VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY-
UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY

March 10, 2025

Data Visualization

Report lab 01

Lecturers and Teaching assistants :

1. Mr. Bui Tien Len
2. Mr. Vo Nhat Tan

Students:

1. **Quach Tran Quan Vinh - 22127460**
2. **Nguyen Hoang Trung Kien - 22127478**

Contents

1 Project and team progress

1.1 Overall progress

#	**Tasks**	**Progress**
1	Data collection	100%
2	Data preprocessing	100%
3	Data visualization	100%

1.2 Team progress

#	**Student's ID**	**Name**	**Tasks**	**Progress**
1	22127460	Quach Tran Quan Vinh	Data collection	100%
2	22127478	Nguyen Hoang Trung Kien	Data preprocessing	100%
3	22127478	Nguyen Hoang Trung Kien	Data exploration	100%
4	22127478	Nguyen Hoang Trung Kien	Solve analysis objectives 1 and 2	100%
5	22127460	Quach Tran Quan Vinh	Solve analysis objectives 3 and 4	100%
6	22127460 and 22127478	Quach Tran Quan Vinh and Nguyen Hoang Trung Kien	Write report	100%

2 Implement steps

2.1 Importing libraries

Libraries to handle and visualize data.

- `matplotlib`, `seaborn` - used to visualize data.
- `pandas` - used to process data in table.
- `numpy` - interact with array more effieciently.
- `re` - process text and string.

Libraries to collect data.

- `requests` - used to request a html request of a webpage, in order to crawl content of the API.
- `sys` - this library we only used `exit` function to stop the program.
- `dotenv` - used to load `.env` file.
- `os` - used along with `dotenv` to extract the API key from `.env` file.

2.2 Data collection

Data are collected using Chartmetric API. The collected data are crawled randomly using Get tracks (with filters) methods on Chartmetric API with default parameters.

2.3 Data exploration

The dataset has 10000 rows and 28 columns. Each row represents a track record and attributes. The columns are as follows:

Column Name	**Description**
`id`	Chartmetric's Track ID
`track_name`	Track's name
`album_name`	Album Name
`release_date`	Track's release date
`artist`	Array of artists related to the track
`genre`	Track genres
`explicit`	Whether the track is explicit
`score`	Chartmetric's track score
`airplay_streams`	Airplay streams
`amazon_playlist_count`	Number of Amazon playlists that include this track
`deezer_playlist_count`	Number of Deezer playlists that include this track
`itunes_playlist_count`	Number of iTunes playlists that include this track

Column Name	Description
shazam_count	Number of Shazam plays
siriusxm_streams	Number of SiriusXM plays
soundcloud_plays	Number of SoundCloud plays
spotify_playlist_count	Number of Spotify playlists that include this track
spotify_playlist_total_reach	Number of Spotify playlists total reach
spotify_plays	Number of Spotify plays
spotify_popularity	Track's Spotify popularity
spotify_ed_playlist_count	Number of Spotify editorial playlists that include this track
spotify_ed_playlist_total_reach	Number of Spotify editorial playlists total reach
tidal_popularity	Track's Tidal popularity
tiktok_posts	TikTok posts featuring this track
tiktok_top_videos_likes	Sum of likes of the top TikTok videos featuring this artist
tiktok_top_videos_views	Sum of views of the top TikTok videos featuring this artist
youtube_likes	Number of YouTube likes
youtube_playlist_count	Number of YouTube playlists that include this track
youtube_views	Number of YouTube views

2.4 Data preprocessing

Firstly, we handled duplicates values. The data has 29 duplicates, so we drop them. Then, we dealt with missing values, here is the number of missing values of each columns:

Column Name	Number of missing values
id	0
track_name	0
album_name	1
release_date	0
artist	0
genre	0
explicit	224
score	0
airplay_streams	1629
amazon_playlist_count	3092
deezer_playlist_count	2717
itunes_playlist_count	1563
shazam_count	1857
siriusxm_streams	5635
soundcloud_plays	7557
spotify_playlist_count	170

Column Name	Missing Values
spotify_playlist_total_reach	170
spotify_plays	216
spotify_popularity	460
spotify_ed_playlist_count	170
spotify_ed_playlist_total_reach	170
tidal_popularity	9971
tiktok_posts	2037
tiktok_top_videos_likes	2804
tiktok_top_videos_views	2805
youtube_likes	1270
youtube_playlist_count	3235
youtube_views	1261

We can see that there are some columns which have high quantity of missing records. We dropped the columns that have over 2000 missing values. That are: amazon_playlist-_count, deezer_playlist_count, siriusxm_streams, soundcloud_plays, tidal_popularity, tiktok_posts, tiktok_top_videos_likes, tiktok_top_videos_views, youtube_playlist_count. Also, Column id is not really meaningful for further analysis so we removed it as well. For categorical columns, there are 2 columns that have missing values. They are album_name and explicit. Explicit can be filled by mode since it only contains 1 of 2 values true/false. However, album_name cannot be determined by mode because it will not be true and reliable based on the data's domain knowledge which misrepresents the data. So the rows with missing value in album_name column were eliminated.
We then checked the distribution of each numeric columns to decide which method of handling missing values would be appropriate.

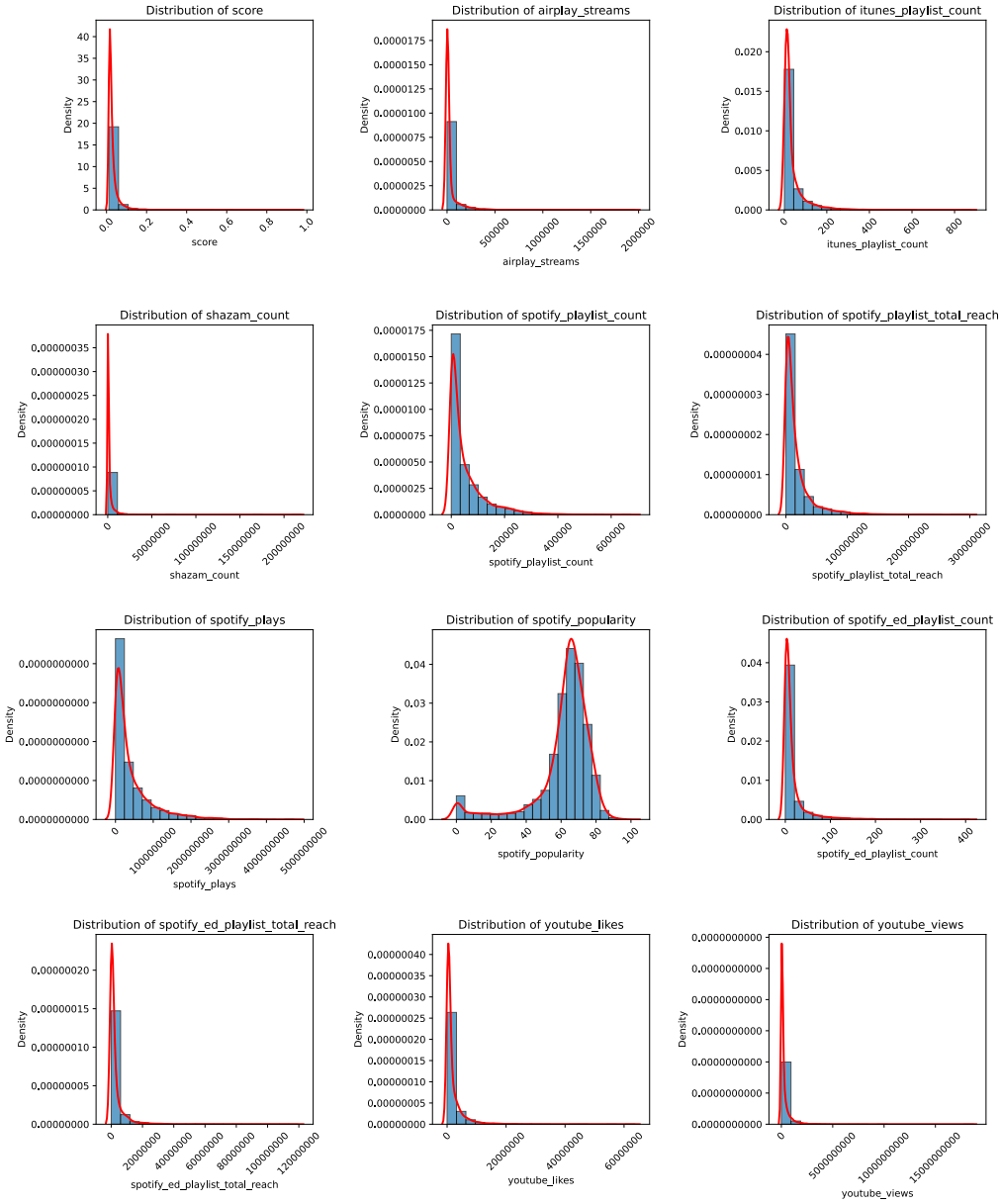

Figure 1: Distribution of numeric columns

From the plot above we can see that the numeric columns have heavily right-skewed distribution because of some outliers. To handle this we filled missing values with median for simplicity since the columns have low number of missing values so use this filling method will not effect the distribution.
Then, we converted data types of each columns into appropriate type. All columns that has float type (except score) were converted into int type since they represent count, likes, views, etc. Column release_date was also converted into datetime type.
The columns artist and genre stores value as list, but the data treated them as string type, for ease use, we removed the characters '[]' and "" in them.

2.5 Data visualization and analysis

2.5.1 Visualization objective 1:

Question: How do Pop tracks perform compared to other genres on Spotify from 2020 to 2024?
Benefits:

- Helps the artists and producers release tracks that can increase their engagement, keep up with the trends and reach a wider audience.
- Enhances playlists, keeps listeners engaged, and recommends the best mix of Pop songs.

Columns used in data: release_date, genre, spotify_plays, spotify_popularity
Visualization used: Pie chart and Line graph.
Reasons:

- For pie chart: It effectively visualizes the proportion of pop tracks relative to all tracks, making it easy to compare their quantity over the 2020-2024 period.
- For line graph: It clearly shows trends and differences in average Spotify plays and popularity across genres, making it easy to compare their performance over time.

Analyzing and Visualizing steps:

First, we identify the Pop tracks in the dataset. We extract the tracks that have Pop in genre column from 2020 to 2024.

We plot the proportion of pop tracks over all tracks in 2020-2024 period using pie chart.

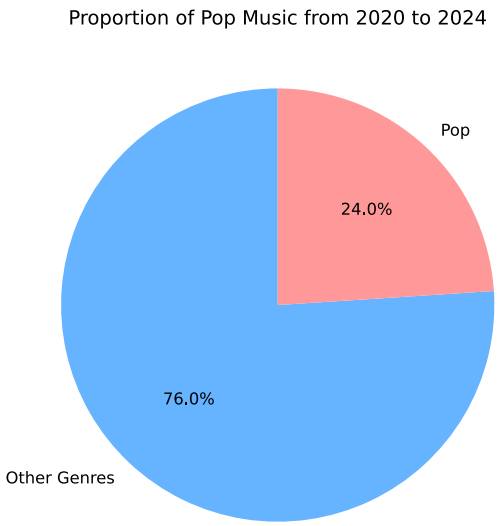

Figure 2: Proportion of Pop Music from 2020 to 2024

We can see that Pop genre has a moderate proportion which is 24% of all tracks in 2020-2024 period. This indicates that while Pop plays a significant role for music genre, it is not the one that dominates the music platform. Music audience tend to listen to various kinds of genre not only Pop.

Next, we analyze Pop music's performance on Spotify by comparing it with other genre performance. We extract the year from release_date column and classify genre as either "Pop" or "Other". Then, we calculate mean of spotify plays and popularity group by year and genre. After that, we divide dataset into pop and other genre dataset.

We visualize the genre comparison, based on average spotify plays and popularity by line chart.

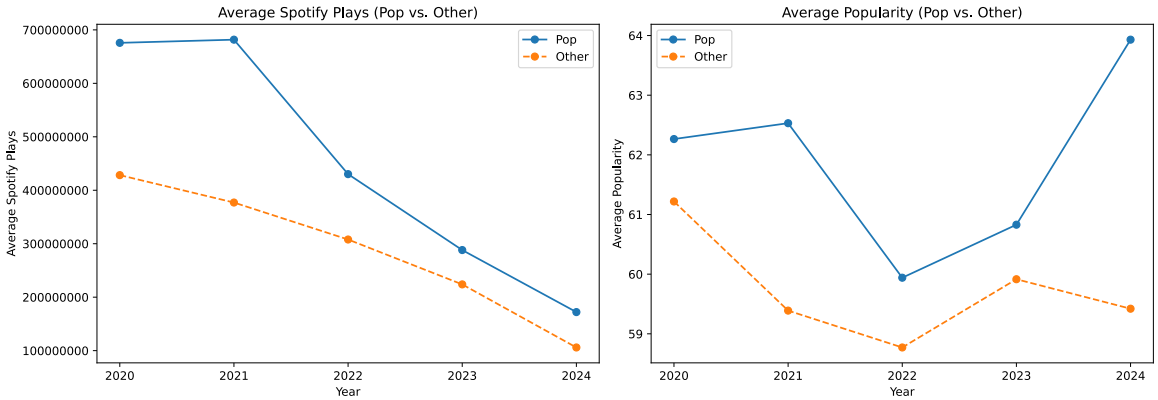

Figure 3: Average Spotify Plays and Popularity of Pop tracks comparing to other genres from 2020 to 2024

From the line graph, we can see that the average Spotify plays and popularity of

Pop songs follow a similar pattern to other genres. However, Pop songs generally have higher plays and popularity than other tracks.

While all genres experience similar trends, Pop consistently attracts more listeners. Its Spotify plays drop but its popularity still tends to rise. Even with fewer streams among Pop tracks, it still can maintain the strong influence.

Conclusion: Pop music holds a steady performance on Spotify. Listeners still explore a wide range of music rather than just sticking to Pop. While Pop songs often get more plays and higher popularity, they follow the same general trends as other genres. However, even when streams drop, Pop's popularity keeps growing, suggesting that its influence is more than just play counts. Whether through the variety of genres, Pop continues to be one of the greatest genres throughout music platform.

2.5.2 Visualization objective 2:

Question: Which season produces the most successful Spotify tracks in terms of YouTube popularity?

Benefits:

- Helps the artists, producers or content creators make strategies to release trending tracks, from that maximizing views and likes count on Youtube platform.
- Viewers can discover the best music videos or playlists for every seasons.

Columns used in data: spotify_popularity, release_date, youtube_likes, youtube_views

Visualizations used: Heatmap

Reasons: It effectively highlights seasonal variations in YouTube engagement, making it easy to compare views and likes across different seasons, also Youtube engagement is based on 2 attributes views and likes.

Analyzing and Visualizing steps:

First, we classify each track based on four seasons in year by adding season column and labeling them based on their month in release_date. Then we identify high Spotify popularity tracks. We decide to choose tracks that have Spotify popularity more than or equal to 70 by filtering the track that has spotify_popularity above or equal 70. As shown above in the preprocessing part, 2 columns youtube_likes and youtube_views have different data range with each other, so we normalize these 2 columns using Min-max scaler to convert them into range [0, 1]. Then we calculate mean of youtube_likes and youtube_views group by season.

We visualize Youtube engagement through four seasons using heatmap.

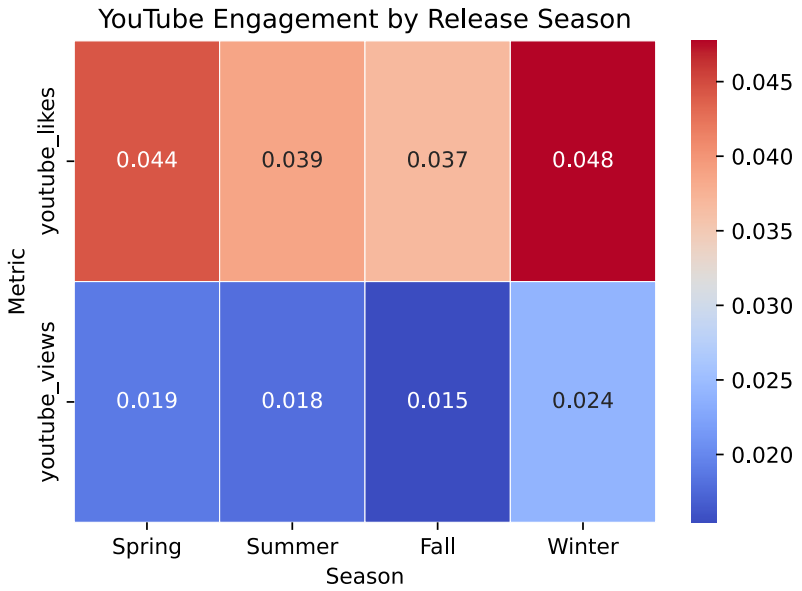

Figure 4: Youtube engagement by Release Season

From the heatmap, we can see that most of the high performing Spotify tracks that have high Youtube engagement, both likes and views, were often released in Winter. Meanwhile, the lowest are those which were in Fall. This shows that Winter releases tend to attract more listeners and viewers, while Fall releases may struggle to gain the same level of attention.

Conclusion:

Winter is the most favorable season for releasing music, as tracks released during this period tend to achieve the highest engagement on YouTube in terms of likes and views. In contrast, Fall sees the lowest engagement.

One possible reason for Winter's high engagement is the presence of major holidays such as Christmas and New Year, when people have more free time to stream and interact with music.

Fall might experience lower engagement due to the people who have academic schedules and work routines resuming after summer, spending less time for music. Additionally, there may be fewer major music events compared to Winter and Summer.

2.5.3 Visualization objective 3:

Question: Which song by the top 10 artists on Spotify has the greatest influence, and is this related to the music genre?

Benefits:

- Provides insights for artists, music producers to review their current work and make some useful strategies for the future plans.
- Helps to understand if the successful artists made huge influence across different platforms.
- Enhances playlist filter by featuring artists with strong cross-platform influence.

Columns used in data: `spotify_plays`, `genre`, `artist`

Visualizations used: Two horizontal bar chart
Reasons:

- For the first chart: It is easier to show top 10 artists by the total `spotify_plays` of all their tracks using horizontal bar chart.
- For the second chart: As the first chart, it is easier to compare tracks based on their number of playbacks.

Analyzing and Visualizing steps:

First, we plot and show the top 10 most influenced artists in this dataset. After that, we analyze a track with highest listens of those artists and its genre to find the patterns.

We visualize the top 10 artists in this dataset.

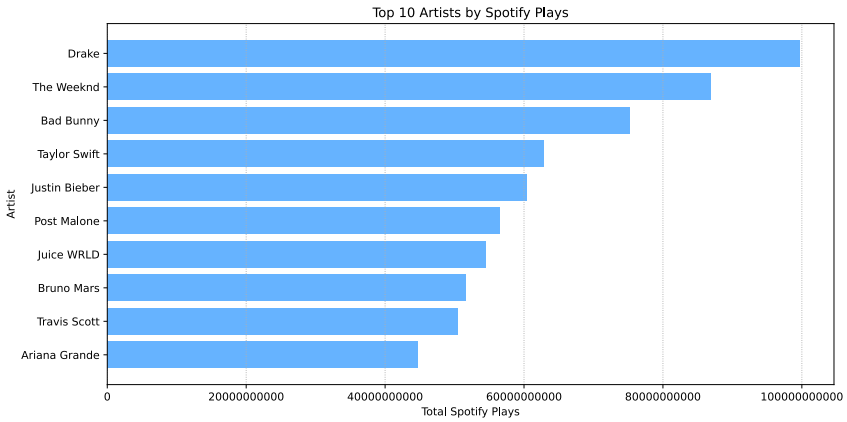

Figure 5: Top 10 artists by Spotify Plays

The chart illustrates top 10 artists along with the total of their tracks' `spotify_plays`. Drake is the most popular artists shown by this dataset, with the total of nearly 100 billions playbacks. While The Weeknd's total playbacks is just around 85 billions.

Now, let's find out top 10 tracks have the highest influence among these artists.

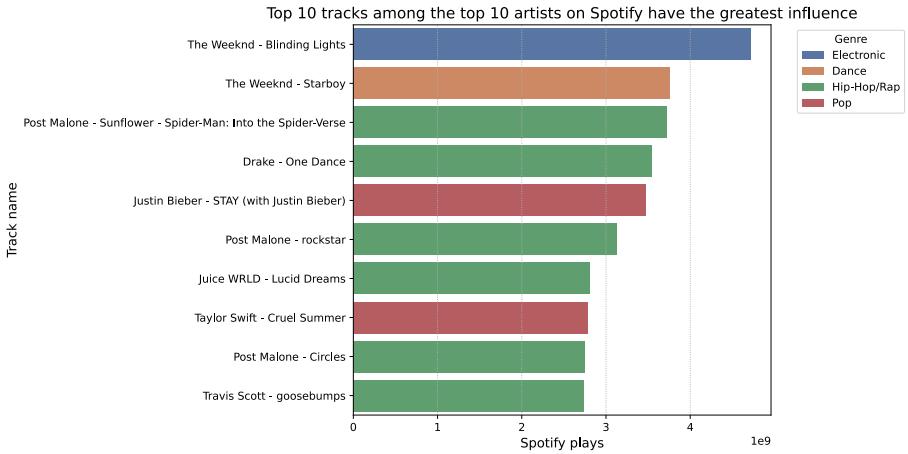

Figure 6: Top 10 artists by Spotify Plays

The second chart focuses on the top 10 most influential songs from these artists, categorized by genre (Electronic, Dance, Hip-Hop/Rap, Pop). "Blinding Lights" by The Weeknd leads with the highest streams in the Electronic category, confirming its global impact. Other hits like "Starboy" (The Weeknd, Dance), "Sunflower" (Post Malone, Hip-Hop/Rap), and "One Dance" (Drake, Hip-Hop/Rap) further demonstrate the dominance of top artists. Notably, Hip-Hop/Rap accounts for 5 out of 10 songs, reflecting its strong presence in modern music trends, while Electronic and Dance are mainly represented by The Weeknd. Pop remains influential, with artists like Justin Bieber and Taylor Swift contributing major hits.
Conclusion: Drake leads in total Spotify plays, likely due to a larger catalog and consistent average streams per song compared to The Weeknd, who ranks second. However, The Weeknd dominates individual track influence with "Blinding Lights" and "Starboy," the top two tracks, showing his strength in creating major hits. Notably, "Blinding Lights," an Electronic track, surpasses Hip-Hop/Rap-heavy tracks like Drake's "One Dance," highlighting its unique appeal in a market where Hip-Hop/Rap occupies half the top 10. Further analysis of track counts and average streams could confirm these trends, but the data reflects a diverse music landscape with standout genre-defying hits.

2.5.4 Visualization objective 4:

Question: How does explicit content affect a track's popularity across platform?
Benefits:

- Helps artists and composers alter the lyrics which is suitable for the audience expectation for maximum reach and engagement.
- Reveals audience preferences for explicit vs. clean content across different genres whether they prefer the tracks that have explicit content or not.

Columns used in data: `spotify_plays`, `genre`, `explicit`, `score`, `spotify_popularity`, etc.
Visualizations used: Pie chart and radar chart
Reasons:

- For the pie chart: We use pie chart to show the overall proportions of explicit and non-explicit tracks.
- For the second chart: This chart is mainly used to compare several attributes between explicit and non-explicit tracks. It helps people gain insights and easily compare the differences.

Analyzing and Visualizing steps:
First, we plot the distribution of explicit and non-explicit tracks in the dataset, then examine whether the group with the larger share has a greater influence on key characteristics.
We visualize the top 10 artists in this dataset.

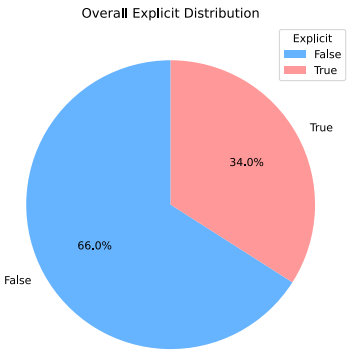

Figure 7: Overall Explicit Distribution

As we can see, non-explicit content accounted for the largest proportion in surveyed tracks at around 66%. Therefore, we can conclude that the majority of tracks is non-explicit content, so that most of the tracks are easily reached by audience of all ages.

Now, let's analyze more on how explicit and non-explicit content affect on popularity (views and likes) across platforms (Spotify, Youtube, Airplay, Itunes and Shazam)

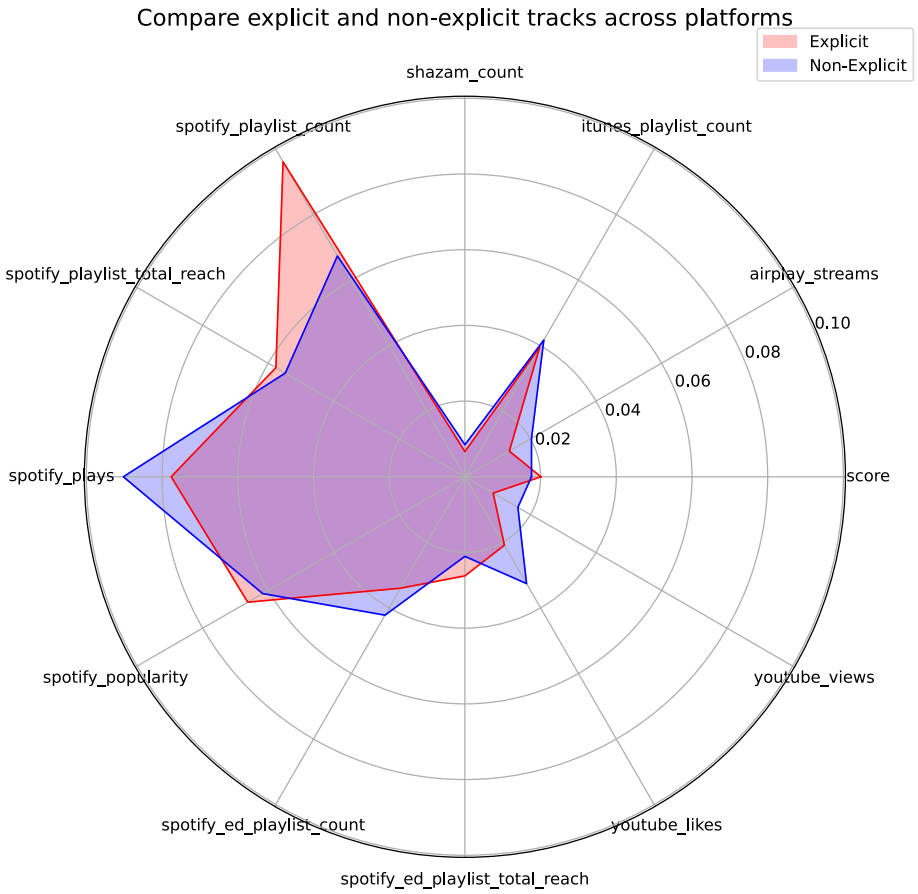

Figure 8: Radar chart for comparing explicit and non-explicit tracks across platforms

Conclusion: The radar chart reveals that non-explicit tracks generally receive higher views, likes, and playbacks across multiple platforms, indicating a broader audience appeal.

However, explicit tracks tend to perform better on Spotify, particularly in playlist placements, suggesting that a significant portion of Spotify's curated and user-generated playlists include explicit content.

This highlights a key insight: while non-explicit tracks gain more visibility and engagement across diverse platforms, explicit tracks benefit from a stronger presence within Spotify's ecosystem, potentially due to user preferences or platform-specific playlist curation strategies.