

linear regrresion

Part One



A book and calculator on a table

Description automatically generated

Group 35

|  |  |  |
| --- | --- | --- |
| 207477936 | 205417439 | 205570203 |

**תוכן עניינים**

[יצירת טבלת משתנים: 2](#_Toc159097879)

[תיאור המשתנים: 3](#_Toc159097880)

[תיאור קשרים בין משתנים: 4](#_Toc159097881)

[ניתוח תיאורי +ניתוח חריגים: 7](#_Toc159097882)

[פונקציית צפיפות והתפלגות מצטברת: 11](#_Toc159097883)

[הצגת תרשים פיזור מורכב: 13](#_Toc159097884)

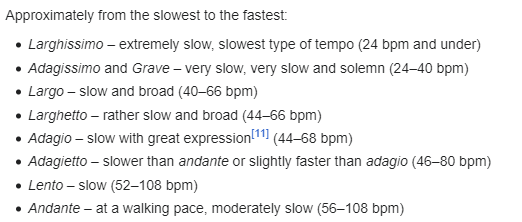
[ייצוג קשרים בעזרת תרשימים: 14](#_Toc159097885)

[טבלאות שכיחות: 15](#_Toc159097886)

יצירת טבלת משתנים:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **סוג המשתנה -מוסבר/מסביר** | **סימון בR (לדוג' X2/age)** | **יחידת מידה** | **סוג המשתנה – רציף / קטגוריאלי** | **הסבר בקצרה על כל משתנה** |
| מסביר | youtube.views..m |  | רציף | מספר הצפיות של השיר (במליונים) מהווה מדד שבעזרתו ניתן לאמוד מהי מידת הפופולריות של השיר. |
| מסביר | energy |  | רציף | מדד תפיסתי, המייצג את עוצמתו של השיר. ערכים גבוהים מעידים שהשיר קצבי ואנרגטי, ונמוכים מעידים שהשיר רך. |
| מסביר | loudness | דציבלים | רציף | מדד אשר נמדד בדציבלים. מעיד כמה השיר רועש/ שקט. |
| מסביר | mode |  | קטגוריאלי | משתנה בינארי המעיד אם השיר מנוגן במיג'ור או מינור. |
| מסביר | speechiness |  | רציף | משתנה שמייצג את כמות המילים שמדוברות בשיר. |
| מסביר | acousticness |  | רציף | מדד תפיסתי המודד עד כמה השיר אקוסטי (0-פחות, 1- יותר). |
| מסביר | liveness |  | רציף | מדד תפיסתי המודד את הנוכחות של קהל בהקלטה (0-ללא נוכחות קהל, 1- נוכחות קהל מרובה). |
| מסביר | valence |  | רציף | מדד תפיסתי, המודד את רמת החיוביות של השיר. רמת חיוביות נמדדת בעזרת האופטימיות של השיר, המילים השמחות והטובות. |
| מסביר | tempo |  | רציף | משתנה רציף, שבעזרתו ניתן למדוד את מקצב השיר. |
| מסביר | duration\_ms |  | רציף | משתנה רציף, הנמדד במילי שניות ומייצג את משך הזמן של השיר. |
| מסביר | gender |  | קטגוריאלי | משתנה בינארי המייצג את המין של הזמר שביצע את השיר. |
| מוסבר | shazam |  | רציף | כמות החיפושים של השיר באפליקציה |

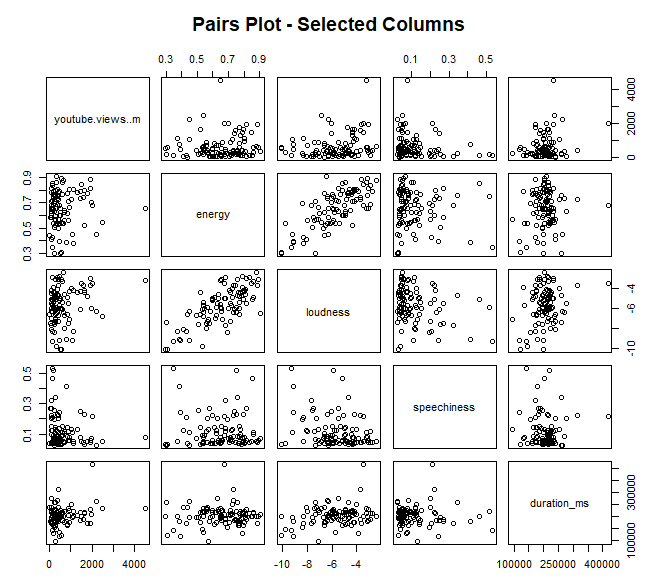
תיאור המשתנים:

1. **YouTube views-** משתנה המתאר את מספר הצפיות של השיר ב-YouTube. בעזרתו, ניתן לאמוד את מידת הפופולריות של השיר. מידת הפופולריות של השיר משפיעה על מספר החיפושים שלו באפליקציה.   
   ככל שמספר הצפיות גבוה יותר עולה מידת הפופולריות שלו עולים מספר החיפושים שלו באפליקציה.
2. **Energy-** מדד תפיסתי של עוצמה (0-עוצמה נמוכה 1-עוצמה גבוהה). ישנם אנשים שמעדיפים שירים מלאי אנרגיה וממריצים, וישנם אנשים שמעדיפים שירים רגועים יותר ורכים. יכול להשפיע על מספר החיפושים של שיר מסוים באפליקציה- אנשים יכולים לחפש שיר בהתאם להעדפה שלהם לסגנון מוזיקה מסיים.   
   ככל שאנרגיית השיר גבוה יותר ההעדפה של המאזין משתנה בהתאם עשוי להשפיע על מספר החיפושים של השיר באפליקציה.
3. **Loudness-** קולניות הרעש הכללי של רצועת השיר בדציבלים. ישנם אנשים שמעדיפים לשמוע שירים קצביים, חזקים ואנרגטיים וישנם אנשים שמעדיפים שירים שקטים ורגועים. משתנה זה עשוי להשפיע על מספר החיפושים של שיר מסוים באפליקציה- אנשים יכולים לחפש שיר בהתאם להעדפה שלהם לסגנון מוזיקה מסוים.   
   ככל שהקולניות של השיר גבוה יותר ההעדפה של המאזין משתנה בהתאם עשוי להשפיע על מספר החיפושים של השיר באפליקציה.
4. **Mode-** האם השיר מנוגן במיג'ור (1) או מנור (0). ישנם אנשים שמעדיפים שירים המנוגנים במינור וישנם אנשים שמעדיפים שירים המנוגנים במיג'ור. יכול להשפיע על מספר החיפושים של שיר מסוים באפליקציה- אנשים יכולים לחפש שיר בהתאם להעדפה שלהם לסגנון מוזיקה מסויים (ניגון שיר לפי מינור או מיג'ור).   
   כאשר הניגון של השיר מתבצע במינור/ מיג'ור ההעדפה של המאזין משתנה בהתאם עשוי להשפיע על מספר החיפושים של השיר באפליקציה.
5. **Speechiness-** כמות המילים שמדוברות בשיר. ישנם שירים בהם יש יותר מילים מדוברות (כגון- שירי ראפ) וישנם שירים שרובם מתמקדים בנגינה ובמלודיה ויש בהם פחות מילים מדוברות. ישנם אנשים שמעדיפים שירים עם הרבה מלל וישנם אנשים שמעדיפים שירים עם פחות מלל.   
   כאשר השיר מכיל בתוכו מספר רב/מועט של מילים ההעדפה של המאזין משתנה בהתאם עשוי להשפיע על מספר החיפושים של השיר באפליקציה.
6. **Acousticness-** עד כמה השיר אקוסטי (0-פחות, 1- יותר). מוזיקה אקוסטית היא מוזיקה המשתמשת אך ורק או בעיקר בכלים המפיקים צליל באמצעים אקוסטיים (כגון- גיטרה, כינור, צ'לו, ויולה או נבל - כל מה שבאמת לא מסתמך על הגברה חשמלית כדי להישמע), בניגוד לאמצעים חשמליים או אלקטרוניים (כגון- סינתיסייזר, מחשב, מכונת תופים, סמפלר ולופר).   
   ישנם אנשים שמעדיפים שירים המבוצעים באמצעות כלים אקוסטיים וישנם אנשים שמעדיפים מוזיקה אלקטרונית- בה הסאונד של הכלים נשמע ממוחשב.   
   כאשר השיר מתנגן בצורה אקוסטית/ אלקטרונית ההעדפה של המאזין משתנה בהתאם עשוי להשפיע על מספר החיפושים של השיר באפליקציה.
7. **Liveness-** נוכחות של קהל בהקלטה. פעמים רבות, מוסיפים לשירים שמע של נוכחות קהל בהקלטת השיר בכדי לייצר למאזין חוויה של נוכחות קהל עימו בשמע השיר, אהדה לשיר או הוספת אנרגיה למקטעים מסוימים שיוצר התוכן מעוניין בכך. נוסף לכך, ישנם שירים שמנוגנים בהופעות, וניתן למצוא שיר מסוים הן מנוגן בהופעה ספציפית שאירעה והן בהקלטה באולפן. לעיתים, נוכחות הקהל בהופעות משפרת את ביצועי הזמר בשיר ועל כן, יש יוצרים שמפרסמים ביצועים שלהם בהופעות, עם השתתפות קהל.   
   ישנם אנשים שמעדיפים את ביצוע של שיר מסוים עם הקלטות קהל או שסתם מעדיפים את הביצוע של הזמר מההופעה, וישנם אנשים שמעדיפים שירים ללא נוכחות קהל וללא כל רעשי רקע.  
   כאשר השיר מכיל בתוכו מספר רב/מועט עם סאונד קהל בהקלטה ההעדפה של המאזין משתנה בהתאם עשוי להשפיע על מספר החיפושים של השיר באפליקציה.
8. **Valence-** האם השיר חיובי (0-לא, 1-כן). לרוב, שירים מעבירים מסר/השקפת עולם/ הבעת רגשות/ חוויה מסוימת שיוצר התוכן רוצה להעביר למאזיניו, ועל כן, שירים עשויים להכיל בתוכנם מידת חיוביות מסוימת אותה יוצרי השיר מעוניינים להעביר.כאשר השיר מכיל מספר רב/מועט של חיוביות בתוכנו ההעדפה של המאזין משתנה בהתאם עשוי להשפיע על מספר החיפושים של השיר באפליקציה.
9. **Tempo-** מקצב השיר. ה-Tempo מחושב על ידי (BPM). בעזרתו מבצעי השיר יכולים לדייק את הביצועים שלהם ולהחזיק קצב יציב או משתנה, בהתאם לרצונם. נוסף על כך, בעזרת Tempo ניתן לשייך את השיר לג'אנר מסוים לדוגמה[[1]](#footnote-1): ****חובבי מוזיקה מסוימים עשויים לחפש שיר בהתאם לג'אנר שלו. עם זאת, יש אנשים שפשוט מעדיפים שירים קצביים יותר וישנם כאלו שמעדיפים קצביים פחות.כאשר השיר מכיל מקצב גבוה/ נמוך ההעדפה של המאזין משתנה בהתאם עשוי להשפיע על מספר החיפושים של השיר באפליקציה.
10. **duration -** משך השיר. משתנה שבודק מה אורכו של השיר. נמדד במילישניות. ישנם שירים שמשכם ארוך יותר/ פחות, בהתאם ליוצר השיר ולכאוות נפשו של מבצע השיר. לעיתים, יש שירים שמשכם ארוך יותר- לדוגמה בשל סולו גיטרה/ תופים, הידברות עם הקהל בהופעה או כל מיני פרמטרים המשפיעים על אורכו של השיר.   
    כאשר משך השיר ארוך יותר/פחות ההעדפה של המאזין משתנה בהתאם עשוי להשפיע על מספר החיפושים של השיר באפליקציה.
11. **Gender-** מין הזמר המבצע. שיר עשוי להתבצע כסולו של זמר אחד, ומינו עשוי להיות זכר, נקבה ואף השיר עשוי להתבצע כדואט של שני המינים. בניסוי זה, ניתן לראות כי מבצע השיר היה או זכר או נקבה (בהתאם לנתוני הניסוי, המשתנה הינו משתנה קטגוריאלי).  
    ישנם אנשים שמעדיפים שירים המבוצעים בהתאם למין הזמר- העדפה לפי קול נשי/גברי .   
    כאשר מגדר זמר השיר זכר/ נקבה ההעדפה של המאזין משתנה בהתאם עשוי להשפיע על מספר החיפושים של השיר באפליקציה (ביצועו של השיר לפי קול נשי או גברי).

תיאור קשרים בין משתנים:  
בחרתנו 5 משתנים מסבירים ורציפים:  
"youtube.views..m", "loudness", "speechiness" "energy", "duration\_ms".   
בעזרת פונקציית ה- plot ניתן לראות את הקשרים של כל 2 משתנים רציפים שבחרנו.  
פלט פונקציית ה- plot הוא גרף של עקומות פיזור בין 2 משתנים שנבחר. בעזרת גרף זה ניתן להסיק על קשרים בין 2 משתנים- האם קיימת בין המשתנים קורלציה. R הינו מקדם המתאם, שבעזרתו ניתן לקבוע האם קיים קשר לינארי בין 2 משתנים X ו-Y.

* ערכו של *R*  נע בין 1 ל- מינוס 1.
* הערך 1 יתקבל כאשר יש קשר ליניארי חיובי מלא בין המשתנים- Y תלוי ב-X.
* הערך 1- יתקבל כאשר יש קשר ליניארי שלילי מלא בין המשתנים- X תלוי ב-Y.
* ערך 0- אין תלות בין המשתנים.
* A screenshot of a graph

  Description automatically generatedניתן גם להסיק על קשר בין שני משתנים אם עוצמתו גבוהה/ נמוכה (לא בהכרח 1 או מינוס 1, אלא מספר קרוב לקצוות אלו).

  
לפי דיאגרמת הפיזור והטבלה לפי ניסוי זה, לא ניתן להעיד על הקשרים בין 2 המשתנים ככלל הואיל ותוצאות הניסוי תלויות בשירים שנבחרו, בתצפיות ובמדגם ספציפי זה. לכן, הקשרים המוצגים בין כל 2 משתנים הם בהתאם לניסוי זה ובניסוי אחר, הקשרים עשויים להיות שונים. בנוסף, יכול להיות שיש קשר בין 2 משתנים מסבירים שנבדקו, פשוט הוא אינו קשר לינארי. לפי מקדם המתאים, ניתן רק לומר האם יש קשר לינארי בין 2 המשתנים.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **משתנה 1** | **משתנה 2** | **מקדם מתאם** | **סוג קשר- סיבתי/ מדגמי** | **הסבר** |
| Energy | Loudness | 0.7392204 | סיבתי | ניתן לראות כי ערכו של מקדם המתאם בין 2 המשתנים המסבירים גבוהה וקרוב ל-1, מה שמעיד על כך שקיימת קורלציה חיובית חזקה בין 2 המשתנים.  ככל שעוצמתו של השיר גבוהה יותר, כך גם יש בו יותר אנרגיה ולהפך.  לדעתנו, הקשר הינו סיבתי הואיל ושירים רועשים יותר יהיו בעלי דציבלים גבוהים יותר- דציבל משמש למדידת עוצמת הקול, לרוב, כאשר שיר אנרגטי יותר הוא קולני יותר. |
| Duration | Speechiness | -0.01815858 | מדגמי | בניסוי זה, אין קורלציה בין המשתנים, וניתן לומר כי לא קיים קשר לינארי בין המשתנים- 2 המשתנים בלתי תלויים. ציפינו לראות קשר יותר חזק בין 2 המשתנים- כאשר יש יותר מילים בשיר, אז הוא עשוי להתבצע במשך ארוך יותר, הואיל וייקח לסולן לבצע את השיר בזמן ממושך יותר.  יכול להיות, שהתוצאות לא יצאו כפי שחשבנו כי נבחרו שירים שכמות המילים בהם מראש מעוטה (המדד לכמות המקסימלית היא 0.53) ומרחב המדגם של סוגי השירים עם כמות המילים לא מגוון דיו (סטיית התקן הינה 0.1051044), מה שמשפיע על מהימנות התוצאות. |
| Energy | YouTube views | 0.07166541 | מדגמי | בניסוי זה, אין קורלציה בין המשתנים, וניתן לומר כי לא קיים קשר לינארי בין המשתנים- 2 המשתנים בלתי תלויים.  ציפינו לראות קשר יותר חזק בין 2 המשתנים כיוון שחשבנו כי אנרגיית השיר עשויה להשפיע על מספר הצפיות ביוטיוב של השיר. פעמים רבות, שירים מושמעים בהתאם לרמת האנרגטיות שלהם. שירים בעליי אנרגיה גבוהה מושמעים במסיבות, אימונים.. ושירים בעליי אנרגיה נמוכה מושמעים בטקסים, שחרור של אימון... על כן, היה לנו הגיוני כי רמת האנרגיה של השיר משפיעה על הצפיות שלו ביוטיוב.  אולם, אנו סבורים כי התוצאות מושפעות מאופן בחירת מרחב המדגם- הן של השירים הנצפו והן של אופי האנשים שצפו בשירים.  על כן, אם נבחר אוכלוסיית מדגם שונה, אנו סבורים כי תוצאות הניסוי יהיו שונים גם הם. |
| Loudness | Speechiness | -0.24505032 | מדגמי | בניסוי זה, יש קורלציה שלילית נמוכה בין המשתנים, וניתן לומר כי קיים קשר לינארי חלש יחסית בין המשנים. ככל שעוצמתי של השיר נמוכה יותר, יש יותר מילים.  ציפינו לראות קשר חיובי יותר חזק בין 2 המשתנים כיוון שברוב השירים שאנו, חבריי הקבוצה מכירים, כאשר כמות המילים המדוברת בהם גבוהה, הם רועשים לרוב ולהפך. (כמו שירי ראפ, שיש בהם הרבה מילים והם לרוב רועשים יותר). תוצאות הניסוי עשויות לנבוע ממרחב המדגם שנבחר בניסוי- הן השירים שנבחרו והן אופי של המאזינים שהשתתפו בניסוי. אנו סבורים כי בניסוי אחר, יתקבלו תוצאות שונות, וכי אם באופן תיאורתי, אם נעשה ניסוי בו ישתתפו כלל השירים הקיימים וכלל האנשים בעולם, עשוי להיות קשר סיבתי חזק בין שני המשתנים. |
| YouTube views | Speechiness | -0.16353343 | סיבתי | בניסוי זה, יש קורלציה שלילית נמוכה בין המשתנים, וניתן לומר כי קיים קשר לינארי חלש יחסית בין המשנים. הקשר הוא קשר רופף: ככל שיש פחות מילים, כך יש יותר צפיות ביוטיוב לשיר מסוים. לא ציפינו לראות תוצאות שונות. יש אנשים שמעדיפים שירים עם יותר מילים ויש כאלה שמעדיפים עם פחות מילים, לכן לא ניתן לומר באופן חד משמעי שככל שיש יותר מילים כך יש יותר צפיות או בכל יחס ישר אחר בין כמות המילים לכמות הצפיות של השיר. |

ניתוח תיאורי + ניתוח חריגים:

|  |  |  |  |
| --- | --- | --- | --- |
| **משתנה** | **ערכים** | **Box plot- ללא שינוי** | **טיפול בחריגים** |
| **YouTube views** | Minimum: 1.216 Q1: 194.7615 Median: 394.2025 Mean: 632.813 Q3: 838.0047 Maximum:4520.946 Skewness: 2.488077 Sd: 691.6722  חריגים:  2489.861 2003.610 1973.177 4520.946 2222.641 1929.154 1919.571 | A diagram of a box plot  Description automatically generated | מרבית השירים נצפו בין 194 מיליון פעמים ל838 מיליון פעמים. לכן שירים בעלי מעל ל1.804 מילארד צפיות נחשבים חריגים. בחרנו שלא להוציא את החריגים מכיוון ובמציאות מצב זה אכן קיים ואפשרי. |
| **Energy** | Minimum: 0.296 Q1: 0.56 Median: 0.676 Mean:0.6569 Q3: 0.7715 Maximum: 0.909 Skewness: -0.5563534 Sd: 0.1455312 |  | אין חריגים |
| **Loudness** | Minimum: -10.109 Q1: -6.631 Median: -5.5665 Q3: -4.34325 Mean: -5.670 Maximum: -2.384 Skewness: -0.455808 Sd: 1.786387  חריגים:  -10.109 | A diagram of a box plot  Description automatically generated | מרבית השירים נמצאים בטווח ביניים שבין 6.631- לבין 4.343-. השיר היחיד המופיע כחריג הוא פשוט שיר שקט ורך מאוד, הדבר נתמך על ידי משתנים אחרים כמו Energy ולכן בחרנו שלא להסיר אותו. |
| **Speechiness** | Minimum: 0.0232 Q1: 0.04545 Median: 0.07495 Mean:0.11558 Q3: 0.1345 Maximum: 0.53 Skewness: 2.018329 Sd: 0.1051044  חריגים  0.270 0.463 0.321 0.516 0.412 0.530 0.342 | A diagram of a box plot  Description automatically generated | 75% מהשירים נמצאים בטווח שבין 0.0232 לבין 0.1345. מה שמצביע על שירים עם מעט מלל. התצפיות החריגות הגיוניות אם לוקחים בחשבון ז'אנרים כמו ראפ אשר יותר מסביר שיופיע בשאזם. לא הסרנו. |
| **Acousticness** | Minimum:0.000282 Q1: 0.0423 Median: 0.1225 Mean: 0.199567 Q3: 0.25325 Maximum: 0.934 Skewness:1.549317 Sd: 0.221515  חריגים  0.580 0.581 0.883 0.652 0.627 0.934 0.779 0.847 0.697 | A diagram of a box plot  Description automatically generated | מרבית השירים בנתונים שלנו מסתמנים כשירים קצביים עם מעט מלל דבר שמתכתב עם אקוסטיות נמוכה אך ייתכן שיופיעו מספר שירים אקוסטיים ושקטים במאגר הנתונים. לא הסרנו חריגים. |
| **Liveness** | Minimum:0.0215 Q1: 0.0952 Median: 0.1185 Mean: 0.1596 Q3: 0.1722 Maximum:0.6360 Skewness:1.935124 Sd:0.1121782  חריגים  0.552  0.396  0.372  0.340  0.294  0.334  0.451  0.297  0.350  0.418  0.409  0.636  0.350  0.308  0.325 |  | במרבית השירים שלנו, ניתן לראות לפי הנתונים שהם עם נוכחות קהל מעוטה בהקלטות השיר. על כן, כשנמצאים שירים בהם הנוכחות קהל גבוהה יותר, הם נחשבים כ'תוצאות חריגות'. אולם, בחרנו שלא להסיר את הנתונים, הואיל והם מעניקים לנו מרחב מדגם מגוון יותר. כך, ההסקות שלנו עבור סוגי השירים השונים יהיה מהימן יותר וייצג לנו את התוצאות עבור משתנה הshazam בצורה מהימנה יותר. |
| **Valence** | Minimum: 0.0796 Q1: 0.351 Median: 0.4795 Mean :0.4904 Q3: 0.6465 Maximum: 0.931 Skewness: 0.0346301 Sd: 0.2035068 |  | אין חריגים |
| **Tempo** | Minimum: 64.934 Q1: 95.29625 Median: 118.0155 Mean :118.68 Q3: 136.7237 Maximum: 198.075 Skewness:0.5950216 Sd: 27.7093 | A graph with a line and a line  Description automatically generated with medium confidence | אין חריגים |
| **Duration** | Minimum: 95467 Q1: 184600 Median: 205047.5 Mean :205090 Q3: 221333.2 Maximum: 417920 Skewness:1.25841 Sd: 40265.2  חריגים 417920 119133 121887 312820 95467 | A graph with numbers and lines  Description automatically generated | מהנתונים ניתן לראות כי אורכם של רוב השירים בטווח של 3:04 - 3:41 דק'. התצפיות החריגות הקיצוניות ביותר הן שיר באורך 7 דק' ושיר באורך 1:35 דק', שני מקרים די שכיחים בז'אנרים מסויימים. החלטנו שלא להסיר את החריגים. |

**משתנים רציפים- פירוש תוצאות סטטיסטיות:**

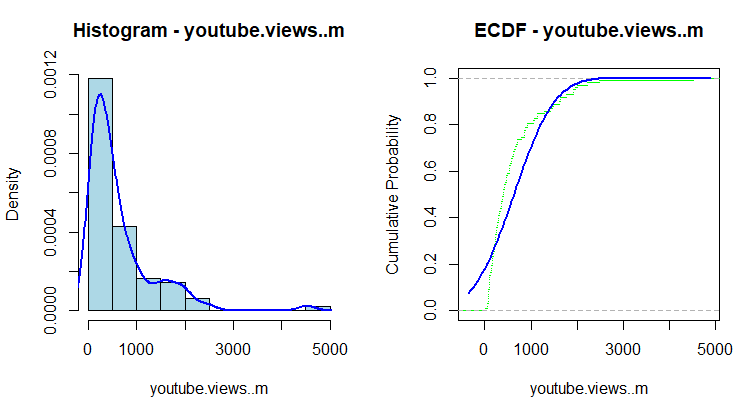
|  |  |
| --- | --- |
| **YouTube views** | 75% מהשירים נצפו בין 1.216 מיליון פעמים ל838 מיליון פעמים. ניתן לראות כי החציון אשר ערכו 394.202 מיליון נמוך משמעותית מהממוצע שערכו 632.813 מיליון. בנוסף לזאת התקבלה לנו אסימטריה חיובית מאוד (גדולה מ1) וסטיית תקן גדולה. מנתונים אלה ניתן לצפות לראות ריכוז גדול של תצפיות בטווח ערכים נמוך ומספר קטן של תצפיות גדולות משמעותית משאר הנתונים. כלומר, נקבל התפלגות בעלת זנב ימני. |
| **Valence** | 75% מהשירים תועדו כחיוביים בערך של בין 0.0796 לערך של 0.6465. ניתן לראות כי החציון אשר ערכו 0.4795 קרוב מאוד לממוצע שערכו 0.4904. בנוסף לזאת, התקבלה לנו אסימטריה חיובית אך בעלת ערך נמוך מאוד (0.034) וסטיית תקן בטווח ערכים הגיוני לערכי המדגם. מנתונים אלה ניתן לצפות לראות פיזור סימטרי של תצפיות לאורך כל טווח הערכים. כלומר, נקבל התפלגות הנראית כמו התפלגות סימטרית. |
| **Speechiness** | 75% מהשירים בעלי ערך המייצג כמות מילים מדוברות בשיר הנע בין 0.0232 לבין 0.1345.  ניתן לראות כי החציון אשר ערכו 0.07495 נמוך משמעותית מהממוצע שערכו 0.11558. בנוסף לזאת התקבלה לנו אסימטריה חיובית מאוד (גדולה מ1) וסטיית תקן גדולה. מנתונים אלה ניתן לצפות לראות ריכוז גדול של תצפיות בטווח ערכים נמוך ומספר קטן של תצפיות גדולות משמעותית משאר הנתונים. כלומר, נקבל התפלגות בעלת זנב ימני. |
| **Energy** | 75% מהשירים בעלי ערך המייצג את האנרגיה בשיר הנמצא בין 0.296 לבין 0.7715.  ניתן לראות כי החציון אשר ערכו 0.676 גבוהה במעט מהממוצע שערכו 0.6569. בנוסף לזאת התקבלה לנו אסימטריה שלילית (0.556-) וסטיית תקן הגיונית ביחס לערכי המדגם.  מנתונים אלה ניתן לצפות לראות פיזור סימטרי של תצפיות לאורך כל טווח הערכים כאשר יהיה ניתן לראות יותר תצפיות ברביע השלישי. כלומר, נקבל התפלגות הנראית כמו התפלגות סימטרית עם הטיה קלה מאוד מה שתראה כמו זנב שמאלי עבה. |

* **mode ~ shazam:**

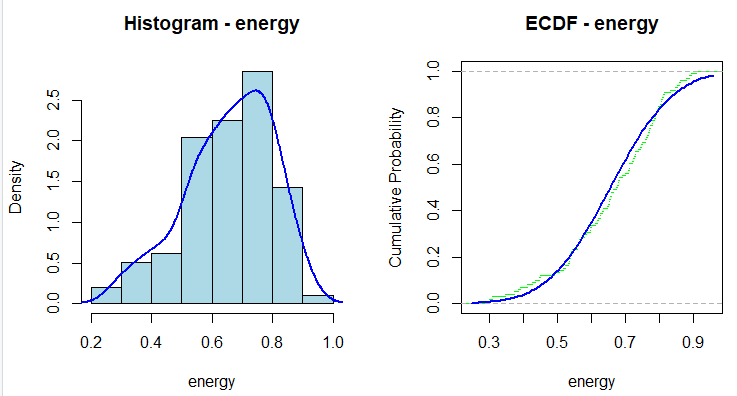
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Shazam (mode=0):** | **Shazam (mode=1):** | **Pie chart- shazam mean by mode** |
| **נתונים** | SD: 4.599655  Skewness: 1.072331  Minimum: 0.889427  Q1: 2.279572  Median: 4.616458  Mean: 5.969863  Q3: 8.688474  Maximum: 19.71998 | SD: 3.460371  Skewness: 1.701955  Minimum: 0.165207  Q1: 2.142121  Median: 4.415885  Mean: 4.823002  Q3: 6.454898  Maximum: 19.74907 |  |
| **הסבר** | 75% מהשירים שהם מינוריים בעלי ערך שאזם הנע בין 0.8894 לבין 6.4548. ניתן לראות כי החציון אשר ערכו 4.6165 נמוך משמעותית מהממוצע שערכו 5.9698. בנוסף לזאת התקבלה לנו אסימטריה חיובית (1.07) וסטיית תקן גדולה (4.5996). מנתונים אלה ניתן לצפות לראות ריכוז גדול של תצפיות בטווח ערכים נמוך עם פיזור די נרחב ומספר קטן של תצפיות גדולות משמעותית משאר הנתונים. כלומר, נקבל התפלגות בעלת זנב ימני. | 75% מהשירים שהם מז'וריים בעלי ערך שאזם הנע בין 0.1652 לבין 8.6884. ניתן לראות כי החציון אשר ערכו 4.4158 נמוך מהממוצע שערכו 4.823 אך בפער קטן ממה שנצפה עבור שירים מינוריים. בנוסף לזאת התקבלה לנו אסימטריה חיובית מאוד (1.7) וסטיית תקן גדולה אך פחות קיצונית מזו של השירים המינוריים. מנתונים אלה ניתן לצפות לראות ריכוז גדול של תצפיות בטווח ערכים נמוך עם פיזור קטן יותר מאשר זה שנראה עבור שירים מינוריים בלבד ומספר קטן של תצפיות גדולות משמעותית משאר הנתונים. כלומר, נקבל התפלגות בעלת זנב ימני. |

פונקציית צפיפות והתפלגות מצטברת:

* **:YouTube views**



* **פונקציית הצפיפות –** ההיסטוגרמה מראה לנו כי רוב השירים עם כמות צפיות נמוכה בצורה יחסית לטווח הקיים בנתונים (נמצאים בטווח שבין 0 ל-M1000) כאשר השיא הוא בערכים שנמצאים בין 0 ל500 מיליון צפיות. נתונים אלו יוצרים התפלגות א-סימטרית עם צידוד חיובי, משמע רק מעט מאוד שירים במאגר הנתונים עם כמות צפיות גבוהה (מעל M2000) ולכן נוצר הזנב הימני.
* **פונקציית ההתפלגות המצטברת –** ניתן לראות בגרף עליה תלולה בטווח ערכים נמוך, דבר המחזק את הטענה כי רוב השירים הקיימים בנתונים שלנו בעלי כמות צפיות נמוכה יחסית. ניתן לראות כיצד בהמשך העלייה הופכת למתונה מאוד עד אשר הגרף מגיע ל1, כלומר הערך אשר ההסתברות לקבל אותו או ערך קטן ממנו הוא 100%. העלייה המתונה באזור המייצג כמות צפיות גבוהה נובעת מכמות מועטה של שירים כאלה בנתונים של שאזם.
* **:Energy**

****

* **פונקציית צפיפות -** התפלגות רמות האנרגיה בין השירים היא אחידה במקצת כאשר השיא נמצא בערכי אנרגיה גבוהים יותר (בין 0.7 ל0.8). המשמעות של התפלגות זו היא שהשירים במערך הנתונים כוללים מגוון רחב, החל משירים עוצמתיים וקצביים ועד שירים רכים ואיטיים. עם זאת, על בסיס ההיסטוגרמה ניתן לראות שקיימת העדפה קלה לשירים קצביים ואנרגטיים יותר.
* **פונקציית התפלגות מצטברת -** ההתפלגות המצטברת של משתנה האנרגיה היא ליניארית יחסית, מה שמצביע על התפלגות אחידה של רמות האנרגיה בין השירים השונים בנתונים. לרוב השירים יש רמת אנרגיה המפוזרת באופן שווה בין נמוך לגבוה. על בסיס הגרף ניתן לראות כי רוב הנתונים נמצאים בטווח שבין 0.4 ל0.8 מכיוון ושם העלייה בפונקציית ההתפלגות המצטברת היא התלולה ביותר.
* **:Loudness**

A diagram and graph of a diagram

Description automatically generated with medium confidence

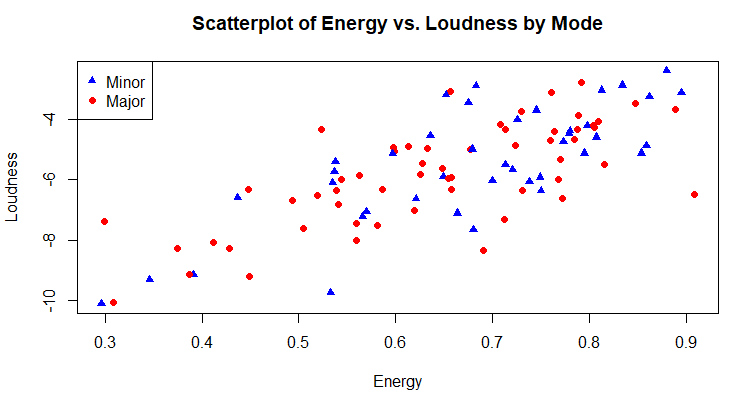
* **פונקציית צפיפות –** רמות עוצמת הקול של השירים נוטות להתקבץ סביב טווח מרכזי שבין 8- לבין 3- כאשר השיא נמצא בערכים שבין 4- ל3-, יש מעט שירים בעלי רמות רעש נמוכות או גבוהות במיוחד בנתונים. התפלגות זו מצביעה על העדפה לטווח עוצמה מסוים, לא רועש ולא שקט מידי אלא טווח ביניים יחסית מאוזן.
* **פונקציית התפלגות מצטברת -** ההתפלגות המצטברת של עוצמת הקול מראה עלייה חלקה ויציבה, המשקפת את האשכול המרכזי שנצפה בפונקציית הצפיפות. לרוב השירים יש את העוצמה שלהם בטווח צר יחסית, כאשר ערכים קיצוניים פחות נפוצים. נתונים אלה מתכתבים היטב עם הנתונים של המשתנה Energy שכן הם מצביעים ומחזקים את אותה המגמה. הרבה שירים מגוונים החל מקצביים ורועשים ועד רכים ושקטים ללא יותר מידי חריגות יוצאות דופן לאף כיוון.
* **Valence**

A graph of a diagram

Description automatically generated

* **פונקציית צפיפות –** רמת החיוביות של השיר מתפלגת בצורה די אחידה גם כן. ניתן לראות כי השיא מגיע עבור ערכים הנמצאים בטווח שבין 0.4 לבין 0.5 וכי המון שירים במאגר הנתונים נמצאים בטווח מצומצם זה. הפיזור רחב ואין זנבות לאף כיוון, דבר המצביע על כך שקיים מגוון רחב של שירים כאשר הרוב נע בטווח הביניים של חיובי עד שלילי אך ללא חריגות יוצאות דופן לשום כיוון.
* **פונקציית ההתפלגות המצטברת –** העלייה החלקה והיציבה בעקומה של פונקציית ההתפלגות המצטברת מצביעה על כך שהשירים מפוזרים באופן שווה על פני טווח הערכים שבין 0 ל-1. מכך נובע שהשירים במאגר הנתונים מגוונים מאוד וקיימים שירים עצובים כמו שקיימים שירים שמחים ומקפיצים. אין מגמה שולטת או טרנד מסוים אלא מגוון אפשרויות שונות.

הצגת תרשים פיזור מורכב:

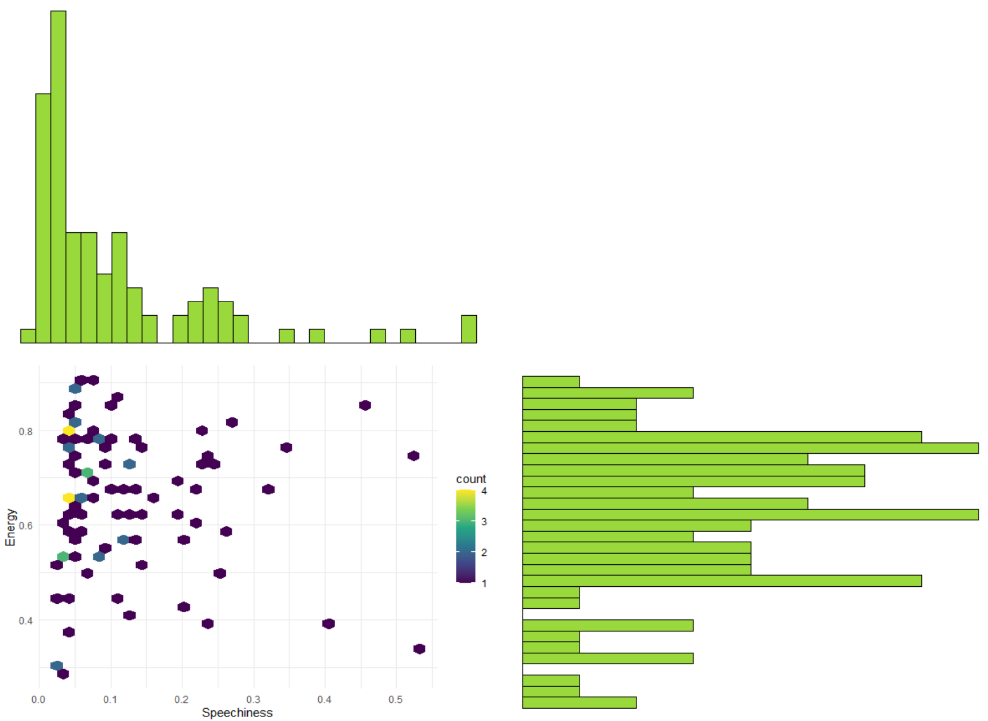
  
הגרף שלמעלה ממחיש את הקשר בין אנרגיה לעוצמת קול עבור שירים, כאשר קיימת הבחנה לפי המצב שלהם (מסווגים כמינור או מז'ור, עם 0 עבור מינור ו-1 עבור מז'ור במערך הנתונים).

* **תוצאות:  
  קשר בין אנרגיה לעוצמת קול -** נראה שיש מתאם חיובי בין אנרגיה לעוצמת קול על פני שירים. זה אומר שככל שרמת האנרגיה של שיר עולה, גם העוצמה שלו נוטה לעלות. קשר זה אינטואיטיבי מכיוון ששירים עם אנרגיה גבוהה יותר לרוב גם רועשים יותר.

**השפעת המצב על אנרגיה ועוצמת קול -** תרשים הפיזור מציג את התפלגות השירים במצבים מינוריים ומז'וריים על פני רמות אנרגיה ועוצמת קול שונות. שני המצבים מפוזרים בכל טווח הערכים של אנרגיה ועוצמה, מה שמצביע על כך שלמצב (מז'ור לעומת מינור) אין השפעה ברורה על רמות האנרגיה או העוצמה של השיר. היעדר הבחנה ברורה בין מצבים מינוריים למז'וריים במונחים של אנרגיה ועוצמה מעידה על כך שתכונות אלו מושפעות יותר מגורמים אחרים (כגון ז'אנר, קצב או נגינה) מאשר מהטונאליות (מז'ור או מינור) של השיר.

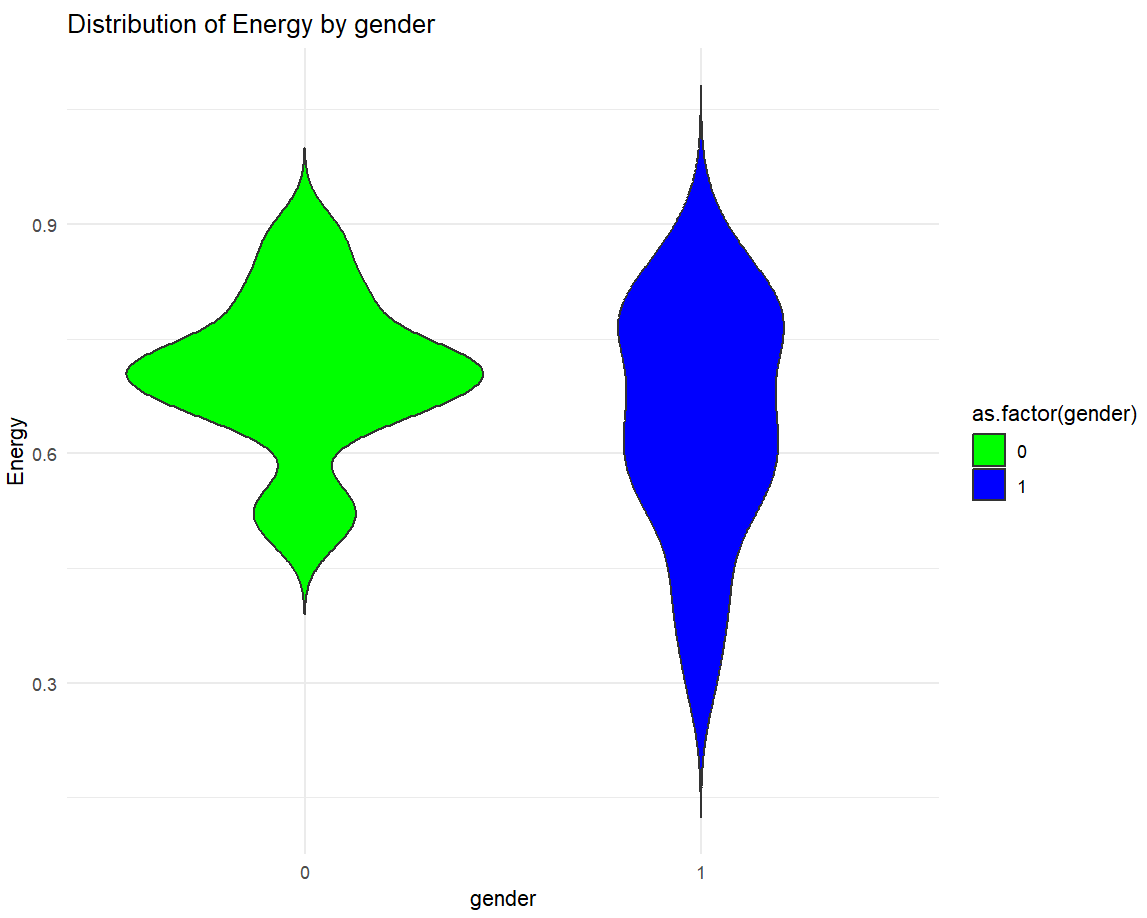
**צפיפות הפיזור -** צפיפות הנקודות מעידה על כך שרוב השירים מתקבצים בטווח בינוני של אנרגיה ועוצמה, עם זאת כן מסתמנת מגמה קלה לעבר שירים קצביים ועוצמתיים יותר שכן הצפיפות הגדולה ביותר של נקודות היא בטווח ערכים זה.   
  
  
  
  
  
  
  
  
  
ייצוג קשרים בעזרת תרשימים:

* **גרף 1:**

 **Speechiness VS Energy**

**סוג גרף –** plot Hexbin

**משמעות הגרף והקשר -** גרף זה מציג את צפיפות הנקודות (שירים) בצמתים של ערכי כמות מילים מדוברות בשיר ורמת האנרגיה. הגרף מראה באמצעות צפיפות גבוהה של נקודות כי רוב השירים מתקבצים בערכים נמוכים של כמות מילים מדוברות, מכך ניתן להסיק כי שירים אינסטרומנטליים או כאלה שמכילים פחות מלל שולטים במאגר הנתונים. קיים פיזור רחב של תצפיות עם רמות אנרגיה שונות עבור שירים בעלי כמות מועטה של מילים מדוברות, דבר המרמז כי שירים עם כמות נמוכה עד בינונית של מילים יכולים להשתנות בצורה משמעותית ברמות האנרגיה ולהיות קצביים ושקטים גם כן. קל לראות כי קיימים פחות שירים עם הרבה מילים מדוברות ואנרגיה גבוהה בבסיס הנתונים שלנו. היחס בין כמות המילים המדוברות בשיר והאנרגיה יכול לספק תובנות לגבי תכונות של שירים וסגנונם. לדוגמה, מוזיקה עם הרבה מילים מדוברות ואנרגיה גבוהה כמו ראפ לא קיימת במאגר הנתונים שלנו כמעט כלל. לעומת זאת, שירים קצביים עם מעט מאוד מלל כמו טראנס לדוגמא שולטים בבסיס הנתונים שלנו. ההבנה הזו יכולה להיות שימושית עבור חלוקת מוזיקה לקטגוריות, בניית מערכות המלצה ומאפשרת בדיקה של טרנדים בסגנון מוזיקלי ואילו ז'אנרים דומיננטיים באפליקציית שאזם.

* **גרף 2:**

|  |  |
| --- | --- |
| **מפתח צבעים** | |
|  |  |
|  |  |

**סוג גרף –** Violin Plot

**משמעות הגרף והקשר –** הגרף מציג את התפלגות הערכים בשירים בקטגוריית האנרגיה שלהם לפי שתי תתי הקטגוריות של :

זמרות:   
זמרים: התפלגות הערכים מוצגת באמצעות צורת הכינור, אחת עבור כל מגדר. בציר ה-X ניתן לראות את התפלגות הערכים לפי מגדר (ירוק- נקבה, כחול- זכר), וציר ה-Y מייצג את ערכי האנרגיה של השירים.

* **חלקים הרחבים של הגרף (Violins)-** לשתי הכינורות יש בסיס רחב, מה שמצביע על מספר ניכר של שירים עם רמות אנרגיה בינוניות לשני המינים. ניתן לומר כי רוב השירים, ללא קשר למין הקשור אליהם, נוטים להיות בעלי רמת אנרגיה מתונה.
* **בגרף הנשי (ירוק)-** שיא ​​הכינור נראה מעט נמוך יותר בהשוואה לכינור הגברי (כחול), מה שמרמז על כך שרמת האנרגיה הנפוצה ביותר לשירי נשים עשויה להיות מעט נמוכה מזו של שירים גברים.
* **בגרף הגברי (כחול)-** הכינור הזכרי מעט יותר מוארך בחלק העליון, מה שמעיד על כך שיש מגוון רחב יותר של רמות אנרגיה עם פחות שירים בעלי אנרגיה גבוהה מאוד.
* **זנבות-** שני הכינורות מתחדדים לכיוון רמות האנרגיה הגבוהות יותר, אך עבור הקטגוריה הנשית, ניתן לראות הפחתה בולטת יותר, מה שמרמז על כך שישנם פחות שירים נשיים עם רמות אנרגיה גבוהות מאוד בהשוואה לזכרים.
* **ההתפלגות** מעידה על כך שקיימת שונות מסוימת ברמת האנרגיה של שירים הקשורים למגדר זכר ונקבה, כאשר שירים זכרים מראים נטייה מעט גבוהה יותר לשונות אנרגטית גדולה יותר.

טבלאות שכיחות: **טבלת שכיחות חד-ממדית עבור משתנה Mode –**

|  |  |  |
| --- | --- | --- |
| **ערך** | **משמעות** | **שכיחות** |
| 0 | מינור | 41 |
| 1 | מז'ור | 57 |

הטבלה מציגה את החלוקה של שירים בין מינור למז'ור, ניתן לראות כי בנתונים שלנו מופיעים יותר שירים מז'וריים מה שיכול להצביע על העדפה מסוימת לשירים עם טונאליות מסוג זה. **טבלת שכיחות חד-ממדית עבור משתנה Gender –**

|  |  |  |
| --- | --- | --- |
| **ערך** | **משמעות** | **שכיחות** |
| 0 | נקבה | 20 |
| 1 | זכר | 78 |

הטבלה מציגה את החלוקה של אמנים בין זכר לנקבה, הטבלה מראה את השכיחות של מין האמנים אך קשה מאוד להסיק מטבלה זו הנחה קונקרטית מכיוון ואני מאמינים שנתונים אלה מאוד תלויי מדגם.

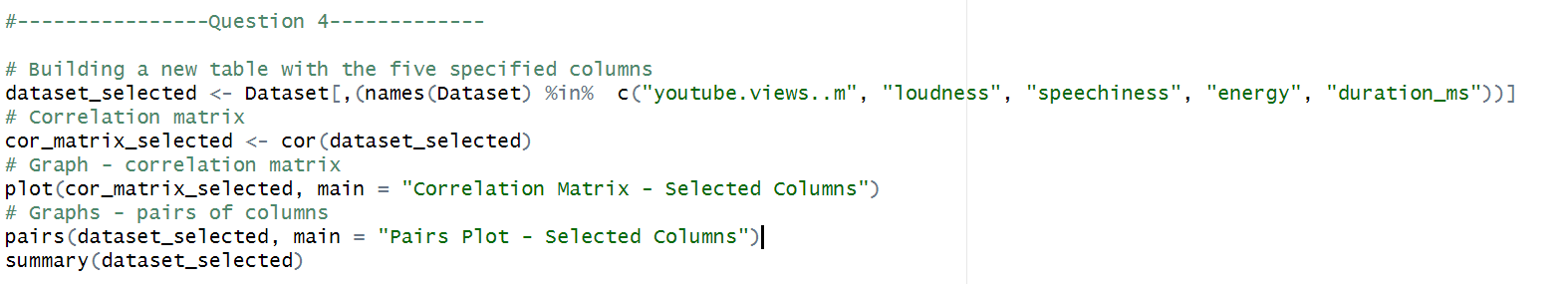
**טבלת שכיחות דו-ממדית עבור המשתנים Mode, Gender –**

|  |  |  |
| --- | --- | --- |
|  | **מינור (0)** | **מז'ור (1)** |
| **נקבה (0)** | 11 | 9 |
| **זכר (1)** | 30 | 48 |

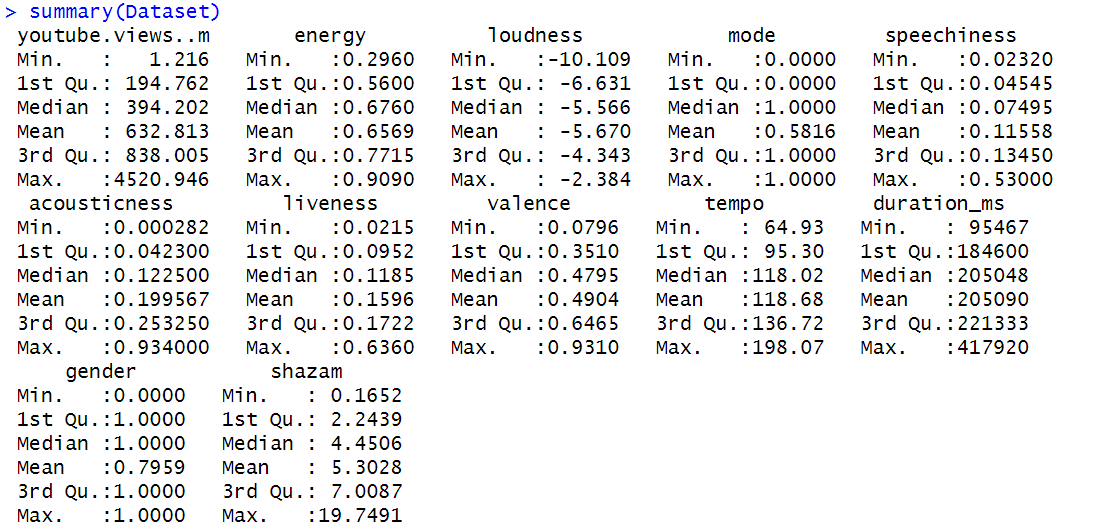
טבלה דו-ממדית מאפשרת לנו לבחון את הקשר שבין mode ל-gender. ניתן לראות כיצד מתפלג משתנה אחד בהינתן ערך של המשתנה השני, לדוגמא רוב האומנים בנתונים שלנו הם זכרים ומתוך האומנים הזכרים ניתן לראות שקיימת עדיפות ברורה לשירים המקוטלגים כמז'וריים. לעומת זאת, עבור האומניות ניתן לראות שהחלוקה בין מינור למז'ור יותר שוויונית ולא ניתן להסיק על העדפה כזאת או אחרת מכיוון ויש יחסית מעט תצפיות בנתונים.

# **נספחים**

* **נספח שאלה 4:**



* **נספח שאלה 5+6:**
* קוד לסיכום נתונים:  
  
* פלט סיכום נתונים עבור כל המשתנים:

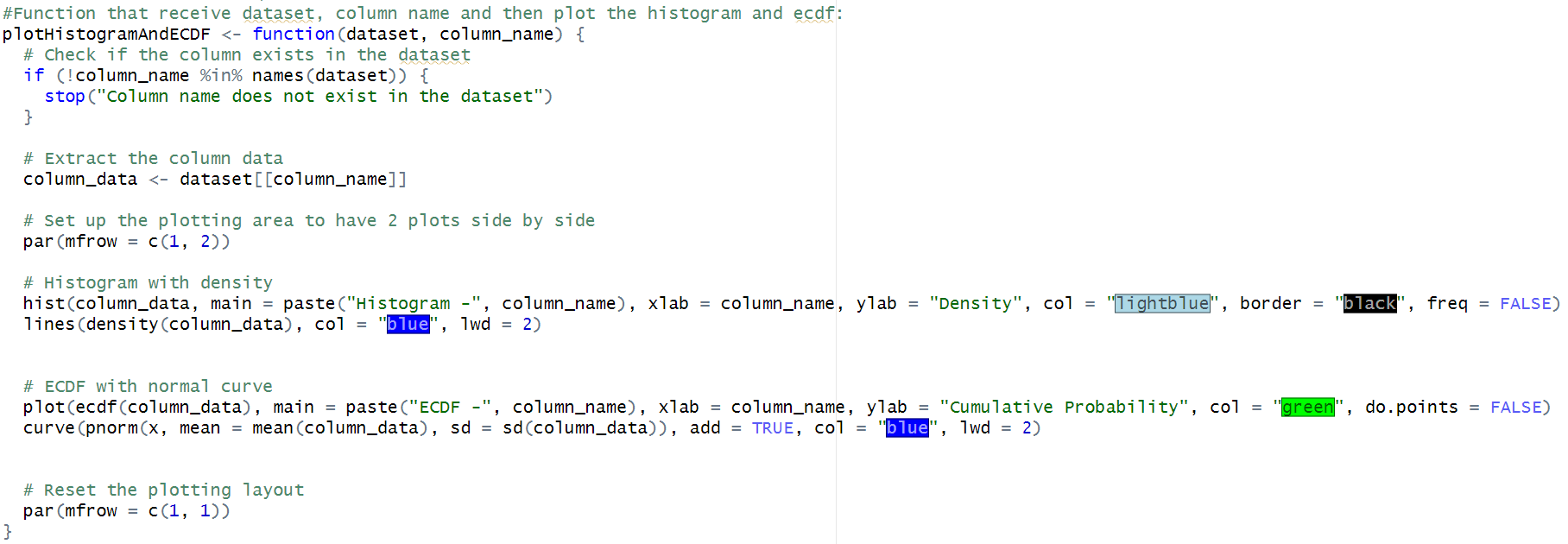


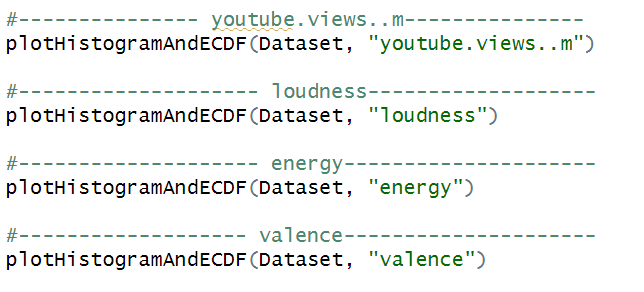
* התפלגות משתנה מוסבר לפי minor & major:

|  |  |  |
| --- | --- | --- |
| **mode ~ shazam** | | |
|  |  |  |
|  |  |  |

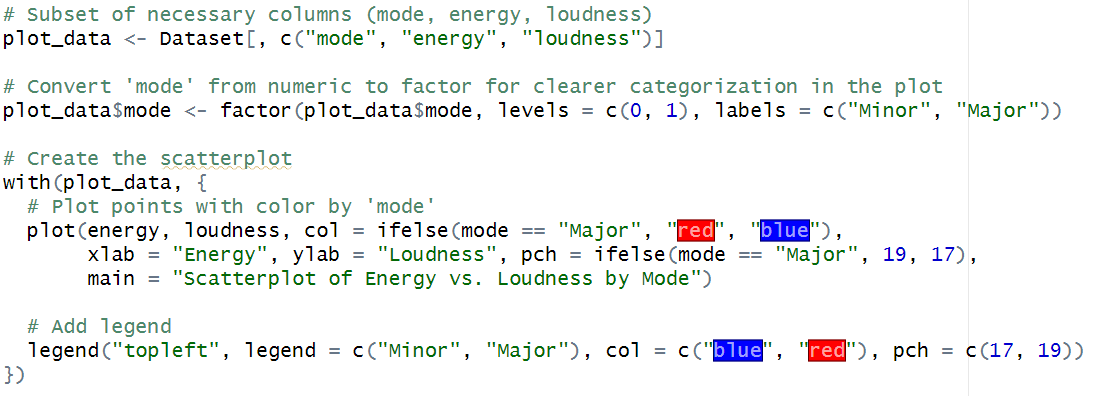
* ניתוח נתונים עבור משתנים רציפים:

|  |  |  |
| --- | --- | --- |
| **YouTube views** | **Energy** | **Loudness** |
|  |  |  |
| **Speechiness** | **Acousticness** | **Liveness** |
|  |  |  |
| **Tempo** | **duration** | **Valence** |
|  |  |  |
|

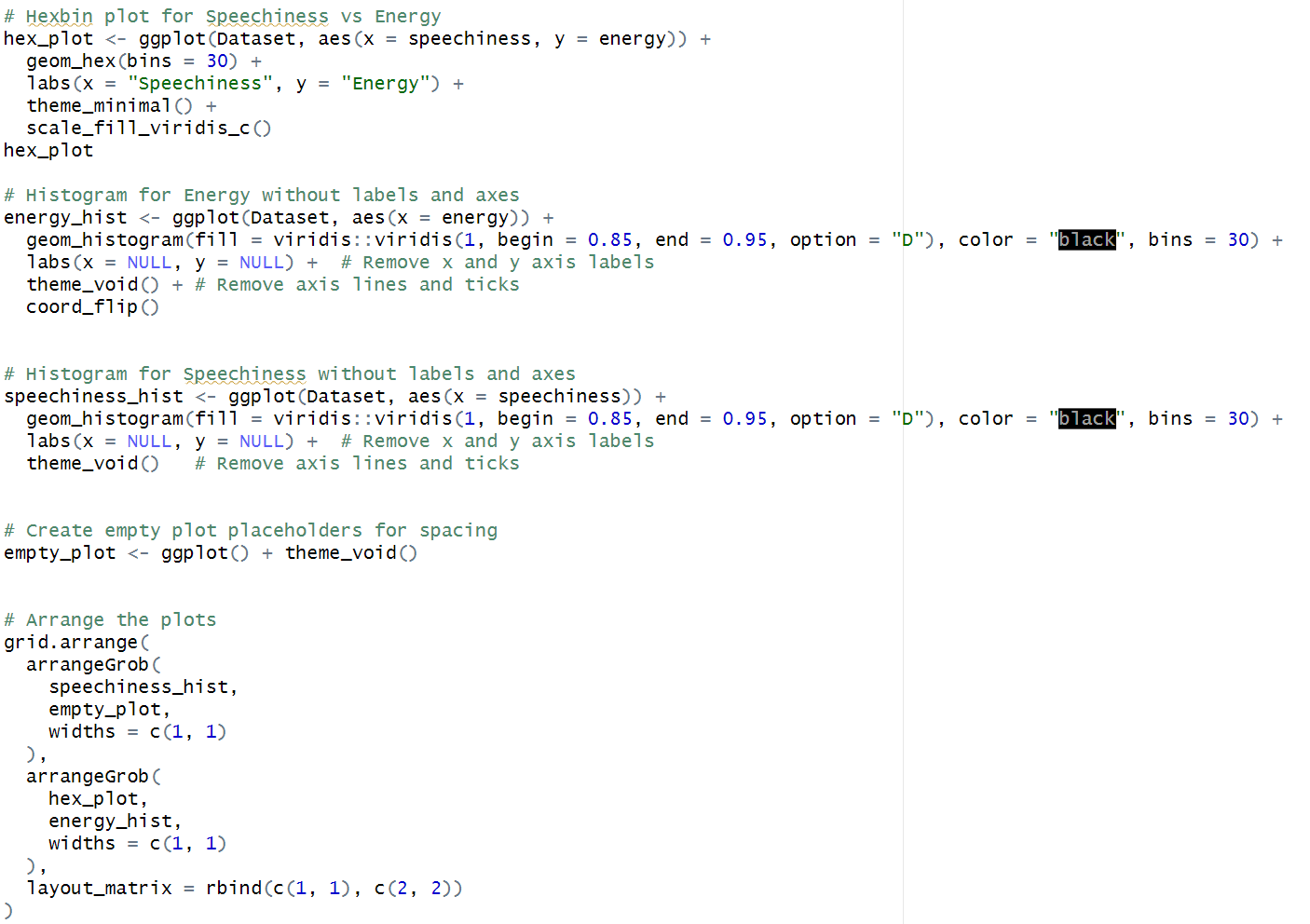
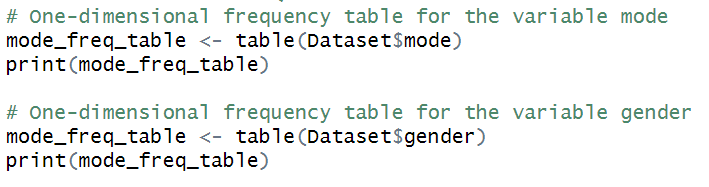
* **נספח שאלה 7:**
* ****קוד ליצירת הפונקציה:
* מימוש הפונקציה עבור ארבעה משתנים מסבירים:

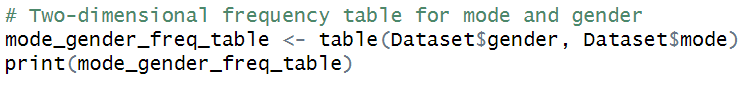
****

* **נספח שאלה 8:**
* תרשים פיזור מורכב: Energy & Loudness by Mode

****

* **נספח שאלה 9:**
* A computer screen shot of a blue and black text

  Description automatically generatedייצוג קשרים בעזרת תרשים: Speechiness VS Energy:  
  - גרף :  
    
  - גרף :
* **נספח שאלה 10:**
* טבלאות שכיחות:  
  - חד מימדית:  
  

- דו מימדית:  


1. <https://en.wikipedia.org/wiki/Tempo>  
    [↑](#footnote-ref-1)