אוניברסיטת בן-גוריון בנגב
המחלקה להנדסת תעשייה וניהול

לימוד מכונה – חלק א'

עומר קידר

CONTENTS

1. DATA COLLECTION AND SENSING

א. ה- Data collection הוא אוסף כל הנתונים שלנו. הדאטה מורכב מישויות (entities), וישויות אלו יהיו מאותו התחום. הישויות צריכות להיות שלמות, ללא מידע חסר. הdata collection צריך להיות מגוון, מייצג, מקיף, אמין, ובמידה וניתן נעדיף שהוא יהיה מתויג (labeled). בהקשר של הדאטה שלנו, הישויות הם משחקי הוידאו השונים. התיוג של סט הדאטה שלנו הוא כמות המכירות באירופה (EU Sales). סוג ה sensing אשר בוצע על הדאטה שלנו הוא סנסינג סטטי. ברור כי פי'צרים כגון פלטפורמה, שנת יציאה, ג'אנר, מפרסם, מפתח, ודירוג גיל הם סטטיים ואינם משתנים בזמן. לעומתם, המידע על מכירות ודירוג הוא מידע שעלול להשתנות עם הזמן, אולם הדאטה שלנו לא מכיל חותמות זמן, ולכן אנו מסיקים מכך שגם פי'צרים אלו עברו תהליך סנסינג סטטי. סוג סנסינג שלא בוצע על הדאטה הוא כאמור סנסינג דינמי. דוגמה ספציפית יכולה להיות אוסף של מכירות לפי חודשים של כל משחק. איסוף המידע של המכירות לפי חודשים מגדיר שינויים בזמן ולכן זהו סנסינג דינמי.

ב. קטגורית הלמידה היא supervised learning והמשימה היא prediction. ההסבר לכך הוא שהדאטה שלנו מתויג (labeled) באופן מלא, והמטרה שלנו היא לבצע חיזוי על ידי זיהוי קשרים בין הפיצ'רים (שהם ערכים מספריים) למשתנה המוסבר (כמות המכירות – גם ערך מספרי). ניתן להשתמש בנתונים כדי לבצע גם סיווג בינארי למשל, אפשר לבנות מודל שיסווג את המשחקים לרבי-מכר או לא באירופה.

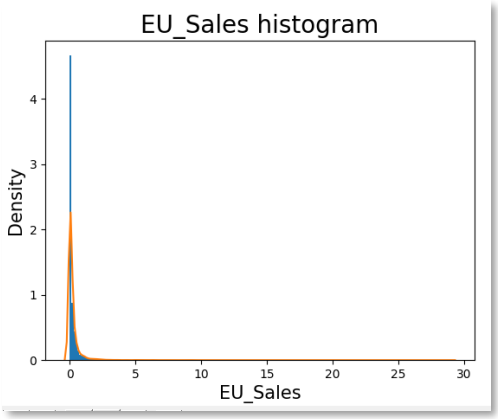
2. DATASET CREATION

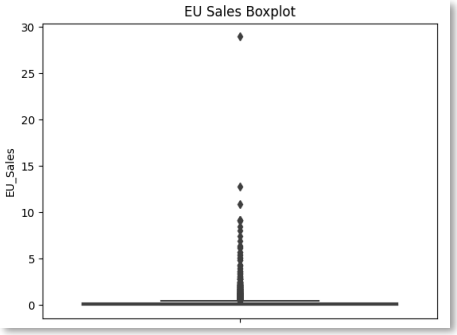
2.1. EXPLORATORY DATA ANALYSIS

בשלב זה נסביר כל אחד מהמשתנים שלנו , סוגם ,מרחב הערכים שלהם ובנוסף נציג את ההסתברויות האפריוריות של משתנים קטגוריאליים ועבור משתנים רציפים נציג היסטוגרמות או תרשימים אחרים אשר ייצגו את הדאטה שלנו בצורה הטובה ביותר.

Name (שם המשחק - משתנה זה לא משפיע על החיזוי הנדרש):

עמודת ה Name מייצגת את שם המשחק שהוא בעצם המזהה לכל דגימה. ניתן לראות בדאטה כי שם של משחק יכול להופיע יותר מפעם אחת אך במידה וזה קורה ה Platform אליה המשחק משויך שונה (בוצעה בדיקה על מנת לבדוק שאכן אין אף דגימה כפולה בדאטה , יוצג בהמשך הדוח ב Pre Processing).

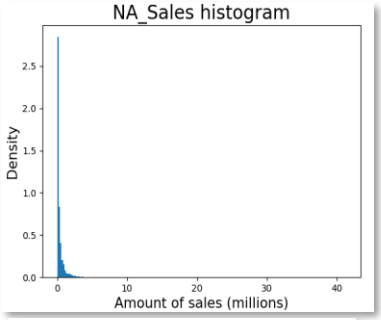


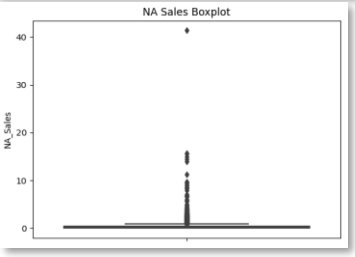


EU_Sales (מכירות באירופה -משתנה המטרה(המוסבר) - **Y**):

משתנה זה הוא משתנה רציף עם טווח ערכים בין 0-28.96. מההיסטוגרמה וה- Boxplot ניתן לראות כי הדאטה לא מאוזנת (ניתן לראות שההתפלגות בגרף ההיסטוגרמה אינה אחידה). אם למשל היינו מחלקים את המשתנה לקטגוריות של על-פי תחומים מסוימים , היינו יכולים לראות שבתחום המכירות הנמוך היינו מקבלים את רוב הדגימות ובשאר היינו מקבלים מספר דגימות קטן מאוד דבר שמעיד על דאטה לא מאוזן. בנוסף על-פי טבלת סיכום המשתנים הרציפים בנספחים , ניתן לראות כי הממוצע הוא 0.235179 והחציון ב 0.06 ובשילוב עם ה Boxplot ניתן לראות שיש מספר תצפיות חריגות מאוד כמו למשל 28.96 .

משתנים מסבירים רציפים:

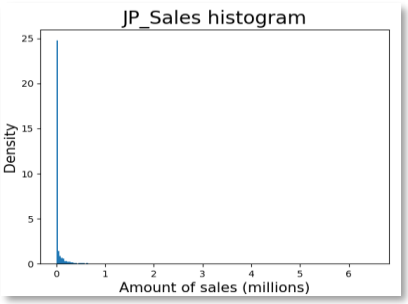


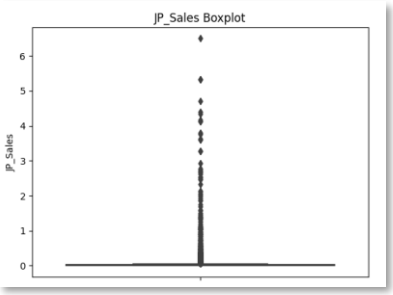


NA_Sales (מכירות בצפון אמריקה):

משתנה זה הינו משתנה רציף בעל טווח ערכים הנע בין 0-41.36 .

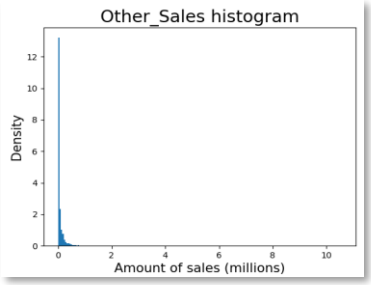
על-פי טבלת סיכום משתנים ניתן לראות כי ממוצע המכירות הוא 0.393794 , כלומר רוב המכירות נמצאות בטווח זה , דבר שבנוסף לתרשים הקופסא מעיד על כך שישנם מספר תצפיות חריגות , כאשר החריגה מבניהם 41.36. יכול להיות שערכים חריגים אלו הוכנסו באופן שגוי לדאטה והם אינם אמיתיים כלל או שאכן אלו נתונים חריגים מאוד אך אמינים. בהמשך הפעולות שנבצע על הדאטה נשקול להסיר תצפיות חריגות אלו אם נבין כי הן עלולות להפריע לנו בחיזוי.

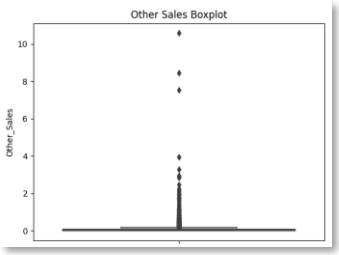




JP_Sales (מכירות ביפן):

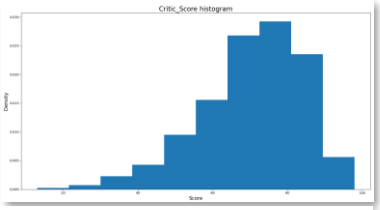
משתנה זה הינו משתנה רציף בעל טווח ערכים הנע בין 0-6.5. על-פי טבלת סיכום משתנים ניתן לראות כי ממוצע המכירות הוא 0.066058. לפי תרשים הקופסא ניתן לראות כי ישנם מספר תצפיות חריגות , כאשר החריגה מבניהם 6.5, ניתן לראות שלעומת המכירות האחרות החריגה כאן היא מתונה יותר ולכן יכול להיות שנשקול לא לנפות את החריגים על מנת שכן תיהיה התייחסות מסויימות לכמויות גדולות של מכירות. בנוסף ניתן לראות מההיסטוגרמה שהמכירות ביפן נמוכות לעומת המכירות באמריקה או בשאר העולם.

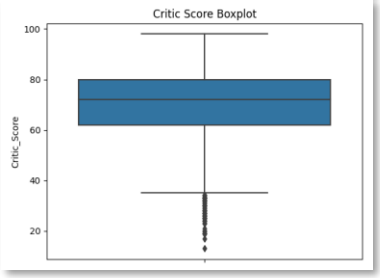




Other_Sales (מכירות בשאר העולם):

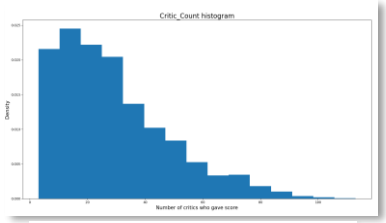
משתנה זה הוא משתנה רציף בעל טווח ערכים הנע בין 0-10.57 . על-פי טבלת סיכום משתנים ניתן לראות כי ממוצע המכירות הוא 0.082245 .לפי תרשים הקופסא ניתן לראות כי ישנם מספר תצפיות חריגות , כאשר החריגה מבניהם 10.57, גם כאן , ננסה להבין את היקפי המכירות בעולם ואם חריגה כזו היא נתון שיכול להיות הגיוני או לא ובמידת הצורך ננפה את החריגים בהמשך.

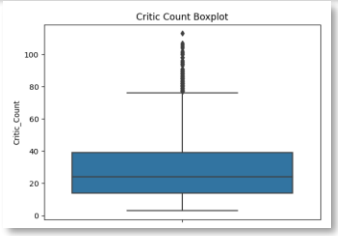




Critic_Score (ציון מבקרים):

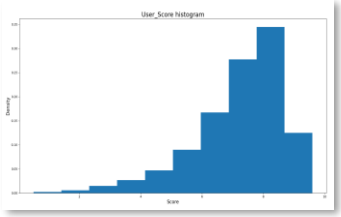
משתנה זה הוא משתנה רציף ובעל טווח ערכים הנע בין 13-98. התפלגות ציון המבקרים מתפלגת נורמלית עם זנב שמאלי כפי שניתן לראות מגרף ההיסטורגרמה . ניתן לראות כי ממוצע הציונים הוא 70.31 , דבר אשר על-פי בדיקה שערכנו במספר אתרים אשר מדרגים משחקים נראה כנתון אמין אשר משקף את המציאות. על-פי תרשים הקופסא ניתן לראות כי יש מספר חריגים אך אין שום חריגה מטווח הערכים המקובל לציונים (0-100) ולכן לדעתינו הנתונים תקינים ואמינים ואין כאן צורך בניפוי חריגים בהמשך.

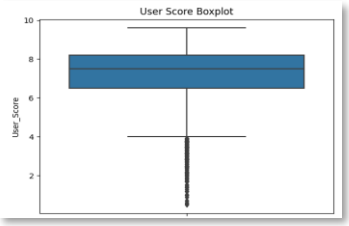




Critic_Count (כמות המבקרים):

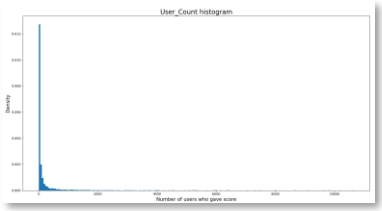
משתנה זה הוא משתנה רציף ובעל טווח ערכים הנע בין 3-113. התפלגות כמות המבקרים נראת כמו התפלגות נורמלית קטומה סביב התוחלת (28.9215). ניתן לראות גם כאן על-פי תרשים הקופסא כי ישנם תצפיות חריגות , אך כל התצפיות הן בעלות ערך חיובי (כלומר תקין , אין כמות שלילית למשל של מבקרים) ובנוסף אין תצפיות חריגות בצורה קיצונית. הערך המקסימלי של מבקרים הוא 113 , כמות מבקרים שנשמעת סבירה והגיונית ולכן אנו משערים כי הנתונים אמינים ואין חריגים שנצטרך לנפות.

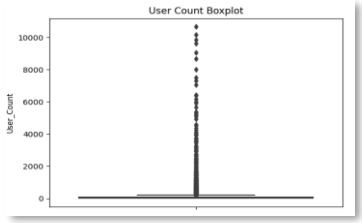




User_Score (ציון משתמשים):

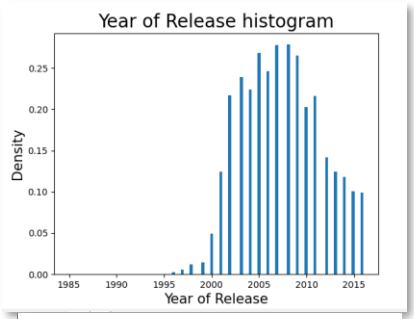
משתנה זה הוא משתנה רציף ובעל טווח ערכים הנע בין 0.5-9.6 .התפלגות ציון המבקרים מתפלגת נורמלית עם זנב שמאלי כפי שניתן לראות מגרף ההיסטורגרמה . ניתן לראות כי ממוצע הציונים הוא 7.1948 , דבר אשר על-פי בדיקה שערכנו במספר אתרים אשר מדרגים משחקים נראה כנתון אמין אשר משקף את המציאות ובנוסף מתאים לדירוג משתמשים שהסקלה שלו בדרך כלל נעה בין 0-10. על-פי תרשים הקופסא ניתן לראות כי יש מספר חריגים אך אין שום חריגה מטווח הערכים המקובל לציונים (0-10) ולכן לדעתינו הנתונים תקינים ואמינים ואין כאן צורך בניפוי חריגים בהמשך.

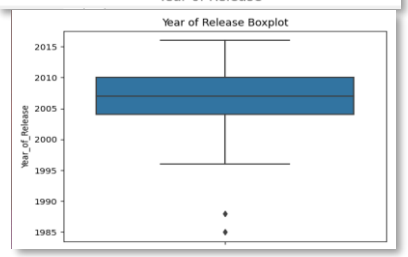




User_Count (כמות המשתמשים המדרגים):

משתנה זה הוא משתנה רציף ובעל טווח ערכים הנע בין 4-10,665 . ממוצע כמות המשתמשים שדירגו את המשחקים הוא 172.35. מכאן ניתן להבין את גרף ההיסטוגרמה המראה את הסיכוי הנמוך להיות בכל כמות מדרגים. טווח הערכים הגדול אינו מפתיע וניתן להעריך כי נובע מהפופולריות השונה של משחקים. כפי שאנו מכירים את עולם המשחקים, יש משחקים נפוצים מאוד אשר מאגדים סביבם קהילת משתמשים גדולה ולעומת זאת משחקים בעלי קהילת משתמשים קטנה מאוד , דבר זה מסביר את השונות הגדולה (585). מכאן אנו מסיקים שנתונים אלו אכן אמינים ומשקפים את המציאות והנתונים החריגים הרבים שאנו רואים בתרשים הקופסא הם הגיוניים.





Year of Release (שנת השחרור של המשחק):

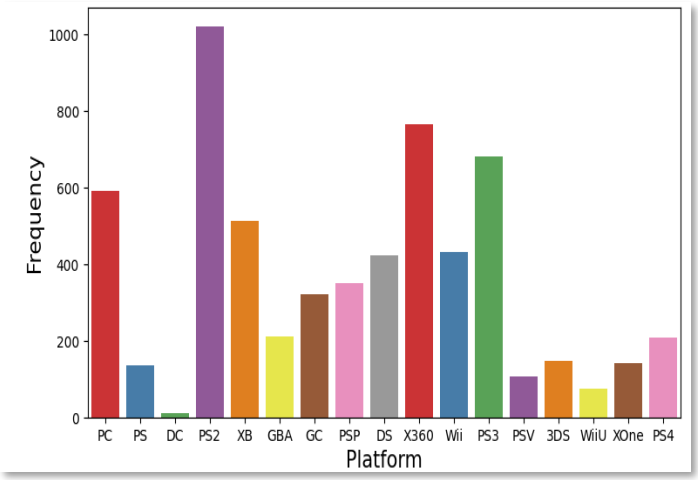
משתנה זה הוא משתנה בדיד אך בשלב זה החלטנו להתייחס אליו כמשתנה רציף ובעל טווח ערכים הנע בין 1985-2016. ניתן לראות מההיסטוגרמה שרוב המשחקים בדאטה שלנו שוחררו לאחר שנת 2000. נוסף על כך אנו מבינים כי ככל הנראה תהיה חשיבות רבה יותר למשחקים ששוחררו מאוחר יותר מכיוון שהם משקפים יותר טוב את המציאות הנוכחית ולכן נשקול בהמשך להפוך את השנים לקטגוריות של שחרור מאוחר ומוקדם וניתן משקל בהתאם. בנוסף מתרשים הקופסא ניתן לראות שיש 2 תצפיות חריגות של שנים שלדעתינו כבר לא רלוונטיות ואינן יכולות לספק מידע אמין לגבי חיזוי עתידי ולכן בהמשך ככל הנראה נסיר אותן מהדאטה או שנשקול החלפה.

משתנים מסבירים קטגוריאליים:

Platform (פלטפורמת המשחק):

משתנה זה הוא משתנה קטגוריאלי אשר מעיד על הפלטפורמה שאליה יצא כל משחק. ישנן 17 קונסולות שונות. לכל קטגוריה חושבה ההסתברות שמשחק יהיה שייך לקונסולה מסויימת (כפי שניתן לראות בתמונה המצורפת). ניתן לראות שאכן הפלטפורמות המוכרות יותר בעולם כמו Wii , PS ו XBOX הן בעלות ההסתברות הגבוה ביותר , דבר שעשוי להעיד על כך שאכן הנתונים אמינים.

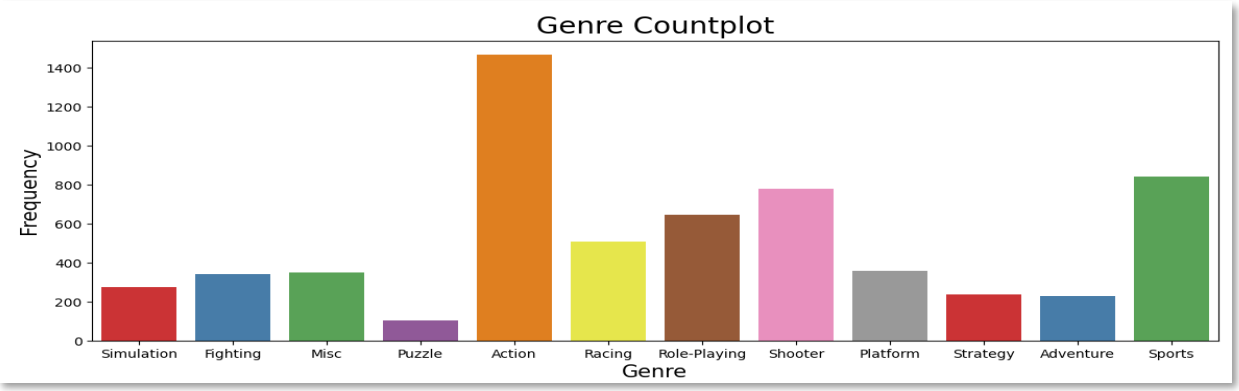
	Platform	prob
0	3DS	0.024096
1	DC	0.002117
2	DS	0.069033
3	GBA	0.034516
4	GC	0.052263
5	PC	0.096223
6	PS	0.022143
7	PS2	0.166070
8	PS3	0.110876
9	PS4	0.033865
10	PSP	0.056985
11	PSV	0.017421
12	Wii	0.070173
13	WiiU	0.012211
14	X360	0.124878
15	XB	0.083686
16	XOne	0.023445



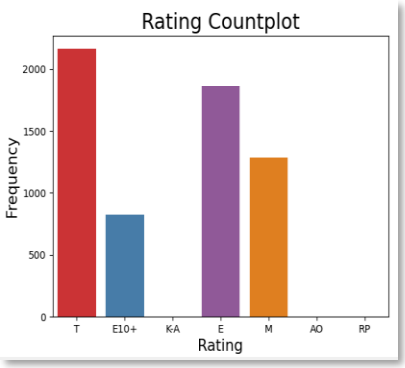
Genre (ז'אנר):

משתנה זה הוא קטגוריאלי ומעיד על הסיכוי של משחק להיות שייך לז'אנר משחקים מסויים. ישנם 12 ז'אנרים שונים של משחקים. ניתן לראות שכפי שאנו מכירים וחקרנו קצת באתרים שונים , הז'אנרים עם ההסתברות הגבוה ביותר הם משחקי פעולה , יריות וספורט שאכן הם המשחקים הפופולריים ביותר בעולם. מהנתונים שהצגנו ניתן להבין כי הדאטה אכן אמין ומשקף את המציאות נכונה.

	Genre	prob
0	Action	0.238684
1	Adventure	0.037121
2	Fighting	0.055519
3	Misc	0.056822
4	Platform	0.058776
5	Puzzle	0.016933
6	Racing	0.083035
7	Role-Playing	0.105015
8	Shooter	0.127157
9	Simulation	0.044937
10	Sports	0.137252
11	Strategy	0.038750



Rating (דירוג **ESRB** -למי מיועד המשחק):

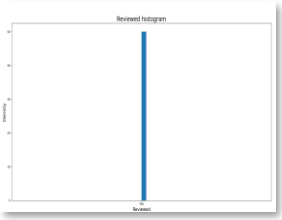


	Rating	prob
0	AO	0.000163
1	E	0.303484
2	E10+	0.134321
3	K-A	0.000163
4	M	0.209541
5	RP	0.000163
6	T	0.352165

משתנה זה הוא משתנה קטגוריאלי אשר מציג את ההסתברות של משחק להיות שייך לקטגוריית גיל מסוימת. במשתנה הדירוג קיימים 7 קטגוריות . ניתן לראות כי רוב המשחקים מתאימים לכולם (E) או לצעירים (T). ולאחר מכן יש עוד קטגוריות גיל ספציפיות יותר שההסתברות שלהם נמוכה יותר. גם כאן יש שיקוף טוב של המציאות והחלוקה לקטגרויות גיל אמינה. בנוסף ניתן לראות כי ישנם קטגוריות אשר אין בכלל או שיש דגימות בודדות השייכות להן כמו K-A שהיא דרך הסימון הישנה של E ו RP אשר מעיד על משחק שלא נקבע לו דירוג, כלומר עדיין לא יודעים למי מיועד המשחק ובנוסף ו AO שמיועד למבוגרים בלבד. את קטגוריות אלו ככל הנראה נרצה להסיר בהמשך משום שנראה שהן לא אינפורמטיביות ולא יועילו למודל שלנו.

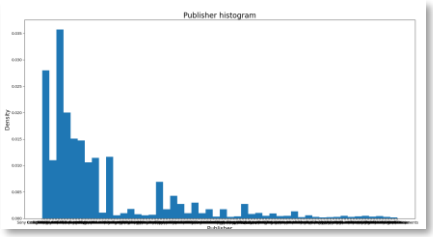
Reviewed (האם נבדק ב ESRB):

	Reviewed	prob
0	YES	1.0



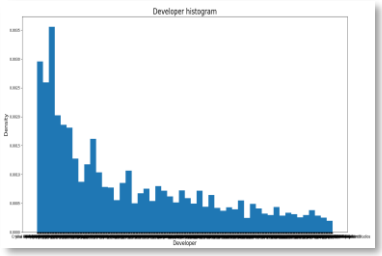
זהו משתנה קטגוריאלי בינארי אשר מעיד האם המשחק נבדק על-ידי ESRB או לא. ניתן לראות כי כל המשחקים בדאטה שלנו נבדקו על-ידי ESRB ולכן משתנה זה נראה כמשתנה לא רלוונטי אשר לא ישפיע לנו על המודל בהמשך ולכן ככל הנראה נסיר אותו.

Publisher (מוציא לאור):



זהו משתנה קטגוריאלי המעיד על ההסתברות של משחק להיות מוצא לאור על-ידי מוציא לאור מסוים . מהניתוח שעשינו ראינו כי ישנם 248 מוציאים לאור, מספר גדול יחסית של מוציאים לאור שונים לדאטה של כ 6000 רשומות , לכן החלטנו שחישוב הסתברות לכל אחד מהם יהיה מיותר. למרות זאת ניתן להבחין מגרף ההיסטוגרמה שהוצאנו כי ישנם מספר מוצאים לאור מצומצם יותר שהינם המוצאים לאור המוכרים ביותר שרוב המשחקים הוצאו על-ידם. דבר זה מתכתב בצורה הגיונית עם המציאות , שכן אנו יודעים כי ישנם מוצאים לאור חזקים מאוד בשוק לעומת הרבה מאוד מוציאים לאור קטנים ולא מוכרים אשר לא מוציאים משחקים רבים. מכאן שאנו מסיקים שהדאטה אמין.

Developer (מפתח):



זהו משתנה קטגוריאלי המעיד על ההסתברות של משחק להיות מפותח על-ידי מפתח כלשהו . מהניתוח שעשינו ראינו כי ישנם 1237 מפתחים , מספר עצום של מפתחים שונים לדאטה של כ 6000 רשומות , ולכן החלטנו שחישוב הסתברות לכל אחד מהם יהיה מיותר. למרות זאת ניתן להבחין מגרף ההיסטוגרמה שהוצאנו כי ישנם מספר מפתחים מצומצם יותר שהינם המפתחים המוכרים ביותר שרוב המשחקים פותחו על-ידם. דבר זה מתכתב בצורה הגיונית עם המציאות , שכן אנו יודעים כי ישנם מפתחים חזקים מאוד בשוק לעומת הרבה מאוד מפתחים קטנים ולא מוכרים אשר לא מפתחים משחקים רבים. מכאן שאנו מסיקים שהדאטה אמין.
בנוסף נגיד שכמות מפתחים גדולה כזו עשויה להיות לא רלוונטים בסופו של דבר למודל ולכן נשקול לבצע פרוצדורות מסויימות כדי להתאים את הפיצ'ר למודל או שנשקול להסיר אותו.

2.2. DATA SET CREATION

שלב קביעת הפי'צרים כבר בוצע על ידי צוות הקורס. בנוסף, גם צורת ההצגה של הdateset כבר נבחרה (הצגה ב- matrix). לכן, נתחיל עם ה- sensed raw data ונבצע שינויים , ניתוחים ומניפולציות על הדאטה שלנו ועל הפיצ'רים על פי כל השלבים וזאת על-מנת שנוכל להזין את המודל שלנו בדאטה הרלוונטית והאינפורמטיבית ביותר שבסופו של דבר תספק לנו את התוצאות הטובות ביותר.

A. Pre-processing

Redundancy in the data:

בשלב זה ביצענו בדיקת כפילויות בדאטה שלנו.על מנת לבדוק האם ישנם כפילויות כתוצאה מטעויות הקלדה או שכפול של שורות בדאטה השתמשנו בקוד הבא :

```
#-----Check if there is duplicate in our data----
checkDuplicates=df.drop_duplicates(subset=None, keep='first')
print(checkDuplicates)
```

מצאנו כי אין שורות שלמות כפולות בדאטה , מה שאישר לנו שהדאטה שלנו תקין ואין בו כפילויות אשר עשויות להטות את תוצאות המודל שלנו.

Missing values:

כדי לבדוק האם יש ערכים חסרים בדאטה סט, הרצנו את הפקודה `df.isnull().values.sum()` בפייתון וראינו כי לא קיימים ערכים חסרים ולכן אין צורך לבצע פעולות של השלמת ערכים בשלב הזה.

Exceptional values:

בשלב זה בדקנו האם ישנם ערכים חריגים בדאטה אשר אנו חושבים כי עלולים לבצע הטעיה למודל שלנו או שאינם רלוונטים לנו. יש מספר פיצ'רים בעלי ערכים חריגים (כפי שכבר הסברנו עליהם בניתוח הראשוני) ובהם החלטנו לטפל.

1. הדבר הראשון שבו בחרנו לטפל הוא שנות שחרור של משחקים ישנים מאוד. קיימים 2 ערכים כאלו , משחקים ששוחררו בשנת 1985 ו 1988. החלטנו להסיר את המשחקים ששוחררו בשנים אלו מהדאטה שלנו משום שאנו חושבים כי משחקים שהוצאו ונמכרו בשנה כה מוקדמת אינם רלוונטים עוד לחיזוי מכירות לשנת 2020 והלאה ולכן החלטנו להסיר אותם.

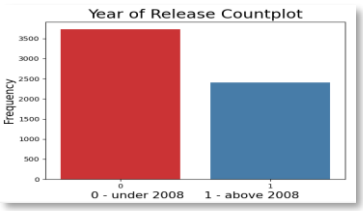
```
#----Exceptional values----
df=df.drop(df[df['Year_of_Release']<1990].index)
print(df)
```

2. בפיצ'ר Genre , כפי שציינו בניתוח הנתונים הראשוני, ישנם מספר קטגוריות שמספר הדגימות השייכות להם כמעט אפסי ובנוסף כפי שהסברנו עליהן , הן קטגוריות לא רלוונטיות כל אחת מהסיבות שהוסברו ולכן החלטנו להסיר אותן וכיוצא מכך הסרנו את הדגימות המשתייכות לקטגוריות אלו. הדגימות שהוסרו הן דגימות השייכות לקטגוריות K-A , RP , AO.

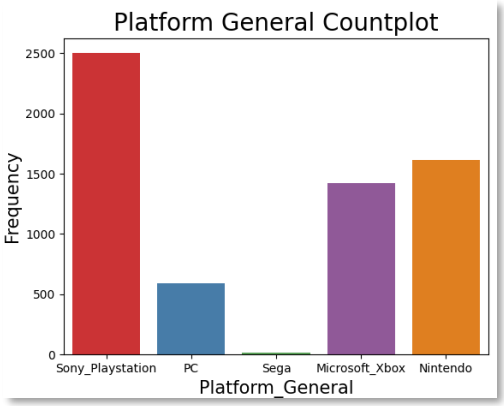
Data type conversions:

חקרנו את הנתונים ובחנו האם יש מקום לבצע המרה של משתנים רציפים לקטגוריאלים או להפך או לבצע איחוד קטגוריות למשתנים קטגוריאלים. בחרנו להמיר את משתנה Year of Release למשתנה קטגוריאלי בינארי - לפני ואחרי שנת 2008 כפי שניתן לראות בקוד הפייתון :

```
#----Change Year of Release to 0/1 - 0-under 2008 , 1-above 2008----
df['Year_of_Release'] = (df['Year_of_Release'] > 2008).astype(int)
```



בחרנו לעשות זאת כיוון שאנו רוצים לייחס חשיבות לרלוונטיות של המשחק (משחקים אשר יצאו מאוחר יותר רלוונטים יותר ממשחקים ישנים ועשויים לספק מידע אמין יותר על מכירות עתידיות).



דבר נוסף שבחרנו לעשות הוא לאחד את כל הפלטפורמות השייכות לאותה חברה יחד , כלומר למשל כל הפלטפורמות אשר שייכות ל Sony Playstation יוכנסו תחת שם החברה , לדוגמא PS, PS1 , 2PS וכו'. אנו חושבים כי האיחוד הזה תחת החברה יתן אינפורמציה טובה יותר מאשר כמות גדולה של פלטפורמות שלחלקן יש מספר משחקים מאוד קטן שיוצא אליהן ושחלקן פלטפורמות ישנות כבר ופחות רלוונטיות.

פעולה אחרונה שביצענו הייתה המרה של הערכים הקטגוריאליים בדאטה לערכים מספריים.

:Proportions in the data

קיבלנו משתנה מטרה רציף ולכן קשה לדבר במונחים של איזון על הדאטה בשלב זה. למרות זאת על פי גרף ההיסטוגרמה שהצגנו בסעיפים הקודמים (בניתוח תחת EU_Sales) , הדאטה לא מאוזן אך מצד שני מהידע שקיים אצלינו , חוסר האיזון הוא די הגיוני משום שכמות מכירות של משחקים שונים הוא לגמרי לא מאוזן ולכן הגיוני שהדאטה שלנו נראת ככה. בהמשך אולי כדאי לשקול את הפיכת המשתנה המוסבר למשתנה קטגוריאלי ולא רציף ואז נוכל לבצע בדיקת איזון מסודרת ובהתאם לכך לבצע מניפולציות מתאימות בדאטה כמו הוספה או הורדת נתונים על מנת לאזן את הדאטה.

B. Segmentation

ב segmentation המטרה היא לבצע פילוח / חלוקה למקטעים על מנת להבדיל את האובייקט המרכזי מרעשי הרקע או מאלמנטים לא הכרחיים למשל כמו תמונות מקבצי וורד וכו'. במקרה שלנו עבור הדאטה סט שקיבלנו, לא מצאנו דרך שבה יישום שלב זה יכול להועיל לנו משום שלא מדובר פה בקבצים או תמונות שמהם אנו צריכים לחלץ מידע כלשהו ובנוסף קיבלנו סט פיצ'רים מוכן ולכן בחרנו שלא לבצע אותו.

C. Feature Extraction

נרצה בשלב זה לחלץ/לייצר את הפי'צרים הטובים ביותר עבור משימת הלמידה שלנו. הפיצ'רים שלנו סטטיים ואינם תלויים בזמן. בנוסף, כמות הפיצ'רים שלנו קבועה ולא משתנה עבור כל ישות. לכן, בחרנו במתודולוגיית Specific – Knowledge based עבור חילוץ הפיצ'רים. למרות שהידע שלנו בנושא יחסית מוגבל, אנחנו מכירים את התחום ומצאנו לנכון כי שיטה זו היא הרלוונטית ביותר עבורנו.

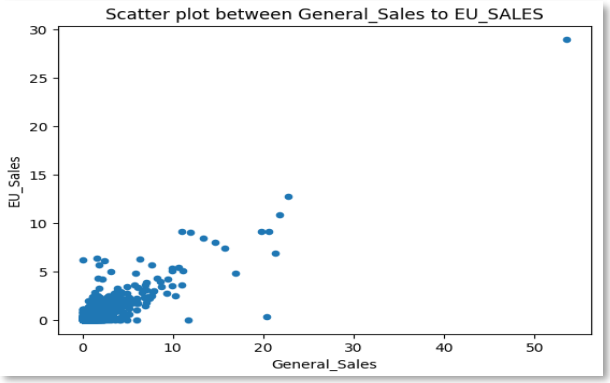
- פיצ'ר ראשון שחילצנו הוא General_Sales. פי'צר זה הוא סכום המכירות בצפון אמריקה, יפן ושאר המכירות בעולם (ללא מכירות באירופה) יחד. כיוון שאנו מנסים לחזות את מכירות המשחקים באירופה , מעניין אותנו כמה הוא נמכר סך הכל בשאר העולם והחלוקה לאיזורים פחות מעניינת. טענה זו גם נתמכת על ידי בדיקת מתאם שעשינו, כאשר המתאם של פיצ'ר זה עם המשתנה המוסבר הוא 0.87, לעומת 0.84, 0.53 ו- 0.71 כאשר כל אחד מהם בנפרד. בדקנו גם האם יהיה נכון יותר להפריד את המכירות של יפן כיוון שהן מסבירות פחות טוב את המכירות באירופה, אך ההתאמה ללא מכירות אלה יורדת ל-0.86. בכל זאת בגלל שההפרדה מורידה את ההתאמה ב 0.01 בלבד , אולי נשקול באימון המודל לבדוק את שתי האופציות של יפן בנפרד ויפן ביחד עם השאר על מנת לראות אם תהיה השפעה שונה על המודל ואולי נקבל תוצאות טובות יותר.
- פיצ'רים נוספים שחילצנו הם שילובים של הפיצ'רים Critic Count עם Critic Score ושל User Count עם User Score. בשני המקרים ביצענו שילוב בין כל שני פיצ'רים ויצרנו פיצ'רים חדשים בשמות Critic_Weight ו User_Weight . יצרנו את הפיצ'רים הללו על-ידי הנוסחה הכללית הבאה :

$$Weight = Score * \sqrt{Count}$$

לנוסחה הזו הגענו לאחר ביצוע מספר נוסחאות ובדיקות של מפת חום לבדיקת המתאם בין הפיצ'ר החדש למשתנה המטרה , בדקנו הכפלה רגילה , הכפלה עם לוגים ומספר דברים נוספים וקיבלנו את התוצאות הטובות ביותר ששיפרו את המתאם מהפיצ'רים הישנים בנוסחה שהצגנו. קיבלנו מתאמים של 0.28 ו 0.32 בהתאמה.

D. Feature Representation

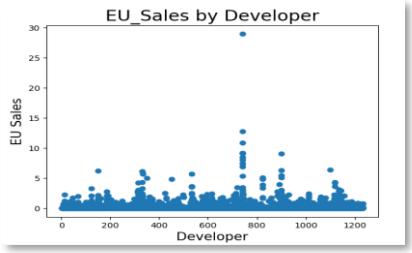
שלב זה הוא השלב בו נחליט כיצד לייצג את פיצ'רים שחילצנו בשלב ה Feature Extraction. הפיצ'רים הראשונים שנייצג יהיו הפיצ'רים שייצרנו מהפיצ'רים Critic Score ו User Score בשילוב עם Critic Count ו User Count. הפיצ'רים המקוריים היו בסקלות אחרות של 0-10 ו 0-100 ולכן גם הפיצ'ר החדש שחילצנו User_Weight ו Critic_Weight הם לא בעלי אותה סקלת ערכים. מכאן שנרצה לבצע נרמול ולהביא את שני הפיצ'רים החדשים שחילצנו לסקלה דומה בין 0-1. נעשה זאת על-ידי חלוקה של כל ערך בערך המקסימלי האבסולוטי על פי הנוסחה הבאה: $X = \frac{x}{\max(|x|)}$



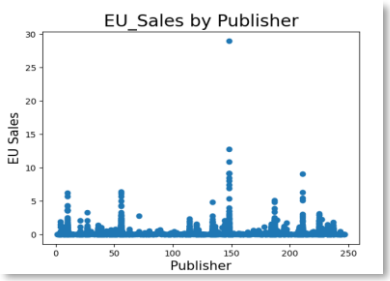
הפיצ'ר השני שנציג הוא האיחוד בין המכירות בצפון אמריקה , יפן ושאר העולם (לא כולל אירופה). החלטנו להציג את הפיצ'ר על-ידי scatter plot ולהראות אותו אל מול המכירות באירופה. ניתן לראות כי נראה שיש קשר לינארי בין הפיצ'ר החדש שחילצנו לבין משתנה המטרה שלנו , ולכן אנו מסיקים כי הפיצ'ר יהיה בעל קורלציה גבוה מאוד ויכול לספק למודל שלנו מידע.

E. Feature Selection

- הדבר הראשון שביצענו הוא ניפוי של הפי'צר name. השם של המשחק הוא רק שם, ולכן הוא אינו מייצג את המכירות הפוטנציאליות של משחק והחלטנו להסירו.
- בנוסף, ניפינו את הפיצ'ר Reviewed. עשינו זאת משתי סיבות עיקריות - (1) כל הערכים בדאטה שלנו היו TRUE, כלומר הפיצ'ר הזה על הדאטה הזה לא מוסיף לנו שום מידע. (2) כאמור אנחנו לא רוצים לעבוד עם דאטה חסר, במידה ומשחק מסוים יהיה FALSE בעמודה זו, סימן שלא יהיה לו RATING.
- כיוון שיש כמות מאוד גדולה של מפתחים ביחס לדאטה (כ-1200 מפתחים על כ-6400 רשומות) ראינו לנכון לשקול את ניפוי פיצ'ר זה. ביצענו מחקר באינטרנט ומצאנו כי רוב המשחקים ידועים בעיקר על ידי מי פרסם אותם, ולא מי פיתח ועיצב. מעבר לכך, ניתן לראות כי ישנו מתאם נמוך עם המכירות באירופה ומהגרף לא נראה כי המפתח נותן מידע רלוונטי לגבי כמות המכירות. לאור זאת, החלטנו לנפות את פיצ'ר Developer.

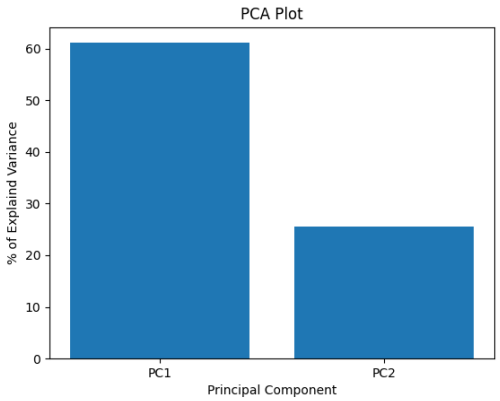


- לבסוף, בחרנו לנפות גם את פיצ'ר Publisher. הסיבה לכך היא שיש לנו כמות מפרסמים גבוהה מאוד (כ-250). לדעתנו כמות כזו לא תניב הסברה טובה של המכירות. טענה זו מחוזקת גם על ידי תרשים הפיזור אשר מראה מתאם נמוך עם המכירות באירופה. כיוון שיש מספר של publishers אשר מפרסמים כמות גבוהה מאוד של משחקים ביחס לאחרים, שקלנו להתייחס רק אליהם, אולם דבר זה יאלץ לא להתייחס לחלק גדול מהדאטה שלנו, ובשלב זה בחרנו לא לעשות זאת.



- בכל הנוגע לשאר המשתנים קטגוריאליים, בחרנו להשאירם. הוצאנו גרפי פיזור לכל אחד מהם(כפי שניתן לראות בנספח) עם המשתנה המוסבר וניתן לראות שכולם מסבירים אותו. משתנה השנה והג'אנר הם בעלי מתאם יחסית נמוך, אולם בשלב זה בחרנו להשאיר אותם, ולבחון האם הוצאתם תשפר את אמינות המודל בהמשך.
- את המשתנים הרציפים שנשארו לנו במודל בחנו תחילה על ידי מפת חום(בנספח). ניתן לראות כי המכירות בשאר העולם מסבירות מצוין את המשתנה המוסבר. לעומת זאת, דירוג הUsers וה-Critics מסבירים פחות טוב. בחרנו להריץ מודל רגרסיה לפנים כדי לקבוע אילו משתנים ישארו במודל. תוצאת הרגרסיה הייתה: **EU_Sales** ~ General_Sales + User_Weight + Critic_Weight
- לסיכום, הפיצ'רים אשר נבחרו למודל הם: General_Sales, User_Weight, Critic_Weight, Rating, Platform_General, Genre, Year_of_Release

F. Dimensionality Reduction



בשלב זה המטרה היא לבצע הורדת מימד. המטרות העיקריות של שלב זה הם צמצום מספר הפיצ'רים ויצירת סט פיצ'רים חדש ואינפורמטיבי יותר ובנוסף יצירת הכמסה – יצירת פיצ'רים חדשים מתוך הקיימים המכילים את המידע וזאת על מנת למנוע את חשיפת הפיצ'רים שלנו. לצורך הורדת המימד בחרנו להשתמש ב **PCA- Principle component analysis** שהיא השיטה המרכזית בעולם להורדה מימד. ביצענו **PCA** באמצעות הפייתון, נתנו לו את שלושת המשתנים הרציפים שנשארו לנו ודרשנו להוריד מימד ל 2. ואכן קיבלנו 2 **PCi** , אחד בעל אחוז שונות של כ 60% , מעיד על פיצ'ר טוב ואיכותי , ואחד עם כ 25% שונות , נמוך יותר אבל עדיין יכול להיות פיצ'ר יותר טוב ממה שהיה לנו לפני. לסיכום , קיבלנו 2 משתנים חדשים במקום ה 3 שהכנסנו ואנו מאמינים כי פיצ'רים אלו יעזרו לנו לשפר את המודל.

3. MODEL TRAINING

א. נבצע cross validation test אשר למדנו בכיתה. נחלק אתה הדאטה לפי שיטת K-fold עם K = 10 שנחשבת לשיטה הפופולארית והטובה ביותר. שיטת leave-one-out נפסלה כי אנו חושבים שגודל הדאטה שלנו לא מספיק גדול והלמידה לא מאוד קשה ולכן ואין סיבה להשתמש בשיטה זו.Holdout נפסלה כיוון שלא ראינו סיבה שלא ניתן להשתמש ב-K Foldאשר נהוגה יותר לשימוש. היתרון של ה K-Fold לעומת holdout הוא שהוא מניב שונות נמוכה יותר בתוצאות. הוא מסתמך פחות על training set ספציפי וכך מרחיב את נכונות המודל ליותר מסט נתונים אחד. היתרון לעומת leave-one-outהוא בזמן החישוב ,חלוקה ל-10 פולדים לעומת N .

ב. מוצג בנספח דוגמא לחלוקה (נספח)

ג. מטרת כל סט נתונים היא לבצע ולידציה למודל שלנו. על ידי כך שאנחנו בודקים כל פעם סט נתונים טיפה שונה מול סט ולידציה שונה, אנחנו יכולים לאמן את המודל שלנו בצורה כללית יותר וכן לבדוק עליו דברים ולבצע שינויים במידת הצורך וכך להקטין את טעות הבדיקה כשנבדוק את המודל על ה- test set האמיתי.

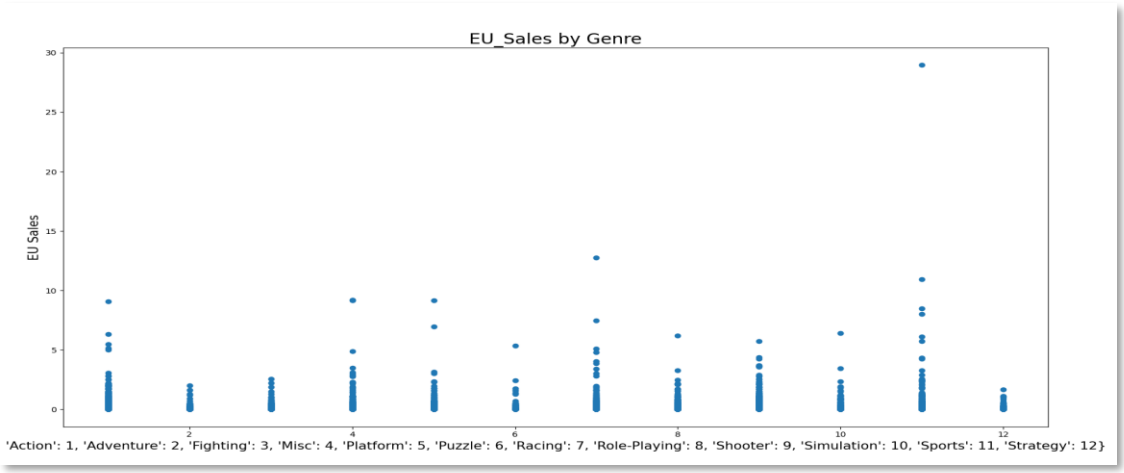
נספחים

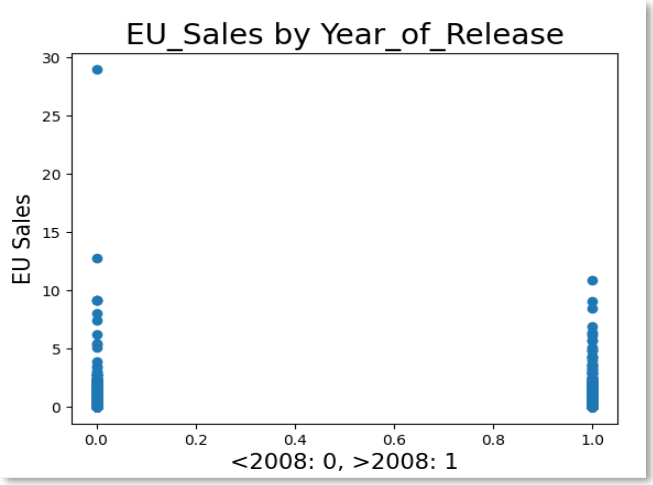
טבלת סיכום משתנים רציפים:

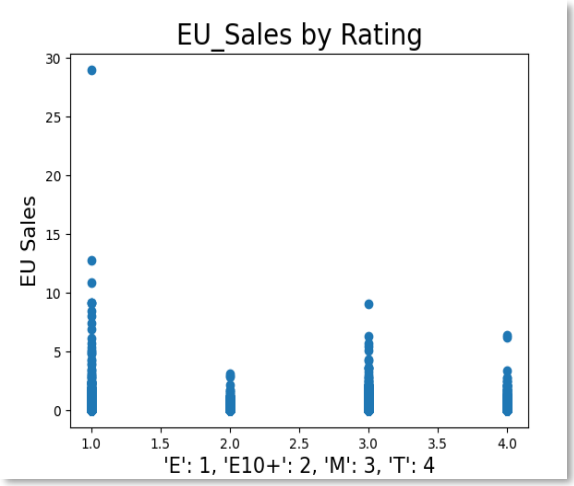
	Year_of_Release	NA_Sales	JP_Sales	Other_Sales	Critic_Score
count	6142.000000	6142.000000	6142.000000	6142.000000	6142.000000
mean	2007.414035	0.393794	0.066058	0.082245	70.310811
std	4.202343	0.982203	0.298954	0.275721	13.812783
min	1985.000000	0.000000	0.000000	0.000000	13.000000
25%	2004.000000	0.060000	0.000000	0.010000	62.000000
50%	2007.000000	0.150000	0.000000	0.020000	72.000000
75%	2010.000000	0.390000	0.020000	0.070000	80.000000
max	2016.000000	41.360000	6.500000	10.570000	98.000000

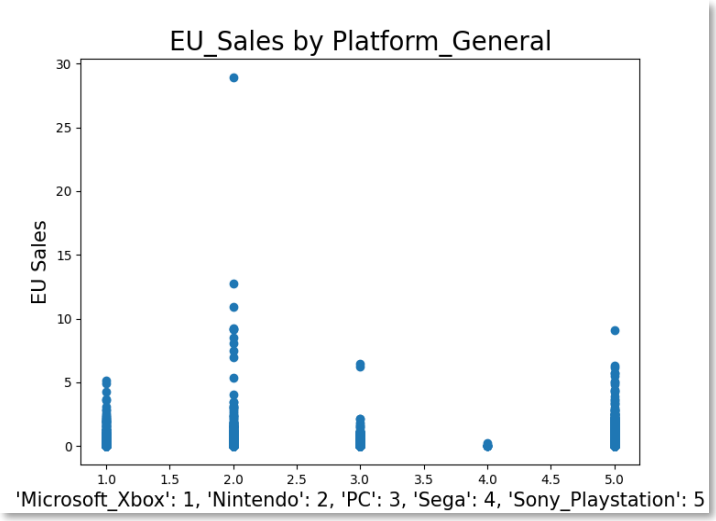
	Critic_Count	User_Score	User_Count	EU_Sales
count	6142.000000	6142.000000	6142.000000	6142.000000
mean	28.921524	7.194888	172.352654	0.235179
std	19.203494	1.427799	585.699762	0.700702
min	3.000000	0.500000	4.000000	0.000000
25%	14.000000	6.500000	11.000000	0.020000
50%	24.000000	7.500000	27.000000	0.060000
75%	39.000000	8.200000	87.000000	0.210000
max	113.000000	9.600000	10665.000000	28.960000

SCATTER PLOTS בין המשתנים הקטגוריאליים למשתנה המטרה :

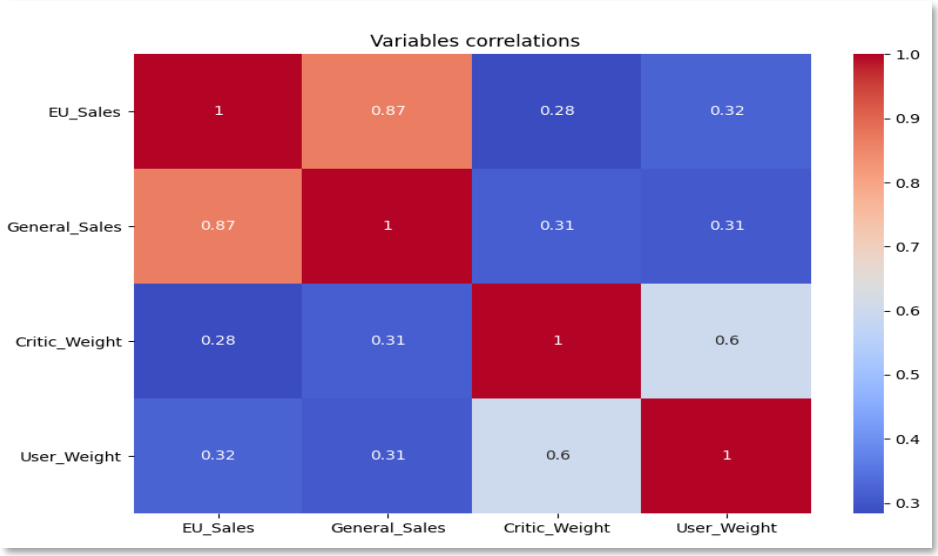








מפת חום : (משתנים רציפים בלבד)



K-FOLD

דוגמא לאחת מ 10 הקבוצות שהקוד שכתבנו חילק את הדאטה , ניתן לראות שהוא בחר קבוצה מ0 עד 613 וקבוצה הבאה מתחילה ב 614.

```
[ 614  615  616 ... 6134 6135 6136] [  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35
  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53
  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71
  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89
  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107
 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125
 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143
 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161
 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179
 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197
 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215
 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233
 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251
 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269
 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287
 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305
 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323
 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341
 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359
 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377
 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395
 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413
 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431
 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449
 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467
 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485
 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503
 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521
 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539
 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557
 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575
 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593
 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611
 612 613]
[   0    1    2 ... 6134 6135 6136] [ 614  615  616  617  618  619  620  621  622  623  624  625  626  627
  628  629  630  631  632  633  634  635  636  637  638  639  640  641
  642  643  644  645  646  647  648  649  650  651  652  653  654  655
```