**חלק ב' – ניתוח פורמאלי של מאגר הנתונים (70%)**

1. תקציר מנהלים:

יש לכתוב תקציר מנהלים המסכם את חלק ב' של הפרויקט (מטרות, צעדים שננקטו, מסקנות ותוצאות) (10 נק'). יש לצרף לסעיף זה את טבלת המשתנים מחלק א' סעיף 2.

לכתוב תקציר מנהלים.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 |

1. עיבוד מקדים:
   1. הסרה של משתנים:

בשלב זה, עבור משתנים שלדעתכם לא קשורים למשתנה המוסבר, או לא מוסיפים עליו מידע שיקלו האם יש להסירם מהמודל ופרטו מדוע אתם חושבים כך. התדיינו על שני משתנים לפחות (10 נק').

בשלב זה, נבחן בנפרד את המשתנים המסבירים הרציפים (לא קטגוריאליים) ואת הקטגוריאליים, ואת הכדאיות בהסרתם מהמודל באמצעות בחינת מקדם המתאם פירסון- מקדם אשר מייצג מדד למתאם לינארי בין 2 משתנים. נוסף על כך, נבחן את הכדאיות בהסרת המתנים בעזרת מבחן P- Value.

|  |  |  |
| --- | --- | --- |
| **משתנה** | **P- Value** | **מקדם מתאם** |
| YouTube views | 0.000533 | 0.3409519 |
| Energy | 0.782288 | 0.05200517 |
| Loudness | 0.134383 | 0.1881835 |
| Speechiness | 0.170348 | -0.1797827 |
| Acousticness | 0.271279 | -0.0309277 |
| Liveness | 0.392333 | -0.01351359 |
| Valence | 0.000574 | -0.163126 |
| Tempo | 0.187857 | -0.1187576 |
| Duration | 0.963993 | 0.1995106 |

בהתאם לנתוני הטבלה, ניתן לראות כי בחרנו נוציא את המשנים- Duration ו- Energy כיוון שהם בעלי מקדם המתאם נמוך - כלומר, הקורלציה הלינארית ביניהם נמוכה מאוד, ובעלי P-Value גבוה מה שמעיד על כך שמשתנים אלו אינם מתאימים למודל.

בנוסף, בסעיף הבא נבחן שינויים עבור המשתנה Liveness הואיל ומקדם המתאם שלו נמוך מאוד ביחס לשאר מקדמי המתאם, וערך ה P- Valueשלו גבוה מאוד, מה שמעיד על חוסר התאמתו למודל.

עבור המשתנים הקטגוריאליים:

|  |  |
| --- | --- |
| **משתנה** | **P- Value** |
| Mode | 0.058689 |
| Gender | 0.259031 |

* 1. התאמת משתנים:

בדקו והחליטו האם יש צורך בהגדרה מחדש של משתנים או בהתאמתם לבסיס הנתונים שלכם. עבור משתנה קטגוריאלי, לעיתים יש צורך לבצע איחוד קטגוריות. עבור משתנה רציף לעיתים יש צורך בהפיכתו למשתנה קטגוריאלי (דיסקרטיזציה), או למשל הגדרת משתנה חדש שהוא הפרש או סכום בין משתנים קיימים. חשבו על 3 התאמות לפחות והחליטו אם יש לבצע אותן. (15 נק')

**דיסקרטיזציה עבור המשתנה Liveness :**

|  |  |  |
| --- | --- | --- |
|  | **לפני הדיסקרטיזציה** | **אחרי הדיסקרטיזציה** |
| **יחידת מידה** |  |  |
| **סוג משתנה** | רציף | קטגוריאלי |
| **מקדם מתאם** | -0.01351359 | -0.06312193 |
| **P- Value** | 0.392333 | 0.128836 |
| **גרף** |  |  |

חילקנו את המשתנה ל-3 קטגוריות שיעזרו לנו לאמוד מהי רמת הנוכחות של הקהל בהקלטת השיר.   
בחרנו לחלק את הקטגוריות כך ש:

ניתן לראות כי ערכו של מקדם המתאם בערך מוחלט גדל (כלומר, "התרחק מהאפס"). על כך, אפשר לומר כי הקשר הלינארי בין 2 המשתנים גדל, והשיפוע השלילי בין שני המשתנים נהפך לתלול יותר- ככל שרמת ה Liveness עולה, כך יורדים החיפושים עבור שיר מסוים בshazam.

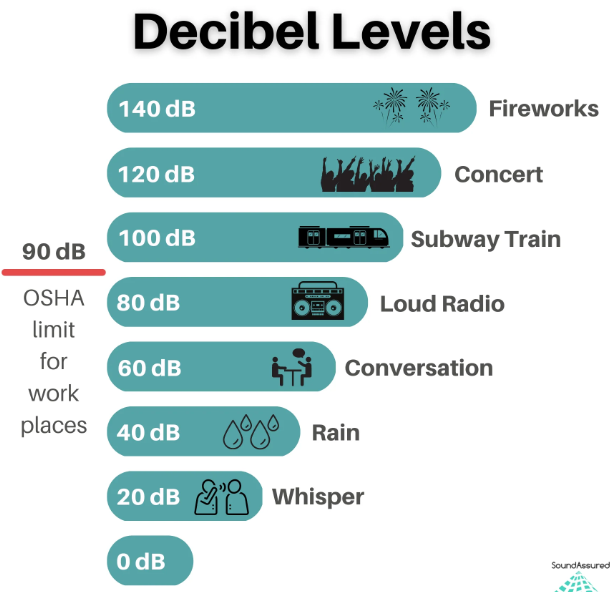
בנוסף, ערך ה P- Value של המשתנה קטן, מכך נובע שהמשתנה מתאים יותר למודל.

**שינוי ערכים עבור המשתנה Loudness:**

החלטנו לבצע התאמה למודל שלנו על המשתנה Loudness. לפני השינוי, ערכי Loudness הוצגו כמספר שלילי הנע בין -10.109 ל -2.384.   
ואולם, המספרים הללו אינם מתיישבים עם ההגיון. הרי, מהמידע שקיבלנו אודות הנתונים, ניתן לראות כי יחידות המידה של המשתנה Loudness הינם בדציבלים. טווח השמיעה של האדם הינו החל מ-0 דציבל ועד ל-140 דציבל (סף הכאב).

על כן, אין זה הגיוני שהמספרים שהוצגו יהיו בדציבלים ולכן בחרנו לבצע שינויים במשתנה בכדי להפוך את הנתונים להגיוניים יותר על ידי הכפלה של כל הערכים במינוס 10. בעקבות ההכפלה, טווח הנתונים של המשתנה תואמים למציאות.

לצורך הבנת המשמעות של מספרי הדציבלים אנו מצרפים תמונה שתמחיש:



|  |  |  |
| --- | --- | --- |
|  | **לפני ההכפלה** | **אחרי ההכפלה** |
| **מקדם מתאם** | 0.1881835 | -0.1881834984 |
| **P-value** | 0.134383 | 0.059437 |

בחרנו לשנות את ערכי המשתנה, הואיל והשינוי מטיב עם התאמת המשתנה למודל. ניתן לראות כי ערך ה-P-value קטן בעקבות השינוי. נוסף על כך, גודלו של מקדם המתאם לא השתנה (בערכו המוחלט) ועל כן, השינוי לא פגם בקורולציה.

לסיכום, ניתן לומר כי השינוי מטיב הן עם ההגיון מבחינת משמעות הערכים לפי יחידות מידתם והן עם התאמת המשנה למודל ועל כן, נשנה את ערכי המשתנה.

**שינויים במשתנים הקטגוריאלים:**

בחרנו שלא לאחד קטגוריות עבור המשתנים הקטגוריאלים שלנו, הואיל והקטגוריות בהם הם כבר תת קטגוריה בעולמות ה- Mode וה- Gender.   
לדוגמא- עבור Mode- לכל אחד מהסולמות יש תווים (לפי הדו- רה- מי)-   
כגון: דו- מג'ור/ רה מינור וכו'..   
מהנתונים הבינארים שיש בידינו, לא נוכל להסיק על התווים של כל משתנה.

לדוגמא- עבור Gender – בעולמינו, ישנם סוגי מגדרים שונים ואנשים המעוניינים שיגדירו אותם בצורה מסויימת.   
נוסף על כך, לפי הנתונים, ניתן לומר כי בכל השירים יש או זמר או זמרת- כי אין שיר עם שני המינים או עם כמה זמרים.

אם היה בידינו מספיק נתונים יכולנו לחלק את הקטגוריות לעוד תתי קטגוריות או לאחד קטגוריות אבל הקטגוריות המשתנים מראש מצומצמות ופרט ל2 תתי הקטגוריות עבור כל אחד מהם, לא נוכל להסיק מהם מידע נוסף/ חדש.

לרשום עוד שינוי ולהסביר למה לא עשינו

* 1. הגדרת משתנה דמה:

בחרו משתנה קטגוריאלי אחד והגדירו בדו"ח את משתני הדמה שנגזרים ממנו. (5 נק').

לתמלל טיפה (?) מרגיש לי ריק אבל לא יודעת מה יש להוסיף פה

|  |  |
| --- | --- |
| **משתנה קטגוריאלי** | **יחידת מידה** |
|  |  |
| Liveness |  |
| Mode |  |

* 1. הגדירו והוסיפו משתני אינטראקציה:

בחרו מדגם של משתני אינטראקציה מסוימים (לפחות שלושה) כאשר עליכם להסביר האם ומדוע, לפי השערתכם, משתנים אלו יכולים לתרום לחיזוי המשתנה המוסבר. הגדירו את משתני האינטראקציה כפי שלמדתם, גם בקוד העבודה וגם בדוח. מצופה להראות תהליך של חשיבה והבנה מדוע יש לאינטראקציה סיבה להוות השפעה ומדוע לא.

לבסוף, התאימו את המודל שמכיל את המשתנים המעובדים, כולל את משתני הדמה והאינטראקציה שהגדרתם. הציגו את מדדי הביצוע שלו. (20 נק').

|  |  |  |
| --- | --- | --- |
| **By Gender categories** | | |
| **Shazam VS Speechness** | **Shazam VS Valence** | **Shazam VS Tempo** |
| A graph with green and red dots  Description automatically generated |  |  |
| לא משאירים משתנה אינטרקציה.  משאירים דמה | משאירים שניהם | משאירים שניהם |

משאירים- שיפועים שונים, חיתוכים שונים. יש בטמפו שיפוע חיובי ושיפוע שלילי- יש קורולציה שלילית.

|  |  |  |
| --- | --- | --- |
| **By Liveness categories** | | |
| **Shazam VS Speechness** | **Shazam VS Valence** | **Shazam VS Tempo** |
| A graph with red and blue dots  Description automatically generated |  |  |
| משאירים- נאחד קטגוריות ב0.5 ו-1 כי יש להם את אותו שיפוע.  ה-p- value נשאר 0.128836 |  | משאירים- נאחד קטגוריות ב0.5 ו-1 כי יש להם את אותו שיפוע |
|  |  |  |
| משאירים שניהם | לא משאירים משתנה אינטרקציה.  משאירים דמה | לא משאירים משתנה אינטרקציה.  משאירים דמה |

משאירים- שיפועים שונים, חותכים שונים. תורם למודל

|  |  |  |
| --- | --- | --- |
| **By mode** | | |
| **Shazam VS Speechness** | **Shazam VS Valence** | **Shazam VS Tempo** |
|  |  | A graph with red and blue dots  Description automatically generated |
| משאירים | משאירים אינטרקציה.  לא משאירים דמה- כמעט נחתכים בx=0. | לא משאירים משתנה אינטרקציה- - אותו שיפוע  משאירים דמה |

משאירים- שיפועים שונים, חותכים שונים. תורם למודל

**להסביר על כל אחד מהגרפים פה ולמה בחרנו להשאיר/להוריד דמה וגם להשאיר/להוריד משתנה אינטרקציה.**

1. התאמת המודל ובדיקת הנחות המודל:
   1. בחירת משתני המודל: (15 נק')

בשלב זה בחרו קריטריון להשוואה בין מודלים (AIC, BIC, R2adj), והריצו שני אלגוריתמים לבחירת משתנים המתבססים על אותו קריטריון. האם שני האלגוריתמים הביאו לאותו המודל? דונו בתוצאות.

\*\* כיוון שסעיף 4 ירד אנו ממליצים לכם לנסות להחזיר משתנים שהורדתם בשלב מוקדם יותר של הפרויקט ולבדוק האם כבר עכשיו החזרתם תשפר את תוצאות ביצועי המודל.

**המודל המלא:**

|  |  |  |
| --- | --- | --- |
|  | **AIC** | **BIC** |
| רגרסיה לפנים | 530.3979 | 545.9077 |
| רגרסיה לאחור | 530.3979 | 545.9077 |
| רגרסיה בצעדים | 530.3979 | 545.9077 |

לתמלל ולהסביר מה עשינו. להסביר על השימוש בשיטה לפי כמו בתרגול 8- השתמשנו ב- **Stepwise Regression**

**המודל הסופי:**

* 1. בדיקת ויזואלית של הנחות המודל: (15 נק')

צרו תרשים שגיאות מנורמלות מול תחזיות כפי שלמדתם בתרגול. בצעו בחינה ויזואלית של השערות הלינאריות ושוויון השונויות והסבירו האם לדעתכם מתקיימות ולמה. צרו תרשים QQplot ובחנו את השערת הנורמאליות.

להוסיף עוד מלל וחירטוטים למה הסקנו את מה שהסקנו.

|  |  |  |
| --- | --- | --- |
| **הנחת הלינאריות** | **תרשים פיזור** | |
| ניתן לראות כי קיימת לינאריות | להוסיף תרשים פיזור- כמו בתרגול 10 | |
| **הרצת רגרסיה לינארית** | |
| לעשות לו summary- כמו בתרגול 10 | |
| **הנחת שוויון שונויות** | **תרשים פיזור של השגיאה המתוקננת מול** | |
| לא ניתן להסיק מהגרף לגבי הנחת שיוויון השונויות. |  | |
| **הנחת הנורמליות של השגיאות** | **QQplot** | **היסטוגרמה** |
| השגיאות מתנהגות בצורה נורמלית |  |  |
| ניתן לראות כי הנתונים באים מהתפלגות נורמלית.  ישנו קו לינארי ישר.  כשאנו בוחנים לינאריות ב QQplot ההסתכלות היא  בעיקר על החלק האמצעי של הגרף, לכן, הגיוני  שבקצוות עשויות להיות סטיות. | יש בגרף ההיסטוגרמה צורה של פעמון-  מעידה על התפלגות נורמלית.  צריך לסדר את הקו מגמה |

* 1. בדיקת פורמלית של הנחות המודל: (10 נק')

בצעו מבחני השערות (עליהם דיברנו בתרגול) לבדיקת הנחת הנורמאליות והנחת הלינאריות. דונו בתוצאות.

|  |  |
| --- | --- |
| **KS- Test** | **Shapiro- Test** |
|  |  |
| ערך הP-value של המבחן KS גדול מ0.05, כלומר, נקבל את השערת האפס ברמת ביטחון של 95%.  ערך הP-value של המבחן Shapiro קטן מ0.05, כלומר, לא נקבל את השערת האפס ברמת ביטחון  של95%. | |

להוסיף עוד חירטוטים ולחפור