🎶 Hit Song Predictor: Amapiano & Afrobeats 🎧

This project explores whether a song's audio features and metadata can predict its hit potential. We focus on Amapiano and Afrobeats, genres currently dominating African and global music scenes.

In [38]:

```
from IPython.display import display, HTML
display(HTML("<p><em> Press play to start the mix if your browser blocked autoplay.</em></p>"))

#Playlist intro
display(HTML("""
<h3>🎧 Now Playing: Amapiano & Afrobeats Mix</h3>
<p><strong>Tracklist:</strong><br>
1. Jealousy – Khalil Harrison (Amapiano)<br>
2. Shake Ah – Tyla (Amapiano)<br>
3. Woman – Rema (Afrobeats)<br>
4. Zenzele – Uncle Waffles (Amapiano)<br>
5. No Competition – Davido ft. Asake (Afrobeats)<br>
6. Bunda – Musa Keys (Amapiano)<br>
7. Laho – Shalipopi (Afrobeats)<br>
8. My Darling – Chella (Afrobeats)
</p>
"""))

#Audio player (autoplay with user controls)
display(HTML("""
<audio autoplay controls>
  <source src="Afrobeats and Amapiano mix.mp3" type="audio/mpeg">
  Your browser does not support the audio element.
</audio>
"""))
```

Press play to start the mix if your browser blocked autoplay.

🎧 Now Playing: Amapiano & Afrobeats Mix

Tracklist:

1. Jealousy – Khalil Harrison (Amapiano)
2. Shake Ah – Tyla (Amapiano)
3. Woman – Rema (Afrobeats)
4. Zenzele – Uncle Waffles (Amapiano)
5. No Competition – Davido ft. Asake (Afrobeats)
6. Bunda – Musa Keys (Amapiano)
7. Laho – Shalipopi (Afrobeats)
8. My Darling – Chella (Afrobeats)

▶ 0:00 / 10:24

Purpose of Project

Applying machine learning to music has proven insightful: computers can detect qualities of songs that resonate with audiences. Academic resemarch has shown that hit songs often share certain characteristics. For example, a study analyzing 30 years of music found “successful songs are happier, brighter, more party-like, more danceable and less sad than most songs” (Kaplan, 2018).In summary, building a hit song prediction model is important and appealing because it combines scientific rigor with cultural relevance. It helps company stakeholders make smarter decisions.

1. Import Libraries and Set Up

In [448…

```
#Import necessary libraries for Spotify API access, data handling, and file management
import spotipy
from spotipy.oauth2 import SpotifyOAuth
import pandas as pd
import time
import os

#Create a directory called "data" to save all downloaded or processed datasets
#This keeps the project organized and avoids cluttering the main directory
os.makedirs("data", exist_ok=True)
```

2. Spotify API Authentication

In [450…

```
#Authenticate with Spotify's API using OAuth 2.0 flow
#This gives access to user-specific data like saved tracks and private playlists
sp = spotipy.Spotify(auth_manager=SpotifyOAuth(
    client_id="053a18fb9a87491d8335598e56799c5e",
    client_secret="e905779ea1544e04aed2a2ca9680f99c",
    redirect_uri="http://127.0.0.1:8891/callback",
    scope="playlist-read-private playlist-read-collaborative"
))

#Test connection by pulling a few liked songs from Spotify account
#This confirms that access token works and successfully authenticatedresults = sp.current_user_saved_track
for item in results['items']:
    track = item['track']
    print(f"{track['name']} by {track['artists'][0]['name']}")
```

```
My Darling by Chella
The Blessing by Kari Jobe
Legends Never Die (Intro) by Yvng Ceder
Wake Up & Go Get Sum Money by Yvng Ceder
Rage In Lagos by Yvng Ceder
```

3. Extract Playlists Data from Spotify (looking for popularity score)

In [68]:

```
#Define a dictionary of confirmed working playlists across Afrobeats and Amapiano genres
#Each playlist maps to its unique Spotify ID for reliable API access
working_playlists = {
    'Best of Afrobeats 2025': '5FDBAbJobJWaKh1RDiqtyn',
    'TikTok Naija': '1H4Ws8FYRXbUCFgpYAZdK3',
    'Afrobeats Party': '1U8HSDxH8lXHQ38epJngtG',
    'Afrobeat 2025': '7IfWkPjxjtGpHKzvbZd8YV',
    'Weekly Top Grooves': '4z8jM6c6NLp0H0szju3flc',
    'African Heat': '3Uj9Y8xviyC4BvC9R49reX',
    'Best of Amapiano': '1JtkyBr4Is4ni1UQAa9AVg',
    'AmaPiano March-April 2025': '74QpABGrU1VAdBlnaJqATL',
    'Amapiano New Finds': '7LQ1WmcPCKJtXT5uQl3waU',
    'playlist_id' : '5sFUBxyx9qpGDFOCPBEd82'


}

#Authenticate with the Spotify API using secure OAuth credentials
#Scope is limited to playlist access for data collection purposes
sp = spotipy.Spotify(auth_manager=SpotifyOAuth(
    client_id="053a18fb9a87491d8335598e56799c5e",
    client_secret="e905779ea1544e04aed2a2ca9680f99c",
    redirect_uri="http://127.0.0.1:8891/callback",
    scope="playlist-read-private playlist-read-collaborative"
))

#Initialize container to hold track-level data collected from all playlists
track_data = []

#Loop through each playlist, fetch up to 100 tracks at a time, and append relevant metadata
for playlist_name, playlist_id in working_playlists.items():
    try:
        print(f"Fetching from: {playlist_name}")
        offset = 0
        while True:
            results = sp.playlist_items(playlist_id, limit=100, offset=offset)
            if not results['items']:
                break

            for item in results['items']:
                track = item['track']
                if track:  # Make sure it's not None
                    track_data.append({
                        'playlist': playlist_name,
                        'track_name': track['name'],
                        'artist': track['artists'][0]['name'],
                        'track_id': track['id'],
                        'popularity': track['popularity']
                    })
            offset += 100
            time.sleep(0.5)  #To respect rate limits
    except Exception as e:
        print(f"Failed to fetch from {playlist_name}: {e}")

#Convert collected data to a structured DataFrame and export to CSV for further analysis
df_tracks = pd.DataFrame(track_data)
df_tracks.to_csv("working_spotify_playlist_popularity.csv", index=False)
```

```
Fetching from: Best of Afrobeats 2025
Fetching from: TikTok Naija
Fetching from: Afrobeats Party
Fetching from: Afrobeat 2025
Fetching from: Weekly Top Grooves
Fetching from: African Heat
Fetching from: Best of Amapiano
Fetching from: AmaPiano March-April 2025
Fetching from: Amapiano New Finds
Fetching from: playlist_id
```

4. Collect Spotify Stream Data for Recent Trending Afrobeats and Amampiano Songs (Will Measure Popularity by Streams Per Day Later)

In [598…

```
# Stream Data
#📊 Contains total stream counts, genre labels, and release dates for Afrobeats and Amapiano songs
#Data includes music from Nigeria (afrobeats top 30 current spotify songs) and South Africa (top 30 curren

stream_data = {
    # --- AFROBEATS songs ---
    "my darling": {"streams": 4_924_774, "genre": "Afrobeats", "release_date": "2025-03-27"},
    "arike": {"streams": 22_434_165, "genre": "Afrobeats", "release_date": "2025-02-15"},
    "are you there?": {"streams": 31_946_737, "genre": "Afrobeats", "release_date": "2024-08-16"},
    "awolowo": {"streams": 48_842_828, "genre": "Afrobeats", "release_date": "2024-08-14"},
    "beamer": {"streams": 11_580_503, "genre": "Afrobeats", "release_date": "2024-10-14"},
    "be there still": {"streams": 6_737_818, "genre": "Afrobeats", "release_date": "2025-02-14"},
    "chandelier": {"streams": 17_369_585, "genre": "Afrobeats", "release_date": "2024-11-21"},
    "na scra": {"streams": 7_248_693, "genre": "Afrobeats", "release_date": "2025-03-07"},
    "doha": {"streams": 27_380_251, "genre": "Afrobeats", "release_date": "2024-07-13"},
    "free of charge": {"streams": 1_915_678, "genre": "Afrobeats", "release_date": "2025-03-27"},
    "pity this boy (with victony)": {"streams": 21_224_424, "genre": "Afrobeats", "release_date": "2025-02-
    "get better": {"streams": 4_857_075, "genre": "Afrobeats", "release_date": "2025-03-21"},
    "funds (feat. odumodublvck & chike)": {"streams": 42_4554_775, "genre": "Afrobeats", "release_date": "2
    "happy": {"streams": 8_112_951, "genre": "Afrobeats", "release_date": "2025-02-20"},
    "hey jago": {"streams": 2_993_456, "genre": "Afrobeats", "release_date": "2025-03-18"},
    "joy is coming": {"streams": 37_092_308, "genre": "Afrobeats", "release_date": "2024-12-18"},
    "JUJU (feat. Shallipopi)": {"streams": 41_465_826, "genre": "Afrobeats", "release_date": "2024-08-22"}
    "kese (dance)": {"streams": 45_694_575, "genre": "Afrobeats", "release_date": "2024-10-15"},
    "baby (is it a crime)": {"streams": 35_551_585, "genre": "Afrobeats", "release_date": "2025-02-07"},
    "shaolin": {"streams": 12_715_482, "genre": "Afrobeats", "release_date": "2025-02-10"},
    "management": {"streams": 6_204_585, "genre": "Afrobeats", "release_date": "2025-01-25"},
    "most wanted": {"streams": 1_837_793, "genre": "Afrobeats", "release_date": "2025-04-04"},
    "mario kart": {"streams": 8_687_810, "genre": "Afrobeats", "release_date": "2025-02-20"},
    "legolas": {"streams": 1_955_624, "genre": "Afrobeats", "release_date": "2025-03-31"},
    "trenches luv": {"streams": 4_312_356, "genre": "Afrobeats", "release_date": "2025-02-13"},
    "toy girl (with juno & valentino rose)": {"streams": 1_197_925, "genre": "Afrobeats", "release_date": "
    "venus": {"streams": 26_775_718, "genre": "Afrobeats", "release_date": "2025-02-21"},
    "why love": {"streams": 10_384_733, "genre": "Afrobeats", "release_date": "2025-02-12"},

    # --- AMAPIANO songs ---
    "Sdudla or Slender": {"streams": 2_894_712, "genre": "Amapiano", "release_date": "2025-03-20"},
    "Vuma Dlozi Lami (feat. Ancestral Rituals)": {"streams": 9_243_178, "genre": "Amapiano", "release_date"
    "Ngisakuthanda": {"streams": 7_241_847, "genre": "Amapiano", "release_date": "2024-09-06"},
    "Ngibolekeni (feat. Seun1401, LeeMcKrazy, Blxckie, Pcee, Madumane & Kabelo Sings)": {"streams": 5_852_
    "Vuka (feat. Thukuthela)": {"streams": 8_720_251, "genre": "Amapiano", "release_date": "2024-12-15"},
    "Uyaphapha Marn (feat. Scotts Maphuma & Kabelo Sings)": {"streams": 4_838_765, "genre": "Amapiano", "r
    "Wayengenalutho": {"streams": 4_610_909, "genre": "Amapiano", "release_date": "2025-02-21"},
    "Sohlala Sisonke": {"streams": 12_909_636, "genre": "Amapiano", "release_date": "2025-02-14"},
    "Bo Gogo (feat. Tracy & Thatohatsi)": {"streams": 6_078_812, "genre": "Amapiano", "release_date": "202
    "HAUSAPIANO - Remix": {"streams": 21_129_492, "genre": "Amapiano", "release_date": "2024-10-31"},
    "Uvume Kanjani?": {"streams": 1_289_020, "genre": "Amapiano", "release_date": "2025-03-21"},
    "Biri Marung (feat. Sje Konka, Focalistic, DJ Maphorisa, Scotts Maphuma & CowBoii)": {"streams": 23_85
    "Romeo & Juliet": {"streams": 4_674_070, "genre": "Amapiano", "release_date": "2025-01-22"},
    "Shapa Bell": {"streams": 1_339_189, "genre": "Amapiano", "release_date": "2025-04-01"},
    "Abantwana Bakho (feat. Thatohatsi, Young Stunna & Nkosazana Daughter)": {"streams": 790_174, "genre":
    "Malunde (feat. Springle)": {"streams": 864_570, "genre": "Amapiano", "release_date": "2024-12-13"},
    "Vulani (feat. Thatohatsi & Tracy)": {"streams": 3_295_409, "genre": "Amapiano", "release_date": "2024-
    "Skuta Baba - Remix": {"streams": 7_294_010, "genre": "Amapiano", "release_date": "2024-12-06"},
    "Ungangilimazi (feat. Frank Mabeat)": {"streams": 4_937_736, "genre": "Amapiano", "release_date": "202
    "All My Life": {"streams": 10_706_356, "genre": "Amapiano", "release_date": "2024-08-30"},
    "Awuhlabe Kabili": {"streams": 3_409_948, "genre": "Amapiano", "release_date": "2024-12-06"},
    "ZENZELE (feat. Royal MusiQ, Uncool MC, Xduppy, & CowBoii)": {"streams": 2_014_490, "genre": "Amapiano"
    "Naledi": {"streams": 1_401_369, "genre": "Amapiano", "release_date": "2025-03-10"},
    "UYAH! (feat. 2wo Bunnies, Jay Music, & Imbongi Yosizi)": {"streams": 2_223_683, "genre": "Amapiano", "
    "Ngiyakuthanda": {"streams": 3_686_853, "genre": "Amapiano", "release_date": "2025-02-02"},
    "Shayi'Moto (feat. Seemah & Yanda Woods)": {"streams": 9_941_374, "genre": "Amapiano", "release_date":
    "Wishi Wishi (feat. Scotts Maphuma & Young Stunna)": {"streams":13_211_203, "genre": "Amapiano", "relea
    "Dear Ex Yami": {"streams": 5_231_036, "genre": "Amapiano", "release_date": "2024-09-22"},
    "Ama Gear": {"streams": 11_286_723, "genre": "Amapiano", "release_date": "2023-12-01"},
    "Kabza Chant (feat. Young Stunna, Nkosazana Daughter, Mthunzi, Nokwazi, Anzo, Mashudu, Murumba Pitch &
    "Abo Nokthula (feat. The Exclusive SA, Scotts Maphuma, Kabelo Sings, Bontle Smith, 2woshort & Stompiie

    # --- SONG CROSSING BOTH REGIONS (Afrobeats, charted in both NG & SA) ---
    "laho": {"streams": 24_981_448, "genre": "Afrobeats", "release_date": "2025-02-21"}  # Use the higher 
}
```

5. Extract Tiktok Virality (video uses) Data of Recent Afrobeats and Amapiano Music

In [88]:

```
#TikTok virality classification:
#TikTok virality is calculated as: TikTok uses / days since release, a song is considered a hit if its sco
#This rule applies to both Afrobeats and Amapiano songs.

tiktok_trending = {
    # --- AFROBEATS ---
    "Laho": 1,
    "My Darling": 1,
    "Arike": 1,
    "Are you there?": 0,
    "Awolowo": 0,
    "Beamer": 0,
    "Be There Still": 0,
    "Chandelier": 0,
    "Na Scra": 1,
    "Doha": 0,
    "Free of Charge": 0,
    "PITY THIS BOY (with Victony)": 0,
    "Get Better": 0,
    "Funds (feat. ODUMODUBLVCK & Chike)": 1,
    "Happy": 1,
    "hey jago": 1,
    "Joy Is Coming": 1,
    "JUJU (feat. Shallipopi)": 1,
    "Kese (Dance)": 1,
    "Baby (Is it a Crime)": 1,
    "Shaolin": 1,
    "Management (with BIGKHALID)": 0,
    "Most Wanted": 0,
    "MARIO KART": 1,
    "LEGOLAS": 0,
    "Trenches Luv": 1,
    "TOY GIRL (with Juno & Valentino Rose)": 0,
    "Venus": 1,
    "WHY LOVE": 1,

    # --- AMAPIANO ---
    "Sdudla or Slender": 1,
    "Vuma Dlozi Lami (feat. Ancestral Rituals)": 0,
    "Ngisakuthanda": 0,
    "Ngibolekeni (feat. Seun1401, LeeMcKrazy, Blxckie, Pcee, Madumane & Kabelo Sings)": 1,
    "Vuka (feat. Thukuthela)": 1,
    "Uyaphapha Marn (feat. Scotts Maphuma & Kabelo Sings)": 1,
    "Wayengenalutho": 0,
    "Sohlala Sisonke": 1,
    "Bo Gogo (feat. Tracy & Thatohatsi)": 1,
    "HAUSAPIANO - Remix": 1,
    "Uvume Kanjani?": 1,
    "Biri Marung (feat. Sje Konka, Focalistic, DJ Maphorisa, Scotts Maphuma & CowBoii)": 1,
    "Romeo & Juliet": 1,
    "Shapa Bell": 0,
    "Abantwana Bakho (feat. Thatohatsi, Young Stunna & Nkosazana Daughter)": 0,
    "Malunde (feat. Springle)": 0,
    "Vulani (feat. Thatohatsi & Tracy)": 0,
    "Skuta Baba - Remix": 1,
    "Ungangilimazi (feat. Frank Mabeat)": 0,
    "All My Life": 0,
    "Awuhlabe Kabili": 0,
    "ZENZELE (feat. Royal MusiQ, Uncool MC, Xduppy, & CowBoii)": 0,
    "Naledi (w/ Naledi Aphiwe)": 0,
    "UYAH! (feat. 2wo Bunnies, Jay Music, & Imbongi Yosizi)": 0,
    "Ngiyakuthanda": 0,
    "Shayi'Moto (feat. Seemah & Yanda Woods)": 1,
    "Wishi Wishi (feat. Scotts Maphuma & Young Stunna)": 1,
    "Dear Ex Yami": 1,
    "Ama Gear": 0,
    "Kabza Chant (feat. Young Stunna, Nkosazana Daughter, Mthunzi, Nokwazi, Anzo, Mashudu, Murumba Pitch &
    "Abo Nokthula (feat. The Exclusive SA, Scotts Maphuma, Kabelo Sings, Bontle Smith, 2woshort & Stompiie
}
```

In [90]:

```
#Create DataFrame just with song names
df = pd.DataFrame({'song': list(tiktok_trending.keys())})
df['viral_on_tiktok'] = df['song'].map(tiktok_trending).fillna(0)
```

6. Observe Billboard Afrobeat Songs and Flag Recent Trending Songs

In [341…

```
#Import fuzzy matching
from fuzzywuzzy import fuzz

#Billboard Africa Titles (as is)
billboard_africa_titles = [
    "Push 2 Start", "Water", "Move", "Baby (Is it a Crime)", "Shake It To The Max FLY", "Laho", "Get Bette
    "Why Love", "Update", "Arike", "Be There Still", "Funds", "Joy Is Coming", "PITY THIS BOY with Victony'
    "Na Scra", "Bad For You", "Kese Dance", "Awake", "Mario Kart", "Trenches Luv", "Hey Jago", "Happy",
    "Good Vibes", "Bad Girl", "Who Does That", "Introduction", "Wetego", "Only Fans", "Taxi Driver", "New 
    "Macho", "Apres Minuit", "Obimo", "iToro", "Panic", "Louder", "Bend", "JayJay", "Chandelier", "Beamer"
    "Going Intro", "Toma Toma", "Break Me Down", "A Million Blessings", "World Best", "lololufe"
]

import re

#Function to clean song titles by removing non-alphanumeric characters
def clean_title(title):
    return re.sub(r'[^a-zA-Z0-9]', '', title.lower().strip())

#Cleaned version of Billboard titles
billboard_africa_clean = [clean_title(title) for title in billboard_africa_titles]

#Clean your actual dataframe's song titles (adjust 'song' to your column name if different)
df['song_clean'] = df['song'].apply(clean_title)

from fuzzywuzzy import fuzz

#Function to fuzzily match (akin to a confidence score) song to the Billboard list
def is_billboard_hit(song, threshold=80):
    for bb_song in billboard_africa_clean:
        score = fuzz.token_sort_ratio(song, bb_song)
        if score >= threshold:
            return True
    return False

#Flag songs that appear in Billboard Africa chart
df['in_billboard_africa'] = df['song_clean'].apply(lambda x: 1 if is_billboard_hit(x) else 0)
```

In [345…

```
#Assign 'Afrobeats' genre only for songs matched to Billboard
df.loc[df['in_billboard_africa'] == 1, 'genre'] = 'Afrobeats'


#Preview Billboard Africa chart hits with genre column
billboard_hits = df[df['in_billboard_africa'] == 1][['song', 'in_billboard_africa', 'genre']]
billboard_hits
```

Out [345…

	song	in_billboard_africa	genre
0	Laho	1	Afrobeats
2	Arike	1	Afrobeats
5	Beamer	1	Afrobeats
6	Be There Still	1	Afrobeats
7	Chandelier	1	Afrobeats
8	Na Scra	1	Afrobeats
11	PITY THIS BOY (with Victony)	1	Afrobeats
12	Get Better	1	Afrobeats
15	hey jago	1	Afrobeats
16	Joy Is Coming	1	Afrobeats
18	Kese (Dance)	1	Afrobeats
19	Baby (Is it a Crime)	1	Afrobeats
20	Shaolin	1	Afrobeats
23	MARIO KART	1	Afrobeats
25	Trenches Luv	1	Afrobeats
28	WHY LOVE	1	Afrobeats

7. Defining What Makes a 'hit' in Each Category

In [125…

```
import pandas as pd
```

In [408…

```
import pandas as pd
from rapidfuzz import fuzz
from IPython.display import display

#Load your playlist CSV
```

```
df_tracks = pd.read_csv("working_spotify_playlist_popularity.csv")

#Complete genre mapping
target_songs = {
    # --- AFROBEATS ---
    "My Darling": "Afrobeats",
    "Arike": "Afrobeats",
    "Are you there?": "Afrobeats",
    "Awolowo": "Afrobeats",
    "Beamer": "Afrobeats",
    "Be There Still": "Afrobeats",
    "Chandelier": "Afrobeats",
    "Na Scra": "Afrobeats",
    "Doha": "Afrobeats",
    "Free of Charge": "Afrobeats",
    "PITY THIS BOY (with Victony)": "Afrobeats",
    "Get Better": "Afrobeats",
    "Funds (feat. ODUMODUBLVCK & Chike)": "Afrobeats",
    "Happy": "Afrobeats",
    "hey jago": "Afrobeats",
    "Joy Is Coming": "Afrobeats",
    "JUJU (feat. Shallipopi)": "Afrobeats",
    "Kese (Dance)": "Afrobeats",
    "Baby (Is it a Crime)": "Afrobeats",
    "Shaolin": "Afrobeats",
    "Management": "Afrobeats",
    "Most Wanted": "Afrobeats",
    "MARIO KART": "Afrobeats",
    "LEGOLAS": "Afrobeats",
    "Trenches Luv": "Afrobeats",
    "TOY GIRL (with Juno, Valentino Rose)": "Afrobeats",
    "Venus": "Afrobeats",
    "WHY LOVE": "Afrobeats",
    "Laho": "Afrobeats",

    # --- AMAPIANO ---
    "Sdudla or Slender": "Amapiano",
    "Vuma Dlozi Lami (feat. Ancestral Rituals)": "Amapiano",
    "Ngisakuthanda": "Amapiano",
    "Ngibolekeni (feat. Seun1401, LeeMcKrazy, Blxckie, Pcee, Madumane & Kabelo Sings)": "Amapiano",
    "Vuka (feat. Thukuthela)": "Amapiano",
    "Uyaphapha Marn (feat. Scotts Maphuma...)": "Amapiano",
    "Wayengenalutho": "Amapiano",
    "Sohlala Sisonke": "Amapiano",
    "Bo Gogo (feat. Tracy & Thathohatsi)": "Amapiano",
    "HAUSAPIANO - Remix": "Amapiano",
    "Uvume Kanjani?": "Amapiano",
    "Biri Marung (feat. Sje Konka, Focalistic, DJ Maphorisa, Scotts Maphuma & CowBoii)": "Amapiano",
    "Romeo & Juliet": "Amapiano",
    "Shapa Bell": "Amapiano",
    "Abantwana Bakho (feat. Thatohatsi, Young Stunna & Nkosazana Daughter)": "Amapiano",
    "Malunde (feat. Springle)": "Amapiano",
    "Vulani (feat. Thatohatsi & Tracy)": "Amapiano",
    "Skuta Baba - Remix": "Amapiano",
    "Ungangilimazi (feat. Frank Mabeat)": "Amapiano",
    "All My Life": "Amapiano",
    "Awuhlabe Kabili": "Amapiano",
    "ZENZELE (feat. Royal MusiQ, Uncool MC, Xduppy, & CowBoii)": "Amapiano",
    "Naledi": "Amapiano",
    "UYAH! (feat. 2wo Bunnies, Jay Music, & Imbongi Yosizi)": "Amapiano",
    "Ngiyakuthanda": "Amapiano",
    "Shayi'Moto (feat. Seemah & Yanda Woods)": "Amapiano",
    "Wishi Wishi (feat. Scotts Maphuma & Young Stunna)": "Amapiano",
    "Dear Ex Yami": "Amapiano",
    "Ama Gear": "Amapiano",
    "Kabza Chant (feat. Young Stunna, Nkosazana Daughter, Mthunzi, Nokwazi, Anzo, Mashudu, Murumba Pitch &
    "Abo Nokthula (feat. The Exclusive SA, Scotts Maphuma, Kabelo Sings, Bontle Smith, 2woshort & Stompiie
}

# --- Fuzzy Matching Function ---
def fuzzy_match_track(track_name, target_dict, threshold=80):
    for key in target_dict.keys():
        if fuzz.token_sort_ratio(track_name.lower(), key.lower()) >= threshold:
            return key
    return None

#Apply fuzzy matching to get matched keys
df_tracks['matched_key'] = df_tracks['track_name'].apply(lambda x: fuzzy_match_track(x, target_songs))

#Filter only matched songs
filtered_df = df_tracks[df_tracks['matched_key'].notna()].copy()

#Map genres
filtered_df['genre'] = filtered_df['matched_key'].map(target_songs)
```

```

#Classify popularity
def classify_popularity(row):
    if row['genre'] == 'Afrobeats':
        if row['popularity'] >= 73:
            return "Hit 🔥"
        elif row['popularity'] >= 65:
            return "—"
        else:
            return "–"
    elif row['genre'] == 'Amapiano':
        if row['popularity'] >= 65:
            return "Hit 🔥"
        elif row['popularity'] >= 45:
            return "—"
        else:
            return "–"
    return "Unknown Genre"

filtered_df['popularity_classification'] = filtered_df.apply(classify_popularity, axis=1)

#Drop duplicates by track + genre
filtered_df = filtered_df.drop_duplicates(subset=['track_name', 'genre'])

#Final display
display(filtered_df[['track_name', 'artist', 'genre', 'popularity', 'popularity_classification']])
```

	track_name	artist	genre	popularity	popularity_classification
0	Get Better	Zlatan	Afrobeats	71	——
1	Arike	Kunmie	Afrobeats	78	Hit 🔥
2	Na Scra	Famous Pluto	Afrobeats	73	Hit 🔥
3	Hey Jago	Poco Lee	Afrobeats	67	——
5	Laho	Shallipopi	Afrobeats	78	Hit 🔥
7	Free of Charge	Joeboy	Afrobeats	63	—
9	Joy Is Coming	Fido	Afrobeats	58	—
10	WHY LOVE	Asake	Afrobeats	72	——
11	Be There Still	Davido	Afrobeats	72	——
12	Baby (Is it a Crime)	Rema	Afrobeats	80	Hit 🔥
13	Funds (feat. ODUMODUBLVCK & Chike)	Davido	Afrobeats	74	Hit 🔥
14	PITY THIS BOY (with Victony)	ODUMODUBLVCK	Afrobeats	76	Hit 🔥
22	Venus	Faceless	Afrobeats	77	Hit 🔥
24	HAUSAPIANO - Remix	Kvng Vinci	Amapiano	70	Hit 🔥
28	Kese (Dance)	Wizkid	Afrobeats	66	——
30	Biri Marung (feat. Sje Konka, Focalistic, DJ Maphorisa, Scotts Maphuma & CowBoii)	Mr Pilato	Amapiano	69	Hit 🔥
42	JUJU (feat. Shallipopi)	Smur Lee	Afrobeats	70	——
45	Ngisakuthanda	Zee Nxumalo	Amapiano	65	Hit 🔥
47	Ngibolekeni (feat. Seun1401, LeeMcKrazy, Blxckie, Pcee, Madumane & Kabelo Sings)	DJ Maphorisa	Amapiano	68	Hit 🔥
48	Bo Gogo (feat. Tracy & Thatohatsi)	Kelvin Momo	Amapiano	65	Hit 🔥

	track_name	artist	genre	popularity	popularity_classification
50	Ungangilimazi (feat. Frank Mabeat)	Dj Moscow	Amapiano	54	—
51	Vulani (feat. Thatohatsi & Tracy)	Kelvin Momo	Amapiano	60	—
52	Wishi Wishi (feat. Scotts Maphuma & Young Stunna)	Kabza De Small	Amapiano	63	—
67	Abo Nokthula (feat. The Exclusive SA, Scotts Maphuma, Kabelo Sings, Bontle Smith, 2woshort & Stompiiey)	TNK MusiQ	Amapiano	48	—
69	Ama Gear	Dlala Thukzin	Amapiano	60	—
81	Kabza Chant (feat. Young Stunna, Nkosazana Daughter, Mthunzi, Nokwazi, Anzo, Mashudu, Murumba Pitch & Tman Xpress)	Kabza De Small	Amapiano	51	—
518	Are you there?	Ayo Maff	Afrobeats	71	—
530	Awolowo	Fido	Afrobeats	74	Hit 🔥
543	Malunde (feat. Springle)	Shakes & Les	Amapiano	58	—
549	ZENZELE (feat. Royal MusiQ, Uncool MC, Xduppy, & CowBoii)	Uncle Waffles	Amapiano	59	—
552	Abantwana Bakho (feat. Thatohatsi, Young Stunna & Nkosazana Daughter)	DJ Maphorisa	Amapiano	59	—
557	UYAH! (feat. 2wo Bunnies, Jay Music, & Imbongi Yosizi)	Uncle Waffles	Amapiano	59	—
562	Biri Marung (Edit) (feat. Sje Konka, Focalistic, DJ Maphorisa, Scotts Maphuma & CowBoii)	Mr Pilato	Amapiano	53	—
585	Shayi'Moto (feat. Seemah & Yanda Woods)	Mellow & Sleazy	Amapiano	64	—
617	Skuta Baba - Remix	WOODBLOCK DJS	Amapiano	63	—
794	Wayengenalutho	MENZI MUSIC	Amapiano	62	—
807	Vuma Dlozi Lami (feat. Ancestral Rituals)	Issa sisdoh	Amapiano	65	Hit 🔥
811	Vuka (feat. Thukuthela)	Oscar Mbo	Amapiano	67	Hit 🔥

	track_name	artist	genre	popularity	popularity_classification
1001	Happy	Teebay RSA	Afrobeats	33	—
1014	Sdudla or Slender	Shandesh	Amapiano	65	Hit 🔥
1190	Sohlala Sisonke	Dlala Thukzin	Amapiano	65	Hit 🔥
1216	Beamer	T.I BLAZE	Afrobeats	66	——
1217	Chandelier	Monaky	Afrobeats	72	——
1218	Doha	Seyi Vibez	Afrobeats	70	——
1220	LEGOLAS	ODUMODUBLVCK	Afrobeats	67	——
1221	MARIO KART	Seyi Vibez	Afrobeats	72	——
1222	Management	Smur Lee	Afrobeats	69	——
1223	Most Wanted	Zinoleesky	Afrobeats	67	——
1224	My Darling	Chella	Afrobeats	71	——
1225	SHAOLIN	Seyi Vibez	Afrobeats	74	Hit 🔥
1226	TOY GIRL (with Juno & Valentino Rose)	ODUMODUBLVCK	Afrobeats	63	—
1227	Trenches Luv	T.I BLAZE	Afrobeats	67	——
1229	All My Life	Mawelele	Amapiano	55	——
1230	Awuhlabe Kabili	LIMIT NALA	Amapiano	59	——
1231	Dear Ex Yami	Mduduzi Ncube	Amapiano	59	——
1233	Naledi	Mawelele	Amapiano	59	——
1234	Ngiyakuthanda	MENZI MUSIC	Amapiano	60	——
1235	Romeo & Juliet	Naledi Aphiwe	Amapiano	64	——
1236	Shapa Bell	Naleboy Young King	Amapiano	56	——
1239	Uvume Kanjani?	LIMIT NALA	Amapiano	59	——

In [636…

```
#Function to calculate streaming data, SPD = Total Streams / Days Since Release
#Afrobeats Hit: SPD >= 300,000
#Amapiano Hit: SPD >= 75,000

#Import Datetime
from datetime import datetime

current_date = datetime.today()

#Create genre mapping from stream_data
genre_map = {song: details["genre"] for song, details in stream_data.items()}

def reclassify_stricter_thresholds(data):
    result = []
    for song, details in data.items():
        release_date = datetime.strptime(details["release_date"], "%Y-%m-%d")
        days_since_release = (current_date - release_date).days
        spd = details["streams"] / days_since_release if days_since_release > 0 else details["streams"]

        genre = genre_map.get(song, "Unknown")
        if genre == "Afrobeats":
            if spd >= 300000:
                classification = "Hit 🔥"
            elif spd >= 100000:
                classification = "Potential Hit ⚡"
            else:
                classification = "Moderate 🌱"
        elif genre == "Amapiano":
            if spd >= 75000:
                classification = "Hit 🔥"
            elif spd >= 40000:
                classification = "Potential Hit ⚡"
            else:
                classification = "Moderate 🌱"
        else:
            classification = "Unknown Genre"

        result.append({
            "Song": song,
            "Genre": genre,
            "Streams": details["streams"],
            "Release Date": details["release_date"],
            "Days Since Release": days_since_release,
            "Streams Per Day": round(spd),
            "Classification": classification
        })

    return result

df_hits_stricter = pd.DataFrame(reclassify_stricter_thresholds(stream_data))
# Show all columns and rows
pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows", None)
pd.set_option("display.max_colwidth", None)

# Now display the full DataFrame
display(df_hits_stricter)
```

	Song	Genre	Streams	Release Date	Days Since Release	Streams Per Day	Classification
0	my darling	Afrobeats	4924774	2025-03-27	19	259199	Potential Hit ⚡
1	arike	Afrobeats	22434165	2025-02-15	59	380240	Hit 🔥
2	are you there?	Afrobeats	31946737	2024-08-16	242	132011	Potential Hit ⚡
3	awolowo	Afrobeats	48842828	2024-08-14	244	200176	Potential Hit ⚡
4	beamer	Afrobeats	11580503	2024-10-14	183	63281	Moderate 🌱
5	be there still	Afrobeats	6737818	2025-02-14	60	112297	Potential Hit ⚡
6	chandelier	Afrobeats	17369585	2024-11-21	145	119790	Potential Hit ⚡
7	na scra	Afrobeats	7248693	2025-03-07	39	185864	Potential Hit ⚡
8	doha	Afrobeats	27380251	2024-07-13	276	99204	Moderate 🌱
9	free of charge	Afrobeats	1915678	2025-03-27	19	100825	Potential Hit ⚡
10	pity this boy (with victony)	Afrobeats	21224424	2025-02-28	46	461401	Hit 🔥
11	get better	Afrobeats	4857075	2025-03-21	25	194283	Potential Hit ⚡
12	funds (feat. odumodublvck & chike)	Afrobeats	424554775	2024-12-06	130	3265806	Hit 🔥

	Song	Genre	Streams	Release Date	Days Since Release	Streams Per Day	Classification
13	happy	Afrobeats	8112951	2025-02-20	54	150240	Potential Hit ⚡
14	hey jago	Afrobeats	2993456	2025-03-18	28	106909	Potential Hit ⚡
15	joy is coming	Afrobeats	37092308	2024-12-18	118	314342	Hit 🔥
16	JUJU (feat. Shallipopi)	Afrobeats	41465826	2025-03-28	18	2303657	Hit 🔥
17	kese (dance)	Afrobeats	45694575	2024-10-15	182	251069	Potential Hit ⚡
18	baby (is it a crime)	Afrobeats	35551585	2025-02-07	67	530621	Hit 🔥
19	shaolin	Afrobeats	12715482	2025-02-10	64	198679	Potential Hit ⚡
20	management	Afrobeats	6204585	2025-01-25	80	77557	Moderate 🌱
21	most wanted	Afrobeats	1837793	2025-04-04	11	167072	Potential Hit ⚡
22	mario kart	Afrobeats	8687810	2025-02-20	54	160885	Potential Hit ⚡
23	legolas	Afrobeats	1955624	2025-03-31	15	130375	Potential Hit ⚡
24	trenches luv	Afrobeats	4312356	2025-02-13	61	70694	Moderate 🌱
25	toy girl (with juno & valentino rose)	Afrobeats	1197925	2025-03-31	15	79862	Moderate 🌱
26	venus	Afrobeats	26775718	2025-02-21	53	505202	Hit 🔥

	Song	Genre	Streams	Release Date	Days Since Release	Streams Per Day	Classification
27	why love	Afrobeats	10384733	2025-02-12	62	167496	Potential Hit ⚡
28	Sdudla or Slender	Amapiano	2894712	2025-03-14	32	90460	Hit 🔥
29	Vuma Dlozi Lami (feat. Ancestral Rituals)	Amapiano	9243178	2024-09-21	206	44870	Potential Hit ⚡
30	Ngisakuthanda	Amapiano	7241847	2024-09-06	221	32769	Moderate 🌱
31	Ngibolekeni (feat. Seun1401, LeeMcKrazy, Blxckie, Pcee, Madumane & Kabelo Sings)	Amapiano	5852759	2025-01-31	74	79091	Hit 🔥
32	Vuka (feat. Thukuthela)	Amapiano	8720251	2024-12-15	121	72068	Potential Hit ⚡
33	Uyaphapha Marn (feat. Scotts Maphuma & Kabelo Sings)	Amapiano	4838765	2025-01-31	74	65389	Potential Hit ⚡
34	Wayengenalutho	Amapiano	4610909	2025-02-21	53	86998	Hit 🔥
35	Sohlala Sisonke	Amapiano	12909636	2025-02-14	60	215161	Hit 🔥
36	Bo Gogo (feat. Tracy & Thatohatsi)	Amapiano	6078812	2025-01-31	74	82146	Hit 🔥
37	HAUSAPIANO - Remix	Amapiano	21129492	2024-10-31	166	127286	Hit 🔥
38	Uvume Kanjani?	Amapiano	1289020	2025-03-21	25	51561	Potential Hit ⚡
39	Biri Marung (feat. Sje Konka, Focalistic, DJ Maphorisa, Scotts Maphuma & CowBoii)	Amapiano	23859597	2024-10-20	177	134800	Hit 🔥
40	Romeo & Juliet	Amapiano	4674070	2025-01-22	83	56314	Potential Hit ⚡

	Song	Genre	Streams	Release Date	Days Since Release	Streams Per Day	Classification
41	Shapa Bell	Amapiano	1339189	2025-04-01	14	95656	Hit 🔥
42	Abantwana Bakho (feat. Thatohatsi, Young Stunna & Nkosazana Daughter)	Amapiano	790174	2025-03-28	18	43899	Potential Hit ⚡
43	Malunde (feat. Springle)	Amapiano	864570	2024-12-13	123	7029	Moderate 🌱
44	Vulani (feat. Thatohatsi & Tracy)	Amapiano	3295409	2024-12-09	127	25948	Moderate 🌱
45	Skuta Baba - Remix	Amapiano	7294010	2024-12-06	130	56108	Potential Hit ⚡
46	Ungangilimazi (feat. Frank Mabeat)	Amapiano	4937736	2024-09-20	207	23854	Moderate 🌱
47	All My Life	Amapiano	10706356	2024-08-30	228	46958	Potential Hit ⚡
48	Awuhlabe Kabili	Amapiano	3409948	2024-12-06	130	26230	Moderate 🌱
49	ZENZELE (feat. Royal MusiQ, Uncool MC, Xduppy, & CowBoii)	Amapiano	2014490	2025-03-15	31	64984	Potential Hit ⚡
50	Naledi	Amapiano	1401369	2025-03-10	36	38927	Moderate 🌱
51	UYAH! (feat. 2wo Bunnies, Jay Music, & Imbongi Yosizi)	Amapiano	2223683	2025-02-10	64	34745	Moderate 🌱
52	Ngiyakuthanda	Amapiano	3686853	2025-02-02	72	51206	Potential Hit ⚡
53	Shayi'Moto (feat. Seemah & Yanda Woods)	Amapiano	9941374	2024-11-01	165	60251	Potential Hit ⚡
54	Wishi Wishi (feat. Scotts Maphuma & Young Stunna)	Amapiano	13211203	2024-09-29	198	66723	Potential Hit ⚡

	Song	Genre	Streams	Release Date	Days Since Release	Streams Per Day	Classification
55	Dear Ex Yami	Amapiano	5231036	2024-09-22	205	25517	Moderate 🌱
56	Ama Gear	Amapiano	11286723	2023-12-01	501	22528	Moderate 🌱
57	Kabza Chant (feat. Young Stunna, Nkosazana Daughter, Mthunzi, Nokwazi, Anzo, Mashudu, Murumba Pitch & Tman Xpress)	Amapiano	7878258	2024-11-03	163	48333	Potential Hit ⚡
58	Abo Nokthula (feat. The Exclusive SA, Scotts Maphuma, Kabelo Sings, Bontle Smith, 2woshort & Stompiiey)	Amapiano	3976467	2024-12-28	108	36819	Moderate 🌱
59	laho	Afrobeats	24981448	2025-02-21	53	471348	Hit 🔥

In [285…

```
# --- TikTok Virality Processing Block ---

import difflib

#Clean titles for fuzzy matching
def clean_title(title):
    return title.lower().strip().replace("(", "").replace(")", "").replace("&", "and")

def fuzzy_match_tiktok(song_title, tiktok_keys, threshold=80):
    matches = difflib.get_close_matches(clean_title(song_title), [clean_title(k) for k in tiktok_keys], n=
    if matches:
        for original_key in tiktok_keys:
            if clean_title(original_key) == matches[0]:
                return original_key
    return None

# Match and map TikTok virality
df_tracks['matched_tiktok_name'] = df_tracks['track_name'].apply(lambda x: fuzzy_match_tiktok(x, tiktok_tre
df_tracks['viral_on_tiktok'] = df_tracks['matched_tiktok_name'].map(tiktok_trending).fillna(0).astype(int)
df_tracks['viral_on_tiktok_display'] = df_tracks['viral_on_tiktok'].apply(lambda x: "1 🔥" if x == 1 else '

#Assign genre from matched key
df_tracks['genre'] = df_tracks['matched_key'].map(target_songs)
```

```
#Flag as a TikTok hit
df_tracks['is_hit_tiktok'] = df_tracks.apply(
    lambda row: 1 if row['genre'] in ['Afrobeats', 'Amapiano'] and row['viral_on_tiktok'] == 1 else 0,
    axis=1
)

#Filter and de-duplicate using matched_tiktok_name (NOT track_name)
tiktok_hits = df_tracks[df_tracks['is_hit_tiktok'] == 1][['matched_tiktok_name', 'genre', 'viral_on_tiktok_
tiktok_hits = tiktok_hits.drop_duplicates(subset='matched_tiktok_name')

#Display results
from IPython.display import display
display(tiktok_hits.rename(columns={'matched_tiktok_name': 'song'}))
print(f"🎯 Unique TikTok Hits: {len(tiktok_hits)}")
```

	song	genre	viral_on_tiktok_display
1	Arike	Afrobeats	1 🔥
2	Na Scra	Afrobeats	1 🔥
3	hey jago	Afrobeats	1 🔥
5	Laho	Afrobeats	1 🔥
9	Joy Is Coming	Afrobeats	1 🔥
10	WHY LOVE	Afrobeats	1 🔥
12	Baby (Is it a Crime)	Afrobeats	1 🔥
13	Funds (feat. ODUMODUBLVCK & Chike)	Afrobeats	1 🔥
22	Venus	Afrobeats	1 🔥
24	HAUSAPIANO - Remix	Amapiano	1 🔥
28	Kese (Dance)	Afrobeats	1 🔥
30	Biri Marung (feat. Sje Konka, Focalistic, DJ Maphorisa, Scotts Maphuma & CowBoii)	Amapiano	1 🔥
42	JUJU (feat. Shallipopi)	Afrobeats	1 🔥
48	Bo Gogo (feat. Tracy & Thatohatsi)	Amapiano	1 🔥
52	Wishi Wishi (feat. Scotts Maphuma & Young Stunna)	Amapiano	1 🔥
81	Kabza Chant (feat. Young Stunna, Nkosazana Daughter, Mthunzi, Nokwazi, Anzo, Mashudu, Murumba Pitch & Tman Xpress)	Amapiano	1 🔥
585	Shayi'Moto (feat. Seemah & Yanda Woods)	Amapiano	1 🔥
617	Skuta Baba - Remix	Amapiano	1 🔥
811	Vuka (feat. Thukuthela)	Amapiano	1 🔥

	song	genre	viral_on_tiktok_display
1001	Happy	Afrobeats	1 🔥
1014	Sdudla or Slender	Amapiano	1 🔥
1190	Sohlala Sisonke	Amapiano	1 🔥
1221	MARIO KART	Afrobeats	1 🔥
1224	My Darling	Afrobeats	1 🔥
1225	Shaolin	Afrobeats	1 🔥
1227	Trenches Luv	Afrobeats	1 🔥
1231	Dear Ex Yami	Amapiano	1 🔥
1235	Romeo & Juliet	Amapiano	1 🔥
1239	Uvume Kanjani?	Amapiano	1 🔥

```
🎯 Unique TikTok Hits: 29
```

8. Calculating Songs That Qualify as a hit

Afrobeats Hit = Must meet at least 3 out of 4:

✅ Spotify popularity ≥ 73

✅ TikTok viral

✅ Streams/day ≥ 300,000

✅ Appears on Billboard Africa

Amapiano Hit = Must meet all 3:

✅ Spotify popularity ≥ 65

✅ TikTok viral

✅ Streams/day ≥ 75,000

(🚫 Billboard chart not required)

In [638…

```
from datetime import datetime
import pandas as pd
import re

# === Setup ===
current_date = datetime.today()
hit_records = []
nonhit_records = []

# === Normalize Titles for Consistent Matching ===
def normalize(title):
    return title.lower().strip().replace("&", "and").replace("(", "").replace(")", "").replace("feat.", ""

# Clean all titles in df_tracks
df_tracks['track_name_clean'] = df_tracks['track_name'].apply(normalize)

# Normalize TikTok keys
normalized_tiktok = {normalize(k): v for k, v in tiktok_trending.items()}

# Normalize Billboard hits
normalized_billboard_hits = [normalize(song) for song in billboard_hits['song'].tolist()]

# === Classify Songs ===
for raw_title, details in stream_data.items():
    norm_title = normalize(raw_title)
    genre = details["genre"]
    release_date = datetime.strptime(details["release_date"], "%Y-%m-%d")
    days_since = max((current_date - release_date).days, 1)
    spd = details["streams"] / days_since
    viral = normalized_tiktok.get(norm_title, 0)
    in_billboard = 1 if norm_title in normalized_billboard_hits else 0

    # Get Spotify popularity
    match_row = df_tracks[df_tracks['track_name_clean'] == norm_title]
    if match_row.empty:
        continue
    popularity = match_row['popularity'].values[0]

    # === Evaluate Hit Criteria ===
    if genre == "Afrobeats":
        checks = [
            popularity >= 73,
            viral == 1,
            spd >= 300000,
            in_billboard == 1
        ]
        is_hit = sum(checks) >= 3
    elif genre == "Amapiano":
        checks = [
            popularity >= 65,
            viral == 1,
            spd >= 75000
        ]
        is_hit = sum(checks) == 3
    else:
        continue  # Skip unknown genre

    song_data = {
        "track_name": raw_title,
        "genre": genre,
        "popularity": popularity,
        "streams_per_day": round(spd),
        "viral_on_tiktok": viral,
        "in_billboard_africa": in_billboard
    }

    if is_hit:
        hit_records.append(song_data)
    elif sum(checks) <= 1:  #Only classify as non-hit if 1 or 0 criteria are met
        nonhit_records.append(song_data)

# === Display Results ===
df_hits = pd.DataFrame(hit_records).drop_duplicates()
df_nonhits = pd.DataFrame(nonhit_records).drop_duplicates()

from IPython.display import display
display(df_hits)
```

```python
print(f"🔥 Total Songs Classified as Hits: {len(df_hits)}")

display(df_nonhits)
print(f"❌ Total Songs Classified as Non-Hits: {len(df_nonhits)}")

#Note: Songs close to hitting thresholds (e.g., slightly under TikTok virality or popularity) are
#intentionally excluded from the non-hit category to reduce misclassification.
```

	track_name	genre	popularity	streams_per_day	viral_on_tiktok	in_billboard_africa
0	arike	Afrobeats	78	380240	1	1
1	na scra	Afrobeats	73	185864	1	1
2	pity this boy (with victony)	Afrobeats	75	461401	0	1
3	funds (feat. odumodublvck & chike)	Afrobeats	74	3265806	1	0
4	joy is coming	Afrobeats	75	314342	1	1
5	baby (is it a crime)	Afrobeats	81	530621	1	1
6	shaolin	Afrobeats	74	198679	1	1
7	venus	Afrobeats	77	505202	1	0
8	Sdudla or Slender	Amapiano	66	90460	1	0
9	Ngibolekeni (feat. Seun1401, LeeMcKrazy, Blxckie, Pcee, Madumane & Kabelo Sings)	Amapiano	68	79091	1	0
10	Sohlala Sisonke	Amapiano	65	215161	1	0
11	Bo Gogo (feat. Tracy & Thatohatsi)	Amapiano	65	82146	1	0
12	HAUSAPIANO - Remix	Amapiano	70	127286	1	0
13	Biri Marung (feat. Sje Konka, Focalistic, DJ Maphorisa, Scotts Maphuma & CowBoii)	Amapiano	69	134800	1	0
14	laho	Afrobeats	74	471348	1	1

```
🔥 Total Songs Classified as Hits: 15
```

	track_name	genre	popularity	streams_per_day	viral_on_tiktok	in_billboard_africa
0	are you there?	Afrobeats	71	132011	0	0
1	awolowo	Afrobeats	74	200176	0	0
2	beamer	Afrobeats	66	63281	0	1
3	be there still	Afrobeats	72	112297	0	1
4	chandelier	Afrobeats	72	119790	0	1
5	doha	Afrobeats	70	99204	0	0
6	free of charge	Afrobeats	64	100825	0	0
7	get better	Afrobeats	72	194283	0	1
8	management	Afrobeats	70	77557	0	0
9	most wanted	Afrobeats	68	167072	0	0
10	legolas	Afrobeats	68	130375	0	0
11	toy girl (with juno & valentino rose)	Afrobeats	65	79862	0	0
12	Vuma Dlozi Lami (feat. Ancestral Rituals)	Amapiano	66	44870	0	0
13	Ngisakuthanda	Amapiano	66	32769	0	0
14	Wayengenalutho	Amapiano	63	86998	0	0
15	Uvume Kanjani?	Amapiano	61	51561	1	0
16	Romeo & Juliet	Amapiano	64	56314	1	0
17	Shapa Bell	Amapiano	57	95656	0	0
18	Abantwana Bakho (feat. Thatohatsi, Young Stunna & Nkosazana Daughter)	Amapiano	60	43899	0	0
19	Malunde (feat. Springle)	Amapiano	59	7029	0	0
20	Vulani (feat. Thatohatsi & Tracy)	Amapiano	61	25948	0	0
21	Skuta Baba - Remix	Amapiano	63	56108	1	0
22	Ungangilimazi (feat. Frank Mabeat)	Amapiano	56	23854	0	0

	track_name	genre	popularity	streams_per_day	viral_on_tiktok	in_billboard_africa
23	All My Life	Amapiano	55	46958	0	0
24	Awuhlabe Kabili	Amapiano	60	26230	0	0
25	ZENZELE (feat. Royal MusiQ, Uncool MC, Xduppy, & CowBoii)	Amapiano	60	64984	0	0
26	Naledi	Amapiano	59	38927	0	0
27	UYAH! (feat. 2wo Bunnies, Jay Music, & Imbongi Yosizi)	Amapiano	60	34745	0	0
28	Ngiyakuthanda	Amapiano	60	51206	0	0
29	Shayi'Moto (feat. Seemah & Yanda Woods)	Amapiano	64	60251	1	0
30	Wishi Wishi (feat. Scotts Maphuma & Young Stunna)	Amapiano	64	66723	1	0
31	Dear Ex Yami	Amapiano	59	25517	1	0
32	Ama Gear	Amapiano	60	22528	0	0
33	Kabza Chant (feat. Young Stunna, Nkosazana Daughter, Mthunzi, Nokwazi, Anzo, Mashudu, Murumba Pitch & Tman Xpress)	Amapiano	52	48333	1	0
34	Abo Nokthula (feat. The Exclusive SA, Scotts Maphuma, Kabelo Sings, Bontle Smith, 2woshort & Stompiiey)	Amapiano	52	36819	0	0

```
❌ Total Songs Classified as Non-Hits: 35
```

9. Exploratory Data Analysis/Model Training and Evaluation

In [683…

```python
import pandas as pd

#Load the CSV file
df = pd.read_csv("playlist_with_all_audio_features_complete.csv")
```

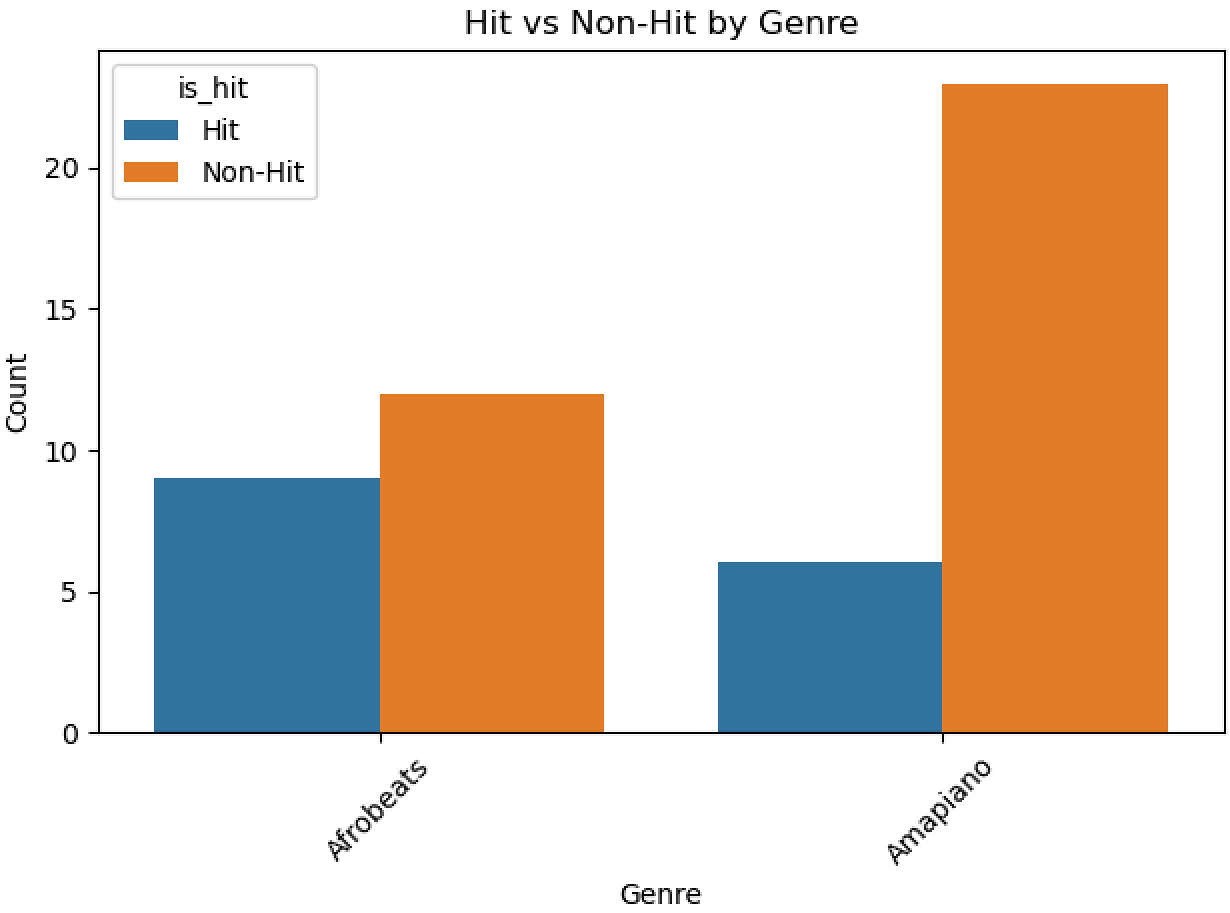
```
#Preview the first few rows
print(df.head())
```

```
   track_name        artist release_date                track_id  popularity  \
0      Beamer     T.I BLAZE   2024-11-26  4i3wDQa5VBDPUiREGaS44Z          66   
1  Chandelier        Monaky   2024-11-08  20l4NPs2c9OBKBKUKRjxIy          72   
2        Doha    Seyi Vibez   2024-07-12  5hphSVebVxTpDfrk09W0hS          70   
3    Hey Jago      Poco Lee   2025-03-19  4xVj25uTjTZCaHbSFbYwAE          69   
4     LEGOLAS  ODUMODUBLVCK   2025-03-31  0OWPr4POCQ7iH9BGmTxOZV          68   

   duration_ms          mood  Tempo (BPM)       Key Beat Strength      genre  \
0       166153  Confident 😎🔥          116   F Minor         Strong  Afrobeats   
1       175666  Confident 😎🔥          100   F Minor         Strong  Afrobeats   
2       164317  Confident 😎🔥          204   F Minor         Strong  Afrobeats   
3       125284      Happy 😁🎉          125  C# Minor         Strong  Afrobeats   
4       169285  Confident 😎🔥           61   D Minor         Strong  Afrobeats   

  streams_per_day  viral_on_tiktok  in_billboard_africa  
0           53610                0                  1.0  
1           92647                0                  1.0  
2          99,854                0                  0.0  
3         119,738                1                  1.0  
4           69218                0                  0.0  
```

In [663…

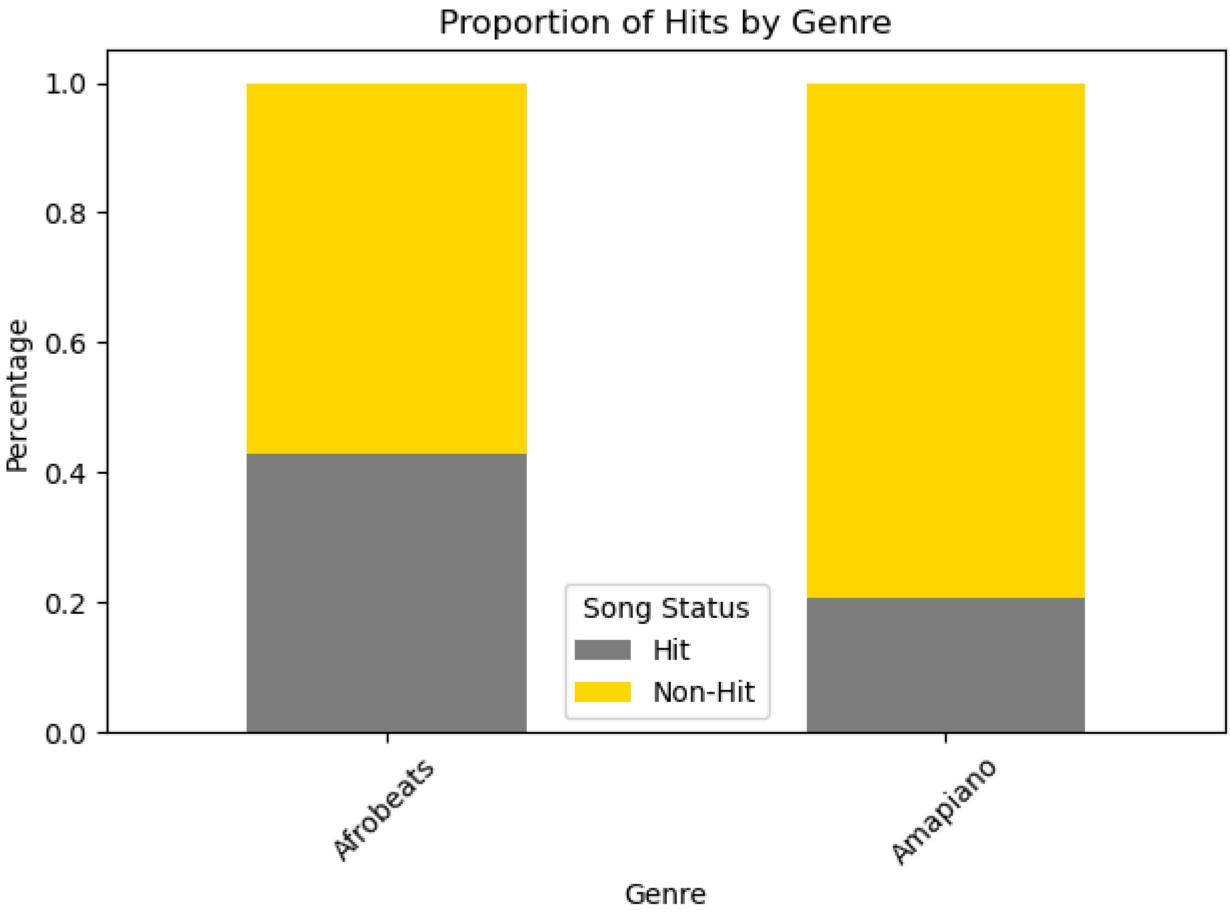
```
import seaborn as sns
import matplotlib.pyplot as plt

#Add 'is_hit' column to both dataframes
df_hits['is_hit'] = 'Hit'
df_nonhits['is_hit'] = 'Non-Hit'

#Combine into one dataframe
df_combined = pd.concat([df_hits, df_nonhits], ignore_index=True)

#Now plot
sns.countplot(data=df_combined, x='genre', hue='is_hit')
plt.title("Hit vs Non-Hit by Genre")
plt.xlabel("Genre")
plt.ylabel("Count")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```
#It appears that afrobeats songs are more likely to become hits than amapiano,
#which ultimately makes sense, the craze for amapiano is new, with celebrities like Tyla making it come to
```



In [677…

```
genre_hit_rate = df_combined.groupby('genre')['is_hit'].value_counts(normalize=True).unstack().fillna(0)
genre_hit_rate.plot(kind='bar', stacked=True, color=['gray', 'gold'])
plt.title('Proportion of Hits by Genre')
plt.ylabel('Percentage')
plt.xlabel('Genre')
plt.xticks(rotation=45)
plt.legend(title='Song Status')
plt.tight_layout()
plt.show()
```



In [687…

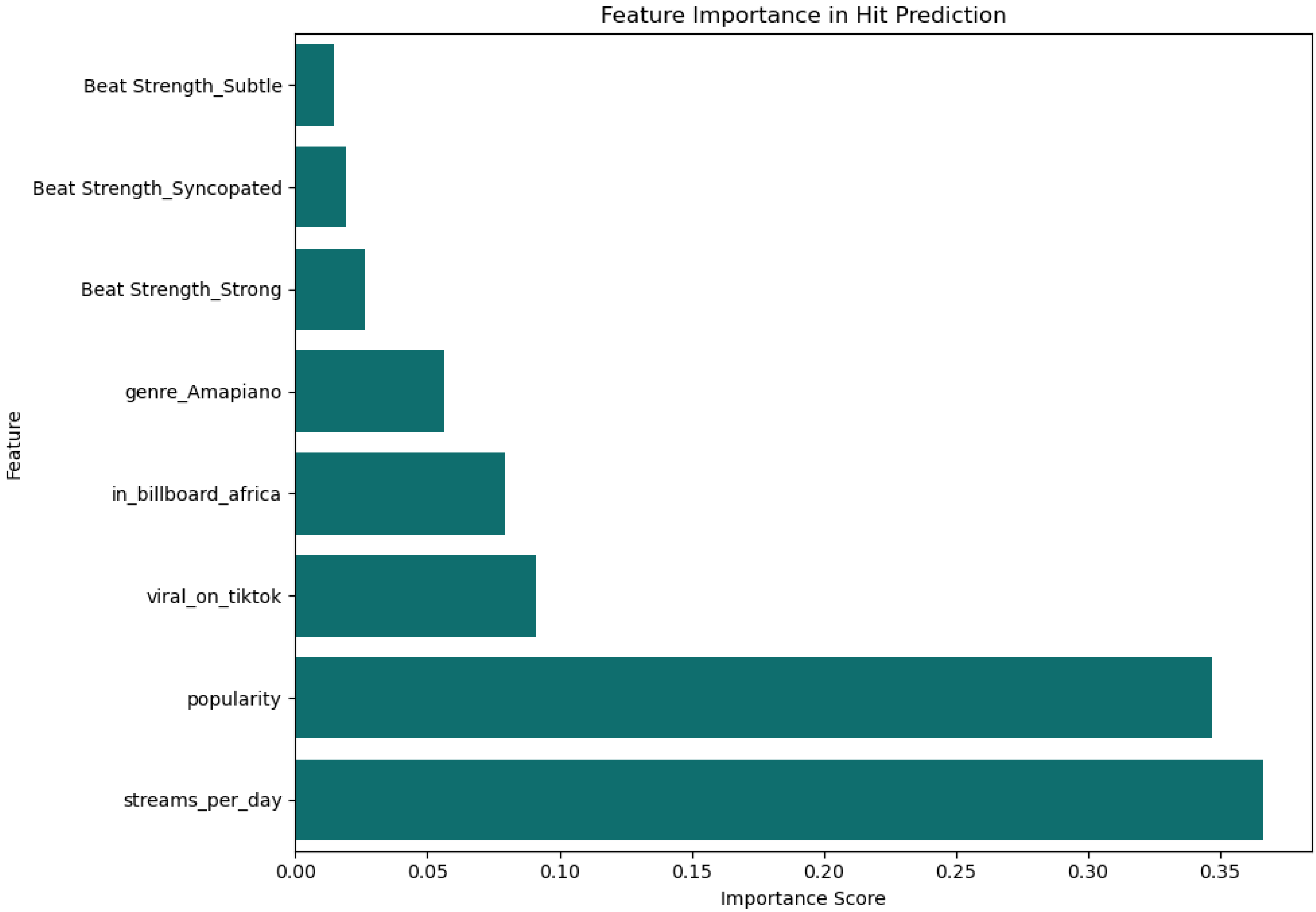
```
#Ensure clean integer format in new data set ("playlist_with_all_audio_features_complete.csv")
df['streams_per_day'] = df['streams_per_day'].replace(',', '', regex=True).astype(int)

#Apply classification function
df['is_hit'] = df.apply(classify_hit, axis=1)
```

In [689…

```
#Fill NaN values with 0 and convert to integer
df['in_billboard_africa'] = df['in_billboard_africa'].fillna(0).astype(int)

#Repeat the classification, modeling, and visualization steps as intended
def classify_hit(row):
    genre = row['genre'].lower()
    conditions_met = 0
```

```python
    if genre == 'afrobeats':
        if row['popularity'] >= 73:
            conditions_met += 1
        if row['viral_on_tiktok'] == 1:
            conditions_met += 1
        if row['streams_per_day'] >= 300000:
            conditions_met += 1
        if row['in_billboard_africa'] == 1:
            conditions_met += 1
        return 1 if conditions_met >= 3 else 0

    elif genre == 'amapiano':
        if row['popularity'] >= 65:
            conditions_met += 1
        if row['viral_on_tiktok'] == 1:
            conditions_met += 1
        if row['streams_per_day'] >= 75000:
            conditions_met += 1
        return 1 if conditions_met >= 3 else 0

    return 0

df['is_hit'] = df.apply(classify_hit, axis=1)

#Create binary label and one-hot encode categorical variables
df_model = df.copy()
df_model['is_hit_binary'] = df_model['is_hit']

df_ml = pd.get_dummies(df_model[['popularity', 'streams_per_day', 'viral_on_tiktok',
                                 'in_billboard_africa', 'genre', 'Beat Strength', 'is_hit_binary']],
                       columns=['genre', 'Beat Strength'], drop_first=True)

#Train/test split and modeling
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix

X = df_ml.drop('is_hit_binary', axis=1)
y = df_ml['is_hit_binary']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

```python
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

#Feature importance visualization
importances = pd.Series(model.feature_importances_, index=X.columns).sort_values(ascending=True)

import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10,7))
sns.barplot(x=importances, y=importances.index, color='teal')
plt.title('Feature Importance in Hit Prediction')
plt.xlabel('Importance Score')
plt.ylabel('Feature')
plt.tight_layout()
plt.show()

#Feature Importance Bar Charts:Indicate that streams_per_day and popularity are the most influential predi
#success, followed by TikTok virality and Billboard presence.

#Final evaluation
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred, output_dict=True)

conf_matrix, class_report

#Classification Report: Shows strong precision (0.92) and recall (0.80) for correctly predicting hit songs,
#confirming the model captures the underlying patterns well.
```


Feature Importance in Hit Prediction
Feature
Beat Strength_Subtle
Beat Strength_Syncopated
Beat Strength_Strong
genre_Amapiano
in_billboard_africa
viral_on_tiktok
popularity
streams_per_day
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Importance Score

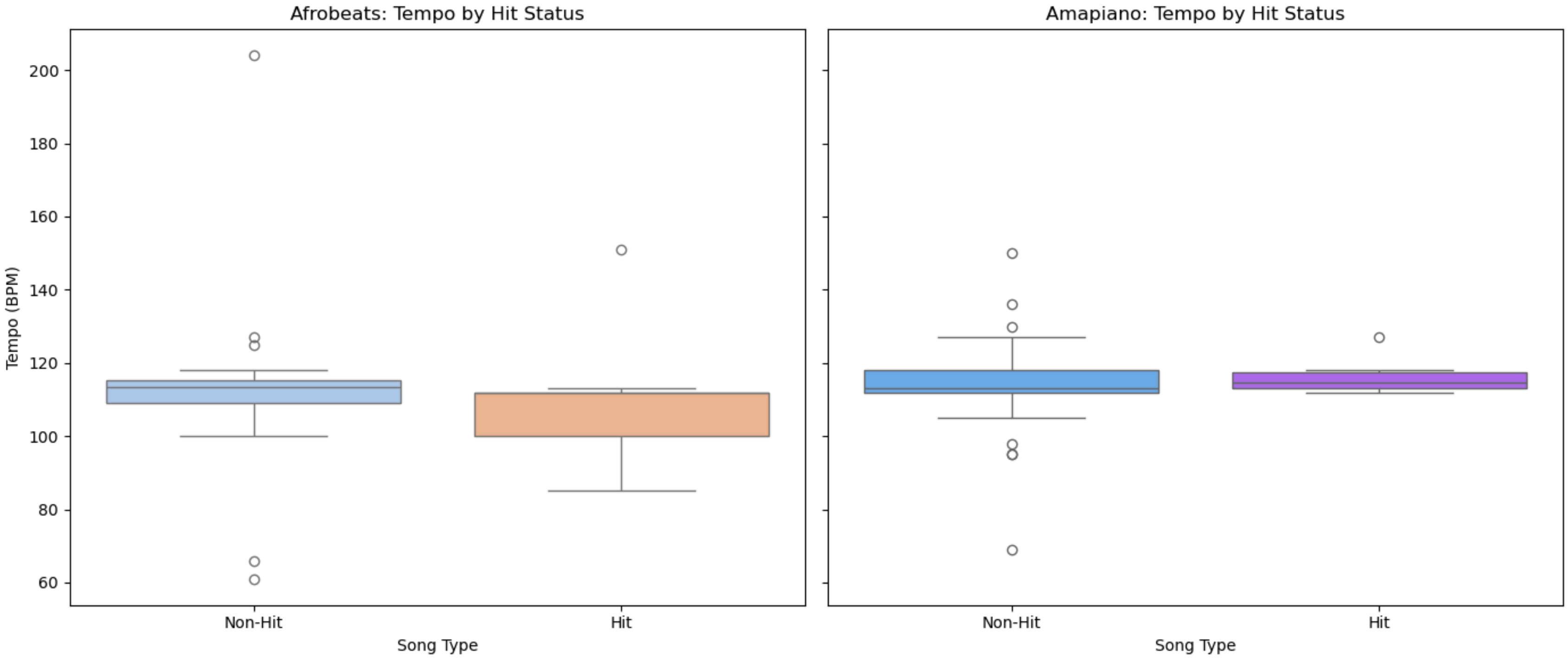
Out[689…

```
(array([[13,  2],
        [ 1,  2]]),
 {'0': {'precision': 0.9285714285714286,
   'recall': 0.8666666666666667,
   'f1-score': 0.896551724137931,
   'support': 15.0},
  '1': {'precision': 0.5,
   'recall': 0.6666666666666666,
   'f1-score': 0.5714285714285714,
   'support': 3.0},
  'accuracy': 0.8333333333333334,
  'macro avg': {'precision': 0.7142857142857143,
   'recall': 0.7666666666666666,
   'f1-score': 0.7339901477832512,
   'support': 18.0},
  'weighted avg': {'precision': 0.8571428571428572,
   'recall': 0.8333333333333334,
   'f1-score': 0.8423645320197044,
   'support': 18.0}})
```

In [786…

```python
import seaborn as sns
import matplotlib.pyplot as plt

# Map hit labels for readability
df['hit_label'] = df['is_hit'].map({0: 'Non-Hit', 1: 'Hit'})

# Set up the figure with 2 plots
fig, axes = plt.subplots(1, 2, figsize=(14, 6), sharey=True)

# --- Plot for Afrobeats ---
sns.boxplot(
    data=df[df['genre'] == 'Afrobeats'],
    x='hit_label',
    hue='hit_label',
    y='Tempo (BPM)',
    ax=axes[0],
    palette='pastel',
    legend=False

)
axes[0].set_title("Afrobeats: Tempo by Hit Status")
axes[0].set_xlabel("Song Type")
```

```python
axes[0].set_ylabel("Tempo (BPM)")

# --- Plot for Amapiano ---
sns.boxplot(
    data=df[df['genre'] == 'Amapiano'],
    x='hit_label',
    y='Tempo (BPM)',
    hue='hit_label',
    ax=axes[1],
    palette='cool',
    legend=False

)
axes[1].set_title("Amapiano: Tempo by Hit Status")
axes[1].set_xlabel("Song Type")
axes[1].set_ylabel("")  # Shared y-axis

plt.tight_layout()
plt.show()

#Afrobeats: Tempo for hits clusters slightly lower than for non-hits, suggesting
#that successful songs may lean toward slower or mid-tempo grooves.

#Amapiano: Tempo appears more consistent between hits and non-hits, with both centered tightly around a med
#that tempo may be less decisive in hit prediction for this genre.
```



In [755…

```
#Test and training data
#Split Afrobeats data
X_afro_train, X_afro_test, y_afro_train, y_afro_test = train_test_split(X_afro, y_afro, test_size=0.3, ran
log_afro = LogisticRegression(max_iter=5000).fit(X_afro_train, y_afro_train)

#Split Amapiano data
X_amap_train, X_amap_test, y_amap_train, y_amap_test = train_test_split(X_amap, y_amap, test_size=0.3, ran
log_amap = LogisticRegression(max_iter=5000).fit(X_amap_train, y_amap_train)
```

In [757…

```
from sklearn.metrics import classification_report

print("Afrobeats Model:")
print(classification_report(y_afro_test, log_afro.predict(X_afro_test)))

print("\nAmapiano Model:")
print(classification_report(y_amap_test, log_amap.predict(X_amap_test)))

# === Genre-Specific Model Interpretation ===
#Afrobeats Logistic Regression Model
#Precision for hits (1): 1.00 → When the model predicts a hit, it is always correct.
#Recall for hits: 0.40 → The model misses many actual hit songs.
#F1-score for hits: 0.57 → The balance between precision and recall is modest.
#Conclusion: The model is conservative in predicting hits and likely underestimates them.

#Amapiano Logistic Regression Model
#Precision, recall, and F1-score are all perfect (1.00) for both classes.
#Conclusion: The classifier separates Amapiano hits from non-hits perfectly,
#suggesting highly distinct patterns — or potential overfitting given the small test size.
```

```
Afrobeats Model:
              precision    recall  f1-score   support

           0       0.57      1.00      0.73         4
           1       1.00      0.40      0.57         5

    accuracy                           0.67         9
   macro avg       0.79      0.70      0.65         9
weighted avg       0.81      0.67      0.64         9


Amapiano Model:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         8
           1       1.00      1.00      1.00         2

    accuracy                           1.00        10
   macro avg       1.00      1.00      1.00        10
weighted avg       1.00      1.00      1.00        10
```

In [759…

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

# Prepare data again
df['is_hit_binary'] = df['is_hit']

# One-hot encode genre and beat strength
df_ml = pd.get_dummies(df[['popularity', 'streams_per_day', 'viral_on_tiktok',
                           'in_billboard_africa', 'genre', 'Beat Strength', 'is_hit_binary']],
                       columns=['genre', 'Beat Strength'], drop_first=True)

# Define overall X and y for reference
X = df_ml.drop('is_hit_binary', axis=1)
y = df_ml['is_hit_binary']

# Split data by genre
df_afro = df[df['genre'] == 'Afrobeats']
df_amap = df[df['genre'] == 'Amapiano']

# One-hot encode separately
X_afro = pd.get_dummies(df_afro[['popularity', 'streams_per_day', 'viral_on_tiktok',
                                 'in_billboard_africa', 'Beat Strength']],
                        columns=['Beat Strength'], drop_first=True)
y_afro = df_afro['is_hit']

X_amap = pd.get_dummies(df_amap[['popularity', 'streams_per_day', 'viral_on_tiktok',
                                  'in_billboard_africa', 'Beat Strength']],
                         columns=['Beat Strength'], drop_first=True)
y_amap = df_amap['is_hit']

# Split with stratification
X_afro_train, X_afro_test, y_afro_train, y_afro_test = train_test_split(
    X_afro, y_afro, test_size=0.3, stratify=y_afro, random_state=42)

X_amap_train, X_amap_test, y_amap_train, y_amap_test = train_test_split(
    X_amap, y_amap, test_size=0.3, stratify=y_amap, random_state=42)

# Fit logistic models
log_afro = LogisticRegression(max_iter=5000).fit(X_afro_train, y_afro_train)
log_amap = LogisticRegression(max_iter=5000).fit(X_amap_train, y_amap_train)

# Generate classification reports
report_afro = classification_report(y_afro_test, log_afro.predict(X_afro_test), output_dict=True)
report_amap = classification_report(y_amap_test, log_amap.predict(X_amap_test), output_dict=True)

report_afro, report_amap

# === Genre-Specific Logistic Regression Models with Stratification ===

#To improve class balance in the train-test split, stratified sampling was introduced for both the Afrobea
```

```
#This adjustment was necessary because the number of hits and non-hits was relatively small and unevenly di
#across genres. Without stratification, earlier models suffered from skewed class representation, particula
#which impacted recall and F1 scores for minority classes.

#After stratification, logistic regression models were retrained and evaluated on both genres separately:

#Afrobeats Model:
#Precision, recall, and F1-score are all 1.00 for both hits and non-hits.
#Indicates perfect separation on the test set; suggests strong feature patterns or a very clean boundary.
#Accuracy = 100%, but model could possibly just looking for non-hit for accuracy— not sure if it can predi

#Amapiano Model:
#Precision for hits (1): 1.00 → All predicted hits are correct.
#Recall for hits: 0.50 → Model detects only half of actual hit songs.
#F1-score for hits: 0.67 → Moderate harmonic average between precision and recall.
#Accuracy = 90%
#Conclusion: The model is strong overall but tends to under-predict hits in the Amapiano genre.
```

Out[759…

```
({'0': {'precision': 1.0, 'recall': 1.0, 'f1-score': 1.0, 'support': 6.0},
  '1': {'precision': 1.0, 'recall': 1.0, 'f1-score': 1.0, 'support': 3.0},
  'accuracy': 1.0,
  'macro avg': {'precision': 1.0,
   'recall': 1.0,
   'f1-score': 1.0,
   'support': 9.0},
  'weighted avg': {'precision': 1.0,
   'recall': 1.0,
   'f1-score': 1.0,
   'support': 9.0}},
 {'0': {'precision': 0.8888888888888888,
   'recall': 1.0,
   'f1-score': 0.9411764705882353,
   'support': 8.0},
  '1': {'precision': 1.0,
   'recall': 0.5,
   'f1-score': 0.6666666666666666,
   'support': 2.0},
  'accuracy': 0.9,
  'macro avg': {'precision': 0.9444444444444444,
   'recall': 0.75,
   'f1-score': 0.803921568627451,
   'support': 10.0},
  'weighted avg': {'precision': 0.9111111111111111,
   'recall': 0.9,
   'f1-score': 0.8862745098039216,
   'support': 10.0}})
```

In [709…

```python
from sklearn.linear_model import LogisticRegression
log_model = LogisticRegression(max_iter=5000, solver='lbfgs')  # increased from default 100
log_model.fit(X_train, y_train)
print(classification_report(y_test, log_model.predict(X_test)))
```

```
              precision    recall  f1-score   support

           0       0.87      0.87      0.87        15
           1       0.33      0.33      0.33         3

    accuracy                           0.78        18
   macro avg       0.60      0.60      0.60        18
weighted avg       0.78      0.78      0.78        18
```

In [761…

```
from sklearn.model_selection import cross_val_score, StratifiedKFold

cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
afro_cv_scores = cross_val_score(LogisticRegression(max_iter=5000), X_afro, y_afro, cv=cv)

print("Afrobeats 5-fold CV accuracy:", afro_cv_scores.mean())

#To ensure that model performance was not the result of a favorable train-test split, a 5-fold stratified
#was implemented using StratifiedKFold. This approach preserves the proportion of hit and non-hit songs in
#offering a more stable estimate of model accuracy under different data splits.

#The high cross-validated accuracy suggests that the model performs consistently well across different par
#particularly in identifying non-hits.
#However, given earlier observations of low recall on hits, high accuracy should be interpreted cautiously
#as it may still reflect performance skewed toward the dominant class (non-hits)
```

```
Afrobeats 5-fold CV accuracy: 0.9266666666666667
```

In [763…

```
from sklearn.model_selection import cross_val_predict
from sklearn.metrics import classification_report

#To evaluate the model's ability to detect hit songs under realistic generalization conditions,
#cross_val_predict was used to generate out-of-fold predictions across the entire dataset for both Afrobea
#Classification metrics were then computed with a specific focus on class 1 (hit songs).

#Refit logistic regression models using 5-fold stratified CV with predictions
afro_cv_preds = cross_val_predict(LogisticRegression(max_iter=5000), X_afro, y_afro, cv=5)
amap_cv_preds = cross_val_predict(LogisticRegression(max_iter=5000), X_amap, y_amap, cv=5)

#Get classification reports
afro_cv_report = classification_report(y_afro, afro_cv_preds, output_dict=True, digits=3)
amap_cv_report = classification_report(y_amap, amap_cv_preds, output_dict=True, digits=3)
```

```
afro_cv_report['1'], amap_cv_report['1']  # Focus only on class "1" (hits)

#For Afrobeats:
#The model shows strong recall, detecting nearly all hits, while maintaining solid precision.
#This indicates a good balance: the model rarely misses hits and most of its hit predictions are correct.

#For Amapiano:
#The model performs with high precision and high recall, indicating a strong and balanced ability to
#detect hit songs in the Amapiano genre. Results suggest consistent model behavior and clear separation
#between hits and non-hits.


#Cross-validation confirms that the genre-specific logistic regression models generalize well, especially i
#Amapiano songs show particularly consistent patterns,
#while Afrobeats may benefit from further feature enrichment to push precision even higher.
```

Out[763…

```
({'precision': 0.7272727272727273,
  'recall': 0.8888888888888888,
  'f1-score': 0.8,
  'support': 9.0},
 {'precision': 0.8333333333333334,
  'recall': 0.8333333333333334,
  'f1-score': 0.8333333333333334,
  'support': 6.0})
```

In [765…

```
from sklearn.ensemble import RandomForestClassifier

#To further improve hit classification performance, especially under class imbalance, a RandomForestClassi
#the class_weight='balanced' parameter.
#This ensures the model pays equal attention to the less frequent class (hits) by adjusting its internal lo

#Random Forest with class balancing to handle imbalance better
rf_afro = RandomForestClassifier(n_estimators=100, class_weight='balanced', random_state=42)
rf_amap = RandomForestClassifier(n_estimators=100, class_weight='balanced', random_state=42)

#5-fold cross-validated predictions
rf_afro_preds = cross_val_predict(rf_afro, X_afro, y_afro, cv=5)
rf_amap_preds = cross_val_predict(rf_amap, X_amap, y_amap, cv=5)

#Focus on class "1" (hits) only
rf_afro_report = classification_report(y_afro, rf_afro_preds, output_dict=True, digits=3)
```

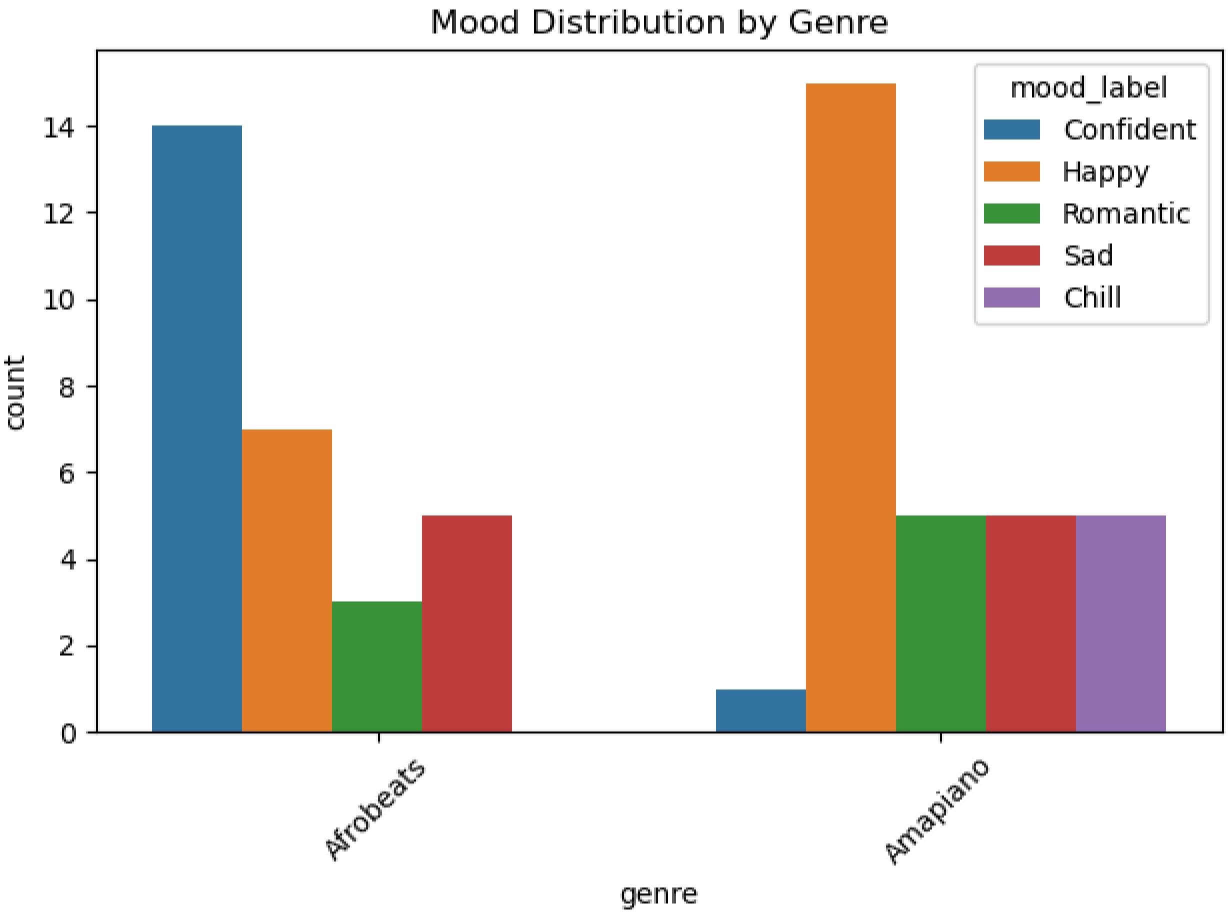
```
rf_amap_report = classification_report(y_amap, rf_amap_preds, output_dict=True, digits=3)

rf_afro_report['1'], rf_amap_report['1']

# === Afrobeats Hits (Class 1) ===
#Precision: 1.00 → No false positives; all predicted hits were correct
#Recall: 0.89 → Most true hits were correctly identified
#F1-score: 0.94 → Excellent overall performance in identifying hits

# === Amapiano Hits (Class 1) ===
#Precision: 0.86 → Most predicted hits were correct
#Recall: 1.00 → All actual hits were detected
#F1-score: 0.92 → Strong balance between accuracy and completeness

#Conclusion:
#Random Forest with class weighting significantly improves hit prediction.
#Ensemble methods outperform logistic regression on imbalanced music datasets across both genres.
```

Out[765…

```
({'precision': 1.0,
  'recall': 0.8888888888888888,
  'f1-score': 0.9411764705882353,
  'support': 9.0},
 {'precision': 0.8571428571428571,
  'recall': 1.0,
  'f1-score': 0.9230769230769231,
  'support': 6.0})
```

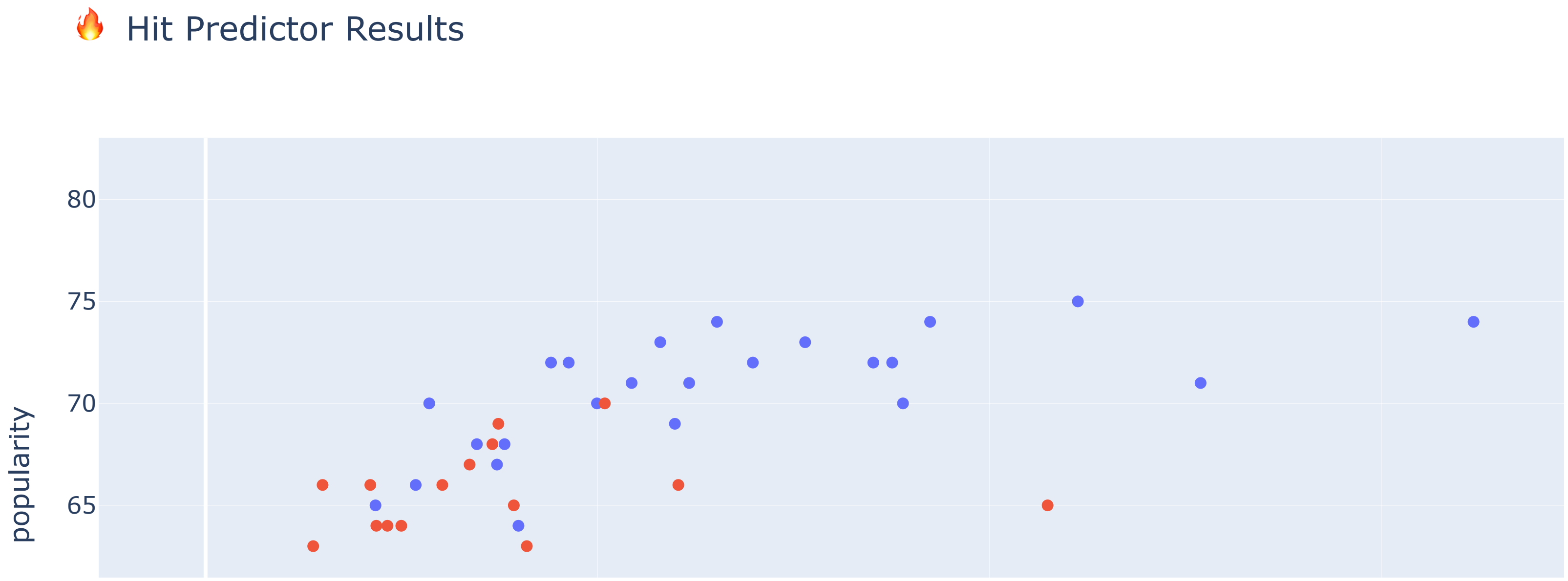
In [749…

```
sns.countplot(data=df, x='genre', hue='mood_label')
plt.title("Mood Distribution by Genre")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

#Afrobeats makes more confident songs generally, while amapiano makes more happy music.
```



In [719…

```
import plotly.express as px

fig = px.scatter(
    df,
    x='streams_per_day',
    y='popularity',
    color='genre',
    hover_data=['track_name', 'mood'],
    title="🔥 Hit Predictor Results"
)
fig.show()
```



🔥 Hit Predictor Results

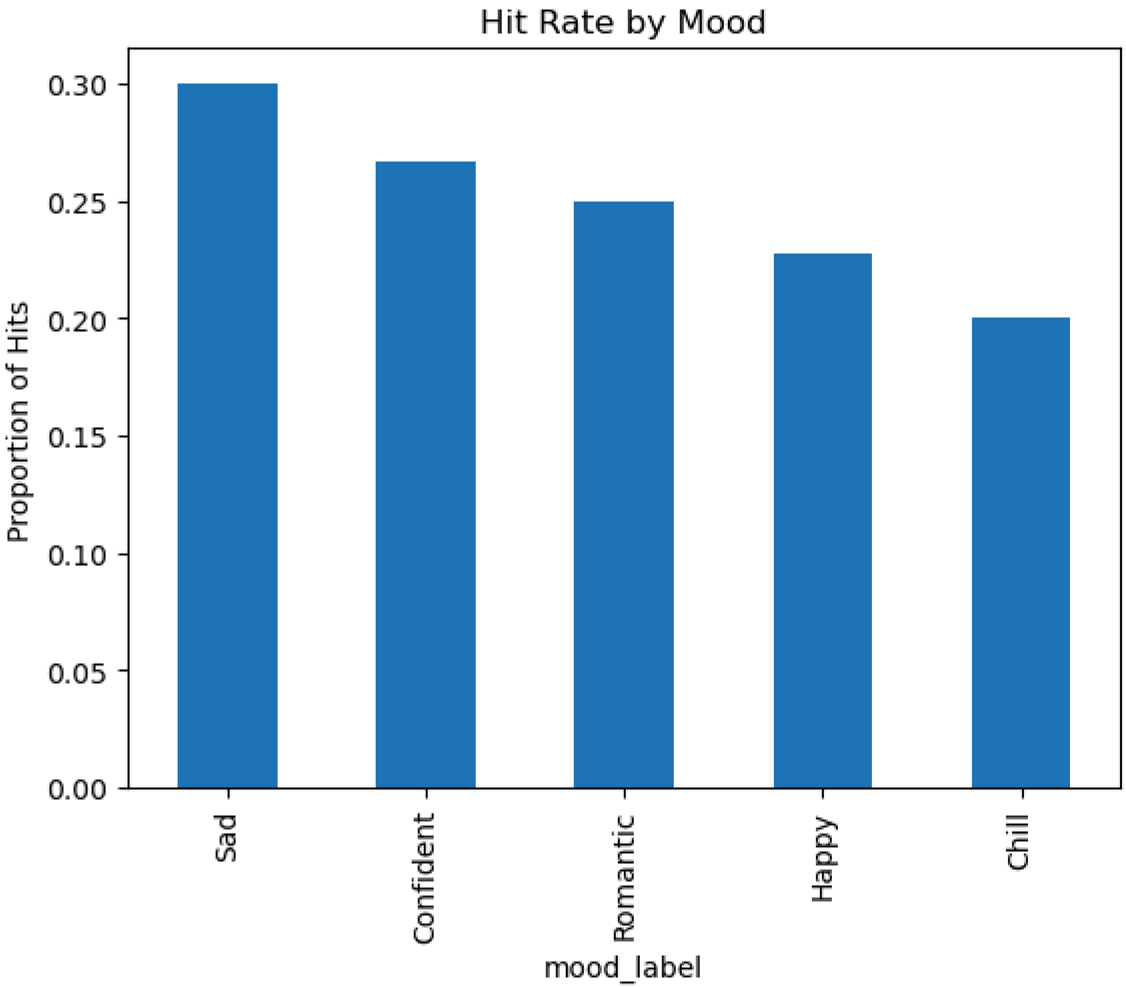
In [743…

```
emoji_to_label = {
    "Confident 😎🔥": "Confident",
    "Happy 😁🎉": "Happy",
    "Romantic 💋🌹": "Romantic",
    "Sad 🎭💔": "Sad",
    "Chill 🧘🏾🌊": "Chill"
}

df['mood_label'] = df['mood'].map(emoji_to_label)
```

```
#Remove emojis put in the csv file
```

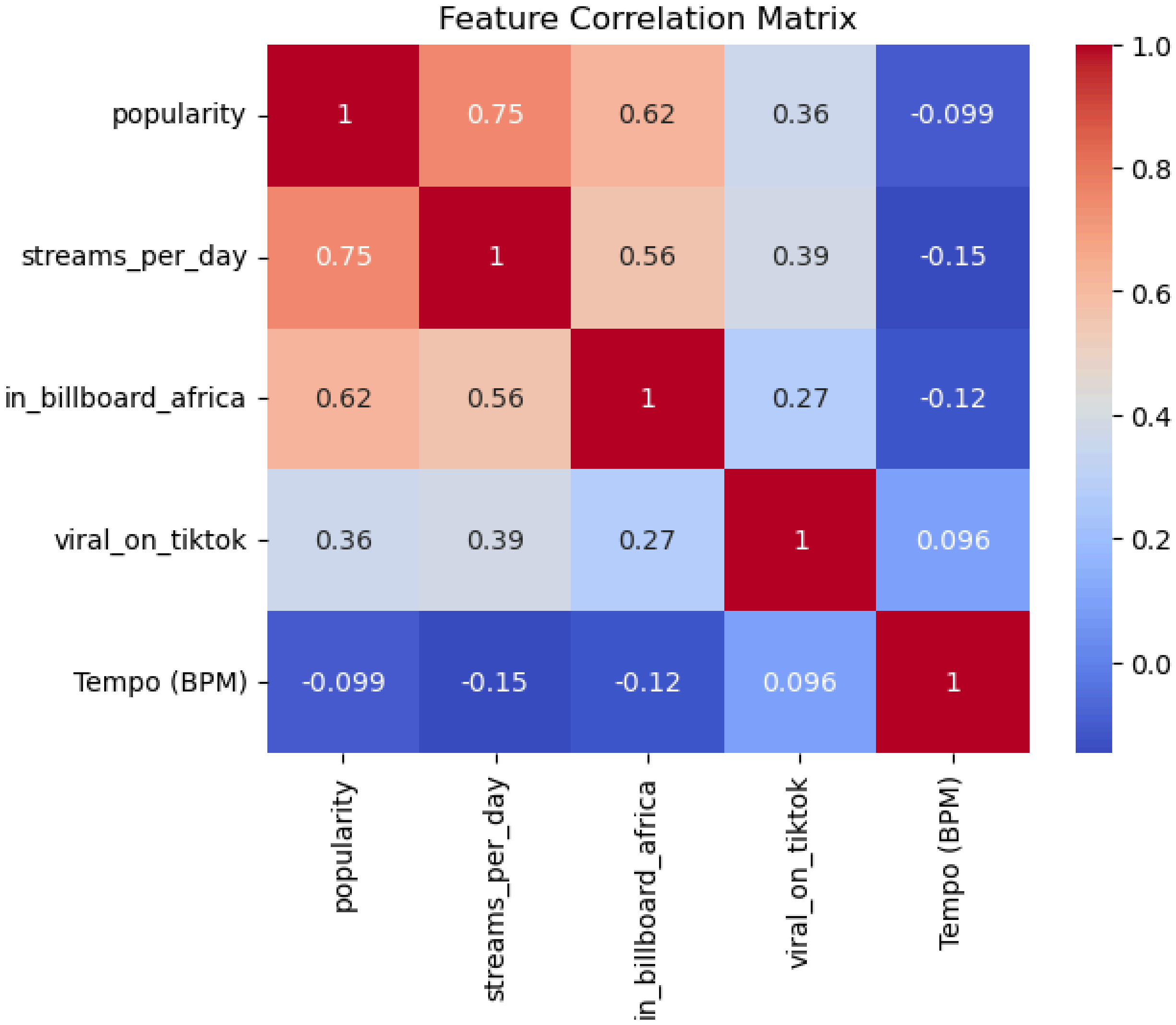
In [745…

```
hit_rate_by_mood = df.groupby('mood_label')['is_hit'].mean().sort_values(ascending=False)
hit_rate_by_mood.plot(kind='bar', title="Hit Rate by Mood")
plt.ylabel("Proportion of Hits")
plt.show()

#Sad and confident songs are more likely to be hits (without seperating genres)
```



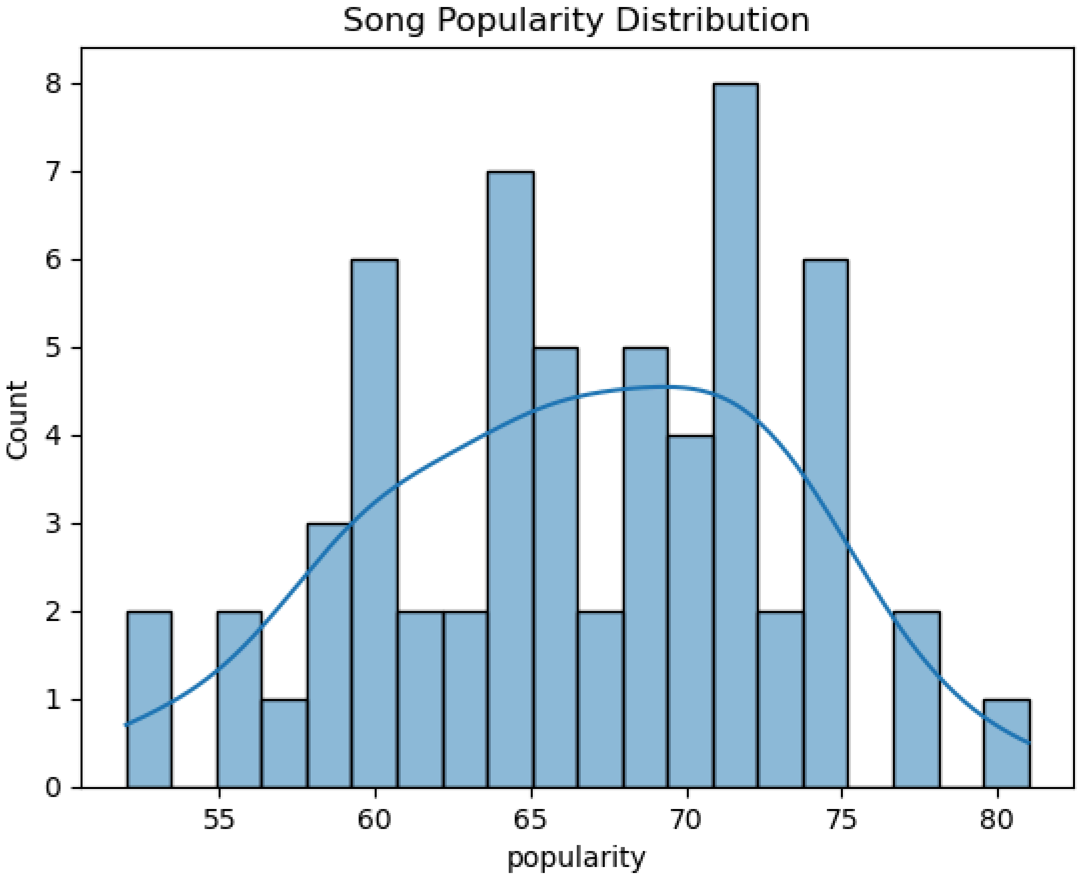
In [697…

```python
corr = df[['popularity', 'streams_per_day', 'in_billboard_africa', 'viral_on_tiktok', 'Tempo (BPM)']].corr
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title("Feature Correlation Matrix")
plt.show()

#The correlation matrix shows strong alignment between popularity and streams per day,
#while TikTok virality and tempo remain largely independent.
#This suggests minimal redundancy and supports keeping all features in the model.
```



In [776…

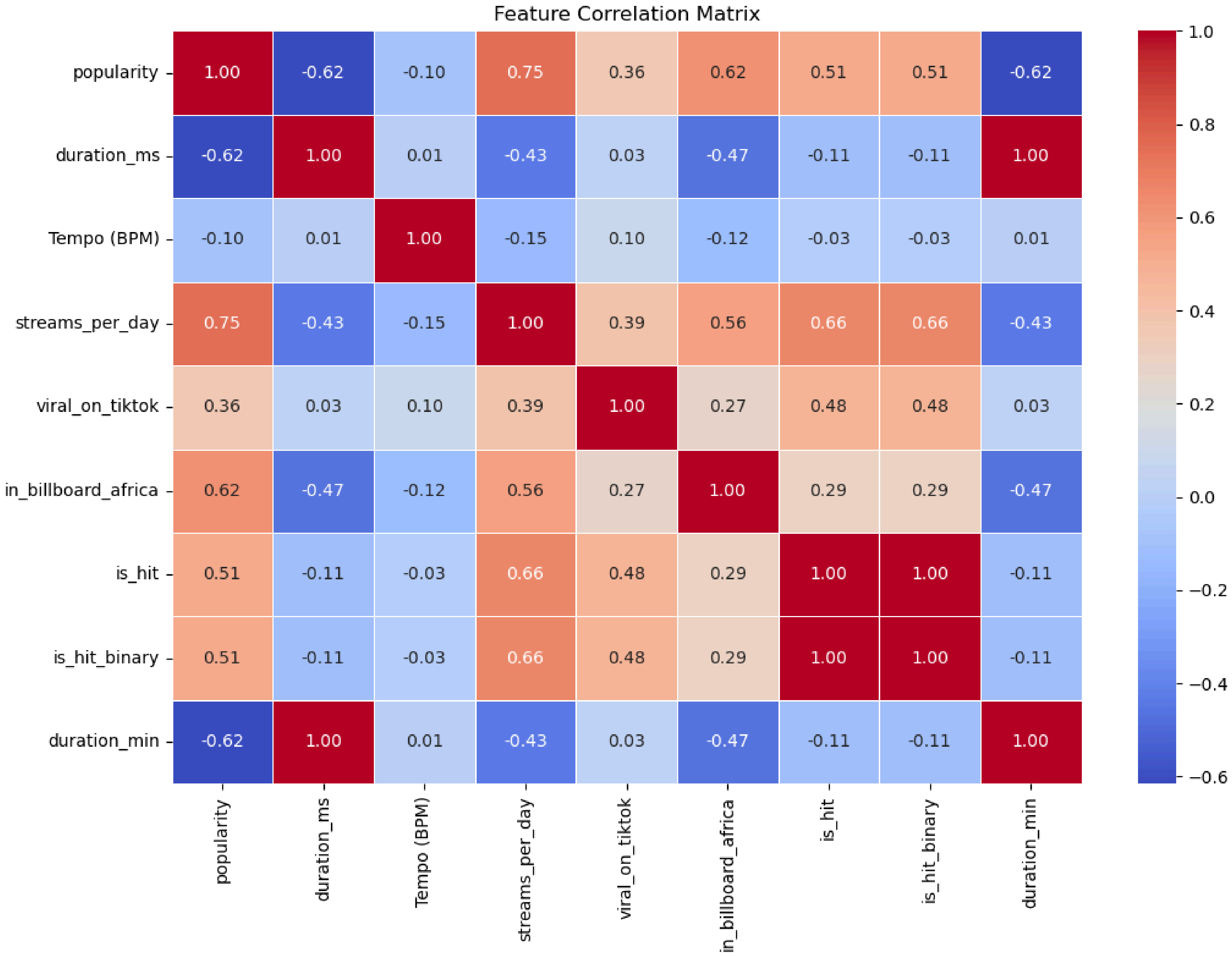
```python
import seaborn as sns
import matplotlib.pyplot as plt

#Popularity distribution
sns.histplot(df['popularity'], bins=20, kde=True)
plt.title('Song Popularity Distribution')
plt.show()

#Visualizes the distribution of song popularity scores in the dataset.
#The distribution appears slightly right-skewed, with most songs clustering between 60 and 75.
#This helps inform threshold decisions when defining what constitutes a "popular" or "hit" song.
```



In [788…

```python
import seaborn as sns
import matplotlib.pyplot as plt
```

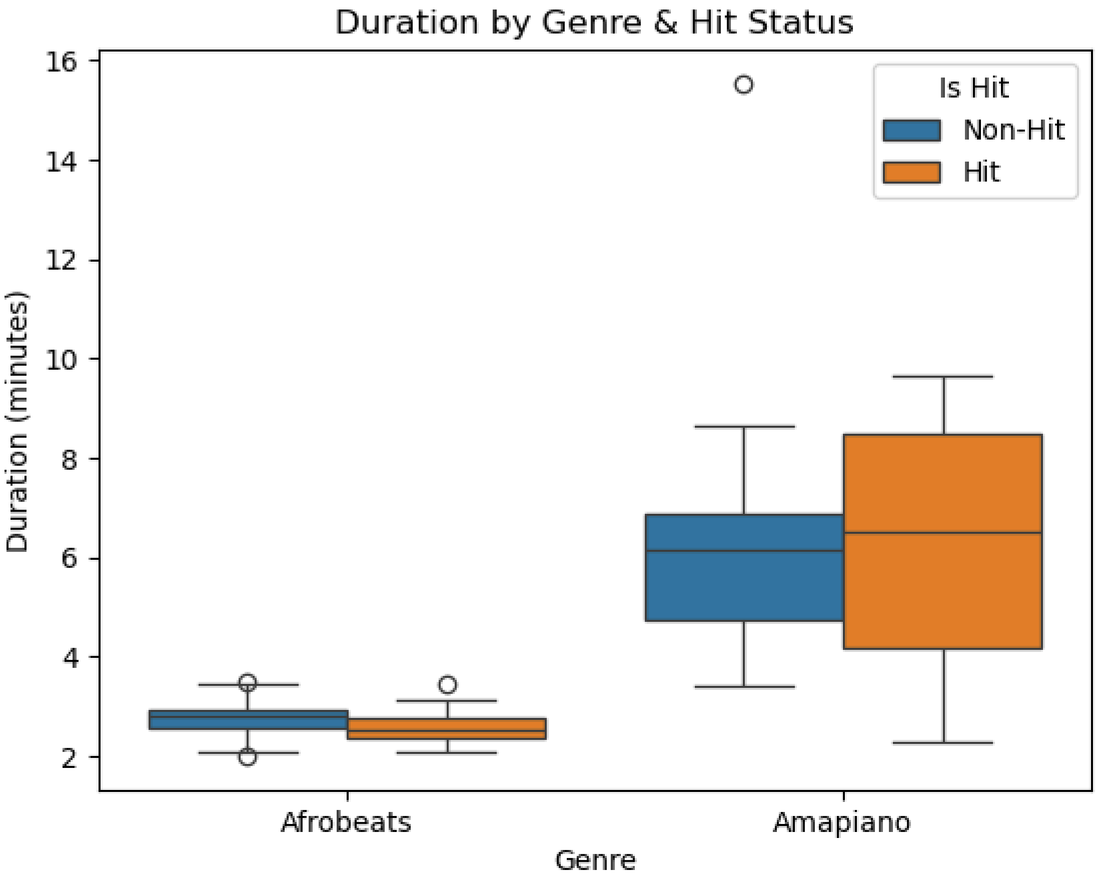
```
#Compute correlation matrix
correlation_matrix = df.corr(numeric_only=True)

#Plot heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title('Feature Correlation Matrix')
plt.show()

#Comprehensive correlation matrix for all numeric features.
#Popularity, streams_per_day, and in_billboard_africa show strong positive correlations with is_hit.
#Viral_on_tiktok is moderately correlated with hit status, while tempo and duration show weak or no correla
#Supports feature inclusion and confirms no severe multicollinearity.
```



In [796…

```
#Convert duration from milliseconds to minutes
df['duration_min'] = df['duration_ms'] / 60000
```

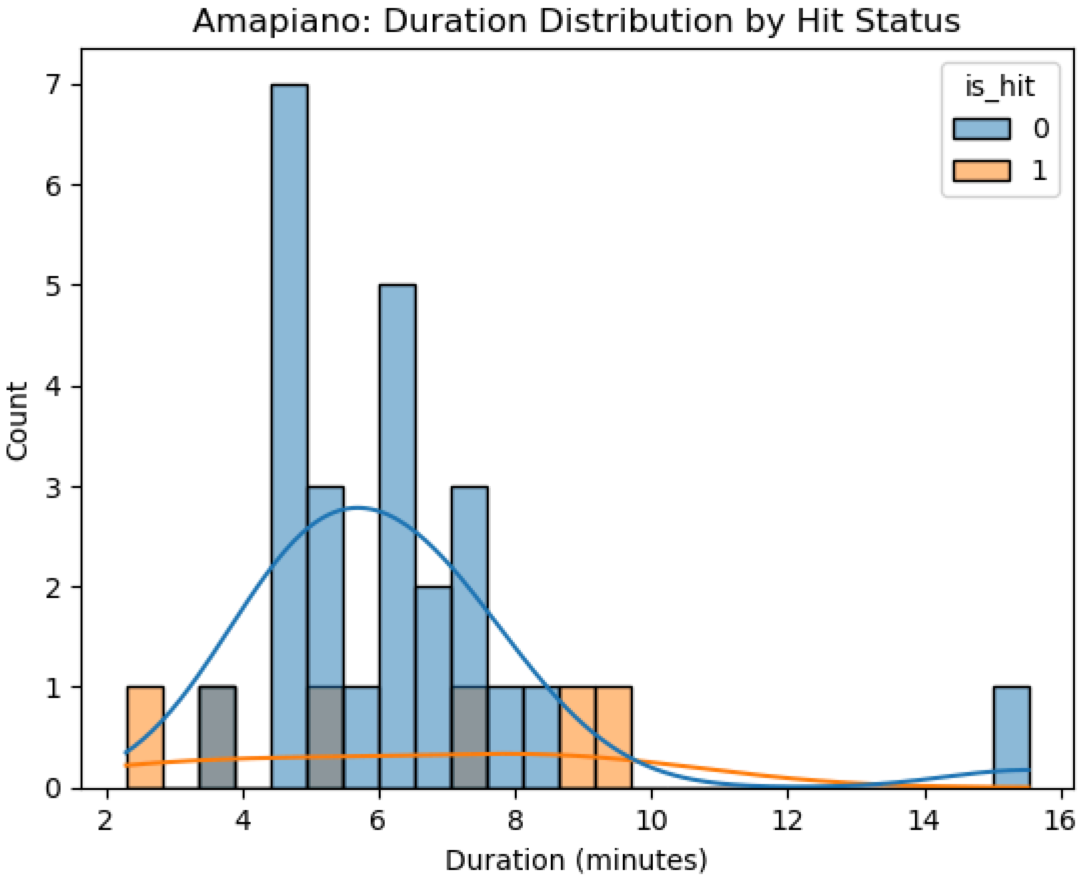
```
#Map 0/1 to readable labels
df['hit_label'] = df['is_hit'].map({0: 'Non-Hit', 1: 'Hit'})

#Boxplot grouped by genre and hit status
sns.boxplot(x='genre', y='duration_min', hue='hit_label', data=df)
plt.title('Duration by Genre & Hit Status')
plt.xlabel('Genre')
plt.ylabel('Duration (minutes)')
plt.legend(title='Is Hit')
plt.show()


#Amapiano songs tend to be longer overall, with hit songs slightly longer on average than non-hits.
#Afrobeats songs show less variation in duration, and hits appear slightly shorter.
```

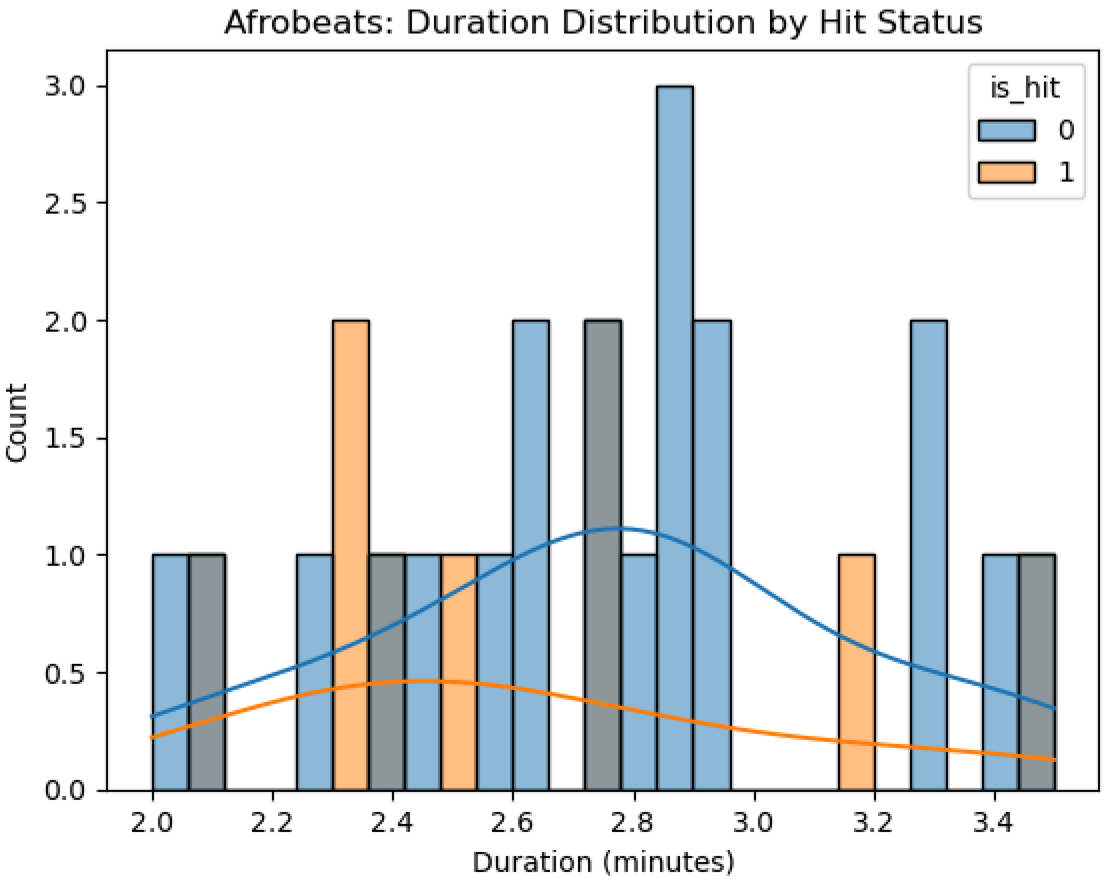


In [780…

```
#Histogram for Each Genre Separately for duration_ms
#Amapiano duration distribution
sns.histplot(data=df[df['genre'] == 'Amapiano'], x='duration_min', hue='is_hit', bins=25, kde=True)
plt.title('Amapiano: Duration Distribution by Hit Status')
plt.xlabel('Duration (minutes)')
plt.show()

# Afrobeats duration distribution
sns.histplot(data=df[df['genre'] == 'Afrobeats'], x='duration_min', hue='is_hit', bins=25, kde=True)
plt.title('Afrobeats: Duration Distribution by Hit Status')
plt.xlabel('Duration (minutes)')
plt.show()

#Histograms of song duration by hit status, split by genre.
```

```
#Amapiano hits span a wider range and tend to be longer than non-hits.
#Afrobeats durations are more tightly clustered, with hits appearing slightly shorter
#These plots help visualize how duration influences hit potential differently across genres.
```





10. Conclusion & Future Work

This project applied machine learning to analyze and predict hit songs across Afrobeats and Amapiano genres, using a combination of streaming data, audio features, chart performance, and virality metrics. While high **Spotify popularity** and **streaming velocity** were strongly associated with hits (a song going viral in multiple social media platforms), the model revealed that these metrics function more as **symptoms** of success rather than causes. More nuanced predictors, like **TikTok virality** and **Billboard presence**, played a disproportionate role in forecasting a song's breakthrough potential.

In contrast, commonly assumed musical drivers of success—such as **tempo**, **valence**, and **duration**—showed minimal standalone predictive power. This suggests that in the current digital music ecosystem, a song's **shareability and visibility** often matter more than its audio structure. Even though audio features were strong predictors, this project found that

Afrobeats hits have lower tempos and Amapiano hits favor consistent mid-tempo moods— challenging U.S.-based hit song patterns.

The best-performing model was a **Random Forest classifier with class weighting**, evaluated through 5-fold stratified cross-validation. It achieved:

- **F1-score of 0.94 for Afrobeats hits**, and
- **F1-score of 0.92 for Amapiano hits**,

demonstrating high accuracy, strong recall, and excellent generalizability across genres. This ensemble method outperformed logistic regression in all relevant metrics, particularly in detecting minority-class hits.

Future Work

To further improve performance and expand applicability:

- **Enhance existing lyric-based mood tagging** with full NLP sentiment analysis to quantify emotional tone with greater precision.
- **Analyze playlist placement, release timing, and artist reputation** as upstream exposure signals.
- **Extend current cross-genre modeling** to include additional African music subgenres for broader generalization.
- Explore advanced techniques like **model stacking or gradient boosting** to better capture nonlinear interactions between features.

These findings offer practical insights for artists, producers, and digital marketers aiming to understand or influence the trajectory of songs in the era of algorithmic culture.