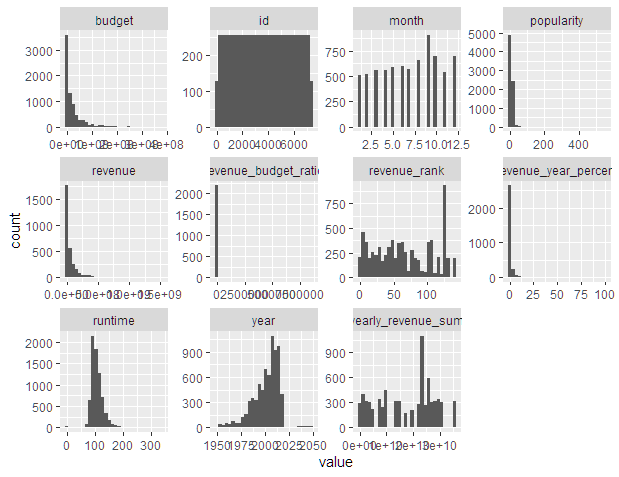
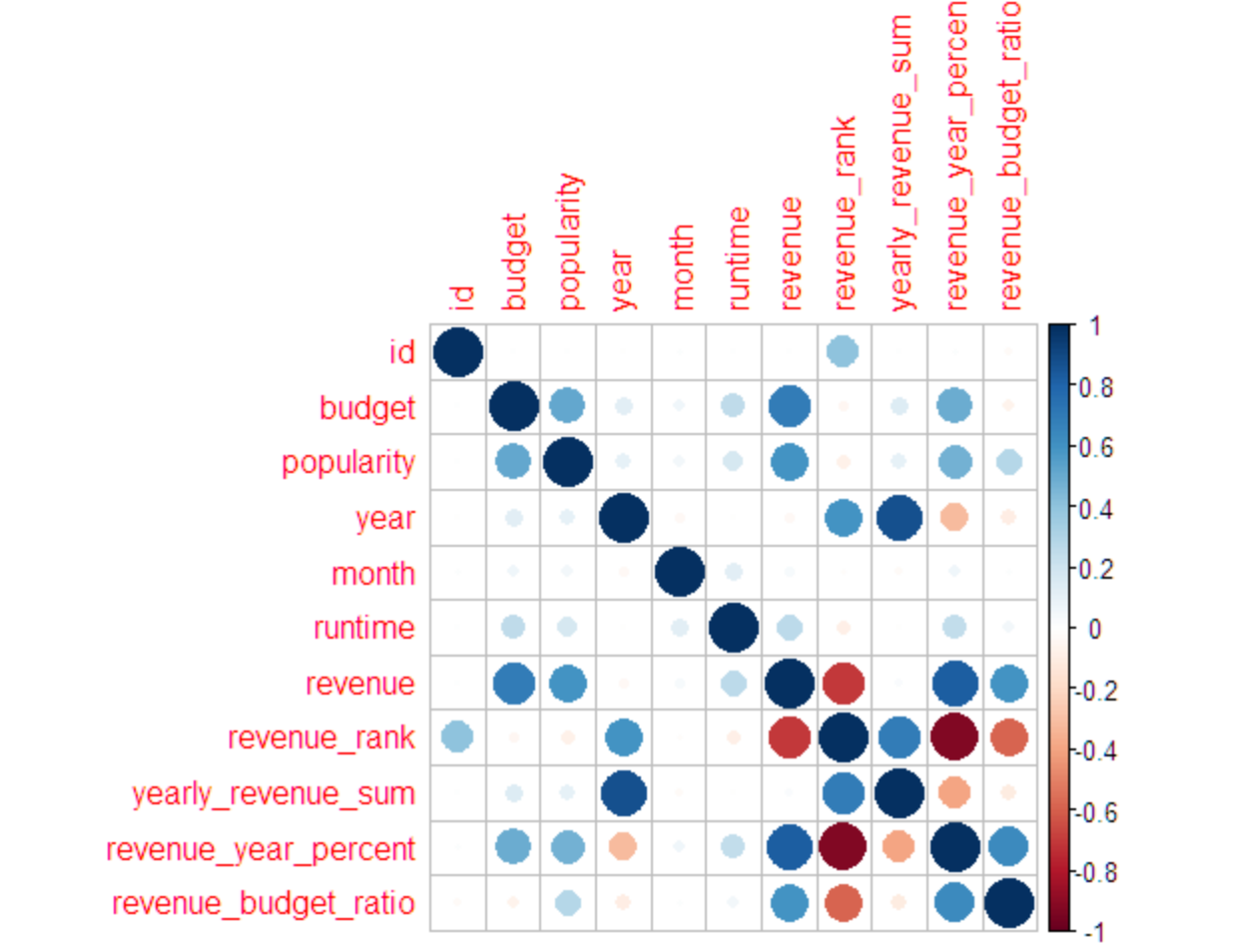
**חלק 2 - בדיקת הנתונים - Exploratory data analysis**

תתארו את הנתונים עם סטטיסטיקה תיאורית (תשתמשו במדדי מירכוז ופיזור)

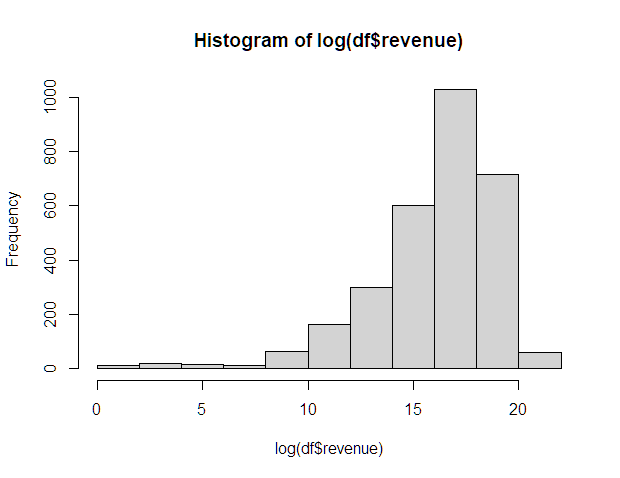
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | budget | popularity | year | month | runtime | revenue | revenue\_rank | yearly\_revenue\_sum | revenue\_year\_percent |
| Min. : 1 | Min. :0.00e+00 | Min. : 0.000 | Min. :1950 | Min. : 1.000 | Min. : 0.0 | Min. :1.000e+00 | Min. : 1.00 | Min. :2.382e+04 | Min. : 0.000 |
| 1st Qu.:1850 | 1st Qu.:0.00e+00 | 1st Qu.: 3.933 | 1st Qu.:1993 | 1st Qu.: 4.000 | 1st Qu.: 94.0 | 1st Qu.:2.380e+06 | 1st Qu.: 28.00 | 1st Qu.:7.852e+09 | 1st Qu.: 0.020 |
| Median :3700 | Median :7.50e+06 | Median : 7.436 | Median :2005 | Median : 7.000 | Median :104.0 | Median :1.681e+07 | Median : 63.00 | Median :2.160e+10 | Median : 0.150 |
| Mean :3700 | Mean :2.26e+07 | Mean : 8.515 | Mean :2001 | Mean : 6.843 | Mean :107.7 | Mean :6.673e+07 | Mean : 66.12 | Mean :1.749e+10 | Mean : 1.195 |
| 3rd Qu.:5549 | 3rd Qu.:2.80e+07 | 3rd Qu.: 10.920 | 3rd Qu.:2011 | 3rd Qu.:10.000 | 3rd Qu.:118.0 | 3rd Qu.:6.892e+07 | 3rd Qu.:106.00 | 3rd Qu.:2.524e+10 | 3rd Qu.: 0.670 |
| Max. :7398 | Max. :3.80e+08 | Max. :547.488 | Max. :2049 | Max. :12.000 | Max. :338.0 | Max. :1.520e+09 | Max. :141.00 | Max. :3.557e+10 | Max. :100.000 |
| NA | NA | NA | NA's :1 | NA's :1 | NA's :6 | NA's :4398 | NA | NA's :20 | NA's :4398 |

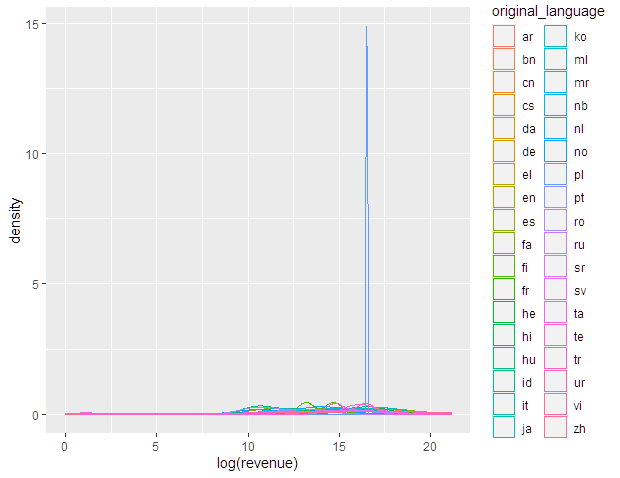
תייצרו גרפים המתארים את ההתנהגות של כל משתנה 

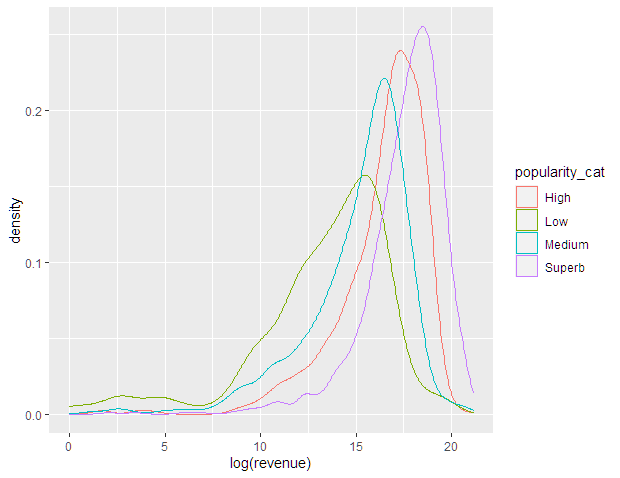
תייצרו מטריצה קורלציות והציגו אותה בגרף

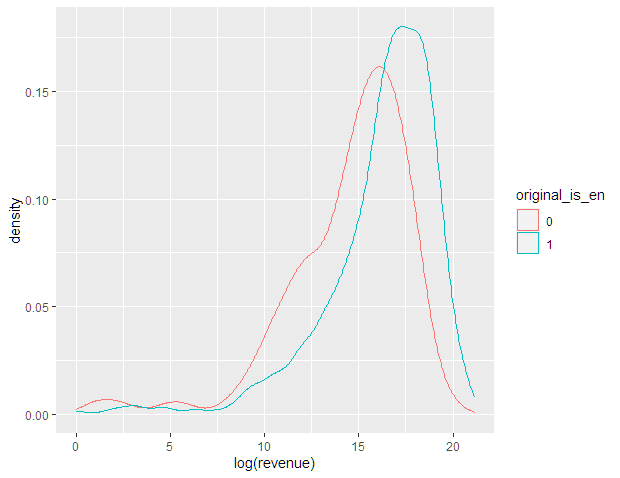


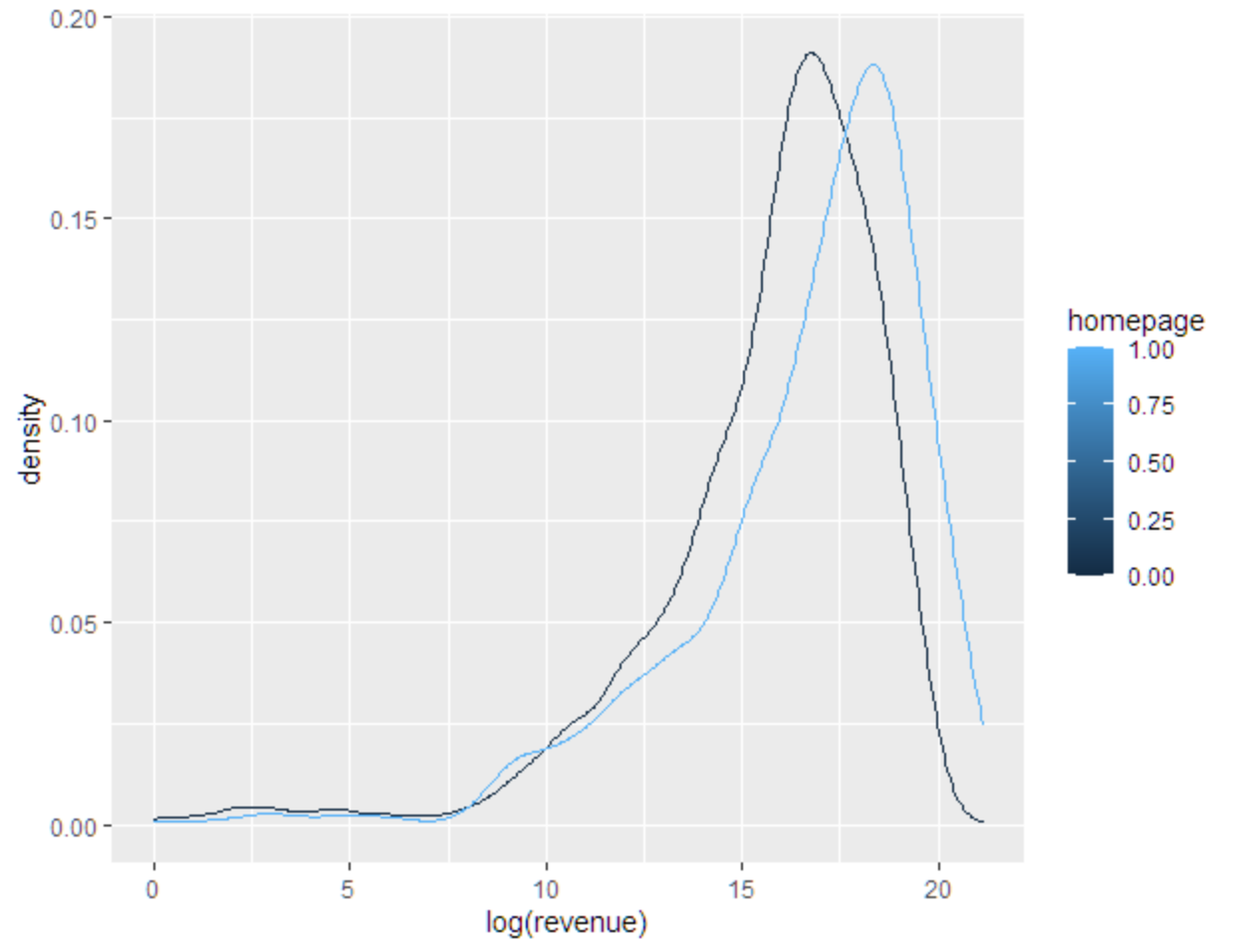
תתארו את משתנה המטרה (revenue) - איך הוא מתפלג? האם יש קטגוריות שמראים שוני גדול התפלגות של משתנה המטרה?

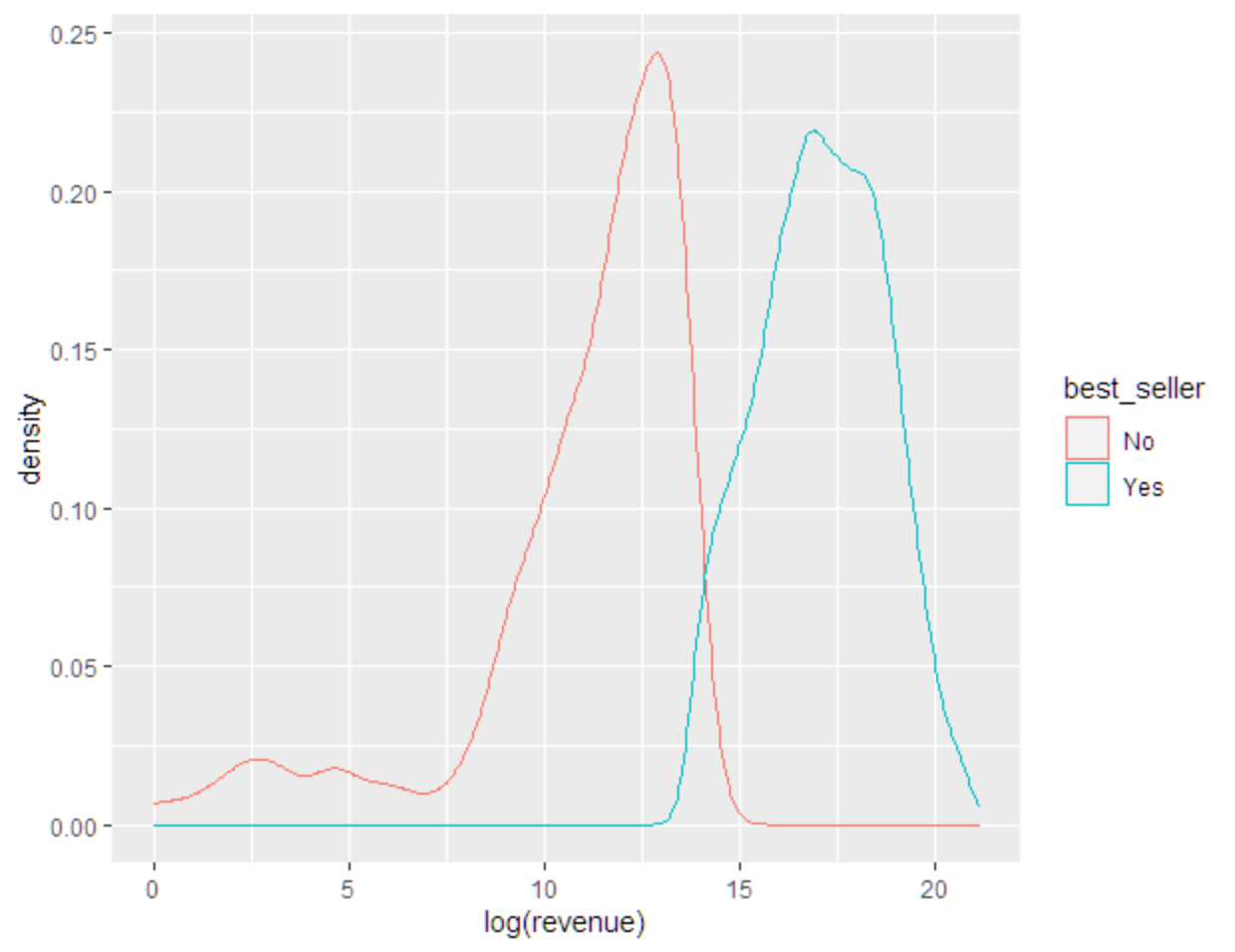




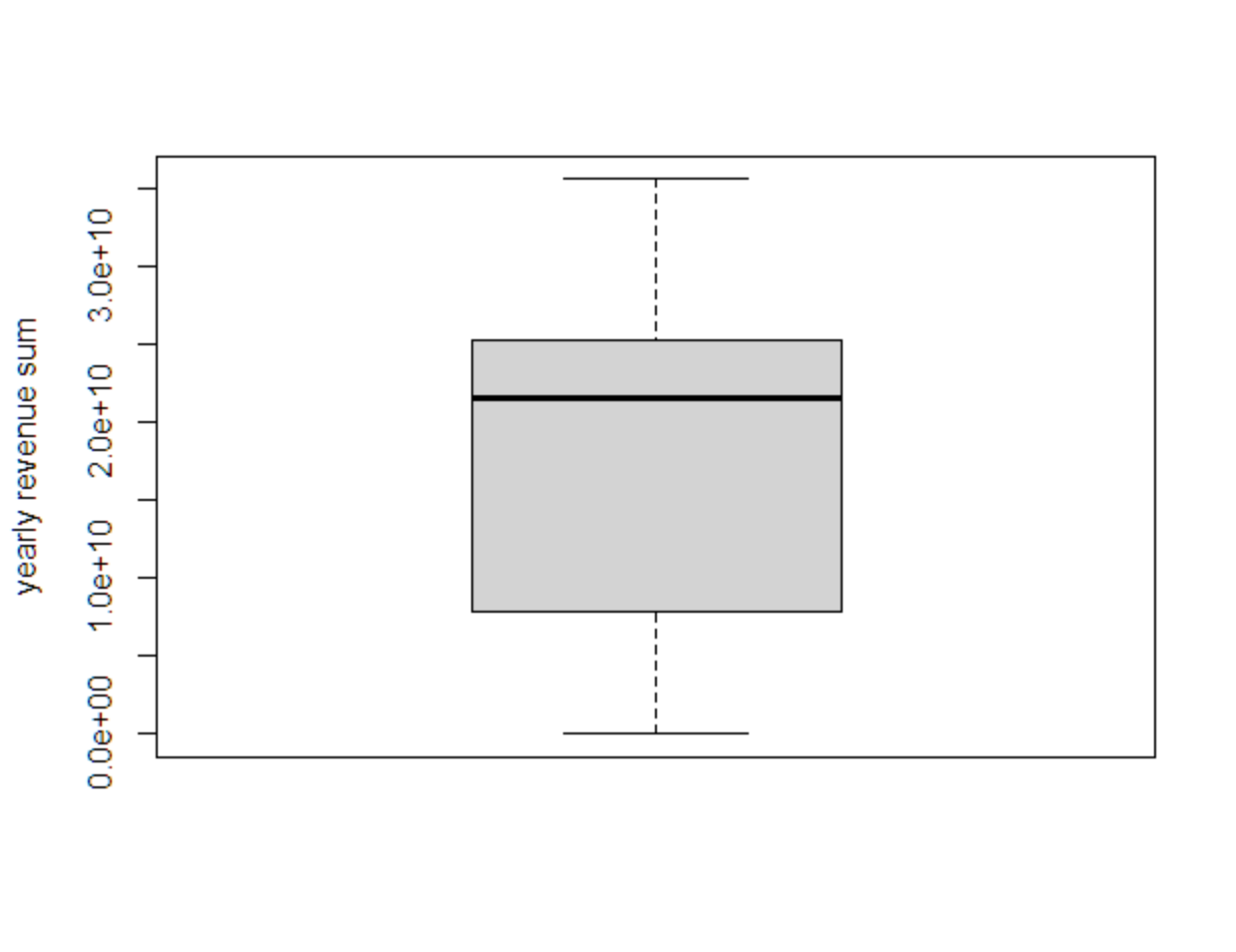
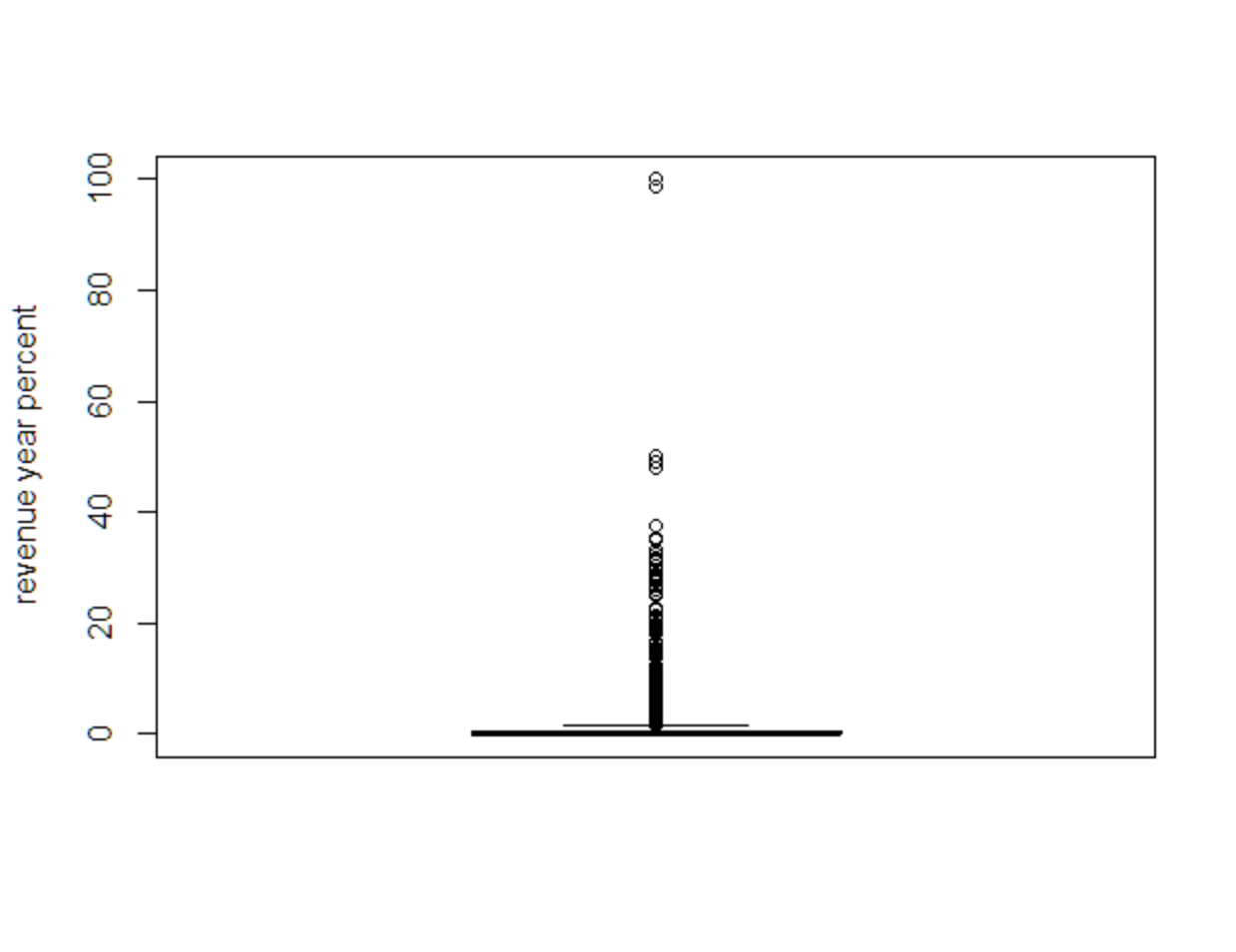
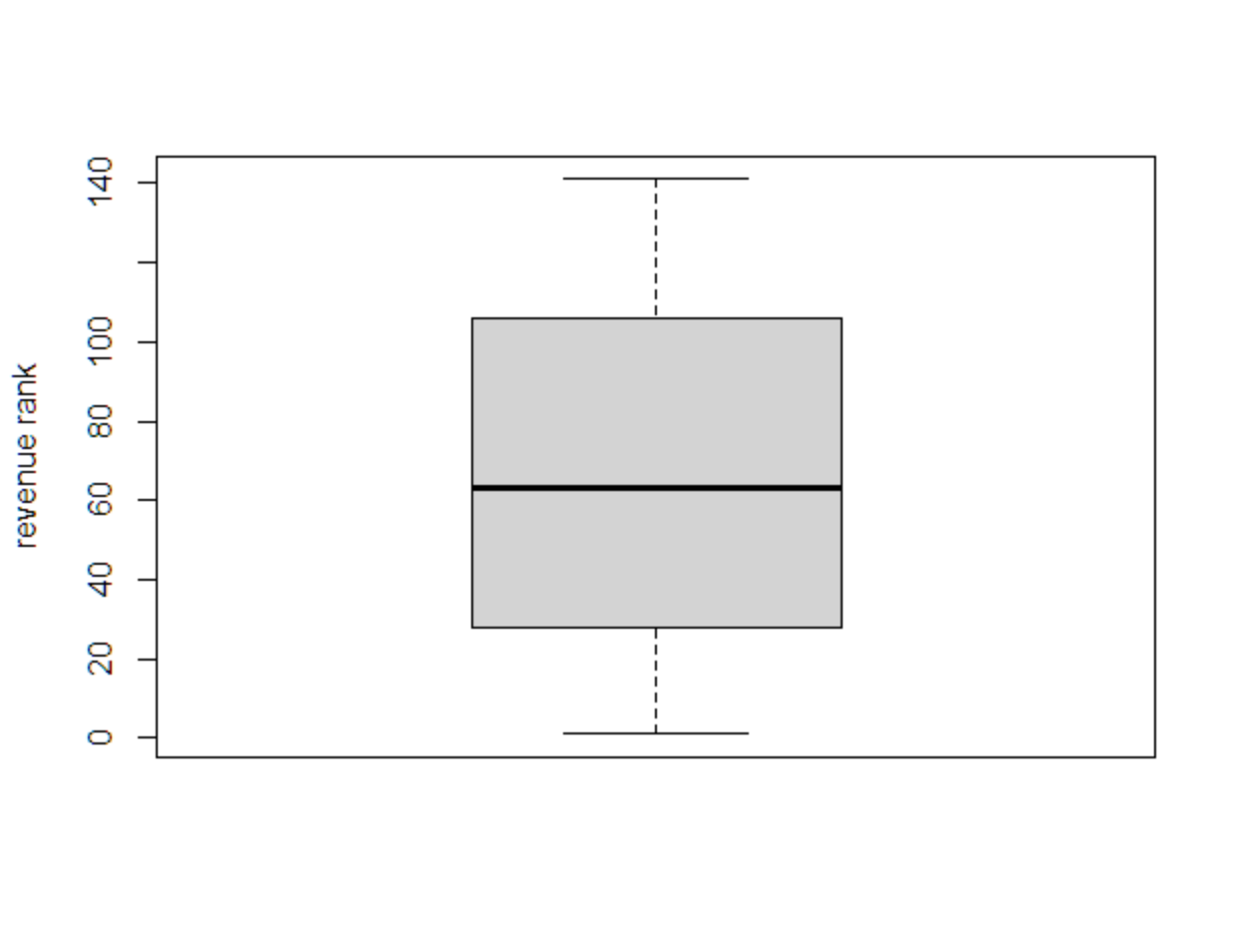
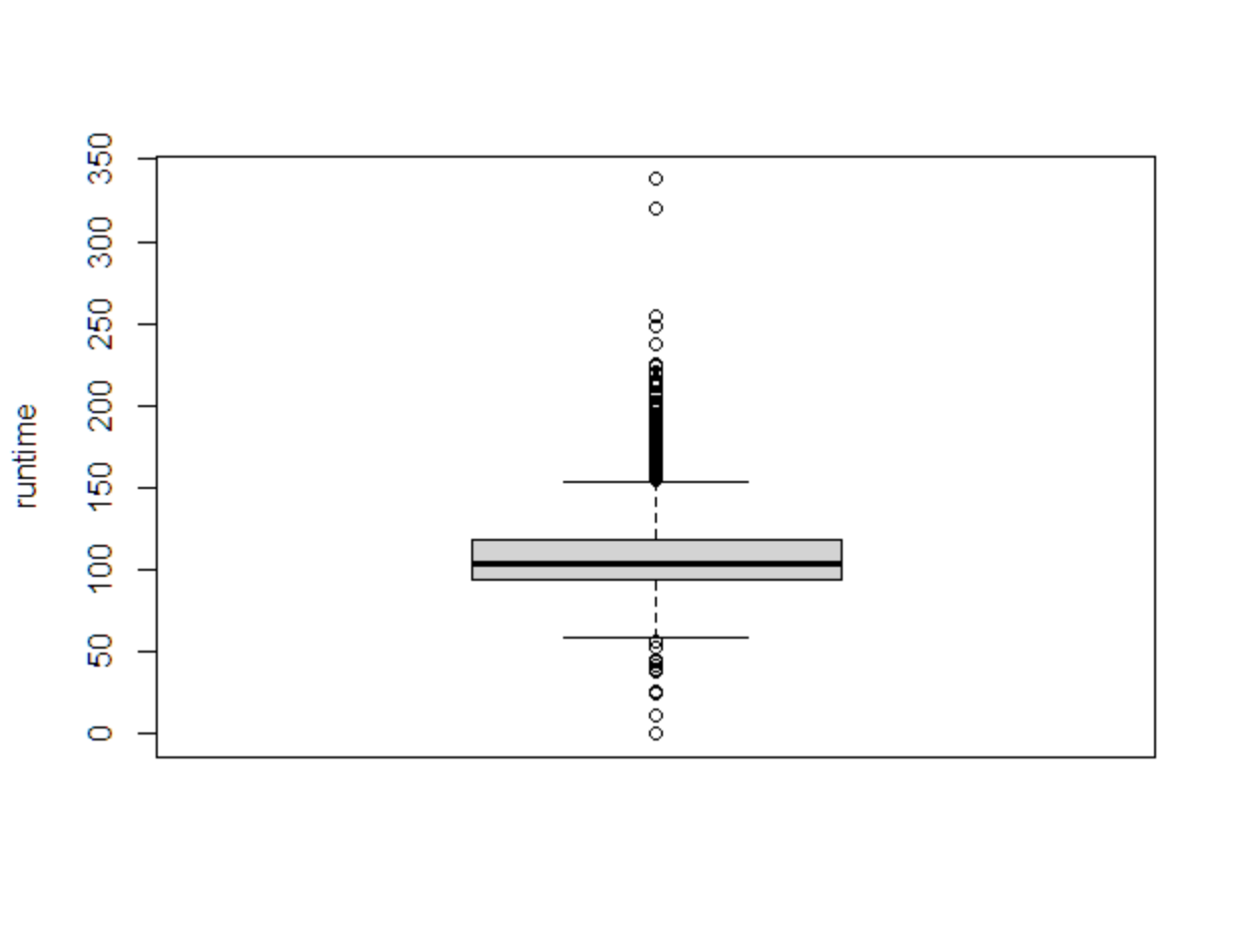
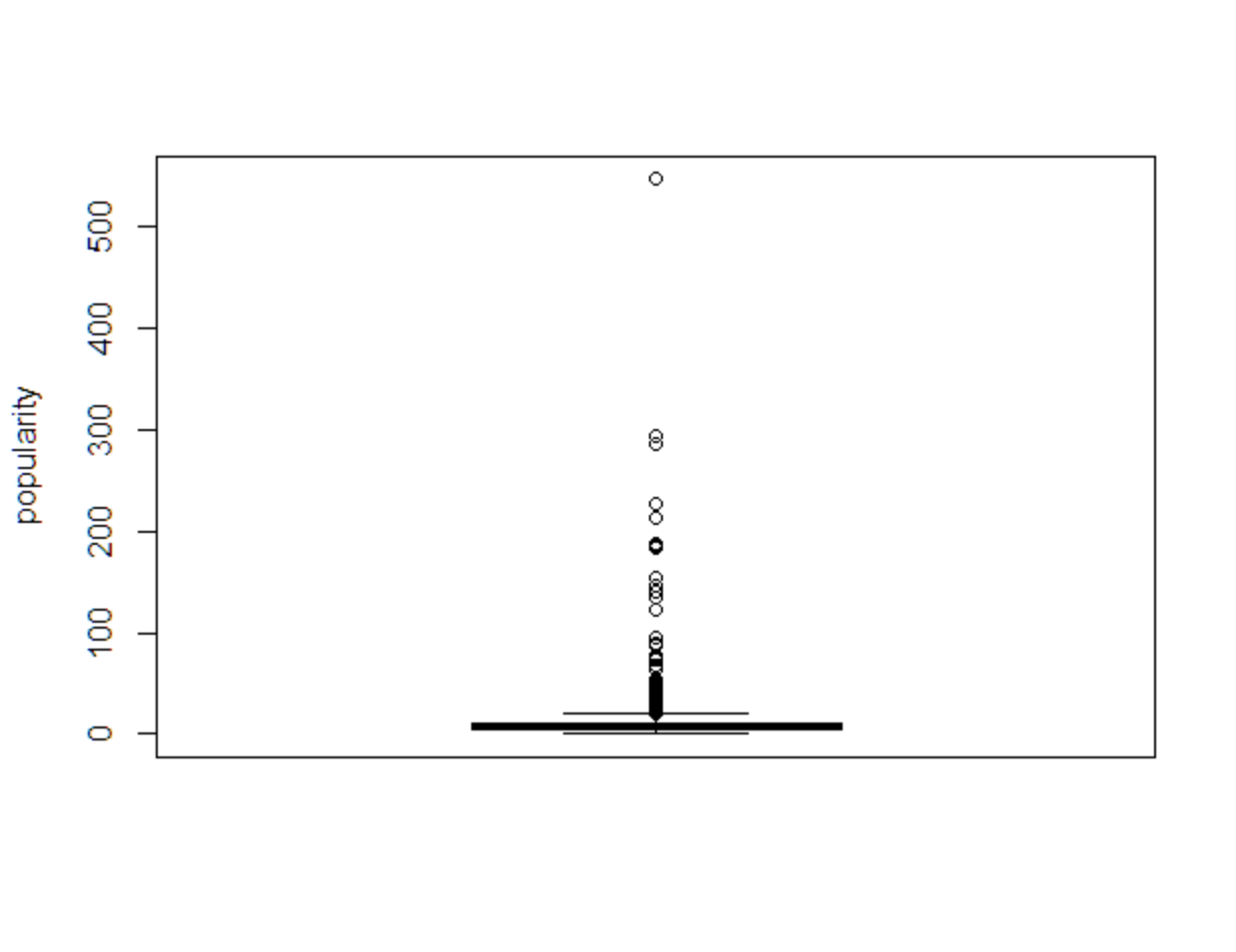
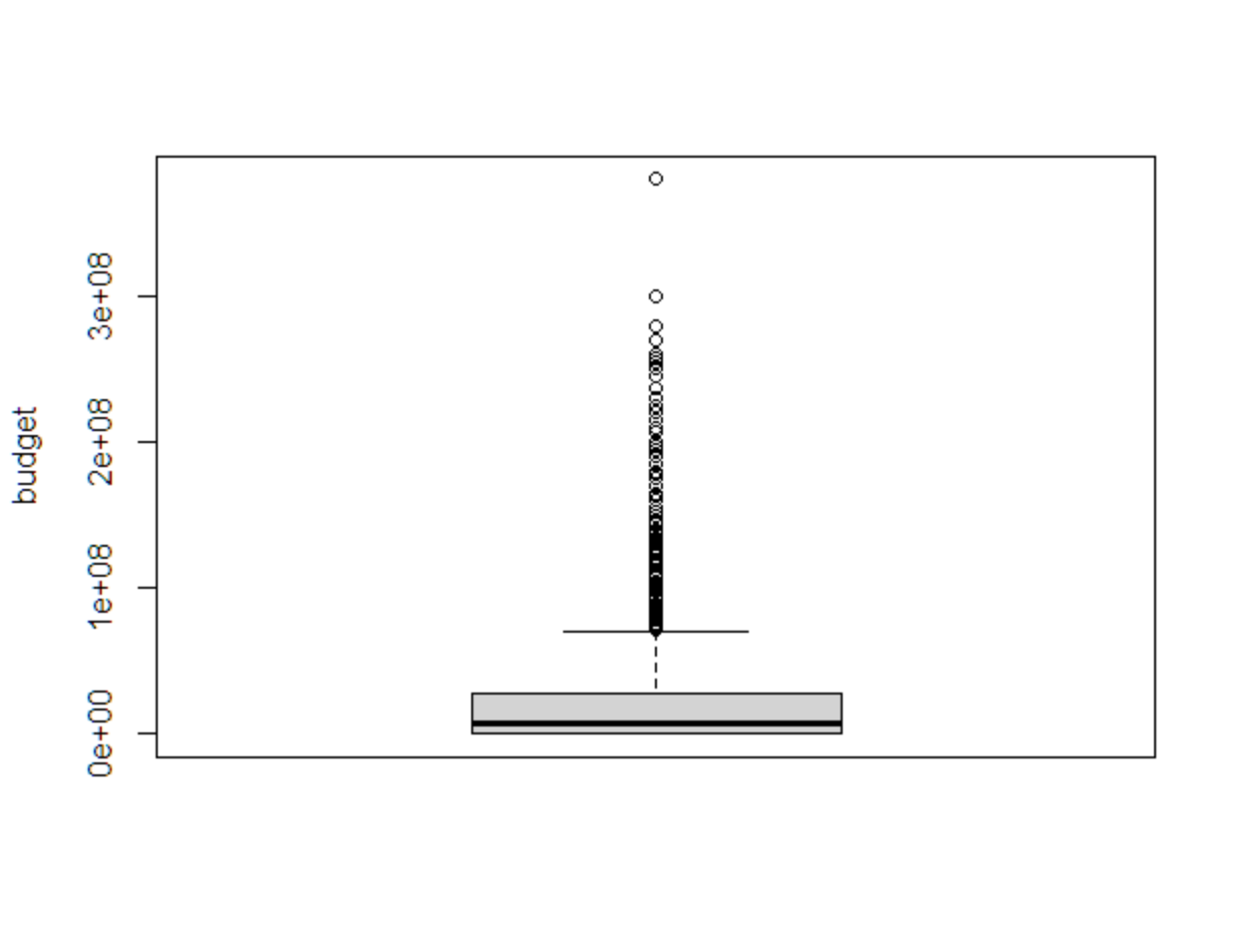








תייצרו גרפים שיכולים לעזור לכם לבדוק האם קיימים נתוני קיצון. תתארו אותם.



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

Missing values:

Year – 1

Month – 1

Runtime -6

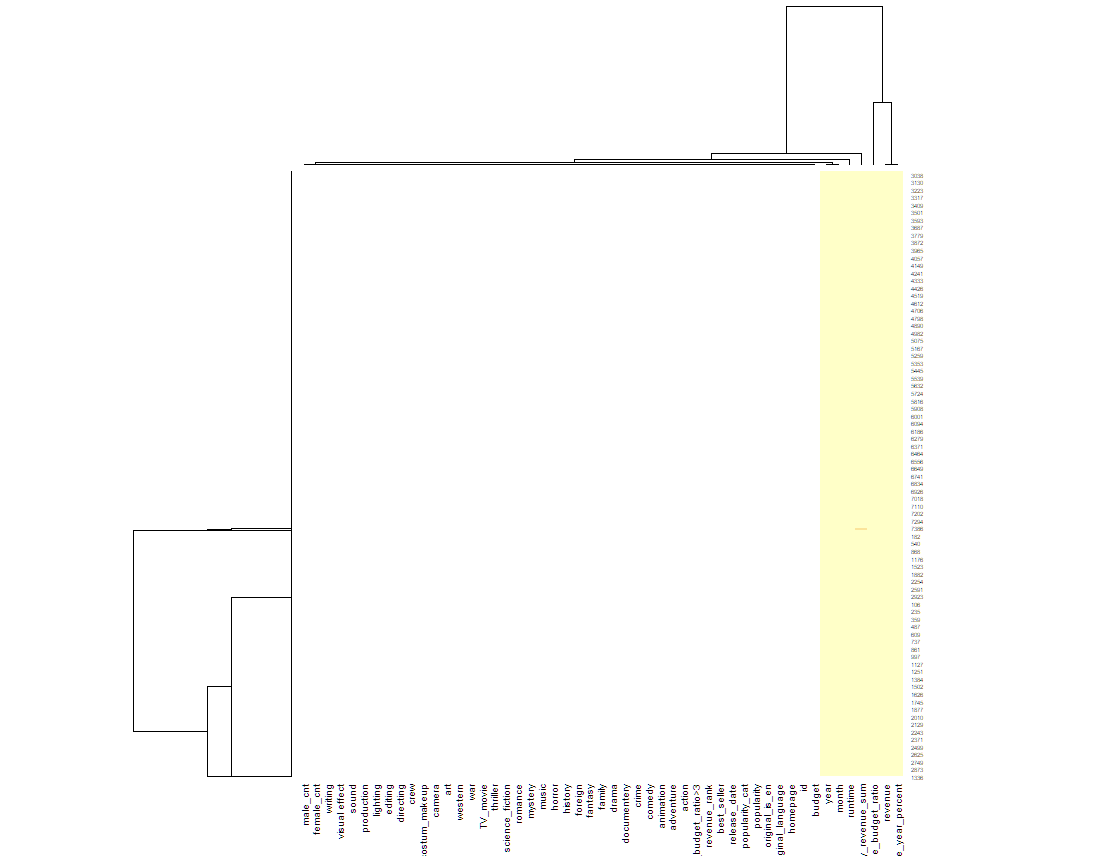
revenue - 4398

yearly\_revenue\_sum – 20

revenue\_year\_percent – 4398

revenue\_budget\_ratio - 5210

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 ?
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?
3. באיזה משתנים הייתם מוחקים את ערכי הקיצון? איך הייתם מוחקים אותם? נמקו.
4. תפעלו על הנתונים לפי מה שהגדרתם בשאלה הקודמת.

var **outlier\_pct distribution\_changed**

[1,] "id" "0" "-"

[2,] "budget" "631" "+"

Correlation - fisher1925: Fisher's z (1925)

z = 15.2657, p-value = 0.0000

Null hypothesis rejected - there is a difference in correlation

**Affect both the distribution and the association – don’t drop it**

[3,] "popularity" "201" "+"

fisher1925: Fisher's z (1925)

z = 0.9092, p-value = 0.3632

Null hypothesis retained – there is no difference in correlation

**Affect the distribution but not the association – drop it**

[4,] "year" "241" "+"

fisher1925: Fisher's z (1925)

z = -0.1970, p-value = 0.8438

Null hypothesis retained – there is no difference in correlation

**Affect the distribution but not the association – drop it**

[5,] "month" "1" "-"

[6,] "runtime" "284" "+"

fisher1925: Fisher's z (1925)

z = -1.1383, p-value = 0.2550

Null hypothesis retained – there is no difference in correlation

**Affect the distribution but not the association – drop it**

**Outcome variables:**

[7,] "revenue" "4713" "+"

[8,] "revenue\_rank" "0" "-"

[9,] "yearly\_revenue\_sum" "20" "-"

[10,] "revenue\_year\_percent" "4778" "+"

var correlation\_changed

[1,] "id" "-"

[2,] "budget" "+"

[3,] "popularity" "-"

[4,] "year" "-"

[5,] "month" "-"

[6,] "runtime" "-"

[7,] "revenue" NA

[8,] "revenue\_rank" "-"

[9,] "yearly\_revenue\_sum" "-"

[10,] "revenue\_year\_percent" "+"

var outlier\_pct distribution\_changed correlation\_changed

1 id 0 - -

2 budget 631 + +

3 popularity 201 + -

4 year 241 + -

5 month 1 - -

6 runtime 284 + -

7 revenue 4713 + <NA>

8 revenue\_rank 0 - -

9 yearly\_revenue\_sum 20 - -

10 revenue\_year\_percent 4778 + +

11 revenue\_year\_percent 4778 + +

>

***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.

var na.count rate

1 revenue 4398 59.4

2 revenue\_year\_percent 4398 59.4

3 yearly\_revenue\_sum 20 0.3

4 runtime 6 0.1

5 runtime\_no\_outliers 6 0.1

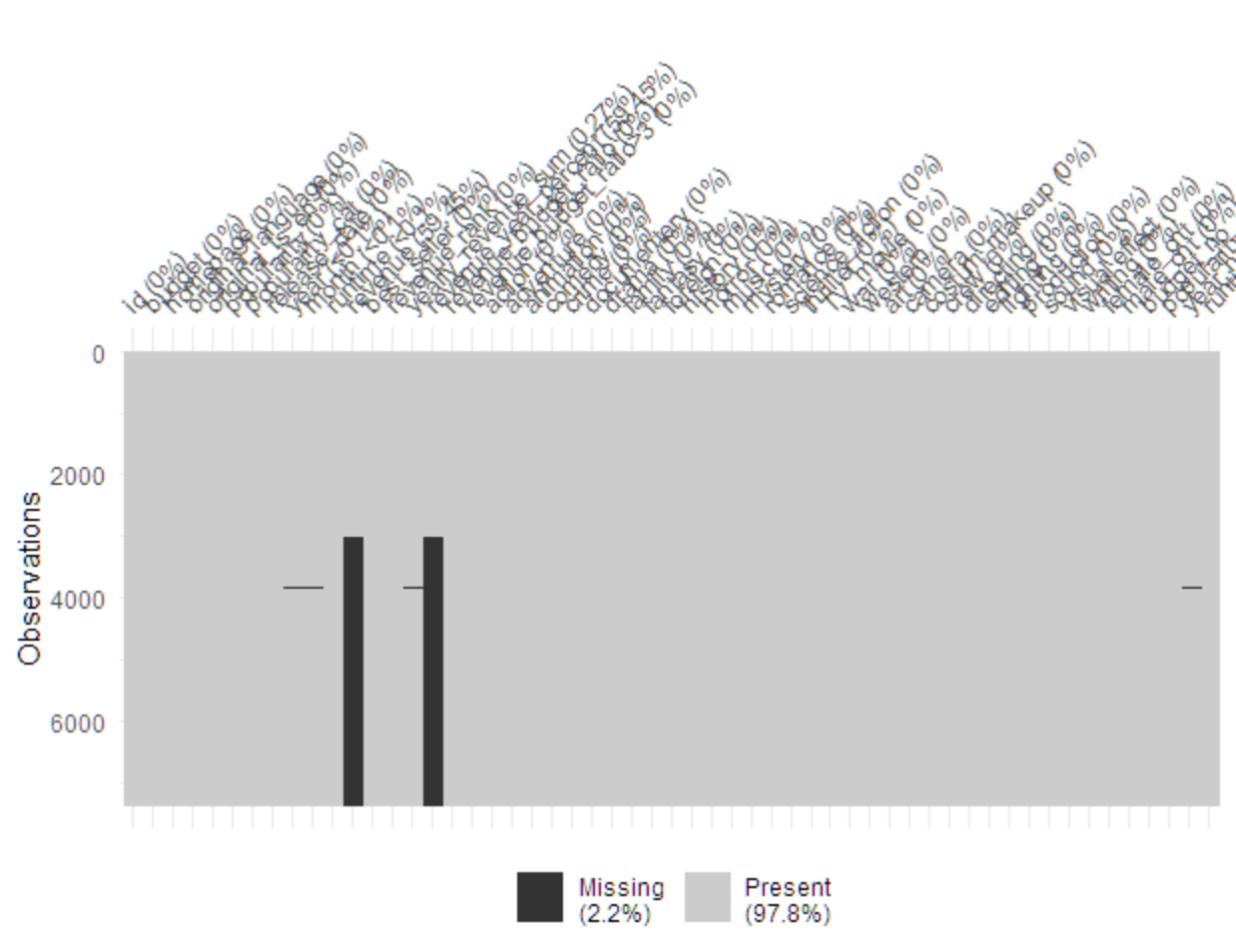
6 year 1 0.0

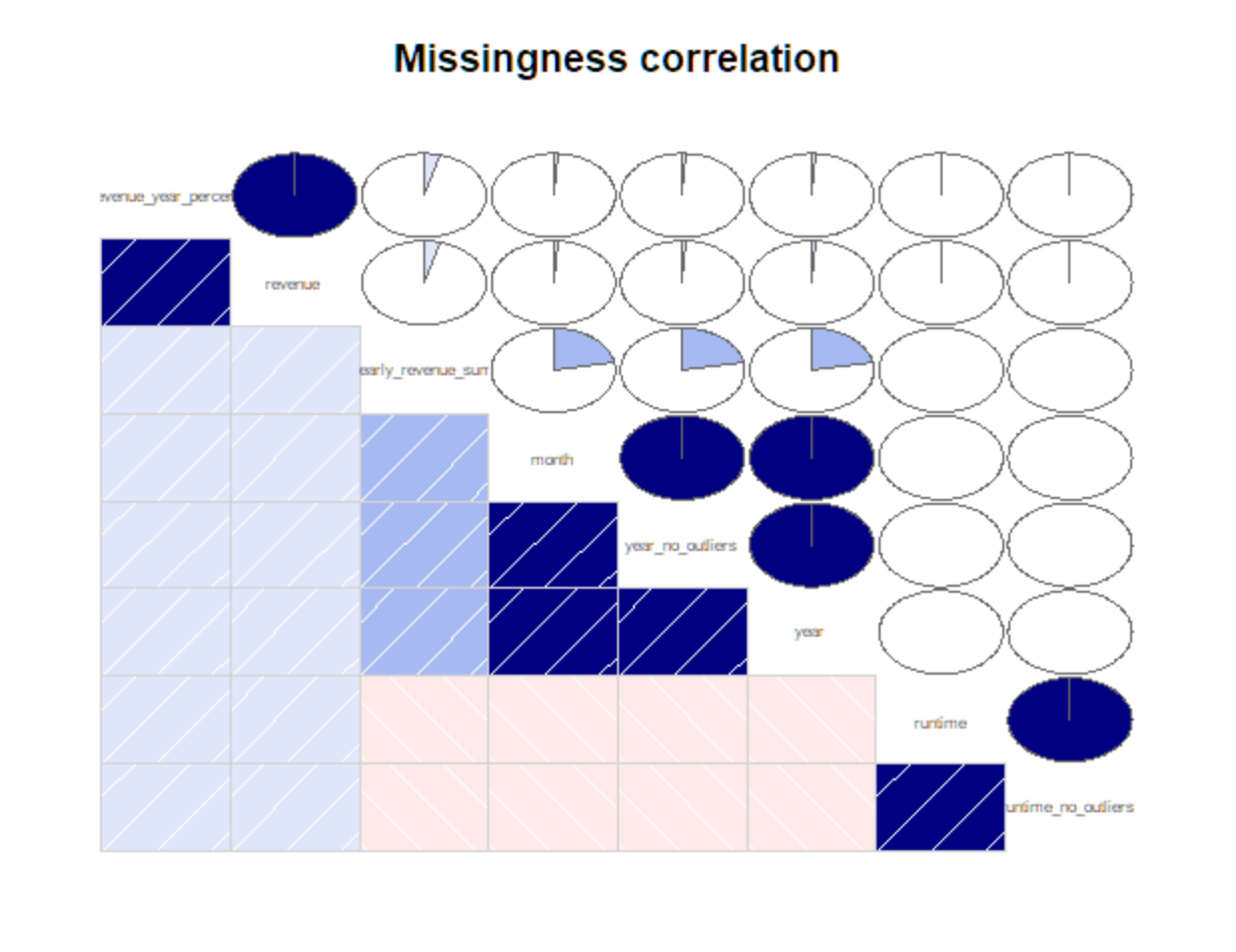
7 month 1 0.0

8 year\_no\_outliers 1 0.0

[[2]]

[1] "This dataset has 2998 (40.5%) complete rows. Original data has 7398 rows."



|  |
| --- |
|  |

m var pvalue

1 revenue\_budget\_ratio (Intercept) 0.9972159

2 revenue\_budget\_ratio revenue\_budget\_ratio 0.9953272

3 revenue (Intercept) 0.9967399

4 revenue revenue 0.9949728

5 revenue\_year\_percent (Intercept) 0.9967399

6 revenue\_year\_percent revenue\_year\_percent 0.9949728

7 yearly\_revenue\_sum (Intercept) 0.9948875

8 yearly\_revenue\_sum yearly\_revenue\_sum 0.9994684

9 runtime (Intercept) 0.9948827

10 runtime runtime 0.9997085

11 year (Intercept) 0.9948809

12 year year 0.9998810

13 month (Intercept) 0.9948809

14 month month 0.9998810

אף משתנה לא יצא סיגניפיקטי

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

**לא הצלחתי לשנות את הנתונים של ה OUTLIERS ל-NULL. זה הקוד שכתבתי**

df < - df %>%

select(popularity, popularity\_no\_outliers)

mutate(popularity\_no\_outliers = if\_else(popularity\_no\_outliers == 0, popularity , 0))

df < - df %>%

mutate(year\_no\_outliers = if\_else(year\_no\_outliers == 0, year , 0))

df < - df %>%

mutate(runtime\_no\_outliers = if\_else(runtime\_no\_outliers == 0, runtime , 0))