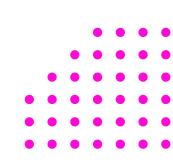# Predicting Hit Songs:

Can a formula for music business success be created?

Solomon Onyeke Ameh

Problem Description

Music production has always been attributed to inspiration or talent, with certain artists and record labels unable to produce record-breaking hit songs consistently. Recently, success in the music industry has been measured more by streaming numbers and popularity on streaming platforms such as Spotify, Apple Music, Deezer, and YouTube Music.

Hit Song = f(A,D)

What if we strip music of all creativity and explore if we can create good music by repeatedly using certain values of music features ?

The Dataset

Energy
Tempo
Danceability
Loudness
Liveness
Valence
Speechiness
Instrumentalness
Mode
Key
Duration_ms
Acousticness

Track Name
Track Artist
Track Album Name
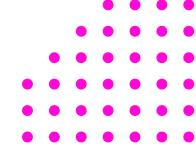Track Album Release Date
Track ID
Track Album ID
Playlist Name
Playlist Genre
Playlist Subgenre
Playlist ID

TRACK POPULARITY
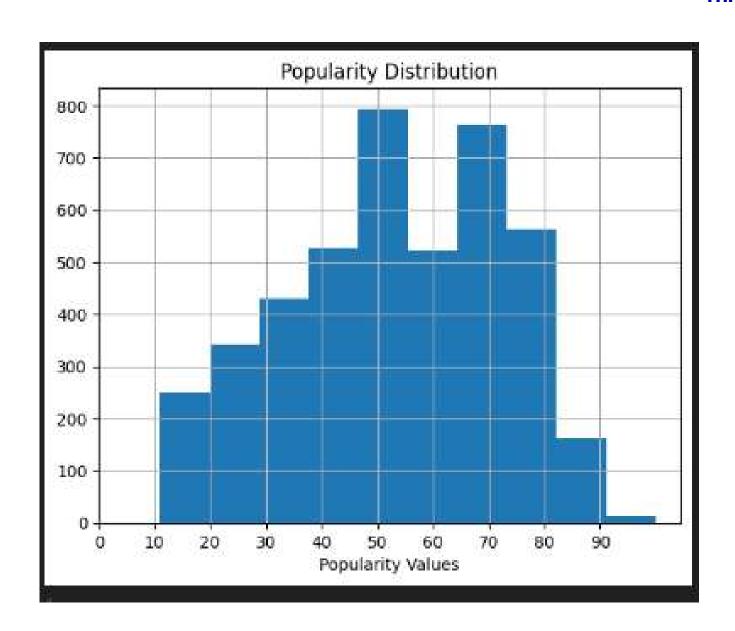
```
RangeIndex: 4831 entries, 0 to 4830
Data columns (total 24 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   energy                    4830 non-null    float64
 1   tempo                     4830 non-null    float64
 2   danceability              4830 non-null    float64
 3   playlist_genre            4831 non-null    object
 4   loudness                  4830 non-null    float64
 5   liveness                  4830 non-null    float64
 6   valence                   4830 non-null    float64
 7   track_artist              4831 non-null    object
 8   speechiness               4830 non-null    float64
 9   track_popularity          4831 non-null    int64
 10  track_album_name          4830 non-null    object
 11  playlist_name             4831 non-null    object
 12  track_id                  4831 non-null    object
 13  track_name                4831 non-null    object
 14  track_album_release_date  4831 non-null    object
 15  instrumentalness          4830 non-null    float64
 16  track_album_id            4831 non-null    object
 17  mode                      4830 non-null    float64
 18  key                       4830 non-null    float64
 19  duration_ms               4830 non-null    float64
 20  acousticness              4830 non-null    float64
 21  id                        4830 non-null    object
 22  playlist_subgenre         4831 non-null    object
 23  playlist_id               4831 non-null    object
dtypes: float64(12), int64(1), object(11)
memory usage: 905.9+ KB
```
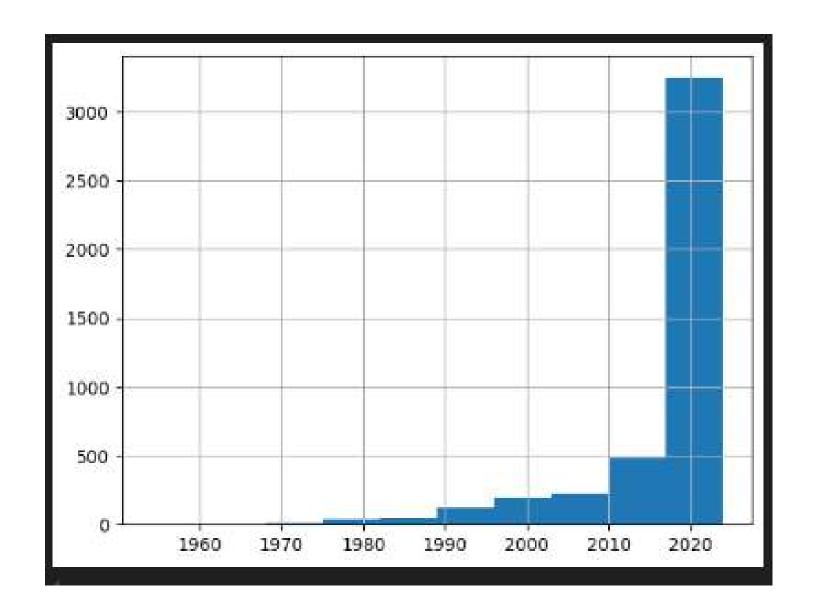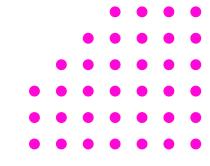
The Dataset

Threshold: 68

The Dataset

Average Track Duration vs. Year
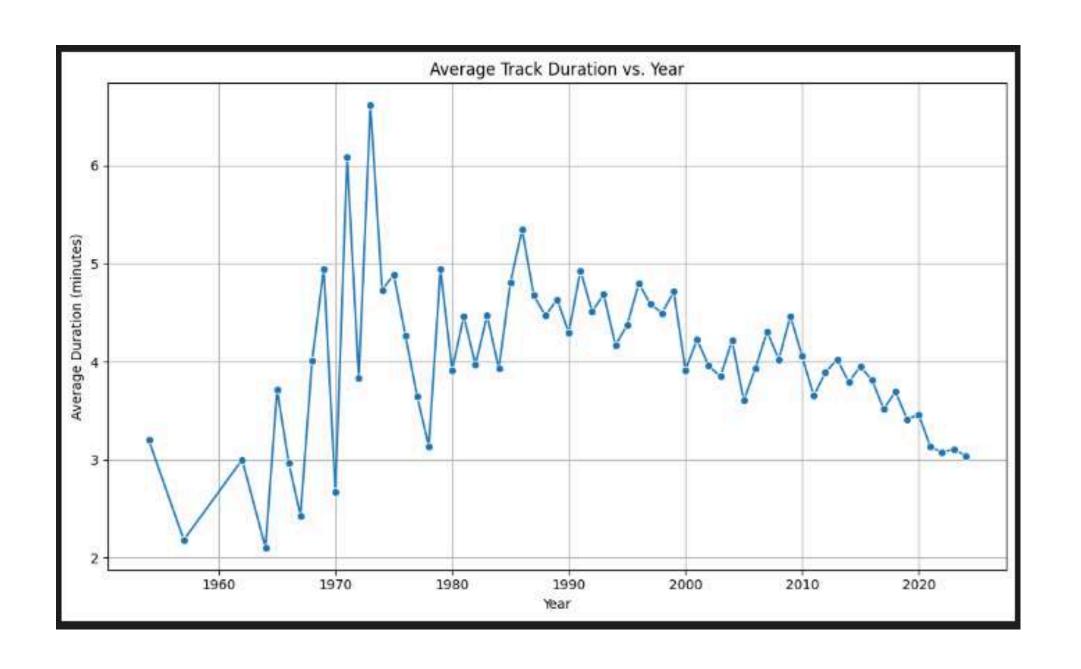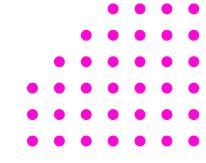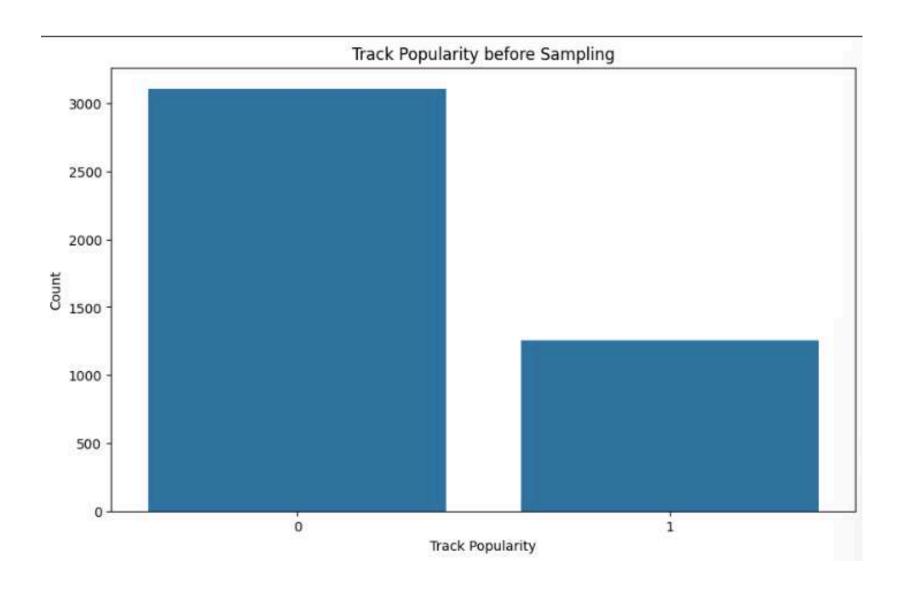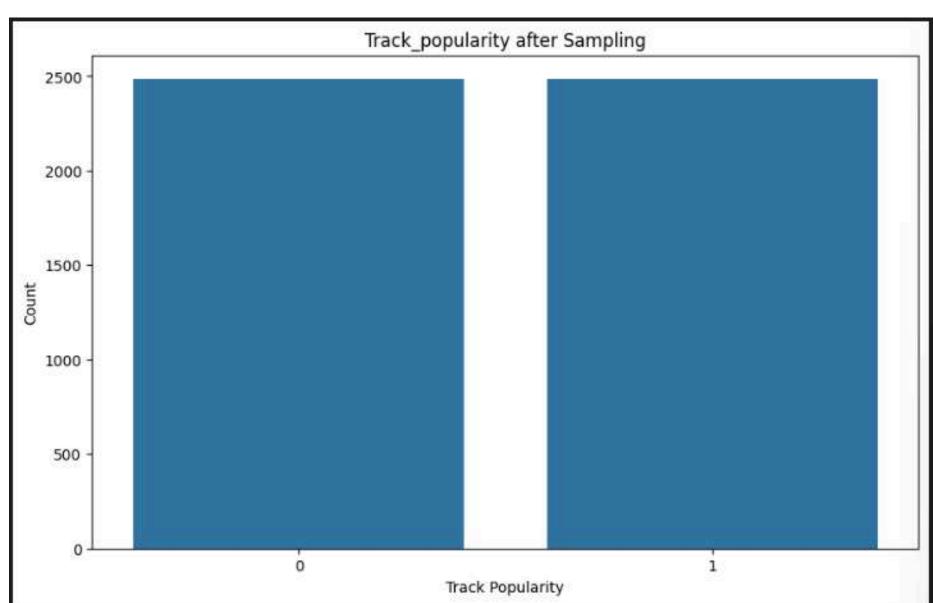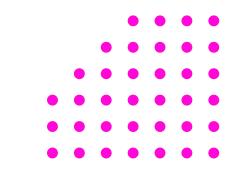
Dataset Imbalance

Original dataset shape: Counter({**0**: 3109, **1**: 1259}) Resampled dataset shape: Counter({**0**: 2487, **1**: 2487})

Results

	precision	recall	f1-score	support
0	0.87	0.88	0.88	622
1	0.70	0.67	0.69	252
accuracy			0.82	874
macro avg	0.79	0.78	0.78	874
weighted avg	0.82	0.82	0.82	874

'**LOGISTIC REGRESSION**': 0.6338672768878718,

'**KNN**': 0.6670480549199085,

'**RANDOM FOREST**': 0.8192219679633868,

'**GRADIENT BOOST**': 0.8100686498855835,
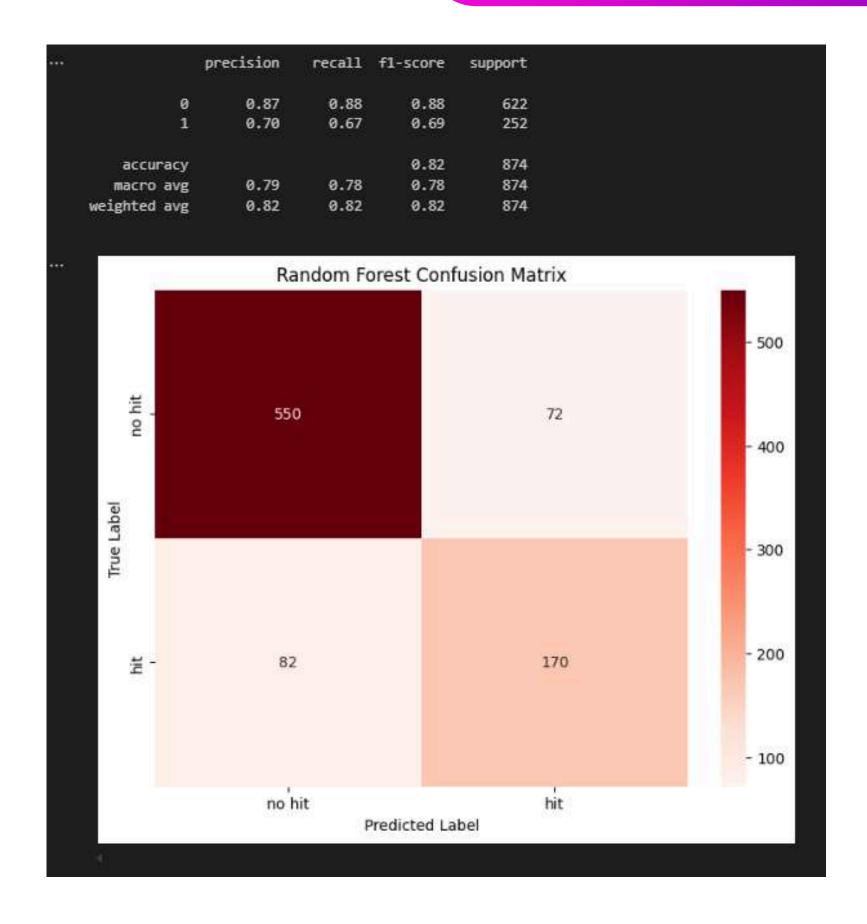
'**ADABOOST**': 0.7700228832951945,
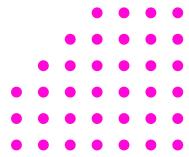
'**EXTRA TREE**': 0.8135011441647597,

'**XGBOOST**': 0.8432494279176201,

'**LIGHTGBM**': 0.8443935926773455,
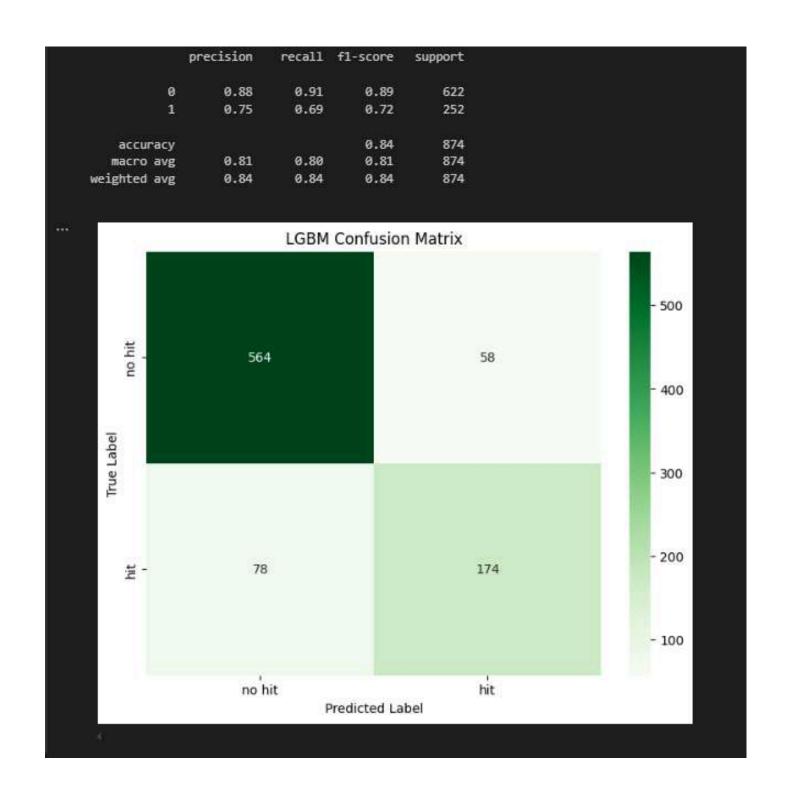
'**DUMMYCLASSIFIER**': 0.7116704805491991

Random Forest Confusion Matrix

	no hit	hit
no hit	550	72
hit	82	170

True Label / Predicted Label
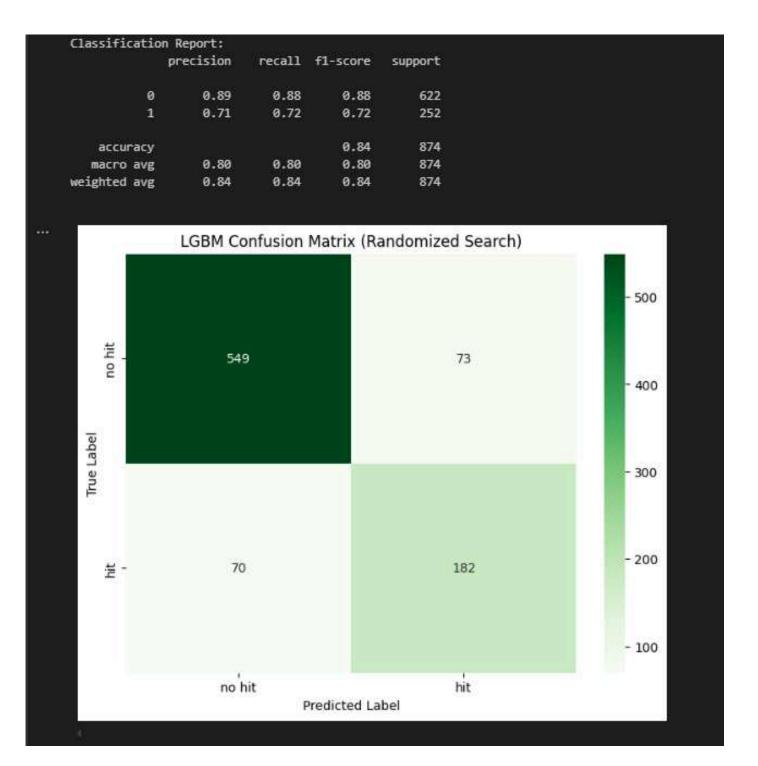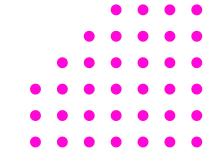
Results Light Gradient-Boosting Machine(LGBM)


```
              precision    recall  f1-score   support

           0       0.88      0.91      0.89       622
           1       0.75      0.69      0.72       252

    accuracy                           0.84       874
   macro avg       0.81      0.80      0.81       874
weighted avg       0.84      0.84      0.84       874
```

LGBM Confusion Matrix


```
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.88      0.88       622
           1       0.71      0.72      0.72       252

    accuracy                           0.84       874
   macro avg       0.80      0.80      0.80       874
weighted avg       0.84      0.84      0.84       874
```

LGBM Confusion Matrix (Randomized Search)
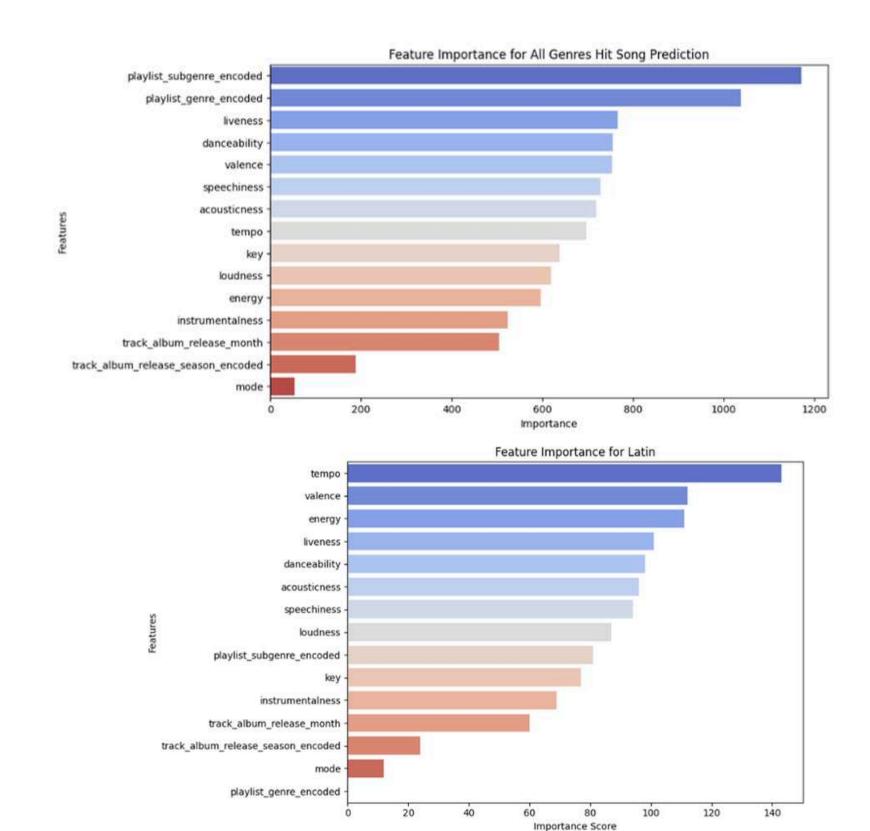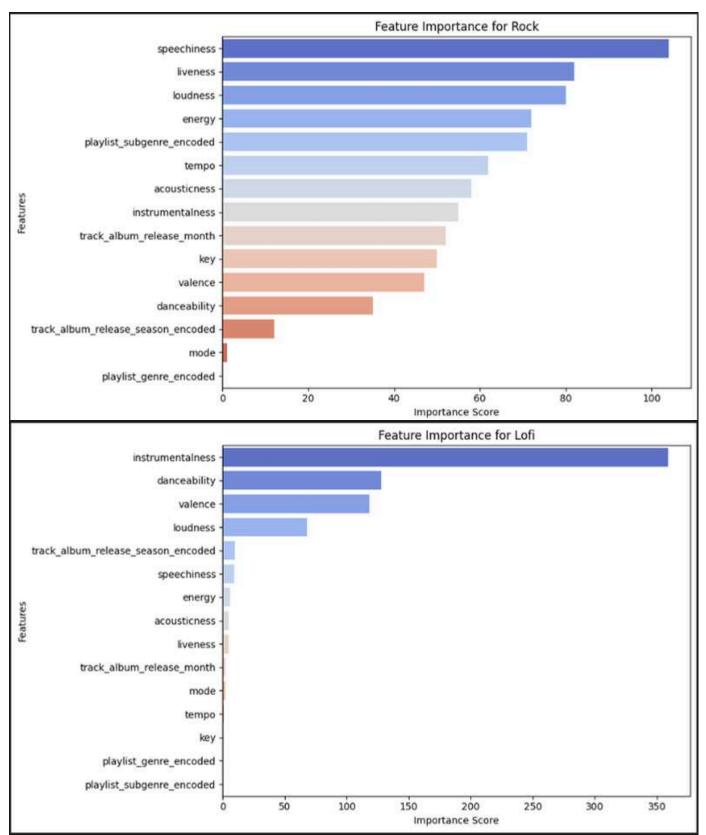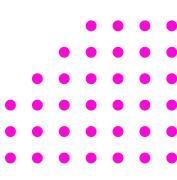
Results

What features most significantly influence the likelihood of a song becoming a hit, and do these features vary across genres?

Feature Importance for All Genres Hit Song Prediction

Feature Importance for Rock

Feature Importance for Latin

Feature Importance for Lofi

Results

How does the release date (e.g., month, season) impact a song's popularity, and are certain time frames consistently better for specific genres?

Average Popularity by Release Month for Rock

Average Popularity by Release Month for R&b

Results
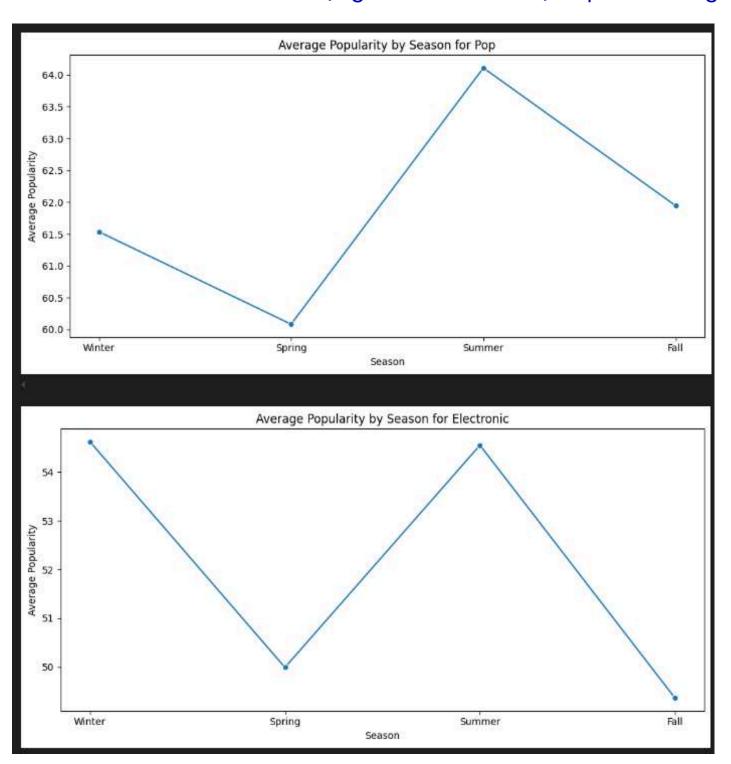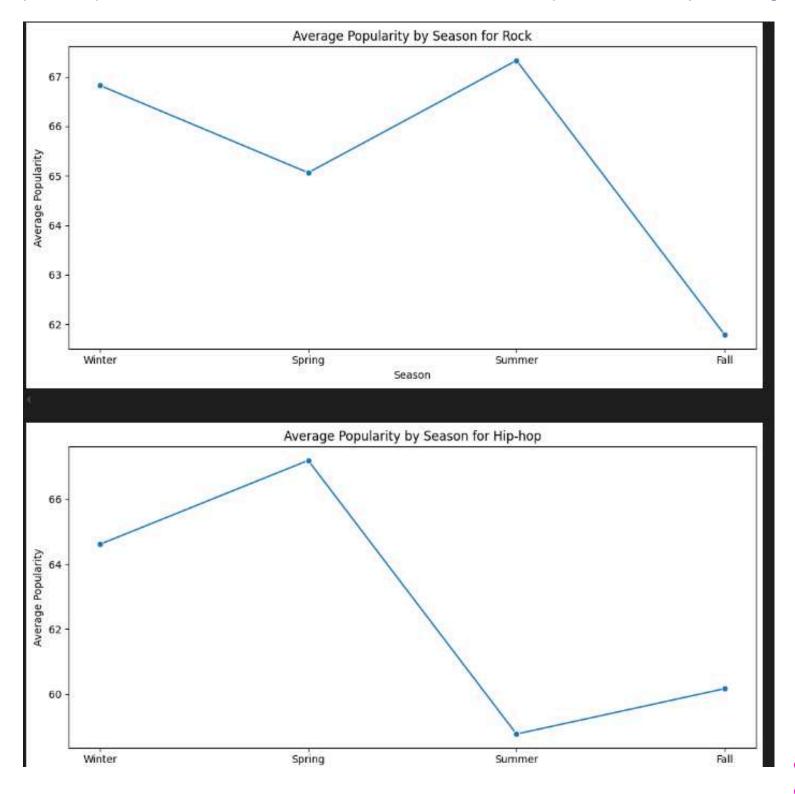
How does the release date (e.g., month, season) impact a song's popularity, and are certain time frames consistently better for specific genres?

Results
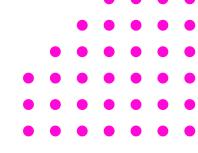
What values of audio features can consistently create a hit song, and how do these values differ across genres?

```
Mode Values for Features in Genre rock:
danceability                                    0.50200
energy                                          0.79500
key                                             0.00000
loudness                                      -18.06400
mode                                            1.00000
speechiness                                     0.02760
acousticness                                    0.00178
instrumentalness                                0.00000
liveness                                        0.10400
valence                                         0.51800
tempo                                          74.98900
playlist_subgenre_encoded                      22.00000
playlist_genre_encoded                         29.00000
track_album_release_season_encoded              2.00000
track_album_release_month                       1.00000
Name: 0, dtype: float64
```
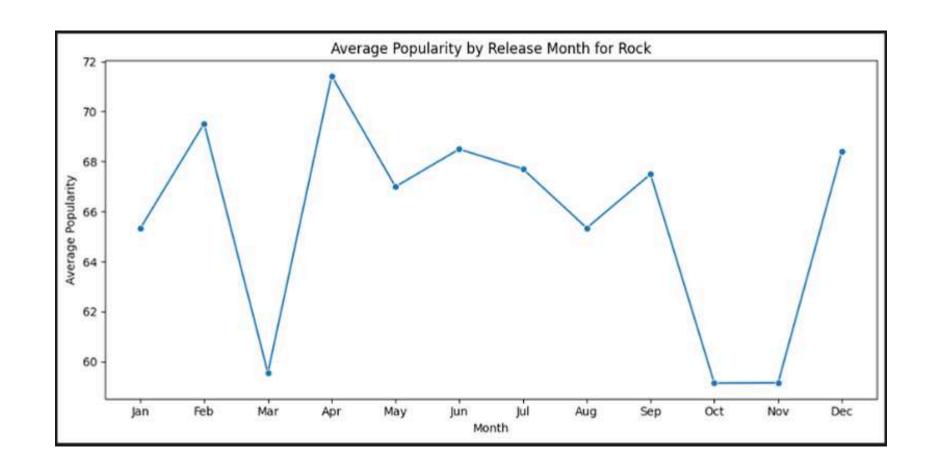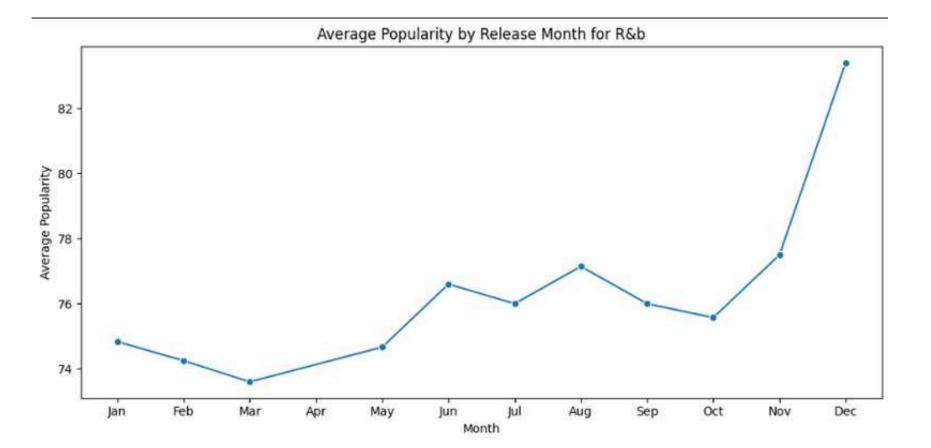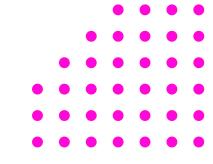
```
Mode Values for Features in High Popularity Class:
danceability                                    0.138000
energy                                          0.001610
key                                             0.000000
loudness                                      -43.643000
mode                                            1.000000
speechiness                                     0.023200
acousticness                                    0.000013
instrumentalness                                0.000000
liveness                                        0.021000
valence                                         0.034800
tempo                                          27.096123
playlist_subgenre_encoded                       0.000000
playlist_genre_encoded                          0.000000
track_album_release_season_encoded              0.000000
track_album_release_month                       1.000000
Name: 0, dtype: float64
```

```
Mode Values for Features in Genre r&b:
danceability                                    0.662000
energy                                          0.500000
key                                             1.000000
loudness                                      -11.733000
mode                                            1.000000
speechiness                                     0.039100
acousticness                                    0.000176
instrumentalness                                0.000000
liveness                                        0.104000
valence                                         0.707000
tempo                                          65.203000
playlist_subgenre_encoded                      58.000000
playlist_genre_encoded                         27.000000
track_album_release_season_encoded              3.000000
track_album_release_month                       8.000000
Name: 0, dtype: float64
```

```
Mode Values for Features in Genre pop:
danceability                                    0.7290
energy                                          0.7110
key                                             1.0000
loudness                                       -8.8240
mode                                            1.0000
speechiness                                     0.0282
acousticness                                    0.1210
instrumentalness                                0.0000
liveness                                        0.1150
valence                                         0.8060
tempo                                         105.0290
playlist_subgenre_encoded                      78.0000
playlist_genre_encoded                         25.0000
track_album_release_season_encoded              0.0000
track_album_release_month                      10.0000
Name: 0, dtype: float64
```

Conclusion

This project set out to understand the factors that make hit songs by analyzing a Spotify dataset of audio features, track popularity and descriptive features. The formula for music success is a combination of all these insights; identifying what genre is most primed for success, optimizing song production to land around the mode values for popular songs in that genre, releasing projects around their peak times and also placing focus on what features make the most impact in that genre. The results from this exploration not only answer the research questions but also provide a solid foundation for further discussions in the music industry.

Future Work

- Role of Artists and Collaborations

- Regional Preferences and Cultural Trends

- Marketing and Release Strategies

- Fan Interaction and Social Media Impact