**עבודת סיכום מס' 2 - EDA וניקוי הנתונים**

***נושאים:***

1. *פרוטוקול שליפה*
2. *בדיקת הנתונים (EDA)*
3. *ניקוי הנתונים*

**הערה**: ניתן לעבוד ב-Python או ב-R. מומלץ לעבוד ב-jupyter notebook. מי שבוחר לעבוד ב-RStudio, בנוסף לקוד יצטרך לכתוב את התובנות ולהעתיק את הגרפים במסמך word.

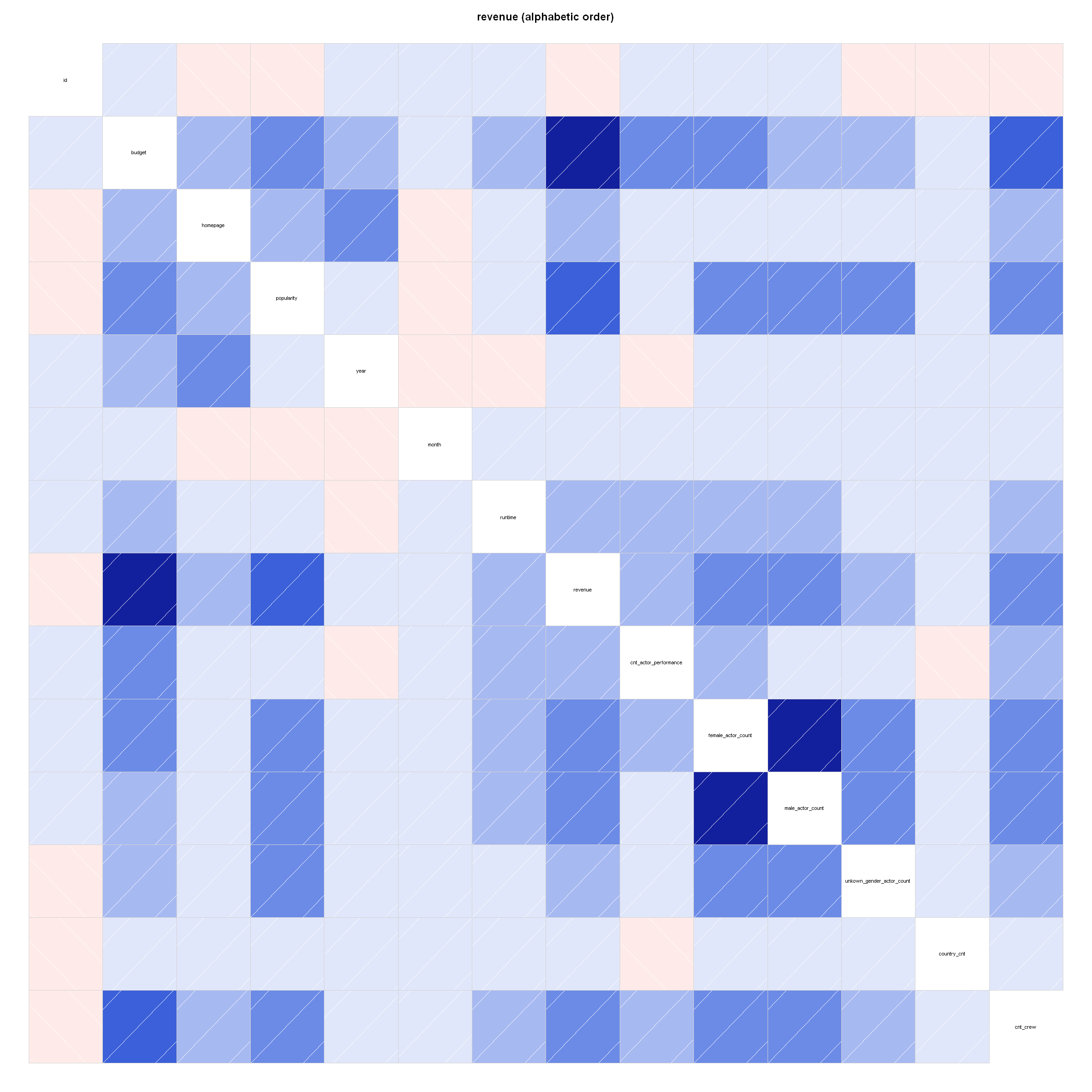
**חלק 1 - פרוטוקול שליפה**

1. להוריד את פרוטוקול השליפה [מהלינק](https://drive.google.com/open?id=15ktj2cdTSK2UT5Enn8iQwzRmtqu9Vgc5eGlY62jX3t4) ולהשלים את כל העמודות. קוד ה-SQL כדי לייצר את ה-VIEW עם הקובץ השטוח נמצא [בלינק](https://drive.google.com/open?id=1hbBQ30GEaoPokOkBCBkMYBrwQVeFABjZ).

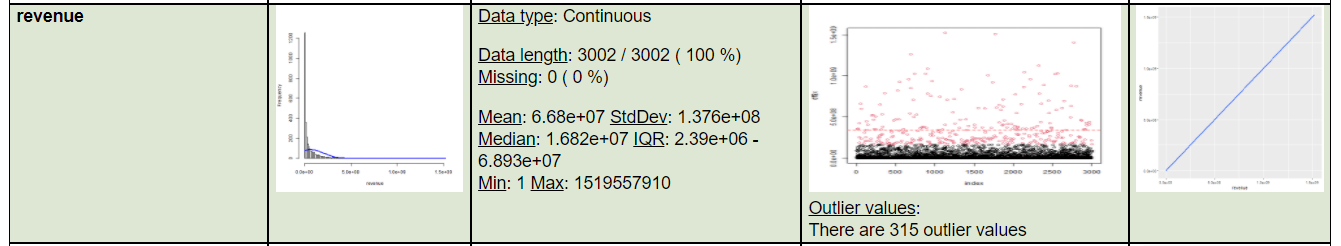
נעשה – הקובץ יושב בתיקייה

**חלק 2 - בדיקת הנתונים - Exploratory data analysis**

1. תתארו את הנתונים עם סטטיסטיקה תיאורית (תשתמשו במדדי מירכוז ופיזור)
   * נעשה בעזרת "מחקר" יש תיאור של כל הנתונים.
2. תייצרו גרפים המתארים את ההתנהגות של כל משתנה.
   * נעשה בעזרת "מחקר" יש גרפים של כל משתנה.
3. תייצרו מטריצה קורלציות והציגו אותה בגרף
   * נעשה בעזרת " corrgram " יש גרפים של כל משתנה
   * בנוסף יש שימוש ב " correlate" בשיטת ספיירמן



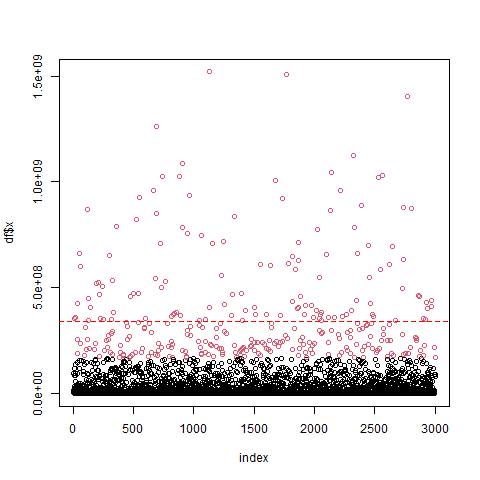
1. תתארו את משתנה המטרה (revenue)
   * איך הוא מתפלג?



* + האם יש קטגוריות שמראים שוני גדול התפלגות של משתנה המטרה?
  + כן – popularity, year, month, runtime, actor\_performance

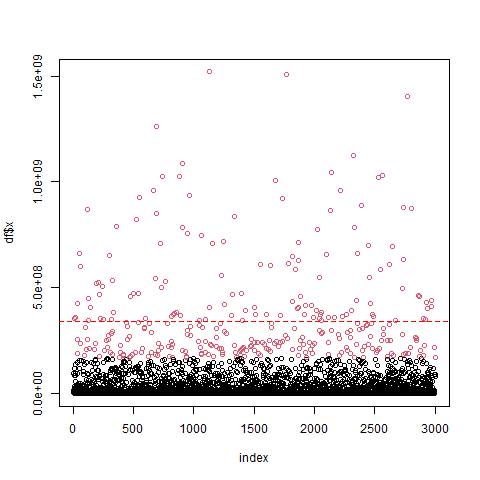
1. תייצרו גרפים שיכולים לעזור לכם לבדוק האם קיימים נתוני קיצון. תתארו אותם.

Revenue



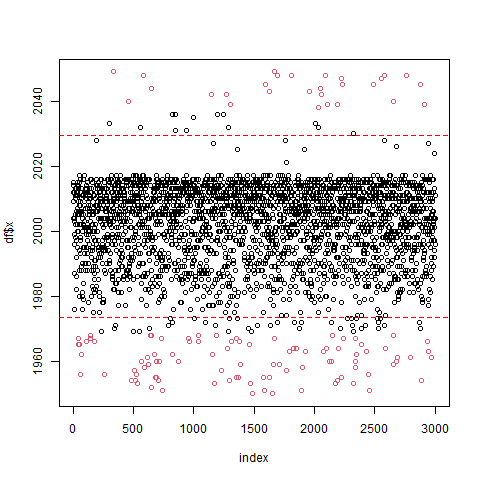
* + רואים בגרף את ערכי הקיצון של revenue הם כולל הכנסות מאוד גבוהות.

Popularity



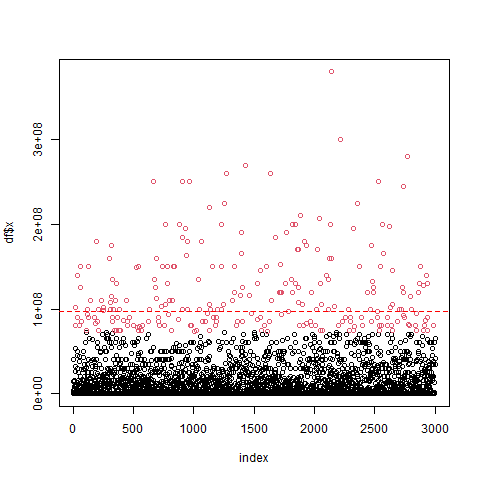
* + בגרף של פופולריות יש ערך אחד מאוד קיצוני שנמצא מעל 500 ועוד כמה ערכים גבוהים אבל יכול להיות שהם תקינים

Year



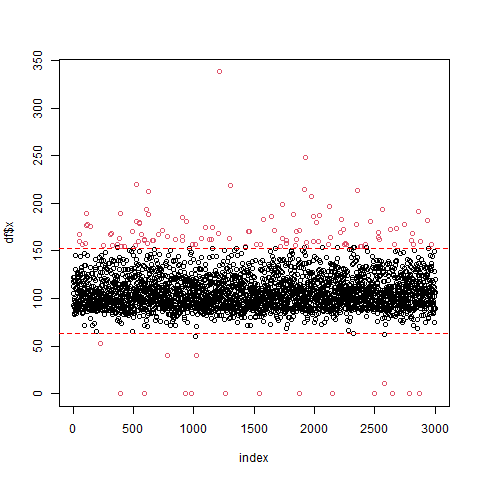
* + בגרף של שנים, משום מה יש שנים מאוחרות שעדיין לא היו

Budget



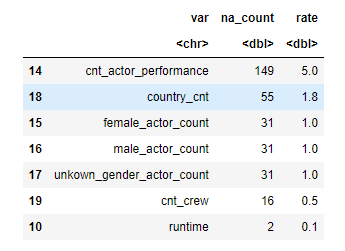
* + בגרף של התקציב, כמו בגרף של הפופולריות, ישנו אחד אחד מאוד קיצוני.

Runtime

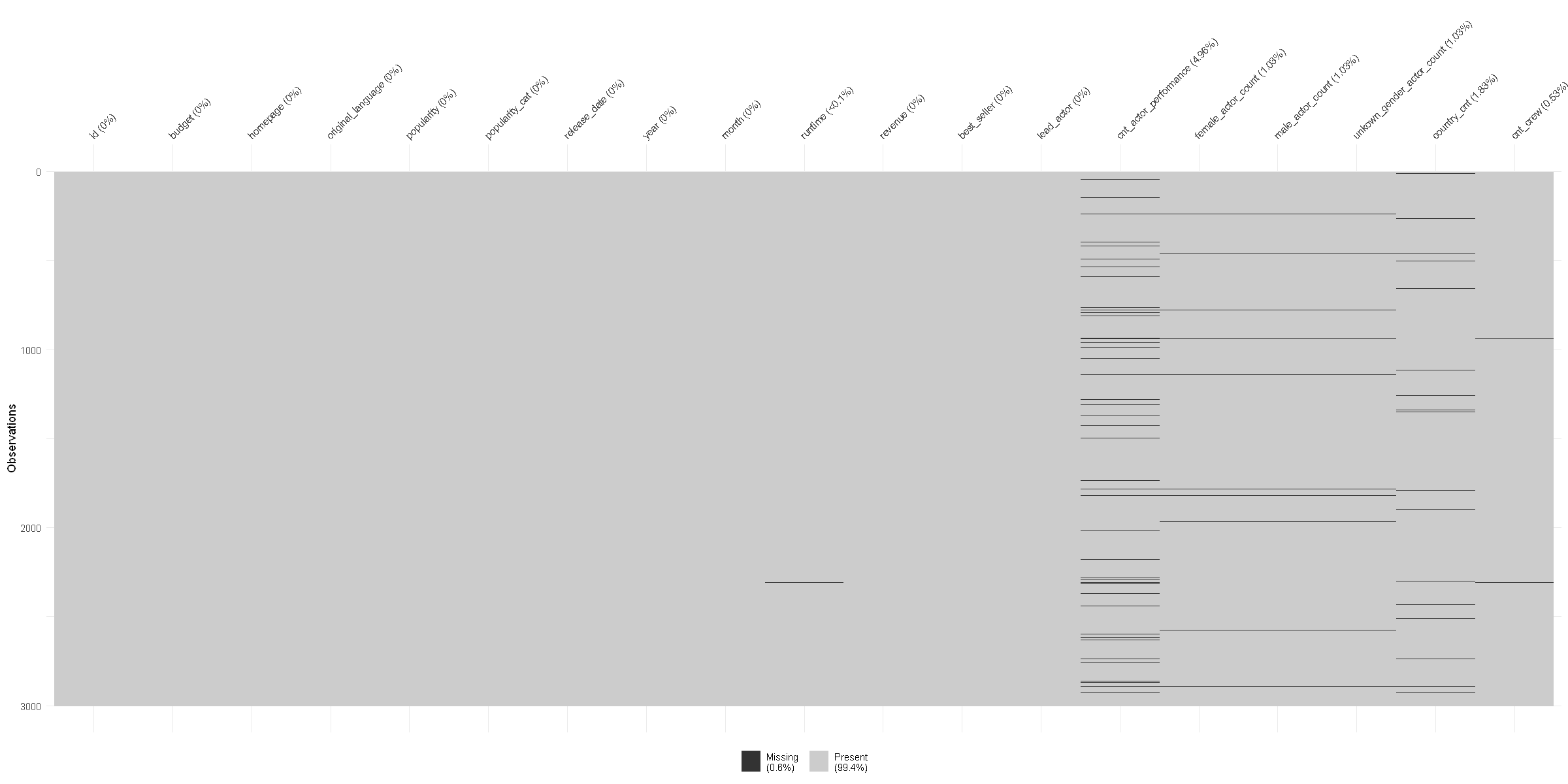


* + בגרף של זמן ההרצה ישנם ערכי קיצור לשני הכיוונים

1. תתארו את הנתונים החסרים: אצל איזה משתנים יש נתונים חסרים? כמה?



1. תייצרו מטריצה של חסרים (תייצרו dataframe עם אותם מימדים מטבלת המקור ותאים שיש חסרים תשימו ערך של אחד ובאלה שיש נתונים ערך אפס). תציגו את המטריצה בגרף heatmap.



**חלק 3 - ניקוי הנתונים :**

***3.1 נתוני קיצון***

1. במשתנים שבהם מוצאים ערכי קיצון תבדקו את ההתפלגות של המשתנה עם ובלי ערכי הקיצון. האם ההתפלגות משתנה ?
   * + אם אני עושה לפי 0.25 עד 0.75 יש שוני בכל ההתפלגויות – אבל לדעתי זה לא מידע רלוונטי – כי הוא עושה את זה אוטומטית על כל המידע. לפי הגרפים שעשינו ב"מחקר" לא כל ה"ערכי קיצון" הם באמת כאלו – לדוגמה חודש.
     + הוספתי דרך אחרת לשנה – ששם הורדתי את השנים שמעל 2020 וזה שינה את התוצאה
2. תייצרו גרף scatter עם המשתנים שבשאלה הקודמת ב-X ומשתנה המטרה ב-Y. האם ערכי קיצון במשתנה ה-X משפיעה על ההתנהגות של משתנה ה-Y? האם רואים שוני עם או בלי ערכי הקיצון?
   * עבדתי על 3 משתנים בנפרד כי לא הצלחתי להשתמש בדוגמה של קוקור – לקחתי את השנה, את הפופולריות ואת זמן ההרצה – בכולם רואים שכאשר מורידים את ערכי הקיצון – התנהגות ערך המטרה משתנה.
3. באיזה משתנים הייתם מוחקים את ערכי הקיצון? איך הייתם מוחקים אותם? נמקו.
   * בפופולריות – הייתי מוחק את ערכי הקיצון כי ישנן מספר נקודות לא הגיוניות
   * בשנים הייתי מוחק את הערכים מעל שנת 2020 כי הם לא הגיוניים
   * בזמן הרצה – הייתי מוריד את הגבוהים ביותר ואת הנמוכים ביותר- הגבוהים ביותר לא הגיוניים והנמוכים ביותר – סביר להניח שמדברים על קטגוריית סרטים שונה.
4. תפעלו על הנתונים לפי מה שהגדרתם בשאלה הקודמת.

***3.2 נתונים חסרים***

1. עבור כל משתנה עם נתונים חסרים, תראו את ההתפלגות **של משתנים אחרים** עם או בלי חסרים. השתמשו במטריצת החסרים שייצרתם בחלק 2, שאלה 7 עבור החיווי של יש/אין חסר. עבור ההתפלגות, תשתמשו בהיסטוגרמה או בגרף density עם קטגוריה/צבע לפי החיווי.
2. תייצרו טבלה של המשתנים שבהם יש חסרים ותתארו מהו מנגנון היצירה של החסרים (מבוסס על התוצאות של השאלה הקודמת).
3. איזה טכניקה imputation מתאימה לכל משתנה? השתמשו בטכניקה הנבחרת כדי להחליף את החסרים.
   * + - לאחר בדיקה – יצא שאין הבדלים בהתפלגויות של המשתנים בהורדה של המשתנים החסרים (הוספתי את הטבלה למטה) בדף העבודה יש את השיטות שבהן השתמשתי.



**חלק 4 - אחרי ניקוי הנתונים, תחזרו על החלק 2 במלואו.**

**על חלק זה חזרתי לאחר שהורדתי את ערכי הקיצון בגרפים – זה מופיע בסוף המחברת – הערה, לא הצלחתי להוריד את הערכים שרציתי לכן זה עשה את זה בצורה גסה.**