מעבדה בניתוח והצגת נתונים

תרגיל בית 1

Box Office Revenue Prediction

מגישים:

משה עבאדי 324658939

רועי רימר 314828732

לינק ל-Github Repository:

https://github.com/moshinhoabadi/Data-Analysis-hw1

1. Exploratory Data Analysis

a. Which features are available in the dataset?

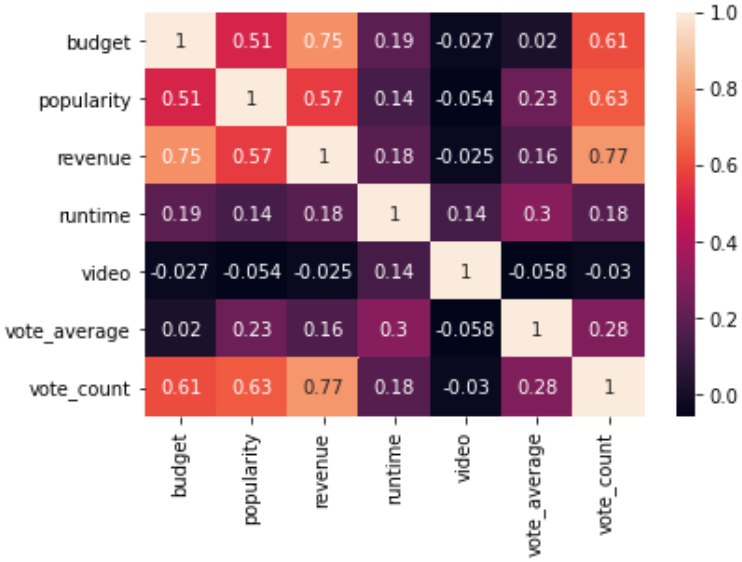
נציג את הפיצ'רים השונים הקיימים בסט הנתונים הנתון:

- backdrop_path – path לתמונה הקשורה לסרט.
- belongs_to_collection – סדרת הסרטים אליה שייך הסרט, אם יש כזו.
- Budget – תקציב הסרט.
- Genres – הז'אנרים (סוגות) השונים אליהם הסרט משתייך.
- Homepage – קישור לאתר הבית של הסרט, אם יש כזה.
- Id – מזהה של הסרט.
- imdb_id – מזהה הסרט באתר IMDB.
- original_language – שפת המקור של הסרט.
- original_title – שם הסרט כפי שמופיע בשפתו המקורית.
- Overview – תיאור קצר של הסרט.
- Popularity – מדד לפופולאריות הסרט.
- poster_path – path לכרזת הסרט.
- production_companies – פרטים על החברות שהשתתפו בהפקת הסרט. פרטים אלו כוללים לוגו, שם, מדינת מקור ו-path ללוגו החברה.
- production_countries – המדינות בהן הופק הסרט.
- release_date – תאריך יציאת הסרט לאקרנים.
- Revenue – הכנסות הסרט (בדולרים).
- Runtime – אורך הסרט (בדקות).
- spoken_languages – שפות בהן השתמשו במהלך הסרט (מידע על כך הסקנו מכאן: https://www.themoviedb.org/talk/555df5c89251416b57003ddd).
- Status – סטטוס לגבי שלב הפקת הסרט. עבור כל הסרטים רשום שהסרט ראה אור ("Released").
- Tagline – המשפט השנון ההוא שמופיע בדרך כלל על הכרזה של הסרט.
- Title – שם הסרט באנגלית. הבחנו כי ערכי שדה זה שונים מן ערכי השדה original_title רק כאשר שם הסרט במקור לא היה באנגלית.
- Video – לא הצלחנו להבין את משמעות השדה הזה. זהו שדה בוליאני שערכו FALSE בכל הרשומות למעט 18 מהן. בנוסף, רוב הרשומות שבהן בשדה זה יש ערך TRUE הן רשומות שמתארות סרט המכסה אירוע האבקות (למשל UFC 97: (Redemption.
- vote_average – ממוצע הדירוג שצופים נתנו לסרט. ערכי שדה זה נעים בן 0 (ציון מינימלי) ל-10 (ציון מקסימלי)
- vote_count – מספר האנשים שדירגו את הסרט.
- Keywords – מילון של ביטויי מפתח הקשורים לסרט.
- Cast – פרטים על השחקנים המשתתפים בסרט. פרטים אלו מכילים לדוגמא את שם השחקן, המגדר שלו ודמותו בסרט.
- Crew – פרטים על רשימת בעלי תפקידים אחרים בסרט. פרטים אלו מכילים למשל את שם האדם, המגדר שלו, תפקידו והמחלקה בתוכה הוא עובד.

b. Feature distribution & comparative analysis

נציג כעת ניתוחים שונים שביצענו על הפיצ'רים, ההתפלגויות שלהם ועל הקשרים או ההבדלים ביניהם:

- <u>מטריצת קורלציה:</u> התחלנו בבניית מטריצת קורלציה של הערכים הנומריים בסט הנתונים. מטריצה זו מוצגת באיור 1.



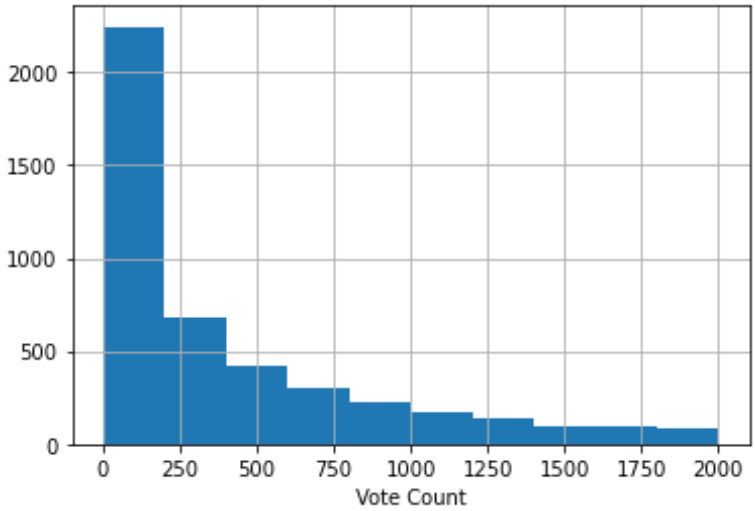
איור 1 – מטריצת קורלציה של הערכים הנומריים בסט הנתונים

הדבר המרכזי שתפס את עינינו במטריצת הקורלציה הוא המתאם הגבוה יחסית שיש בסט האימון בין השדה revenue (אותו נרצה לחזות) לבין השדות:
vote_count (0.77), budget (0.75), popularity (0.57).
נראה כי יש קורלציה חיובית יחסית גבוהה בין הכנסות הסרט לתקציב שלו, למידת הפופולריות שלו, ולכמות האנשים שדירגה אותו. כמות האנשים שדירגה את הסרט מהווה גם היא מעין מדד לפופולריות הסרט, שכן סביר שככל שהסרט יותר פופולרי כך יותר אנשים ייכנסו לאתר על מנת לדרג אותו. ניתן לראות גם עדות לכך במטריצה, שכן יש קורלציה יחסית גבוהה (0.63) בין השדות popularity ו-vote_count.

כתוצאה מסקירת מטריצת הקורלציה, בחרנו להתמקד תחילה בשלושת השדות בעלי הקורלציה הגבוהה לשדה revenue, ולבחון את התפלגותן ואת היחס ביניהן לבין השדה revenue.

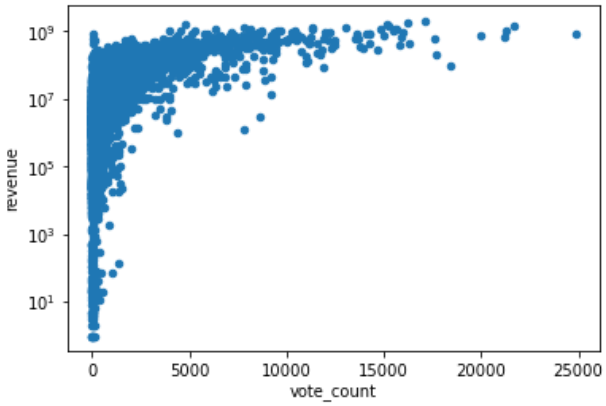
הערה: אנחנו לא בטוחים אם זה מפתיע כל כך, אך על סמך הקורלציה נראה כי אין הרבה קשר בין revenue לבין דירוג הסרט הממוצע (vote_average)...

- <u>השדה vote_count</u>: איור 2 מציג את התפלגות השדה vote_count עד לערך 2000 (ההגבלה נעשתה לשם בהירות האיור).



איור 2 – התפלגות השדה vote_count בקטע [0,2000].

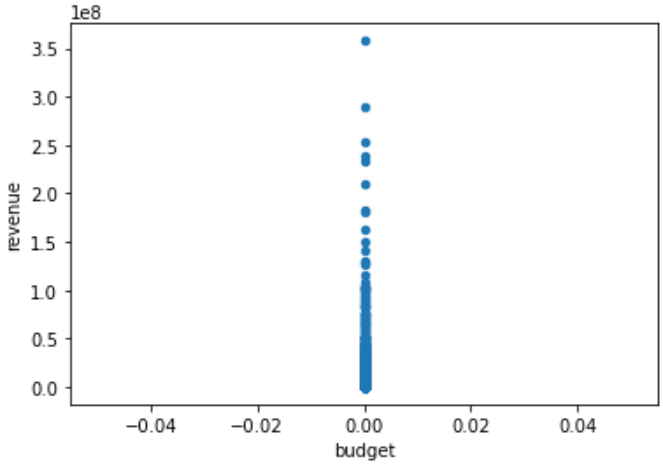
ניתן לראות מגמת ירידה ברורה בכמות הסרטים ככל שערך ה-vote_count גדל. מגמה זו המשיכה גם לאחר הערך 2000.
כעת נרצה לזהות את מגמת השדה revenue כתלות בשדה vote_count. ה-scatter_plot הבא (איור 3) מציג את ערכי ה-revenue וה-vote_count של כל הרשומות בסט הנתונים:



איור 3 – scatter_plot של ערכי vote_count ו-revenue (המוצג בסקאלה לוגריתמית).

ניתן לראות מגמה ברורה של עלייה בהכנסות ככל שיותר משתמשים דירגו את הסרט.
ה"קפיצה" הגדולה ביותר נעשית בתחילת הגרף, עד אשר מגיעים לסף כלשהו של אנשים שדירגו את הסרט, ומשם העלייה מתמתנת יחסית, אך עדיין קיימת. נראה כי הפרשים ברמת החשיפה לסרט (שמתבטאת כאן בכמות המדרגים) לא משפיעים על ההכנסה כאשר נשווה בין סרט בעל חשיפה רבה לסרט בעל חשיפה רבה מאוד, אך הם כן נעשים משמעותיים אם נשווה בין סרטים להם כמות מדרגים יחסית קטנה.

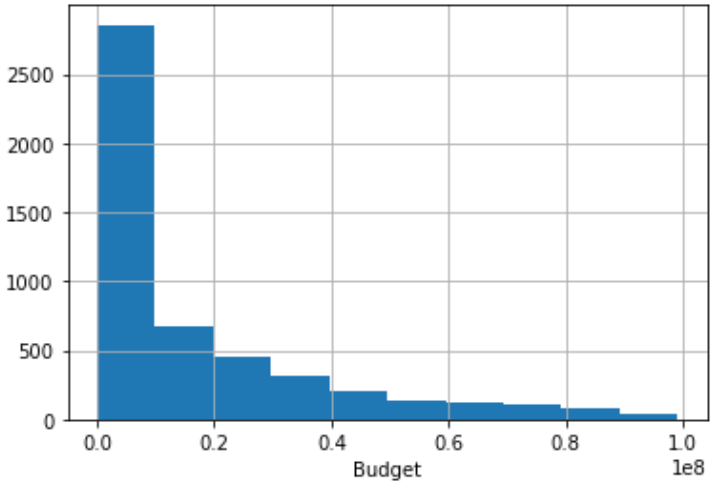
- <u>השדה budget</u>: שמנו לב כי בשדה זה, 1487 מן הערכים (כ-28.5%) הם 0. כדי לבדוק האם זה הגיוני, בדקנו את ערכי revenue עבור רשומות אלו (איור 4).



איור 4 – תצוגת ערכי ה-revenue של הסרטים להם ערך השדה budget הוא 0.

על פי איור 4, נראה כי לרוב הסרטים האלו יש הכנסות של מיליוני ועשרות מיליוני דולרים על פי סט הנתונים הנ"ל. אם כן, נראה כי רוב מוחלט של המקרים בהם הוזן ערך budget של 0 הן טעות. נתייחס לערכים אלו כאל ערכים חסרים.

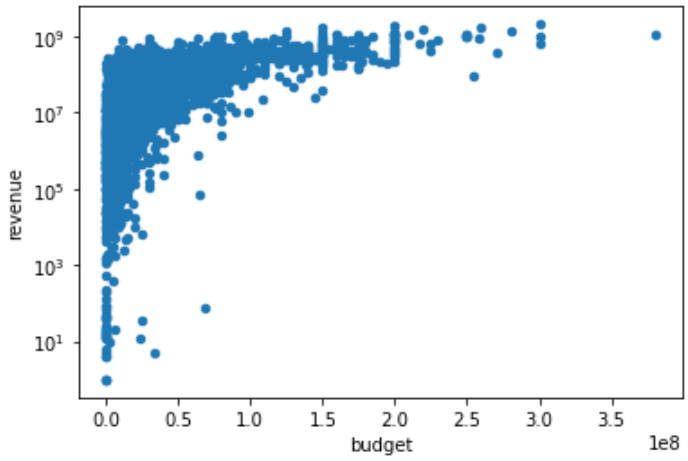
- באיור 5 מוצגת ההתפלגות של השדה budget עד לערך של מיליארד (ההגבלה נעשתה גם כאן לשם בהירות האיור).



איור 5 – התפלגות השדה budget בקטע [0,1,000,000,000].

ההיסטוגרמה שמוצגת באיור 5 דומה מאוד במגמתה לזו של השדה vote_count (איור 2). גם כאן, ניתן לראות מגמת ירידה בכמות הרשומות ככל שערך התקציב עולה. ירידה זו מתחילה באופן חד ומתמתנת ככל שערך התקציב גבוה יותר.

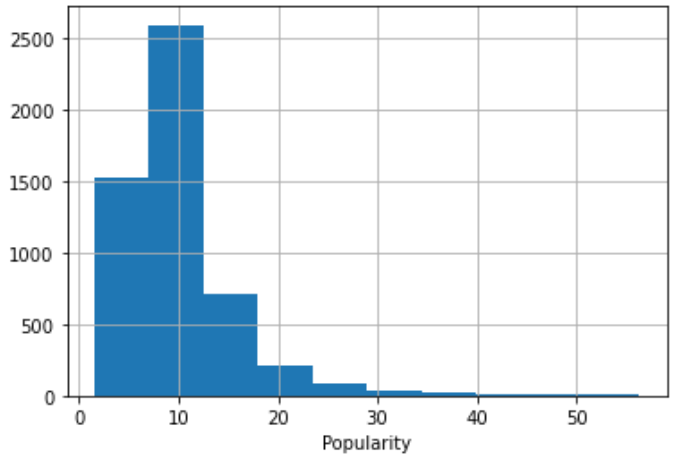
באיור 6 מוצג scatter plot של ערכי ה-revenue וה-budget של כל הרשומות בסט הנתונים:



איור 6 - scatter_plot של ערכי budget ו-revenue (המוצג בסקאלה לוגריתמית).

גם במקרה זה ניתן לראות מגמה דומה מאוד בין גרף זה למקבילו עבור השדה vote_count (איור 3). הפער בין הכנסת סרט אחד לסרט אחר עם תקציב הגדול ממנו בערך קבוע יהיה משמעותי יותר אם לסרט הזול מביניהם יש תקציב קטן מאשר המקרה בו יש לו תקציב גדול.

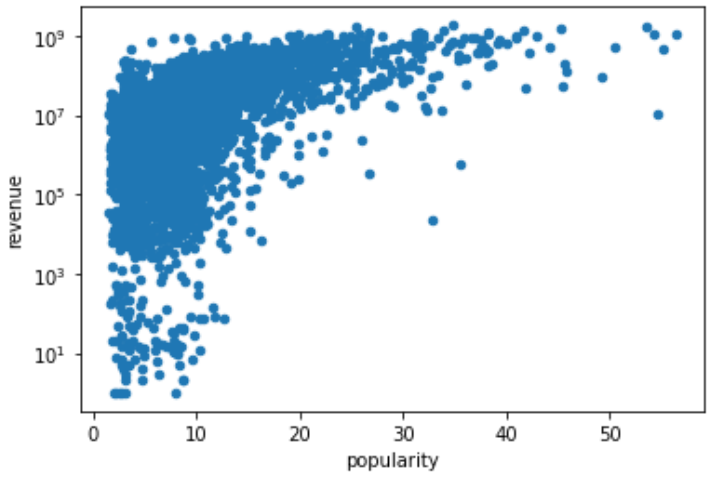
- <u>השדה popularity</u>: איור 7 מציג את התפלגות השדה popularity עד לערך 60 (ההגבלה נעשתה לשם בהירות האיור).



איור 7 - התפלגות השדה popularity בקטע [0,60].

כמו באיורים 2 ו-5, גם במקרה זה המגמה הכללית היא של ירידה בכמות הרשומות ככל שערכו של ה-popularity עולה. עם זאת, בערכים הנמוכים ישנה דווקא מגמה של עלייה. נראה אם כן כי סף ה-popularity השכיח ביותר אינו דווקא הסוף הנמוך ביותר, אלא מעט יותר גבוה.

באיור 8 ניתן לראות scatter plot של ערכי ה-revenue וה-popularity של כל הרשומות בסט הנתונים:

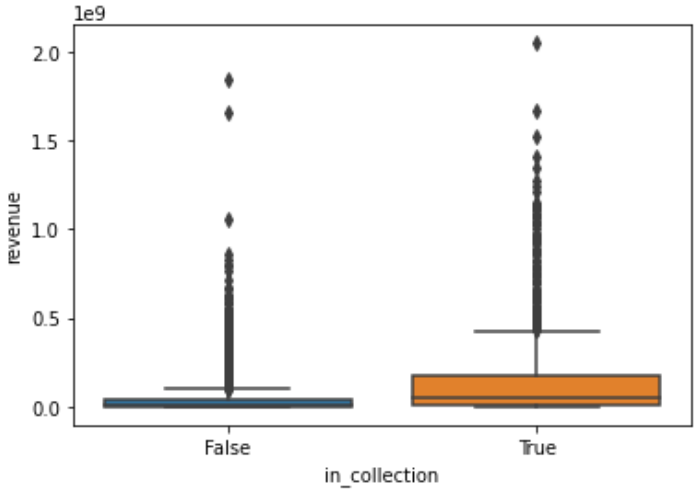


איור 8 - scatter_plot של ערכי popularity ו-revenue (המוצג בסקאלה לוגריתמית).

כמו באיורים 3 ו-6, גם כאן ניתן לראות מגמה ברורה של עלייה בהכנסות הסרט ככל שמד הפופולריות שלו גבוה יותר. כמו באיורים 3 ו-6, גם כאן הגרף מתחיל בעלייה חדה יותר לעומת המגמה שמתרחשת בערכים הגבוהים יותר, אך כאן גם העלייה המתרחשת בערכים הנמוכים יותר של popularity היא מתונה יותר לעומת העלייה המתרחשת באזורים המקבילים עבור השדות vote_count ו-budget.

לאחר שניתחנו את שלושת השדות הנ"ל, עברנו לנתח פיצ'רים אחרים, בפרט כאלו שאינם רציפים. נציין כעת כמה מן הממצאים היותר מעניינים שמצאנו:

- <u>השדה belongs_to_collection</u>: רצינו להשוות בין התפלגות הכנסת הסרטים ששייכים לסדרת סרטים כלשהי לבין אלו שלא, על מנת לראות האם יש טעם להוסיף למודל הרגרסיה שנבנה פיצ'ר בוליאני שמורה על כך. באיור 9 ניתן לראות boxplots עבור הסרטים ששייכים ושאינם שייכים לסדרת סרטים:

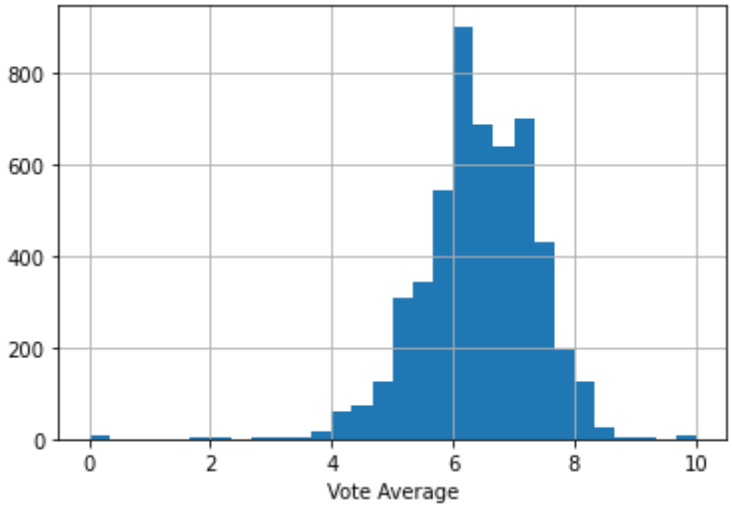


איור 9 – boxplots של revenue עבור סרטים ששייכים לסדרת סרטים (True) ושאינם שייכים (False).

כפי שניתן לראות באיור 9, נראה כי שתי ההתפלגויות נבדלות גם בתוחלת וגם בשונות. נציין בפרט כי ההכנסה החציונית לסרטים שאינם שייכים לסדרת סרטים היא כ-11.3 מיליון דולר, ואילו ההכנסה הממוצעת לסרטים שכן שייכים לסדרה היא כ-55 מיליון דולר. כלומר, ההבדל

ניכר גם במבט על ההכנסה החציונית (מדד הרגיש פחות לערכים חריגים). נסיק מכך כי נרצה לשלב באלגוריתם שנבנה שדה שיסמן האם הרשומה שייכת לסדרת סרטים או שלא.

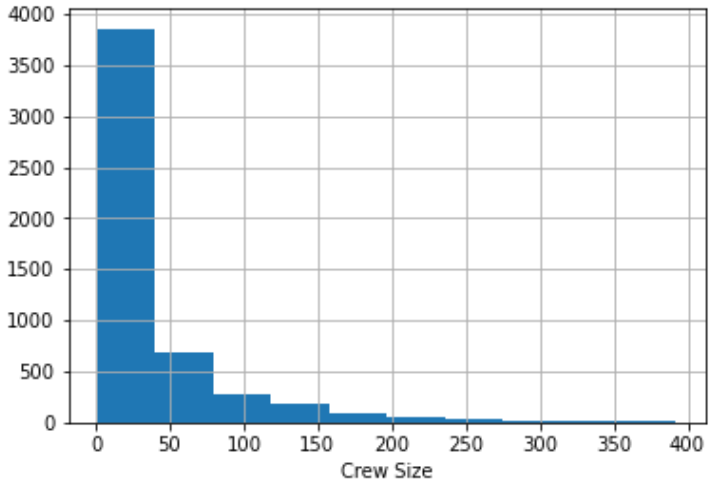
- <u>התפלגות השדה vote_average</u>: התעניינו לבדוק מהי ההתפלגות של ממוצע ציוני הגולשים לסרטים השונים (איור 10).



איור 10 – התפלגות השדה vote_average

מן האיור ניתן להסיק שמרבית הסרטים מקבלים ציון ממוצע של בין 5 ל-8, כלומר רובם אינם מדורגים כמאוד טובים או איומים ונוראיים, אלא בסקאלה יותר מצומצמת ומאופקת. חלק הארי של הסרטים נמצא בדירוג שהוא לפחות 6, כך שנראה שרוב הסרטים מדורגים בחצי העליון של סקאלת הציונים.

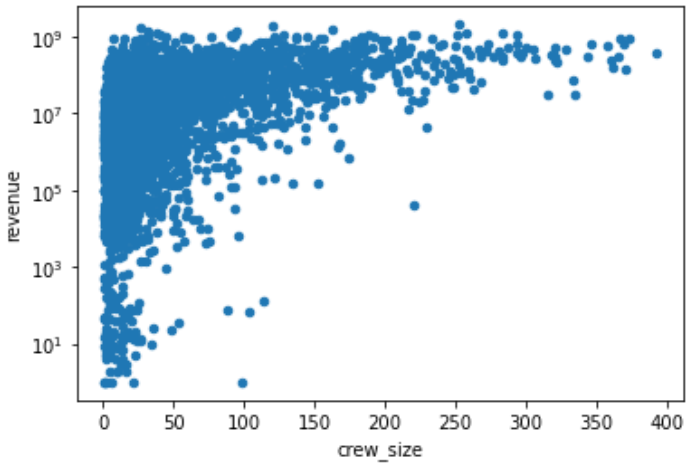
- <u>גודל צוות ההפקה</u>: בשדה crew מופיעים רשימת בעלי תפקידים (מלבד השחקנים) שהשתתפו בהפקת הסרט. במקום לנסות ולהתמקד בתפקידים או באנשים ספציפיים, התעניינו לבדוק האם יש השפעה לגודל הצוות על הכנסת הסרט. ייתכן למשל שסרט בעל צוות הפקה גדול יותר זהו סרט יותר מושקע, ולכן גם הכנסתו תהיה גבוהה יותר מסרט שנוצר על ידי צוות קטן של אנשים.
באיור 11 ניתן לראות את התפלגות גודל ה-crew עד לערך 400 (ההגבלה נעשתה לשם בהירות האיור).



איור 11 – התפלגות גודל צוות ההפקה בקטע [0,400].

ניתן לראות כי במרבית הסרטים צוות ההפקה שצוין בסט הנתונים אינו עולה על 50 אנשים, אך ישנם גם סרטים עם כמות של כמה מאות אנשים. המגמה היא של ירידה בכמות הרשומות ככל שגודל הצוות עולה.

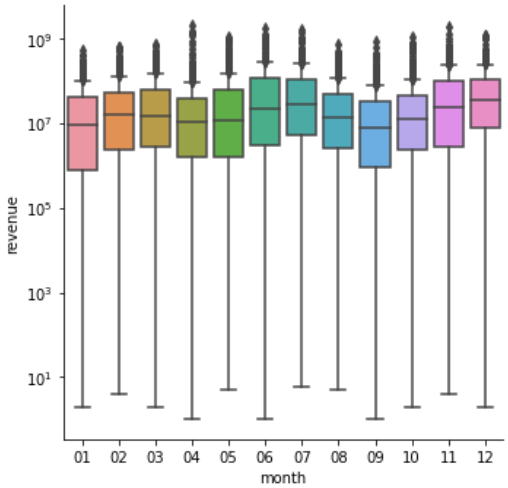
בדומה לניתוחים בשדות קודמים, נציג גם scatter plot של ערכי ה-revenue וגודל הצוות של הרשומות בסט הנתונים (איור 12):



איור 12 - scatter_plot של ערכי גודל הצוות (crew_size) ו-revenue (המוצג בסקאלה לוגריתמית) עבור רשומות בהן גודל ההפקה הוא בקטע [0,400].

באיור 12 ניתן לראות מגמה מאוד דומה לזו שראינו באיורים 3,6,8. אם נשווה בין סרטים בעלי צוות הפקה גדולים, סביר כי הפער בהכנסה לא יהיה משמעותי גם אם לאחד הסרטים צוות הפקה גדול יותר, כך הפער עשוי להיות משמעותי בהרבה אם הוא יימדד בין שני סרטים בעלי צוות הפקה קטן יחסית.
הקורלציה שנמדדה בין גודל צוות ההפקה להכנסת הסרט היא של כ-0.49, כלומר נראה כי ישנו מתאם לא זניח בין שני השדות.

- פילוח הכנסה לפי חודשי השנה – כחלק מניתוח שדה תאריך יציאת הסרט, בדקנו את התפלגות הכנסות הסרט לפי החודש בו הוא ראה אור. באיור 13 ניתן לראות boxplots מתאימים עבור כל החודשים.



איור 13 – boxplots של התפלגות ההכנסות לפי החודש בשנה בו יצאו לאור. ציר ה-y הוא בסקאלה לוגריתמית.

ישנם boxplots שנראו לנו חריגים יותר ביחס לשאר (בממוצע ההכנסות שלהם למשל), ואלו ה-boxplots שמייצגים את חודשים יוני ויולי, ואלו שמייצגים את חודשים נובמבר ודצמבר. אנו חושבים שהסיבה לכך היא שסרטים מרכזיים נוהגים לצאת בחודשי הקיץ או בסמיכות לחג המולד, נוסף על כך שעשוי להיות גידול במספר צופי הקולנוע בתקופות אלו עקב החופשות השונות.
בעקבות האבחנה הזו, החלטנו להוסיף שדה קטגורי (שלושה ערכים אפשריים) של התקופה בשנה בה יצא הסרט: קיץ (יוני-יולי), חורף (נובמבר-דצמבר) או אחר.

c. Missing data

נציין כעת שדות שונים בדאטה אשר להם ערכים חסרים, או שאנו חושדים שהערכים שהוזנו אינם נכונים.

- Backdrop_path – כ-11% מהערכים ריקים. עם זאת, ייתכן שערכים אלו ריקים בכוונה (כלומר לא קיים path) ואין כאן באמת מחסור בנתונים.
- Belongs_to_collection – כ-80% מהערכים ריקים. גם כאן, ייתכן שהסיבה לכך היא פשוט שרוב הסרטים הללו לא שייכים לסדרת סרטים.
- Budget – אמנם אין ערכים ריקים, אך כפי שהסקנו בניתוח הנתונים, לכ-28.5% מן הרשומות יש ערך של 0 בשדה זה, ולכן סביר להניח שברוב המקרים זהו נתון שגוי והערך האמיתי חסר.
- Homepage – כ-66% מהערכים ריקים. ייתכן שבחלק מן המצבים פשוט לסרט אין אתר בית.
- Imdb_id - כ-0.2% מן הערכים ריקים.
- Overview – כ-0.1% מן הערכים ריקים.
- Poster_path – כ-3% מן הערכים ריקים. ייתכן שהסיבה לכך היא שאין לסרטים אלו פוסטר, ואז זהו לא חוסר במידע.
- Runtime – שדה זה ריק עבור ארבע רשומות (כ-0.007%).
- Tagline – שדה זה ריק עבור כ-20% מן הרשומות.

2. Feature Engineering

נציין כעת את הפיצ'רים השונים בהם השתמשנו עבור המודלים שלנו, ובנוסף לכל פיצ'ר:

a) נסביר מדוע חשבנו להשתמש בו.
b) במקרה שהוא אינו שדה מקורי של סט הנתונים ,נציין את משמעותו וכיצד ייצרנו אותו.
c) נציין איך טיפלנו בערכים חסרים, אם ישנם.

- **Budget** – ראינו בשלב ניתוח הנתונים שלשדה זה יש קורלציה חיובית די גבוהה עם ה-revenue, ולכן בחרנו להכניסו למודל.
  - כדי לטפל בערכים החסרים שאיתרנו (הרשומות בהן מופיע 0 בשדה זה), החלטנו לייצר <u>פיצ'ר נוסף בשם **budget_0**</u>, ולהכניסו גם למודל. זהו שדה בוליאני שאומר האם היה 0 בערך ה-budget ברשומה זו או לא. כך נסייע למודלים למשקל באופן שונה את הרשומות עם הערכים החסרים במהלך החיזוי. עם זאת, לא שינינו ערכים בשדה המקורי.

- **Popularity** – הסיבה לשימוש בשדה זה היא גם כן קורלציה חיובית גבוהה עם שדה ה-revenue, כפי שגילינו בשלב ניתוח הנתונים.
- **Vote_count** – גם כאן, הסיבה לשימוש בשדה היא הממצאים אודותיו משלב ניתוח הנתונים, ובפרט הקורלציה הגבוהה שמצאנו בינו לבין revenue.
- **Runtime** – אמנם הקורלציה בין שדה זה ל-revenue הייתה נמוכה, אך חשבנו ששדה זה יוכל לסייע בתהליך החיזוי, בעיקר במקרה בו אורך הסרט הוא חריג.
  - השלמת ערכים חסרים נעשתה על ידי Mean Imputation, כלומר השמת הערך הממוצע של השדה על פני כל הרשומות ללא ערך חסר בשדה זה.
- **Vote_average** – אמנם גם בין שדה זה הייתה קורלציה די נמוכה עם revenue, אך הכנסנו אותו מאותה סיבה כמו runtime.
- **Crew_size** – שדה זה נוצר על ידי ספירת כמות המופעים במילון שמופיע בכל רשומה תחת השדה crew. הכנסנו שדה זה למודל מכיוון שחשבנו שהוא יכול לסייע בתהליך החיזוי בעקבות מה שראינו אודותיו בשלב ניתוח הנתונים.
- **Cast_size** – שדה זה נוצר על ידי ספירת כמות המופעים במילון שמופיע בכל רשומה תחת השדה cast. ביצענו על שדה זה ניתוח הדומה לזה שעשינו ל-crew_size (לא מוצג במסמך זה), ובעקבותיו החלטנו שכדאי להוסיף את השדה הנ"ל למודל.
- **Genre** – פירקנו שדה זה למספר שדות כמספר הז'אנרים שמופיעים בשדה זה. כל שדה חדש מהווה אינדיקטור עבור אחד הז'אנרים, כך שלכל סרט מסומן 1 בשדות הרלוונטיים לז'אנרים אליהם משתייך סרט זה, ו-0 עבור שאר הז'אנרים. בניתוח הנתונים שביצענו ראינו שההכנסה עבור כל ז'אנר מתפלגת באופן קצת שונה, ולכן חשבנו להשתמש בשדה זה, אך פירקנו אותו לאינדיקטורים מכיוון שזהו פיצ'ר קטגוריאלי. נציין שלכל סרט יכולים להיות מספר ז'אנרים ולכן יופיע להם הערך 1 ביותר מעמודה אחת.
- **Lang** – השתמשנו בשדה זה גם כן כי ראינו שישנה התפלגות קצת שונה של ה-revenue עבור הערכים התדירים ביותר בשדה. מכיוון שזהו שדה קטגוריאלי, פירקנו גם אותו לאוסף של אינדיקטורים, אך השמנו אינדיקטורים ספציפיים רק ל-15 השפות הנפוצות ביותר בסט האימון, ואילו את שאר השפות קיבצנו יחד תחת אינדיקטור בשם lang_other. גם בשדה זה לכל סרט יכולות להיות מספר שפות ולכן יופיע להם הערך 1 ביותר מעמודה אחת.
- **Prod_comp –** השתמשנו בשדה זה מסיבות זהות לשימוש בשדות genre ו-lang. מכיוון שזהו גם כן שדה קטגוריאלי, פירקנו גם אותו לאינדיקטורים כך שיהיו אינדיקטורים עבור 20 חברות ההפקה הנפוצות ביותר בסט האימון, ואילו את שאר החברות קיבצנו יחד תחת אינדיקטור בשם comp_other. וגם בשדה זה לכל סרט יכולות להיות מספר חברות הפקה ולכן יופיע להם הערך 1 ביותר מעמודה אחת.
- **Prod_cntr** - השתמשנו בשדה זה מסיבות זהות לשימוש בשדות genre ו-lang. מכיוון שזהו גם כן שדה קטגוריאלי, פירקנו גם אותו לאינדיקטורים כך שיהיו אינדיקטורים עבור 20 המדינות הנפוצות ביותר בסט האימון, ואילו את שאר המדינות קיבצנו יחד תחת אינדיקטור בשם cntr_other. כמו בשדות הקודמים, גם בשדה זה לכל סרט יכולות להיות מספר מדינות מקור לסרט ולכן יופיע להם הערך 1 ביותר מעמודה אחת.
- **Is_collection** – בעקבות השוני שראינו בהתפלגות ה-revenue בין סרטים ששייכים לסדרת סרטים לבין אלו שלא רשום ששייכים לסדרת סרטים (איור 9), החלטנו על סמך השדה belongs_to_collection לייצר שדה בשם is_collection אותו נכניס למודל. זהו שדה בוליאני שמקבל True אם הסרט שייך לסדרת סרטים, ו-False אחרת. אנו מניחים שהסרטים שלהם ערך ריק בשדה belongs_to_collection אינם שייכים לסדרת סרטים.

- **month_cat** – שדה זה נוצר על סמך פיצ'ר תאריך יציאת הסרט (release_date). זהו שדה קטגוריאלי עם שלושה ערכים אפשריים: summer אם הסרט יצא בחודשים יוני/יולי, winter אם יצא בנובמבר/דצמבר, ו-other אחרת. לאחר מכן פירקנו שדה זה לאינדיקטורים כפי שעשינו עבור שדות קטגוריאליים קודמים.
בחרנו להשתמש בשדה זה בעקבות הניתוח שעשינו לפי חודשי השנה בשלב ניתוח הנתונים (איור 13 והפסקה שדנה בה).
- **Year_Cat** – בניתוח שעשינו על התפלגות ה-revenue לפי שנת יציאת הסרט (לא מוצג בשלב ניתוח הנתונים), היה נראה לנו כי ניתן לחלק את השנים לחמש קבוצות, כאשר בכל קבוצה התפלגות ההכנסה נראית שונה. לכן יצרנו שדה קטגוריאלי עם חמישה ערכים (לפי שנת יציאת הסרט): עד שנת 1960, 1960-1980, 1980-1995, 1995-2010, מ-2010 ואילך. לאחר מכן, פירקנו את השדה לאינדיקטורים כפי שעשינו עם שדות קטגוריאליים קודמים.

3. Prediction

בשלב זה בחרנו להתנסות עם שלושה מודלים הפועלים באופן די שונה אחד מהשני, ולראות עם מי מהם נצליח להגיע לתוצאות הטובות ביותר עבורנו.

נציין את שלושת המודלים המרכזיים בהם ניסינו להשתמש. עבור כל מודל, נציין את ערכי ההיפרפרמטרים המרכזיים, איזו רגולריזציה ביצענו (אם בכלל). בנוסף, נדבר על תהליך האימון והואלידציה שביצענו בשלושת המודלים האלו.

ההיפרפרמטרים בכל שלושת המודלים נבחרו על ידי הרצת שלוש שיטות tuning שונות, ובחירת ההיפרפרמטרים שנתנה השיטה עבורה שגיאת הואלידציה הייתה הקטנה ביותר. השיטות שניסינו הן:

- Random Search
- Bayesian Search
- Tree-Structured Parzen Estimator Search

Linear Regression (Elastic Net) – ניסינו מודל רגרסיה לינארית, אשר לו הוספנו רגולריזציות מסוג Lasso (נורמת L1) ו-Ridge (נורמת L2), כך שהמודל שהשתמשנו בו הוא בעצם Elastic Net. שיטת ה-tuning בה בחרנו בסוף להשתמש עבור מודל זה הייתה Random Search. עבור שיטה זו התקבלה שגיאת הואלידציה הנמוכה ביותר.

אימון המודל נעשה תוך כדי שימוש בכל הפיצ'רים שצוינו בסעיף Feature Engineering. הואלידציה נעשתה על ידי תהליך של 10 fold cross-validation.

<u>ההיפרפרמטרים וערכם</u>: $alpha = 2.44, l1\_ratio = 0.512$.

<u>שגיאת ואלידציה</u>: 2.608.

<u>שגיאת מבחן</u>: 2.550.

K Neighbors Regressor – מודל זה חוזה את ערך ה-revenue לרשומה חדשה על סמך ממוצע של K השכנים הקרובים ביותר מסט האימון. הממוצע יכול להיות רגיל, או משוקלל כאשר לשכנים קרובים יותר יש משקל גדול יותר.

התוצאות שקיבלנו משיטות ה-tuning שניסינו היו די דומות, ובחרנו ללכת עם ההיפרמטרים שקיבלנו משיטת Random Search, מכיוון שבה ההיפרפרמטר של מספר השכנים קיבל את ערך גבוה יותר מאשר בשיטות האחרות (5 שכנים), וחשבנו כי ערך זה יסייע לרגולריזציה של המודל.

אימון המודל נעשה על כל הפיצ'רים שצוינו בסעיף Feature Engineering. הואלידציה נעשתה על ידי תהליך של 10 fold cross-validation.

<u>ההיפרפרמטרים וערכם</u>: $leaf\_size = 30, metric = "manhattan", n\_neighbors = 5$.

<u>שגיאת ואלידציה</u>: 2.289.

<u>שגיאת מבחן</u>: 2.238.

Extra Trees Regressor (אותו בחרנו להגיש) – זהו מודל הדומה מאוד ל-Random Forest אך שונה ממנו בשני גורמים עיקריים:

1. תת הקבוצה מסט האימון בה משתמשים לגידול כל עץ מוגרלת <u>בלי החזרה</u>.
2. ביצוע ה-splits בכל עץ נעשית בצורה רנדומלית, ולא על סך סמך תת הקבוצה (היא משמשת רק לקביעת הערכים בעלים).

בגלל הנקודות האלו, מודל זה פחות יקר מבחינה חישובית לעומת Random Forest, וגם נוהג להתמודד טוב יותר עם פיצ'רים רועשים.

על סמך התוצאות שקיבלנו משיטות ה-tuning השונות, החלטנו כי הבחירה הטובה ביותר התקבלה מ-Bayesian Search אז בחרנו להשתמש בהיפרפרמטרים שקיבלנו משיטה זו.

אימון המודל נעשה על כל הפיצ'רים שצוינו בסעיף Feature Engineering. הואלידציה נעשתה על ידי תהליך של 10 fold cross-validation.

ההיפרפרמטרים וערכם: $bootstrap = False, criterion = \text{mae}, max_{depth} = 10,$
$max\_features = 0.756, min\_samples\_split = 10, n\_estimators = 230$

היפרפרמטרים כמו max_{depth} (העומק המקסימלי של כל עץ), ו-
$min\_samples\_split$ (מספר הדגימות המינימלי הדרוש כדי לפצל צומת פנימי) מסייעים לרגולריזציה של המודל בכך שהם מונעים ממנו להיות עמוק ומסובך מדי, ובכך הוא נהיה יותר מוכלל ופחות מותאם ספציפית לנתונים מסט האימון.

<u>שגיאת ואלידציה</u>: 2.180.

<u>שגיאת מבחן</u>: 2.136.

בחרנו להגיש את מודל ה-Extra Trees Regressor, מכיוון שעבורו קיבלנו את שגיאת הואלידציה (והמבחן) הנמוכה ביותר.