סטטיסטיקה תיאורית

ברצונינו לבחון קשר בין כמות ההאזנות כתלות מכמות פלאיליסטים בהם שיר מופיע בSpotify.
הנתונים נלקחו מאתר Kaggle.com קישור: Most streamed Spotify songs 2024.

בהתחלה היו בקובץ 4600 תצפיות (שירים עם כמות האזנות הכי גדולה), לאחר ניקוי הנתונים שאינם נומרים מספר התצפיות ירד ל-4467. אחרי סריקה ראשונית בנתונים שמנו לב כי מספרים הם מאוד גדולים, לכן לשם הנוחות החלטנו לעשות טרנספורמציה ליניארית. חילקנו כמות האזנות במיליון וכמות פלאיליסטים ב100. אחר כך בנינו רגרסיה ליניארית ראשונית. על ציר ה-x כמות הפלאיליסטים שבהם מופיע שיר (נמדד במאות) ועל ציר ה-y כמות האזנות של השיר (נמדד במיליונים).

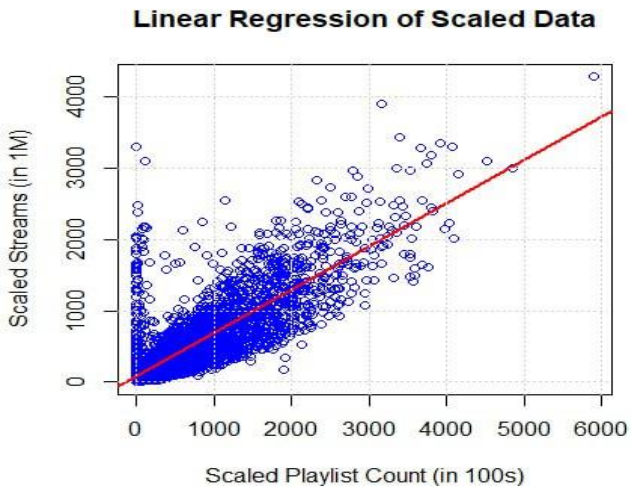
נציג נתוני של מדדי מיקום ופיזור של שני משתנים כנ"ל

מדדי מיקום ופיזור של כמות האזנות (נמדד במיליונים):

ממוצע = 443.839, חציון = 239.635 סטיית תקן = 530.635

באופן דומה נציג כמות פלאיליסטים בו מופיע שיר (נמדד במאות):

ממוצע = 602.161 חציון = 331.51 סטיית תקן = 712.869




```
Residuals:
    Min      1Q  Median      3Q     Max
-1051.1  -110.9   -64.6    48.8  3220.6

Coefficients:
                            Estimate
(Intercept)                78.503081
Spotify.Playlist.Count.scaled  0.606708

Multiple R-squared:  0.6644
```

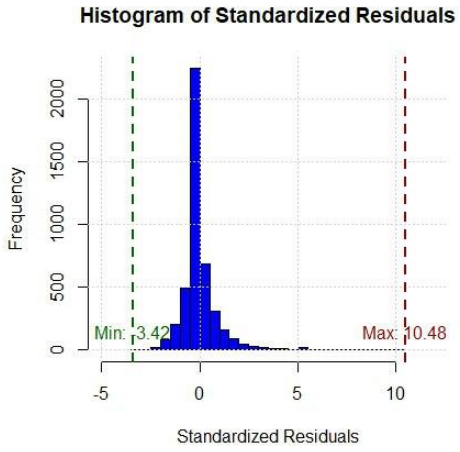
כעת נבנה משוואת הרגרסיה: $y = 0.607x + 78.5$. משמעות השיפוע היא שעל כל עלייה ביחידה אחד (כלומר, במאה) בכמות פלאיליסטים בו מופיע שיר, נצפה לראות עלייה ב 0.6 יחידות (כלומר, ב600000) בכמות האזנות. משמעות החותך היא שמצופה ששיר שלא מופיע באף פלאיליסט יהיה בעל של 78.5 מיליוני האזנות.

מקדם המתאם = 0.8151. מדובר במקדם שמעיד על קורלציה חזקה בין 2 המשתנים, אך לא מושלמת. וזה הגיוני כי קל וחומר כי ככל שהשיר מופיע יותר ברשימות השמעה, כך יהיו לו יותר ההאזנות.

נתבונן בטיב הקשר R^2 שמעיד על אחוז שונות המוסברת של המשתנה המוסבר. במודל שלנו שונות המוסברת של כמות האזנות על ידי כמות הפלאיליסטים בהם מופיע שיר שווה ל 66.44% של השונות של כמות ההאזנות.

נשים לב כי יש לנו מספר רב של תצפיות חריגות. נרצה לחשב סטיית תקן של הטעויות (RMSE)

קיבלנו ש - Residual standard error: 307.4 ובשביל לבחון האם ישנן תצפיות חריגות נרצה לתקנן את הערכים של השגיאות לפי סטיית תקן. קיבלנו ציוני תקן של כל אחד מהשגיאות. כעת, נבנה היסטוגרמה של ערכים מתוקננים.

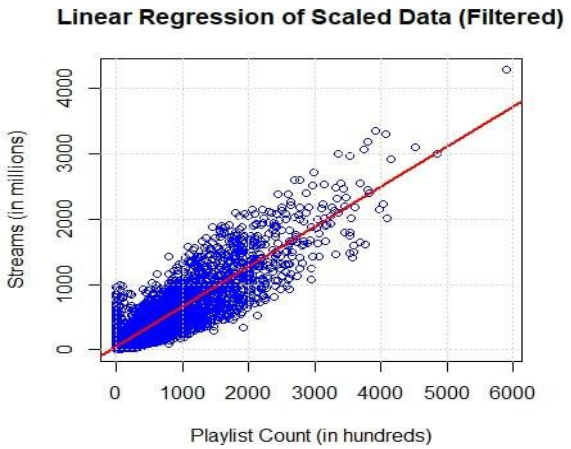


משום שיש לנו כמות רב תצפיות ציירנו קוים שמסמנים ציון תקן מינימלי ומקסימלי של השגיאות כדי שיהיה קל לראות.

הוחלט למחוק את התצפיות שנמצאים במרחק גדול מ3 סטיית תקן מהממומע. אחרי מחיקה נותרו 4376 תצפיות

```
[1] "Number of removed observations: 91"
```

לאחר מחיקת התצפיות החריגות נרצה לבנות רגרסיה מחדש ולהשוואת קורלציה, R^2, ומקדמי הרגרסיה לפני ואחרי



השוואה בין המודלים

```
Residuals:
    Min      1Q  Median      3Q     Max
-880.47  -82.28  -34.08   66.30  944.49

Coefficients:
                                Estimate
(Intercept)                    45.460346
Spotify.Playlist.Count.scaled   0.612394
```

<u>נציג נתוני של מדדי מיקום ופיזור לאחר מחיקה</u>

מדדי מיקום ופיזור של כמות האזנות (נמדד במיליונים): ממוצע = 423.723, חציון = 228.855 סטיית תקן = 485.175

באופן דומה נציג כמות פלאיליסטים בו מופיע שיר (נמדד במאות): ממוצע = 601.350 חציון = 336.205 סטיית תקן = 706.523

משוואת רגרסיה החדשה שהתקבלה היא $y = 0.612x + 45.46$. הירידה המשמעותית בחותך ב 33.04 מיליוני ההזאנות לאחר הסרה של התצפיות, מצביעה על כך שהתצפיות שהוסרו השפיעו על התחזיות בערכים נמוכים של x. (רואים שרוב התצפיות חריגות נמצאות קרוב ל x=0). לעומת זאת, השינוי הקטן בשיפוע (מ-0.607 ל-0.612) מצביע על כך שהתצפיות חריגות השפיעו באופן זניח על המגמה הכללית. מקדם המתאם החדש הוא 0.8918 שמעיד על קשר עוד יותר חזק בין המשתנים. אחוז שנות המוסברת גדל. R^2 שווה ל- 0.7953, כלומר 79.53%. גידול בשנות המוסברת מעיד לנו על כך שהמודל הזה יותר טוב לחיזוי.

<u>בחרנו את 5 תצפיות</u> (שירים) אקראיות וחזינו את כמות השמיעות לשיר זה לפי המודל החדש שלנו.

Track	Spotify.Playlist.Count.scaled	Spotify.Streams.scaled	Predicted	Residual
BBY BOO	11.59	6.172030	52.55799	-46.38596
The Painter	160.04	107.754524	143.46786	-35.71334
Tell Em	1569.28	499.262863	1006.47778	-507.21492
Hijo Mio	3.89	1.612356	47.84256	-46.23020
Double Fantasy (with Future)	366.31	180.940067	269.78634	-88.84627