פרויקט מדעי הנתונים

HIT FACTOR

ליאור בצלאל ועמית פומפס

שאלת המחקר
מה הם מאפייניו של להיט?

הקדמה
למה הנושא מעניין?

למה הנושא מעניין?

- הכנסות: להיטים מושמעים ונצפים על ידי עשרות מיליונים בכל ברחבי העולם. בכך, יוצר השיר יכול להרוויח הכנסות רבות כגון מכירות תקליטים, מהשמעות ברדיו ובאפליקציות , הופעות וכדומה.
- הגשמה מקצועית: הגעה למעמד של כותב להיט מביאה את כותבו לידי מימוש עצמי.

ידוע שאין תבנית מסויימת לכתוב להיט,אך מחקר על הידע הקיים בתחום וניתוחו, יכול להתוות כיוון כללי

מקורות הנתונים והרכשה

Billboard HOT 100

- ארגון תקשורת אמריקאי המתמחה בכתיבת ופרסום חדשות, עדכוני מוזיקה ודירוגים של הקטעים המוזיקליים הפופולריים ביותר.
החברה הוקמה בשנת 1894 ומתמחה בפרסום מידע על מכירות אלבומים סינגלים הופעות והאזנות למוזיקה דרך מגוון מקורות

בילבורד ידועה בפרסום דירוגי המכירות התקופתיים של אלבומים וסינגלים והם נחשבים לסטנדרט בתעשיית המוזיקה בארצות הברית.

Selenium WebDriver

Spotify®

- פלטפורמה להאזנה למוזיקה בסטרימינג, המאפשרת למשתמשים להאזין לשירים, אלבומים, פלייליסטים, פודקאסטים ועוד מכל סוגי המוזיקה ברשת. זהו ארגון גלובלי המספק שירותי סטרימינג מוזיקה באינטרנט.
- נוסדו בשנת 2006 והם פועלים ביותר מ-90 מדינות ברחבי העולם

API

Spotify for Developers

popularity integer

The popularity of the track. The value will be between 0 and 100, with 100 being the most popular.
The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are.

danceability number<float>

Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

duration_ms integer

The duration of the track in milliseconds.

energy number<float>

Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.

ממשק לקבלת אינפורמציה משרתי ספוטיפי באמצעות שאילתות לAPI של ספוטיפי.

עיקר שימושינו בממשק:

- מתן קונטקסט מוסקלי לשירים שאת שמותיהם קיבלנו מהרשימות של בילבורד.
- הוספת שירים שאינם נכנסו לרשימות של בילבורד ומאפייניהם.

ניקוי נתונים

ניקוי נתונים

ניקוי IQR

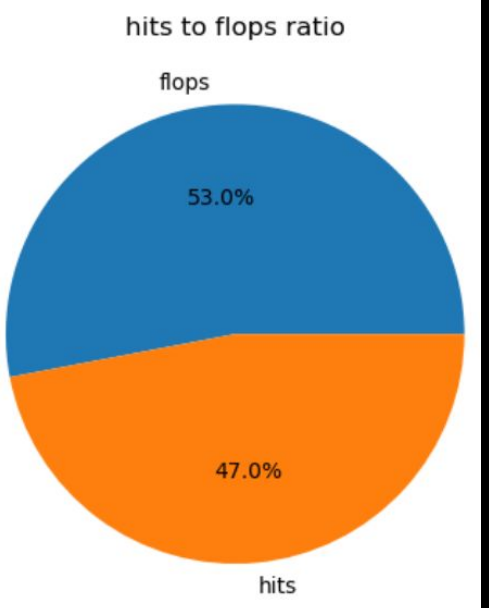
Data cleaning

```
all_outlier_rows = []
for col in num_vars:
        Q1 = np.percentile(data[col], 25)
        Q3 = np.percentile(data[col], 75)
        IQR = Q3 - Q1
        IQR_range = 1.5 * IQR
        col_outlier = data[(data[col] < Q1 - IQR_range) |
                           (data[col] > Q3 + IQR_range )].index
        all_outlier_rows.extend(col_outlier)
```

ויזואליזציה EDA
לאחר ניקוי IQR

ויזואליזציה ו-EDA

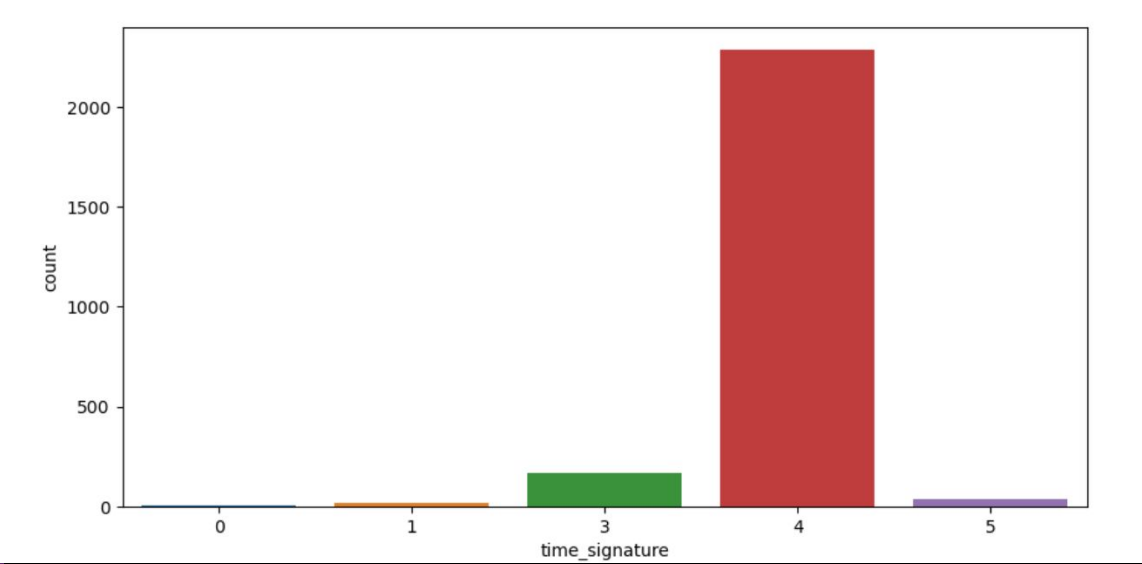
טבלת חישוב קורלציות

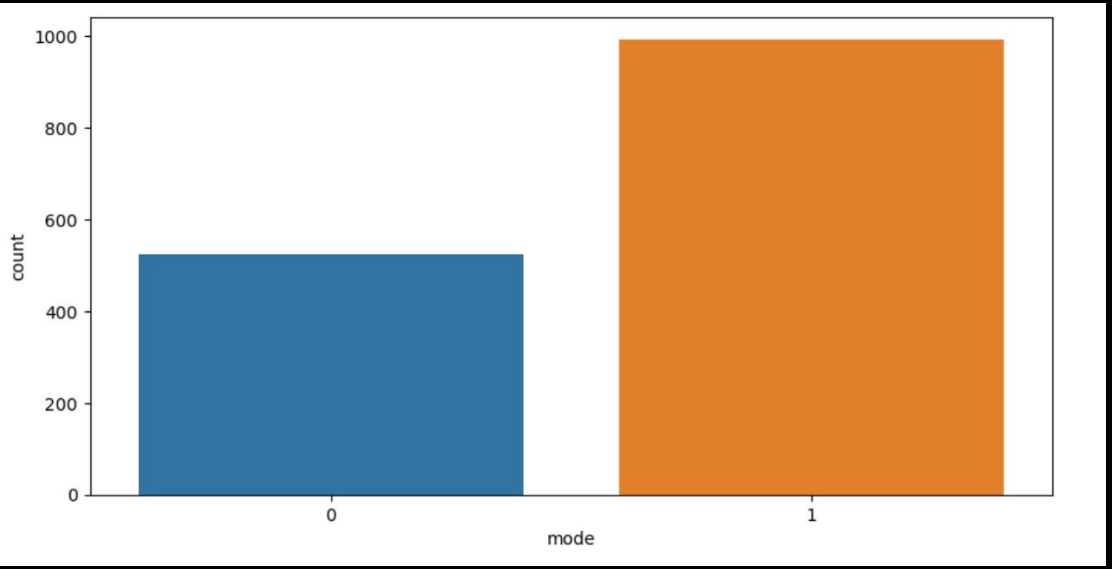


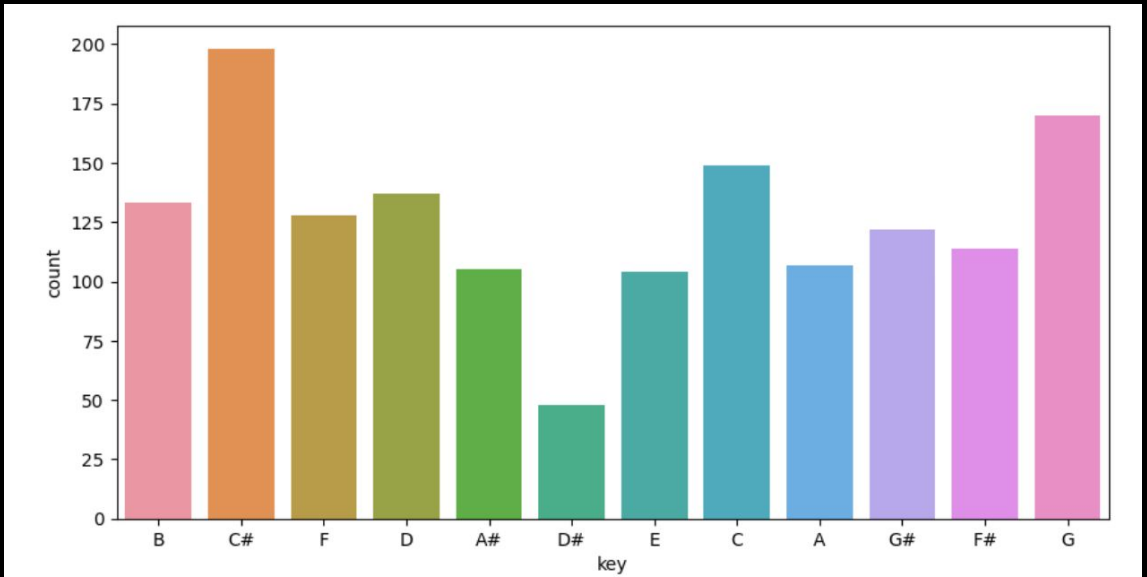
	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature	tracks_popularity	entered_biilboard
danceability	1.00	0.06	-0.02	0.07	-0.16	0.32	-0.22	-0.00	-0.02	0.28	-0.11	-0.20	0.13	0.02	0.05
energy	0.06	1.00	0.03	0.73	-0.02	0.03	-0.57	-0.02	0.07	0.44	0.11	0.02	0.18	-0.12	0.15
key	-0.02	0.03	1.00	0.03	-0.09	0.04	-0.00	0.01	0.00	0.03	0.02	-0.03	-0.01	-0.01	-0.02
loudness	0.07	0.73	0.03	1.00	0.01	-0.07	-0.44	-0.09	0.04	0.35	0.07	0.03	0.15	-0.07	0.16
mode	-0.16	-0.02	-0.09	0.01	1.00	-0.13	0.10	0.01	0.02	0.01	-0.02	0.05	0.01	-0.08	0.00
speechiness	0.32	0.03	0.04	-0.07	-0.13	1.00	-0.11	-0.08	0.04	-0.00	0.07	-0.12	0.06	-0.01	-0.07
acousticness	-0.22	-0.57	-0.00	-0.44	0.10	-0.11	1.00	0.01	-0.08	-0.13	-0.10	-0.04	-0.19	0.12	-0.16
instrumentalness	-0.00	-0.02	0.01	-0.09	0.01	-0.08	0.01	1.00	-0.08	-0.04	0.00	0.04	0.02	0.09	-0.08
liveness	-0.02	0.07	0.00	0.04	0.02	0.04	-0.08	-0.08	1.00	0.03	-0.01	-0.01	-0.04	-0.02	-0.01
valence	0.28	0.44	0.03	0.35	0.01	-0.00	-0.13	-0.04	0.03	1.00	0.02	-0.17	0.04	-0.04	0.09
tempo	-0.11	0.11	0.02	0.07	-0.02	0.07	-0.10	0.00	-0.01	0.02	1.00	-0.01	-0.07	0.01	-0.02
duration_ms	-0.20	0.02	-0.03	0.03	0.05	-0.12	-0.04	0.04	-0.01	-0.17	-0.01	1.00	-0.01	-0.10	0.19
time_signature	0.13	0.18	-0.01	0.15	0.01	0.06	-0.19	0.02	-0.04	0.04	-0.07	-0.01	1.00	-0.10	0.11
tracks_popularity	0.02	-0.12	-0.01	-0.07	-0.08	-0.01	0.12	0.09	-0.02	-0.04	0.01	-0.10	-0.10	1.00	-0.27
entered_biilboard	0.05	0.15	-0.02	0.16	0.00	-0.07	-0.16	-0.08	-0.01	0.09	-0.02	0.19	0.11	-0.27	1.00

1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

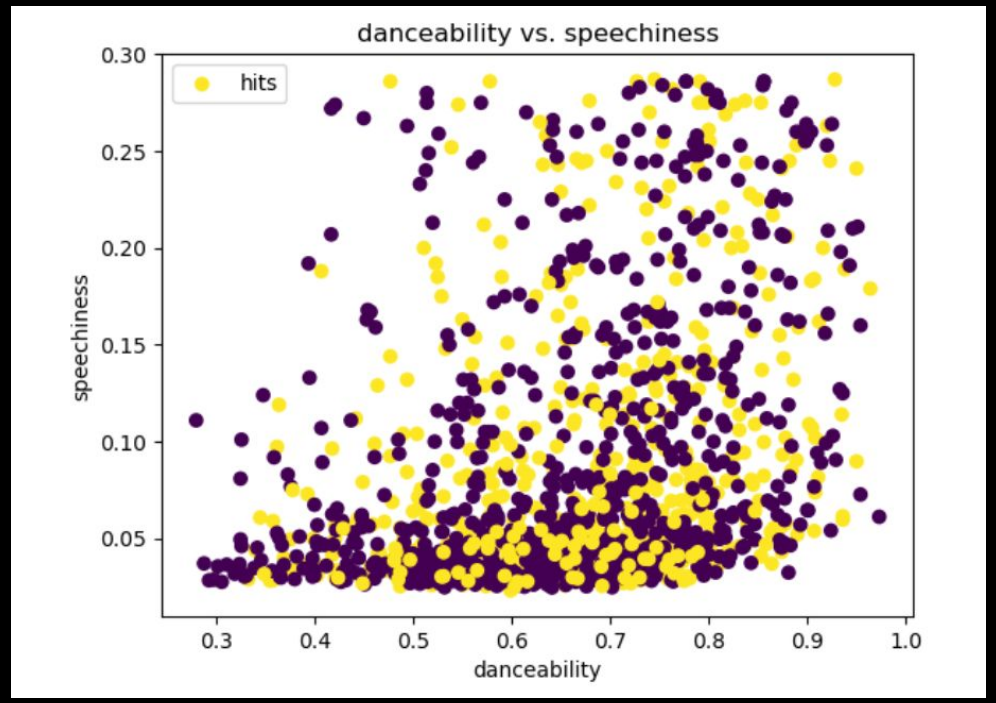
ממה מורכב המידע שלנו

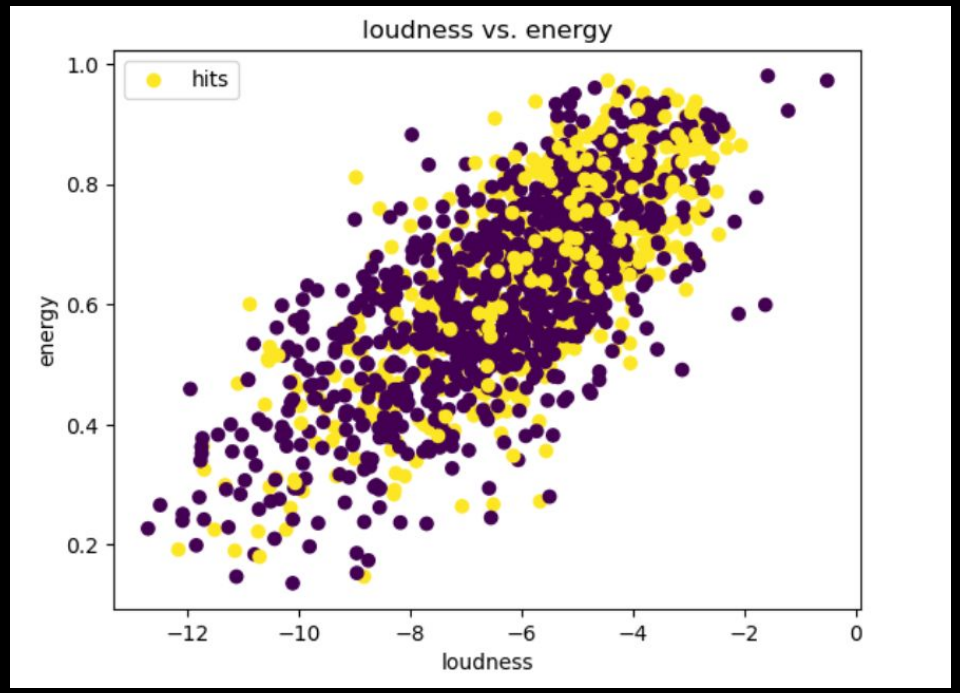






מה הם מאפייניו של להיט?





מה הם מאפייניו של להיט?

Machine learning

- בשלב למידת המכונה אימנו כמה מודלים של כמה אלגוריתמים שונים ובחרנו במודל שהפיק את תוצאות החיזוי הטובות ביותר מבין כל המודלים.
- בחירתנו הסופית הייתה KNN.

```
lrm = LogisticRegression()
lrm.fit(X_train_normalized,yTrain)
y_pred = lrm.predict(X_test_normalized)
recall score is:  0.6822429906542056
f1 score is:  0.6774941995359629
accuracy is:  0.6945054945054945
precision score is:  0.6728110599078341
```

```
decisionTree = tree.DecisionTreeClassifier(max_depth = 2, min_samples_split = 15)
Accuracy on test data =  0.6615384615384615
```

```
#-----------NAIVE-BAYSE
clf_naive_bayse = GaussianNB()
naive_scores = cross_val_score(clf_naive_bayse, XTrain, yTrain, scoring = 'recall')
print(naive_scores.mean())
0.5844848484848485
```

KNN- Machine learning algorithm

- אלגוריתם מבוסס השוואות המשייך דוגמא חדשה ל-k המופעים הקרובים אליה ביותר. מכאן שמו K- NEAREST NEIGHBORS.
- במקרה שלנו k = 3 היה מספר השכנים שמצא את התוצאה האופטימלית ביותר.

	k	train_accuracy	test_accuracy
0	3	0.782075	0.591209
1	7	0.716981	0.606593
2	9	0.709434	0.621978
3	11	0.700943	0.630769

KNN K = 3 Data without outliers

- תחילה אימנו מודל על בסיס קבוצת המידע שעברה ניקוי של ערכי קיצון.
- על מנת לשפר את תוצאות החיזוי של האלגוריתם ניסינו לבצע מניפולציות על המידע על ידי בחירת תכונות דומיננטיות(Feature selection) אך ללא הצלחה מרובה. התוצאה הגבוהה ביותר הייתה התוצאה שקיבלנו לאחר האימון הראשוני של המודל.
- בשלב זה עלה אצלנו חשד שאולי המידע שניקנו מהווה גורם משמעותי עבור מודל הלמידה.

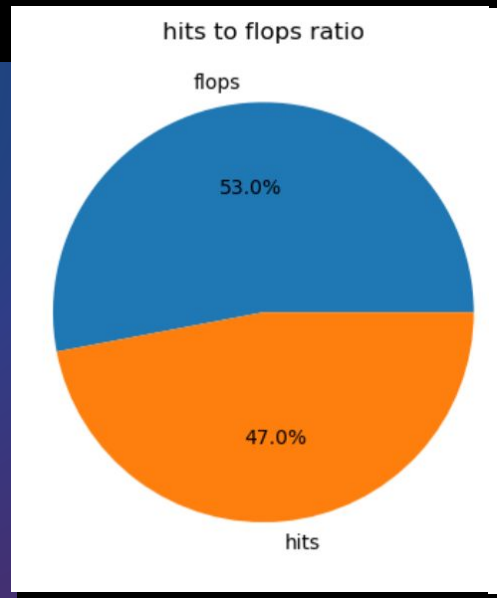
```
k_s=[]
train_accuracies=[]
test_accuracies=[]
# Create an instance of the StandardScaler class
scaler = StandardScaler()

# Normalize the input features
X_train_normalized = scaler.fit_transform(XTrain)
X_test_normalized = scaler.transform(XTest)
for k in [3, 7, 9, 11]:
    clf = KNeighborsClassifier(n_neighbors=k)
    clf.fit(X_train_normalized, yTrain)
    y_pred_train=clf.predict(X_train_normalized)
    y_pred=clf.predict(X_test_normalized)
    k_s.append(k)
    train_accuracies.append(metrics.accuracy_score(y_true = yTrain, y_pred = y_pred_train))
    test_accuracies.append(metrics.accuracy_score(y_true = yTest, y_pred = y_pred))

df=pd.DataFrame({"k":k_s,"train_accuracy":train_accuracies,"test_accuracy":test_accuracies})
df
```

	k	train_accuracy	test_accuracy
0	3	0.782075	0.591209
1	7	0.716981	0.606593
2	9	0.709434	0.621978
3	11	0.700943	0.630769

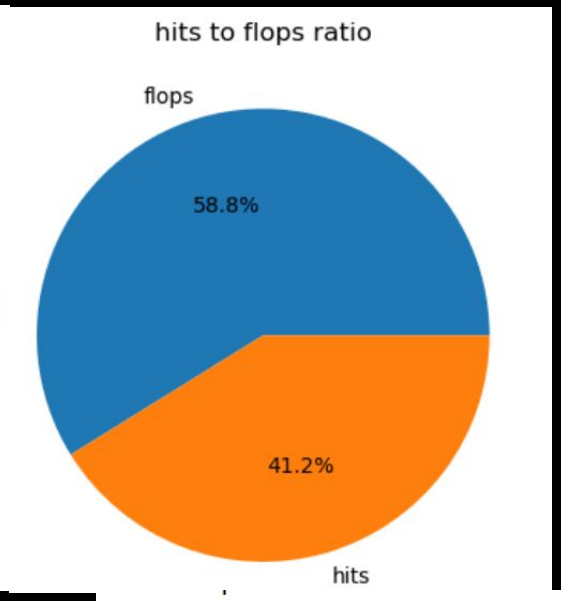
ויזואליזציה EDA
ללא ניקוי IQR

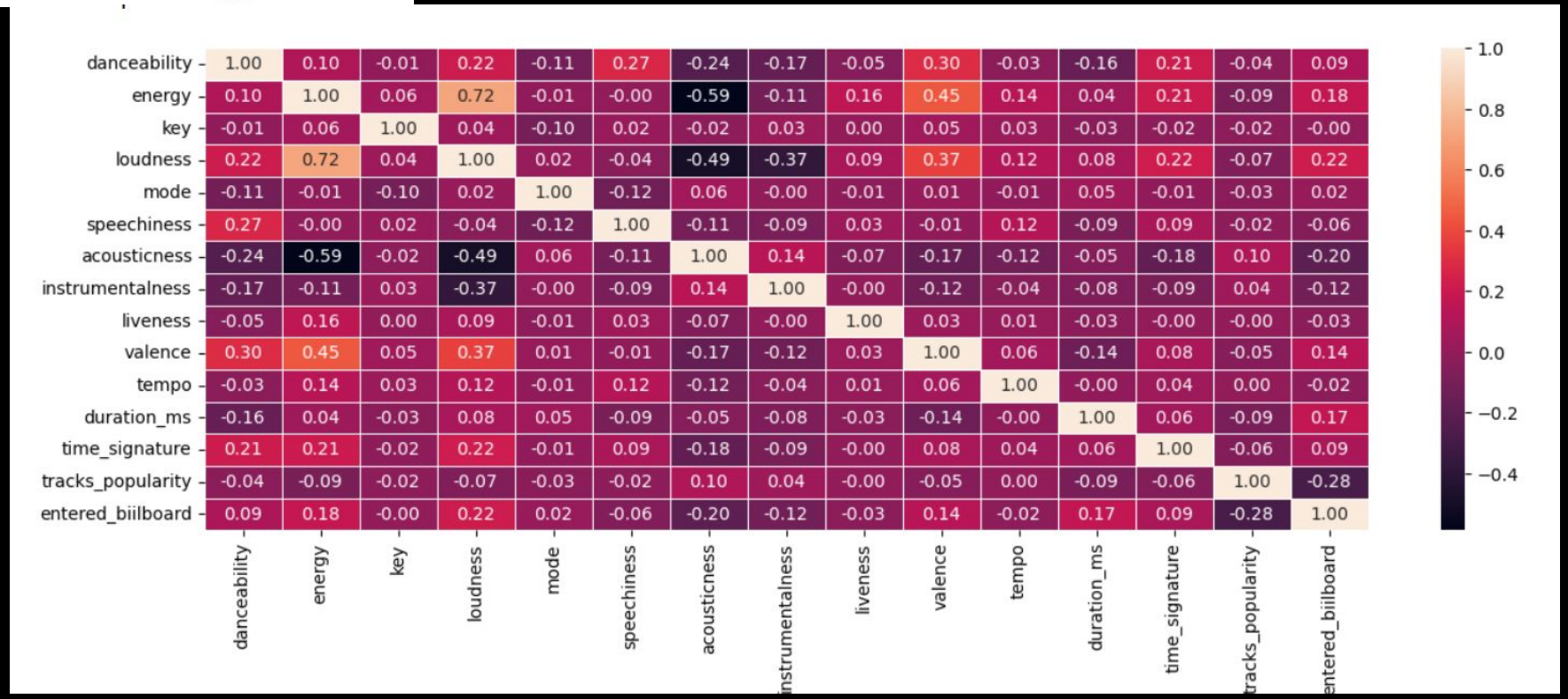
ויזואליזציה ו-EDA

עם ניקוי IQR

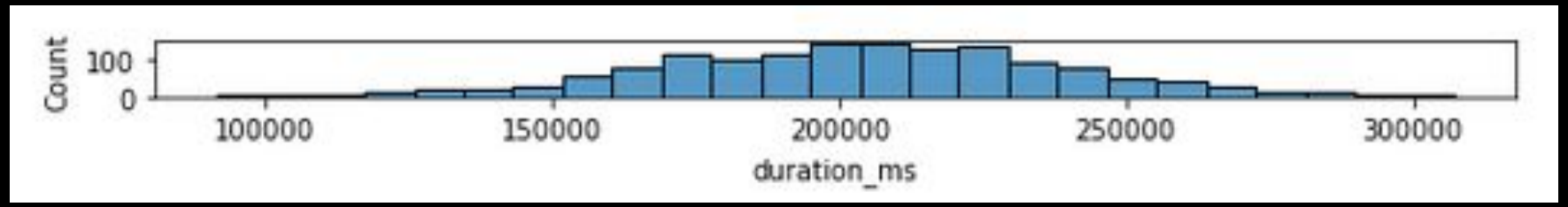


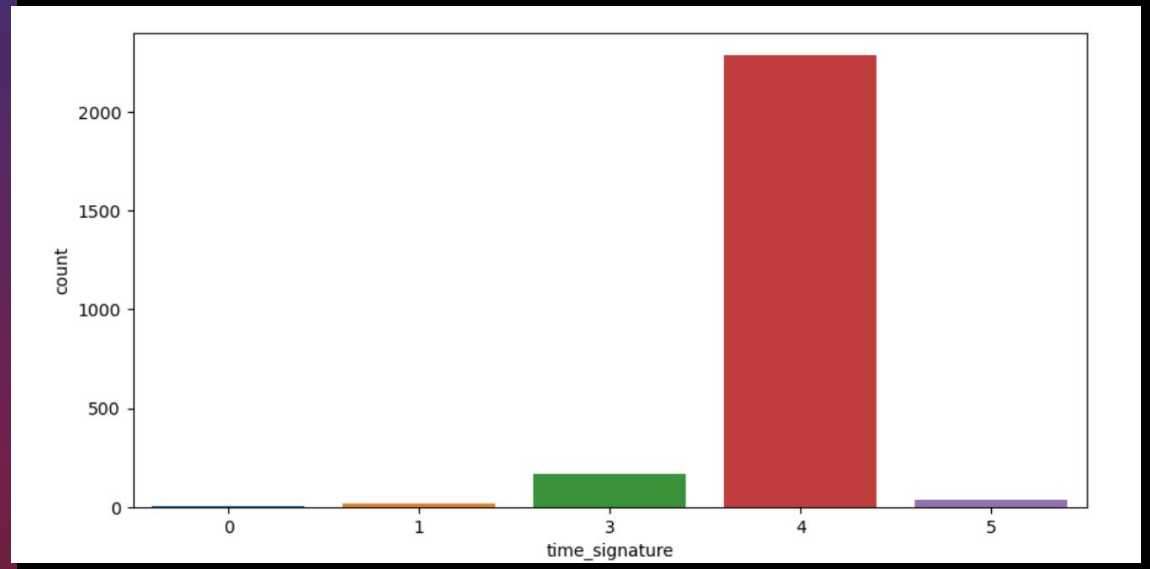
ללא ניקוי IQR

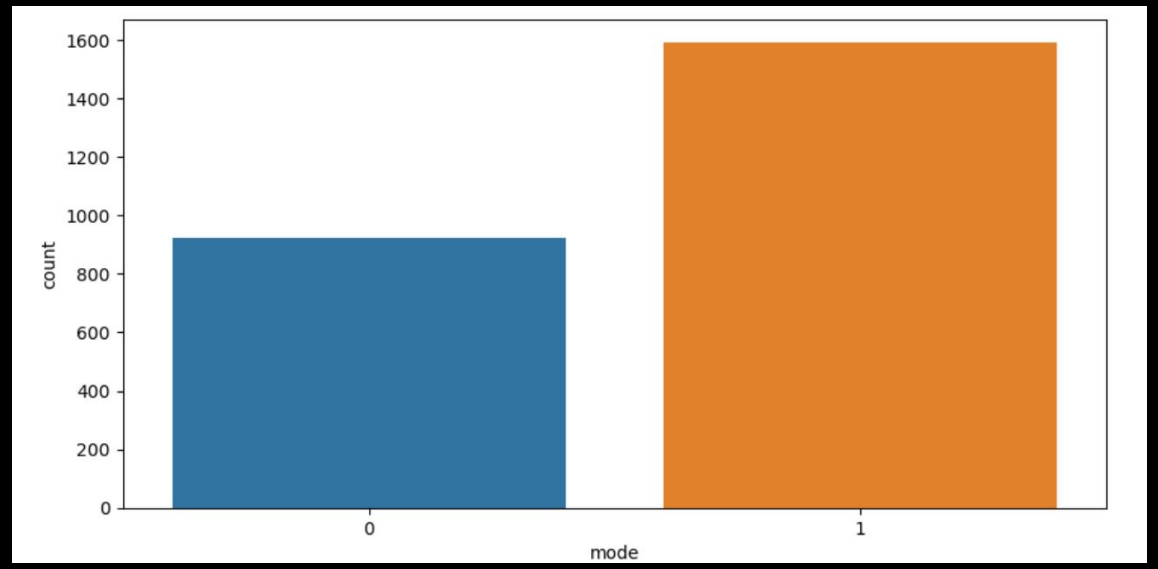




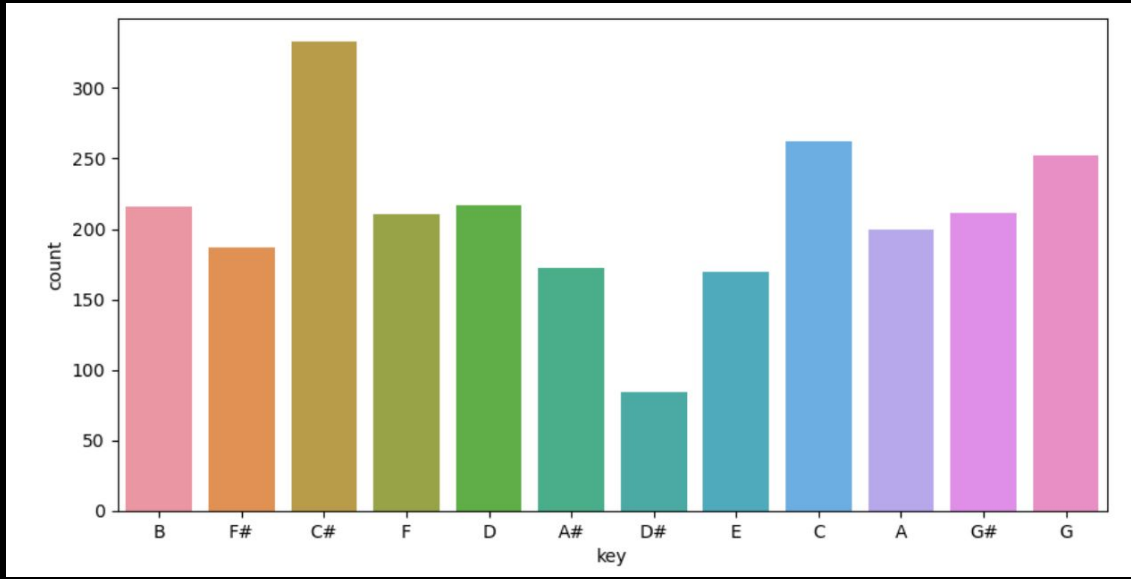
מה הם מאפייניו של להיט?



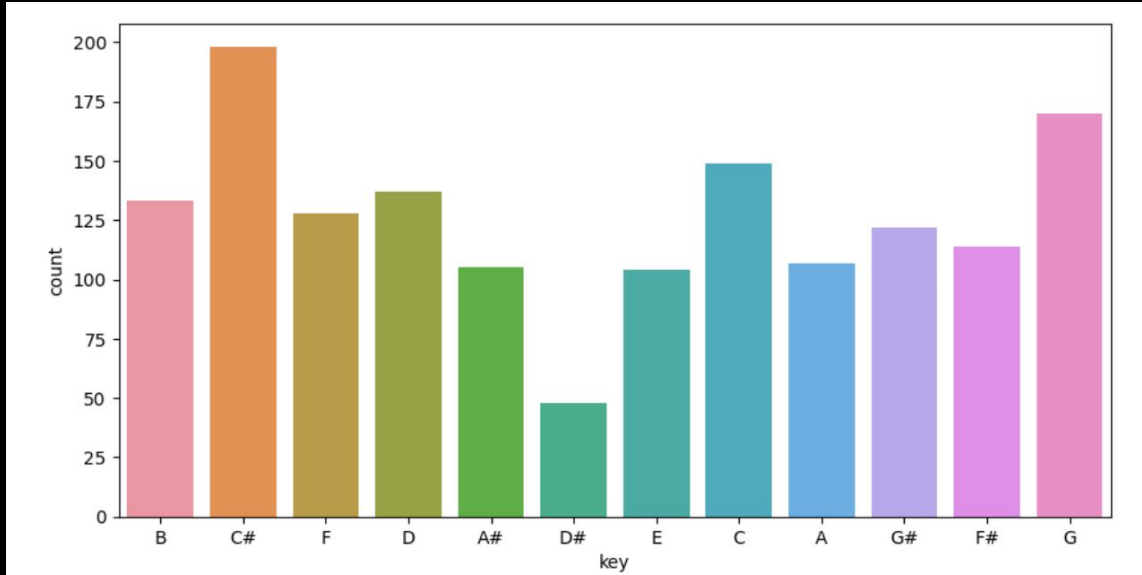




מה הם מאפייניו של להיט?

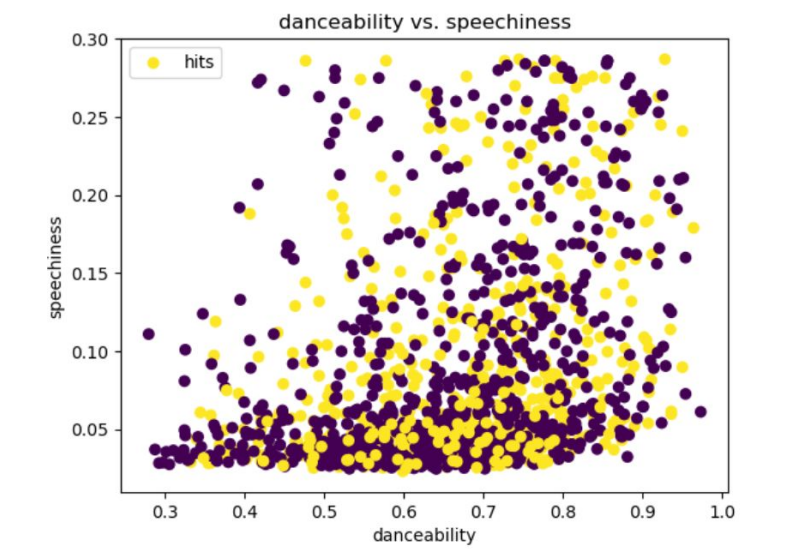


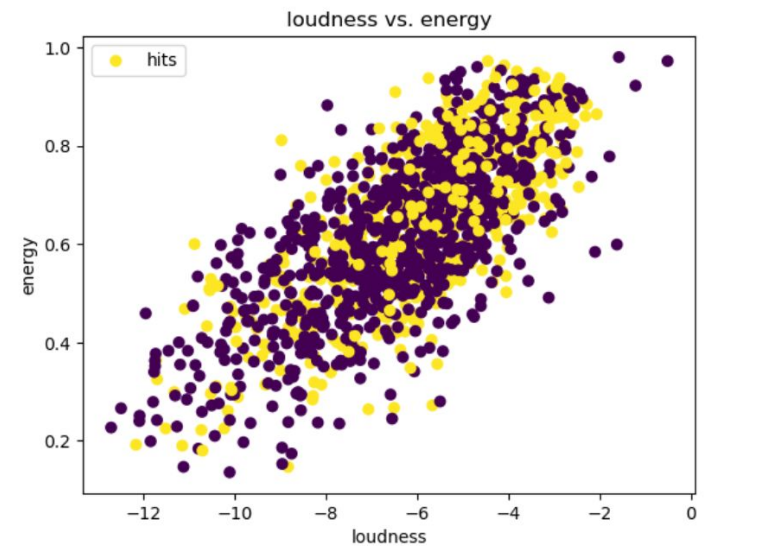
ללא ניקוי IQR



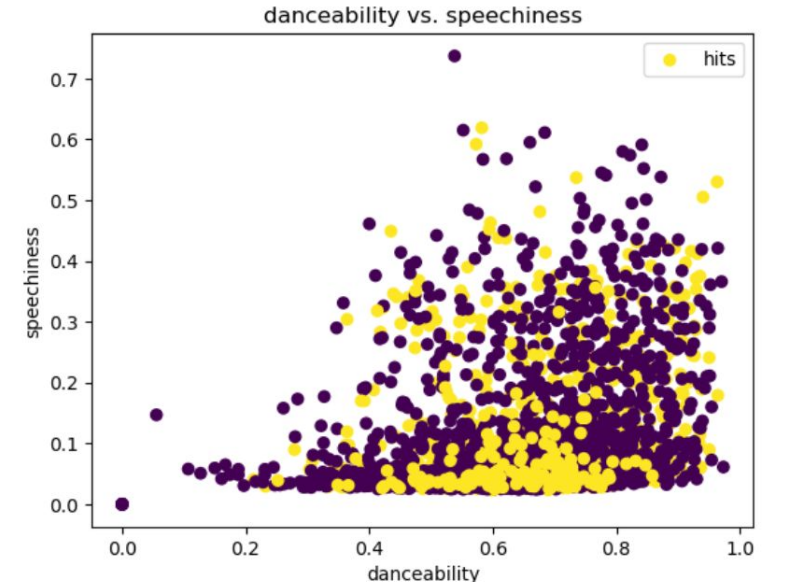
עם ניקוי IQR

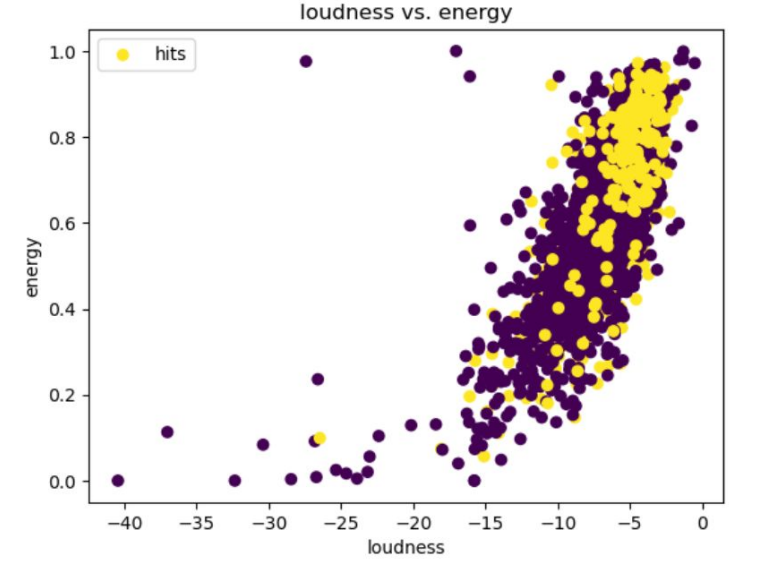
מה הם מאפייניו של להיט?





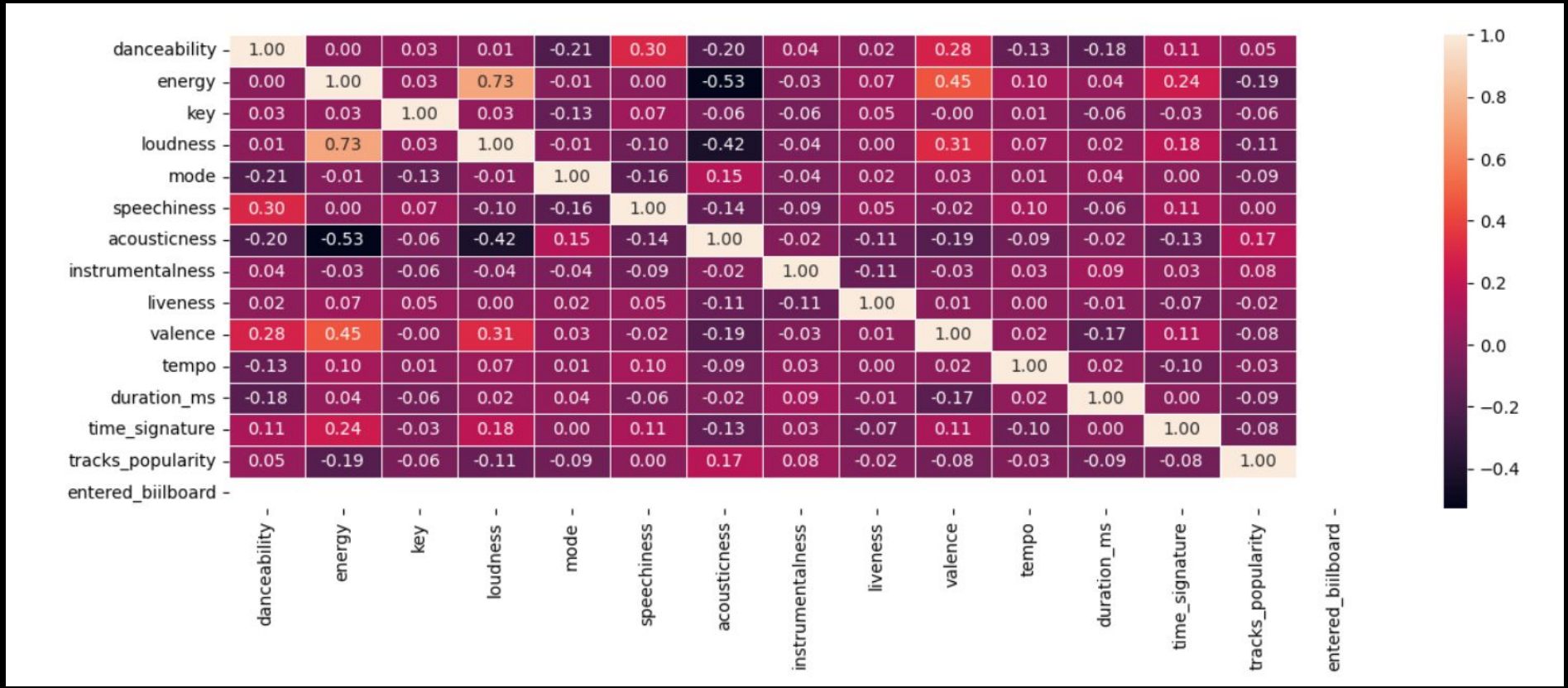
עם ניקוי IQR





ללא ניקוי IQR

להיטים בלבד ללא ניקוי



KNN K = 3 Raw data

- לאחר ההרצה הראשונית של אלגוריתם למידת המכונה חשדנו כי במהלך שלב ניקוי **הנתונים** איבדנו נתונים קריטיים המשפיעים על דיוק החיזוי של המודל אותו אימנו
- המשכנו עם בחירת ערך ה k=3 אשר נתן את התוצאות המיטביות
- אימנו מודל חדש על בסיס כל הנתונים שלנו, ללא ניקוי ערכי קיצון, באמצעותו קיבלנו את התוצאה הטובה ביותר.
- גם כאן על מנת לשפר את תוצאות החיזוי של האלגוריתם ניסינו לבצע מניפולציות על המידע על ידי בחירת תכונות דומיננטיות(Feature selection) אך ללא הצלחה מרובה. התוצאה הגבוהה ביותר הייתה התוצאה שקיבלנו לאחר האימון הראשוני של המודל.

k	train_accuracy	test_accuracy
3	0.793633	0.633113

סיכום

- מסקנתנו הראשית ממחקר זה היא שעל מנת ששיר יהיה להיט עליו לא להיות קיצוני במאפייניו (למשל אורך השיר, מקצב ממוצע, איזון בין כמות מילים למוזיקה).
- ככול ששיר רקיד יותר עם אנרגיה גבוה, הסיכוי שלו להיות להיט עולים משמעותית.
- מסקנה נוספת היא ששלב ניקוי הנתונים הוא שלב קריטי שלעיתים עלול להשפיע לרעה על אימון המודל, שלב שבו קיימת אופציה לאיבוד נתונים יקרים אשר תורמים להבנת התמונה הכוללת
- אנו סבורים כי תוצאות החיזוי הנמוכות יחסית, נבעו מהעובדה שהשירים שבחרנו באמצעות הSpotify’s API היו שירים אשר דומים מאוד באופיים ובמאפייניהם לשירים שנכנסו לרשימת בילבורד.