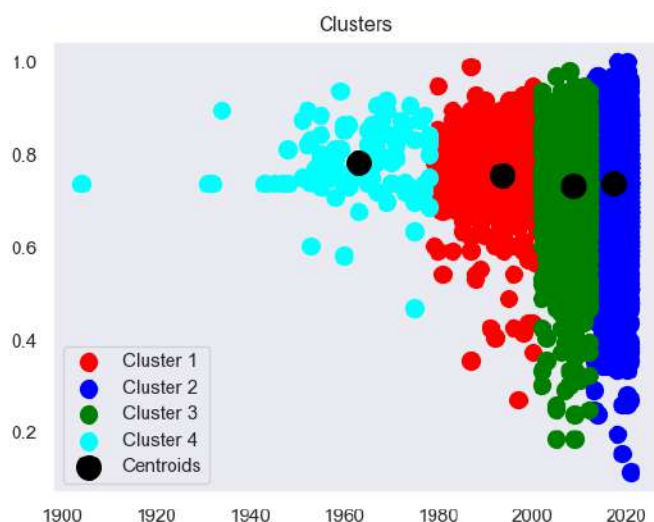
שלב 7:



בשיטת המרפק אני מחפש את המספר האופטימלי של האשכולות.

WCSS הוא סכום המרחק בריבוע בין כל נקודה לבין מרכז הצביר (האשכול).

ניתן לראות מהגרף שהערך האופטימלי הוא $K=4$.

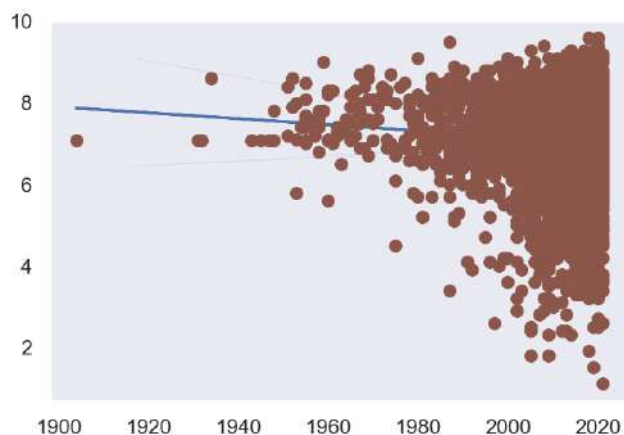


ניתן לראות שמדובר בויזואליזציה של האשכולות.

המרכזים של כל קלסטר הוא בצבע שחור.

ניתן לראות שלכל קלסטר יש צבע שונה, בקלסטר 2 הכחול ניתן לראות שיש טווח גדול יותר של דירוג. בכל הקלסטרים למעט קלסטר 4 תכלת, רואים שיש צפיפות גדולה. כל המרכזים של הקלסטרים נמצאים מ-1960.

שלב 8:



לקו רגרסיה יש שיפוע שלילי שמצביע על ירידה בדירוגי IMDb לאורך השנים.

האזור המוצל סביב קו הרגרסיה מציין את השונות הפוטנציאלית בדירוגי IMDb.

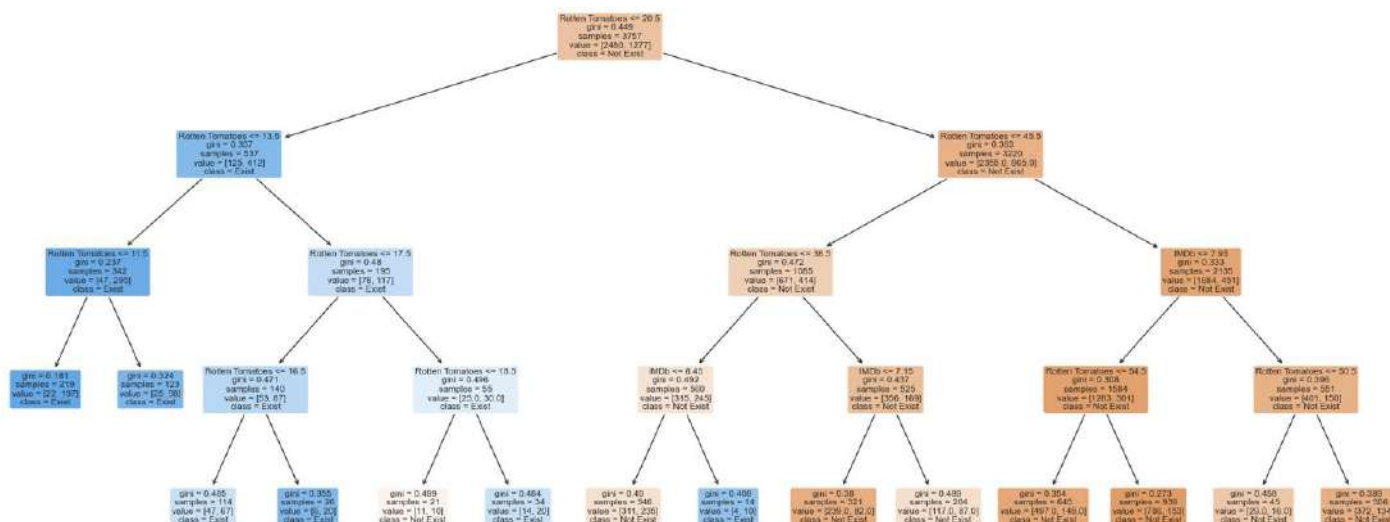
נקודות הנתונים מקובצות על קו הרגרסיה, וזה מצביע על קשר ליניארי חזק מכיוון שיש צפיפות וקורלציה גבוהה יותר מאזור שנת 2000.

```
Significance: Yes
R-squared: 0.005327300810787139
Adjusted R-squared: 0.0051419350450047085
```

ניתן לראות את מדדי הרגרסיה והרגרסיה יצאה 0.0053 ולכן רגרסיה זאת מובהקת, כיוון שהיא קטנה יותר מ-0.05. מדד הרגרסיה Adjusted R-squared יצא 0.00514.

שלב 9:

Gini



בעץ החלטה על בסיס מודל החלטה של ג'יני לוקחים 70 אחוז מהנתונים לאימון ו-30 אחוז לבדיקה, מדדתי כ-3757 ערכים, על פי דירוגי Rotten Tomatoes ו-IMDb ועל פי classes שמחלקים ל-Exist ל-Not Exist ב-Prime Video. בחלק העליון של העץ נתוני הג'יני הוא 0.449 שזה מדד גבוה. כמות הרמות שיש בעץ היא 4 ובעומק זה הגעתי לרמת הדיוק הגבוהה ביותר במדד זה. העץ החלטה מתחלק לנתונים שהם מתחת 20.5 ומעל 20.5. כשמסתכלים על צד שמאל של העץ, מדד הג'יני משתפר ונהיה 0.357 על בסיס 537 ערכים. ברמה 2 הגעתי למדד ג'יני טוב יותר שהוא 0.237 על בסיס 342 ערכים. ברמה 3 הגעתי למדד ג'יני הכי טוב שהוא 0.181 על בסיס 219 ערכים ושהוא נמצא ב-Class Exist. כשמסתכלים על צד ימין של העץ שהמדדים פחות טובים כיוון שמדד הג'יני הכי טוב שהגעתי אליו בצד זה הוא 0.273 שנמצא ב-Class – Not Exist.

רמת הדיוק שהגעתי במודל זה היא 71.5% שמצביע על רמת דיוק גבוהה יחסית.

F1 score – במקרה הנ"ל הציון הוא 0.403.

Recall score – במקרה הנ"ל הציון הוא 0.279.

Precision – במקרה הנ"ל יצא 0.72. מציין את חלקן של התחזיות החיוביות האמיתיות מתוך כל התחזיות החיוביות.

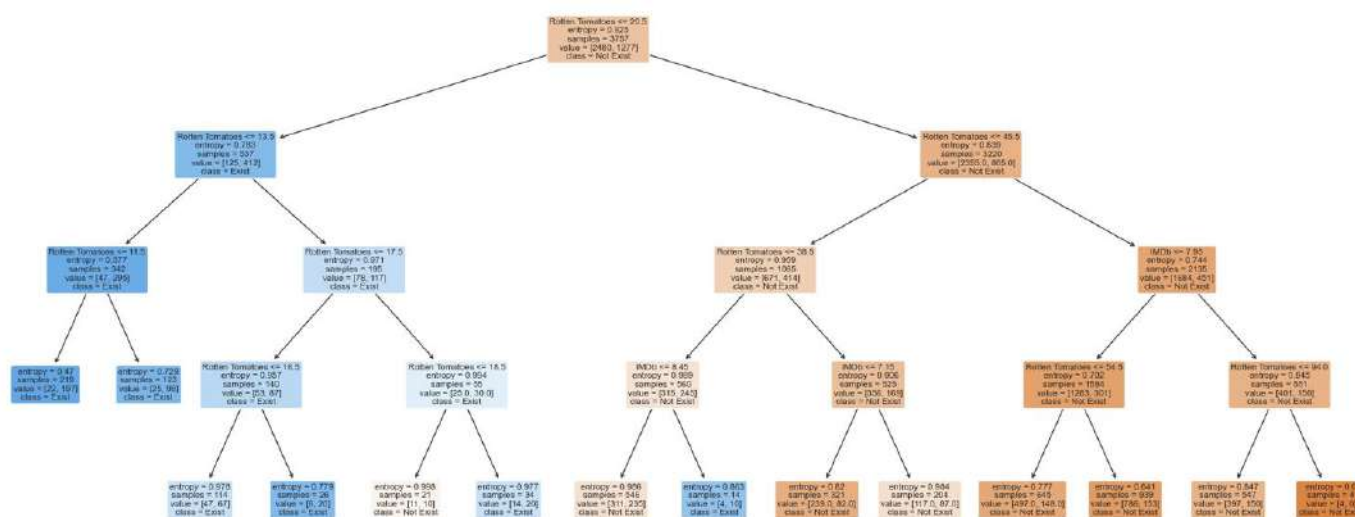
Confusion Matrix – מטריצה זאת נותנת פירוט מפורט של סיווגים נכונים ושגויים, מראה 155 חיוביות אמיתיות, 997 שליליות אמיתיות, 60 חיוביות שגויות ו-399 שליליות שגויות.

ROC AUC – במקרה הנ"ל הציון הוא 0.702 וזה מצביע על ביצועים יחסית טובים של המודל.

```
gini - Model Performance:
Accuracy = 0.7150837988826816
F1_score = 0.40312093628088425
Recall_score = 0.27978339350180503
Precision = 0.7209302325581395

Confusion Matrix:
[[997 60]
 [399 155]]
ROC AUC 0.7021378876938684
```

Entropy



בעץ החלטה על בסיס מודל החלטה של entropy לוקחים 70 אחוז מהנתונים לאימון ו-30 אחוז לבדיקה, אני מודד כ-3757 ערכים, על פי דירוגי Rotten Tomatoes ו-IMDb ועל פי classes שמחלקים ל-Exist ל-Not Exist ב-Prime Video. בחלק העליון של העץ נתוני entropy הוא 0.925 שזה מדד שמצביע על אי סדר גבוה. כמות הרמות שיש בעץ היא 4 ובעומק זה הגעתי לרמת הדיוק הגבוהה ביותר במדד זה. העץ החלטה מתחלק לנתונים שהם מתחת ל-20.5 ומעל ל-20.5 ב-Rotten Tomatoes. כשמסתכלים על צד שמאל של העץ, מדד entropy משתפר ונהיה 0.783 על בסיס 537 ערכים. ברמה 3 הגעתי למדד entropy הכי נמוך בצד שמאל של העץ שהוא 0.47 על בסיס 219 ערכים שהוא Class-Exist. כאשר מדד ה-entropy יותר נמוך זה מצביע על פחות אי סדר בנתונים. כשמסתכלים על צד ימין של העץ ניתן לראות בצד זה שמדד ה-entropy הכי נמוך הוא 0.0 על בסיס 4 ערכים שנמצא ב-Class-Not Exist.

```
entropy - Model Performance:
Accuracy = 0.7150837988826816
F1_score = 0.40312093628088425
Recall_score = 0.27978339350180503
Precision = 0.7209302325581395

Confusion Matrix:
[[997 60]
 [399 155]]
ROC AUC 0.7014701713520659
```

רמת הדיוק שהגעתי במודל זה היא 71.5% שמצביע על רמת דיוק גבוהה יחסית.

F1 score – במקרה הנ"ל הציון הוא 0.403.

Recall score – במקרה הנ"ל הציון הוא 0.279.

Precision – במקרה הנ"ל יצא 0.72. מציין את חלקן של התחזיות החיוביות האמיתיות מתוך כל התחזיות החיוביות.

Confusion Matrix – מטריצה זאת נותנת פירוט מפורט של סיווגים נכונים ושגויים, מראה 155 חיוביות אמיתיות, 997 שליליות אמיתיות, 60 חיוביות שגויות ו-399 שליליות שגויות.

ROC AUC – במקרה הנ"ל הציון הוא 0.701 וזה מצביע על ביצועים יחסית טובים של המודל.

לבסוף אחרי שביצעתי את שני עצי החלטה אלה, מדד הג'ני מראה רמת דיוק גבוה יותר במעט ממדד ה-entropy. לעומת זאת בשאר המדדים קיבלתי מדדים זהים בשני המודלים.

שלב 10:



בגרף Training Loss per Epoch אני מודד את ביצועי המודל בכל epoch, במודל לוקחים 75 אחוז מהנתונים לאימון ו-25 אחוז לבדיקה וניתן לראות את נתוני ה-AUC וה-VAL_AUC נמצאים בטווח של בין 0.6 לבין 0.7 על בסיס 50 epochs.

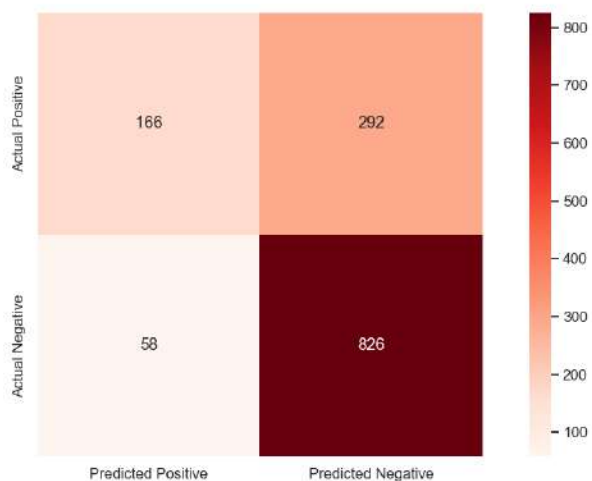
ה-AUC מייצג את רמת הדיוק על גביי נתוני האימון. ה-VAL_AUC מייצג את רמת הדיוק על גביי נתוני האמת.

שניהם עולים בהתאמה בכל שכמות ה-epoch עולה, כאשר נתוני VAL_AUC גבוהים במעט מה-AUC.

שניהם מגיעים לרמות דיוק יחסית גבוהות באזור ה-0.7, אם זאת בשלב מסוים בקושי יכולים לראות שיפורים (לקראת ה-epochs האחרונים).

ניתן לראות שה-loss (מייצג את ההפסד של המודל) וה-val_loss יורדים בהתאמה בכל שכמות ה-epoch עולה, מתחיל באזור ה-0.65 ויורד לכיוון 0.6. ניתן לראות שגם לקראת epoch ה-50 עדיין ה-loss נמצא בשיפוע יחסית נמוך שלילי.

כתוצאה מכך ניתן לראות שככל שכמות ה-epoch גדלה, כמות ההפסד במודל יורדת.



ניתן לראות שהרשת העצבית הפשוטה הזאת מספקת ביצועי מודל עבור AUC 0.73. $(166+826)/1342$

ניתן לראות שחיזוי שלילי עם תוצאה שלילית אמיתית קיבל תוצאה של 826 לעומת חיזוי שלילי עם תוצאה נכונה אמיתית שקיבל 292.

חיזוי נכון עם תוצאה נכונה אמיתית קיבל 166 לעומת חיזוי נכון עם תוצאה שלילית אמיתית שקיבל 58. ניתן לראות מהתוצאות שקיבלנו שמודל זה מבנא יחסית נכון ברוב המצבים.