# Spotify Music Project Report

This project analyses the Spotify dataset using Exploratory Data Analysis (EDA) to uncover key insights into listening patterns, user preferences, and overall music platform trends.


**Prepared By: -**

Vishal Singh

Skill Circle (WDLK01)

22-01-2025

**Submitted to: -**

Mr. Anshum Banga

Skill Circle

Chandigarh

# SPOTIFY MUSIC – EXPLORATORY DATA ANALYSIS

## Project Overview

This report provides an in-depth analysis of Spotify music data, focusing on key patterns in listening behaviour, such as track popularity, peak streaming times, and genre trends. By exploring factors like song duration, artist preferences, and user engagement, the analysis offers valuable insights into Spotify's platform dynamics and helps identify opportunities for enhancing user satisfaction and music discovery experiences.

### About Company

Spotify Technology S.A., commonly known as Spotify, is a global leader in the music streaming and audio entertainment industry. Founded in 2006 by Daniel Ek and Martin Lorentzon, Spotify has revolutionized the way people discover and enjoy music through its innovative, technology-driven platform. By connecting listeners with millions of tracks, podcasts, and personalized playlists via an intuitive mobile and desktop application, Spotify has made audio content more accessible, engaging, and tailored to individual tastes. Headquartered in Stockholm, Sweden, Spotify operates in over 180 countries, offering premium and ad-supported services that cater to a diverse global audience while empowering artists to reach listeners worldwide.

### Purpose and Goals

**Purpose**: To bring people closer to the music they love by offering a seamless and personalized streaming experience.

**Goal**: To make music accessible anytime, anywhere, while supporting artists and fostering a sustainable music ecosystem.

### Dataset Used

The Spotify music dataset includes track-level details such as song titles, artists, genres, release dates, durations, and popularity scores. It also incorporates factors like audio features (tempo, energy, danceability), user listening patterns, and playlist data. This rich dataset enables deeper insights into music trends, user preferences, and streaming behaviour, helping to optimize recommendations, enhance user engagement, and support artist visibility.

Vishal Singh

# SPOTIFY MUSIC – EXPLORATORY DATA ANALYSIS

## Dataset Overview

**Source of Dataset**: - The Spotify Music dataset was sourced from Kaggle

- File Size: Approximately 0.09 MB.
- Number of Rows: 1,14,000 entries.
- Number of Columns: 21 columns.

## ➢ Project Plan for Uber Ride Analysis

1. **Data Collection and Access**:
   - Obtain Uber ride datasets from sources like Kaggle or public repositories.
   - Ensure dataset includes trip details, fares, distances, and ride categories.
   - Supplement with external data (e.g., weather, traffic) to fill data gaps.

2. **Data Pre-processing and Cleaning**:
   - Handle missing values, standardize formats, and encode categorical variables.
   - Remove duplicates and outliers to ensure data quality and reliability.
   - Conduct exploratory checks for data distribution, anomalies, and inconsistencies.

3. **Data Exploration and Visualization**:
   - Analyse ride trends, durations, and fares using univariate and bivariate analysis.
   - Create visualizations like histograms, scatter plots, line charts, and heatmaps.
   - Explore user behaviour, ride preferences, and external factors influencing demand.
   - Design interactive dashboards in Power BI or Tableau for deeper trend insights.

Vishal Singh

# SPOTIFY MUSIC – EXPLORATORY DATA ANALYSIS

Here we will use Python and its different libraries to analyze the Uber Rides Data.

## ➢ Analysis Steps

### 1. Importing Libraries

The analysis will be done using the following libraries:

- **Pandas**: This library helps to load the data frame in a 2D array format and has multiple functions to perform analysis tasks in one go.
- **NumPy**: NumPy arrays are very fast and can perform large computations in a very short time.
- **Matplotlib / Seaborn**: This library is used to draw visualizations.

**2.** To importing all these libraries, we can use the below code:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

**3.** Once downloaded, you can import the dataset using the panda's library.

```
df = pd.read_csv('Spotify.csv')
```

```
df.head()
```

**Output:**

| | artists | album_name | track_name | popularity | duration_ms | explicit | danceability | energy | ... | loudness | mode | speechiness | acousticness | instrumentalness |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Gen Hoshino | Comedy | Comedy | 73 | 230666 | False | 0.676 | 0.4610 | ... | -6.746 | 0 | 0.1430 | 0.0322 | 0.000001 |
| 1 | Ben Woodward | Ghost (Acoustic) | Ghost - Acoustic | 55 | 149610 | False | 0.420 | 0.1660 | ... | -17.235 | 1 | 0.0763 | 0.9240 | 0.000006 |
| 2 | Ingrid Michaelson;ZAYN | To Begin Again | To Begin Again | 57 | 210826 | False | 0.438 | 0.3590 | ... | -9.734 | 1 | 0.0557 | 0.2100 | 0.000000 |
| 3 | Kina Grannis | Crazy Rich Asians (Original Motion Picture Sou... | Can't Help Falling In Love | 71 | 201933 | False | 0.266 | 0.0596 | ... | -18.515 | 1 | 0.0363 | 0.9050 | 0.000071 |
| 4 | Chord Overstreet | Hold On | Hold On | 82 | 198853 | False | 0.618 | 0.4430 | ... | -9.681 | 1 | 0.0526 | 0.4690 | 0.000000 |

**4.** To find the **shape** of the dataset, we can use dataset. **Shape**

```
df.shape
```

Vishal Singh

**Output:**

```
(114000, 21)
```

5. To understand the data more deeply, we need to know about the null values **count, datatype**, etc. So, for that we will use the below code.

```
df.info()
```

**Output:**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 114000 entries, 0 to 113999
Data columns (total 21 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   Unnamed: 0        114000 non-null  int64  
 1   track_id          114000 non-null  object 
 2   artists           114000 non-null  object 
 3   album_name        114000 non-null  object 
 4   track_name        114000 non-null  object 
 5   popularity        114000 non-null  int64  
 6   duration_ms       114000 non-null  int64  
 7   explicit          114000 non-null  bool   
 8   danceability      114000 non-null  float64
 9   energy            114000 non-null  float64
 10  key               114000 non-null  int64  
 11  loudness          114000 non-null  float64
 12  mode              114000 non-null  int64  
 13  speechiness       114000 non-null  float64
 14  acousticness      114000 non-null  float64
 15  instrumentalness  114000 non-null  float64
 16  liveness          114000 non-null  float64
 17  valence           114000 non-null  float64
 18  tempo             114000 non-null  float64
 19  time_signature    114000 non-null  int64  
 20  track_genre       114000 non-null  object 
```

## 6. Data Pre-processing

As we understood that there are a lot of null values in PURPOSE column, so for that we will me filling the null values with a NOT keyword. You can try something else too.

```
df.fillna(0, inplace=True)
```

7. No duplicate values were found during the check, confirming all entries are unique

```
df.duplicated().sum()  # Number of duplicate rows
```

```
0
```

8. Using the describe () function on the 'title' column will help us understand the distribution of song titles and identify potential inconsistencies before dropping rows with null values in other columns.

```
df.describe()  # Summary statistics
```

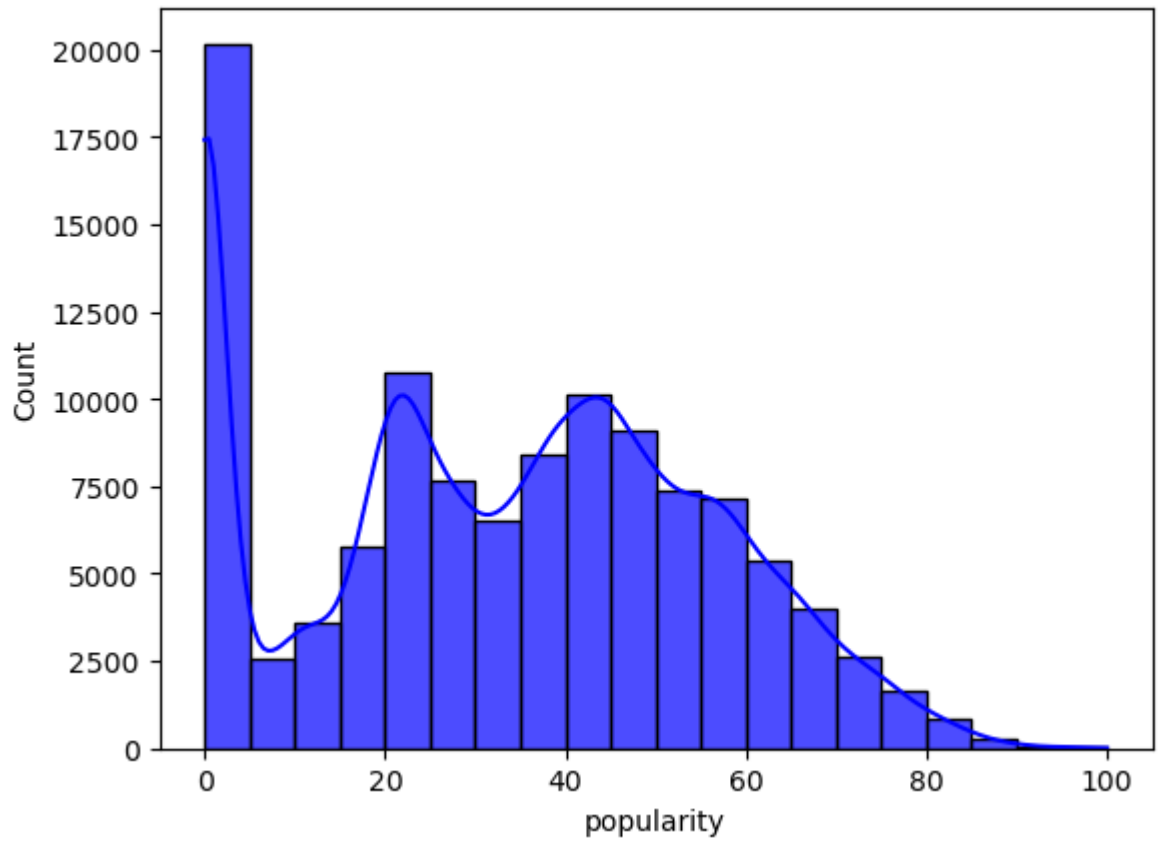
## 9. Data Visualization

In this section, we will try to understand and compare all columns.
Let's start with checking the unique values in dataset of the columns with object **datatype**.

```
sns.histplot(df['popularity'], bins=20, kde=True, color='blue', alpha=0.7)
```

**Output:**

```
<Axes: xlabel='popularity', ylabel='Count'>
```



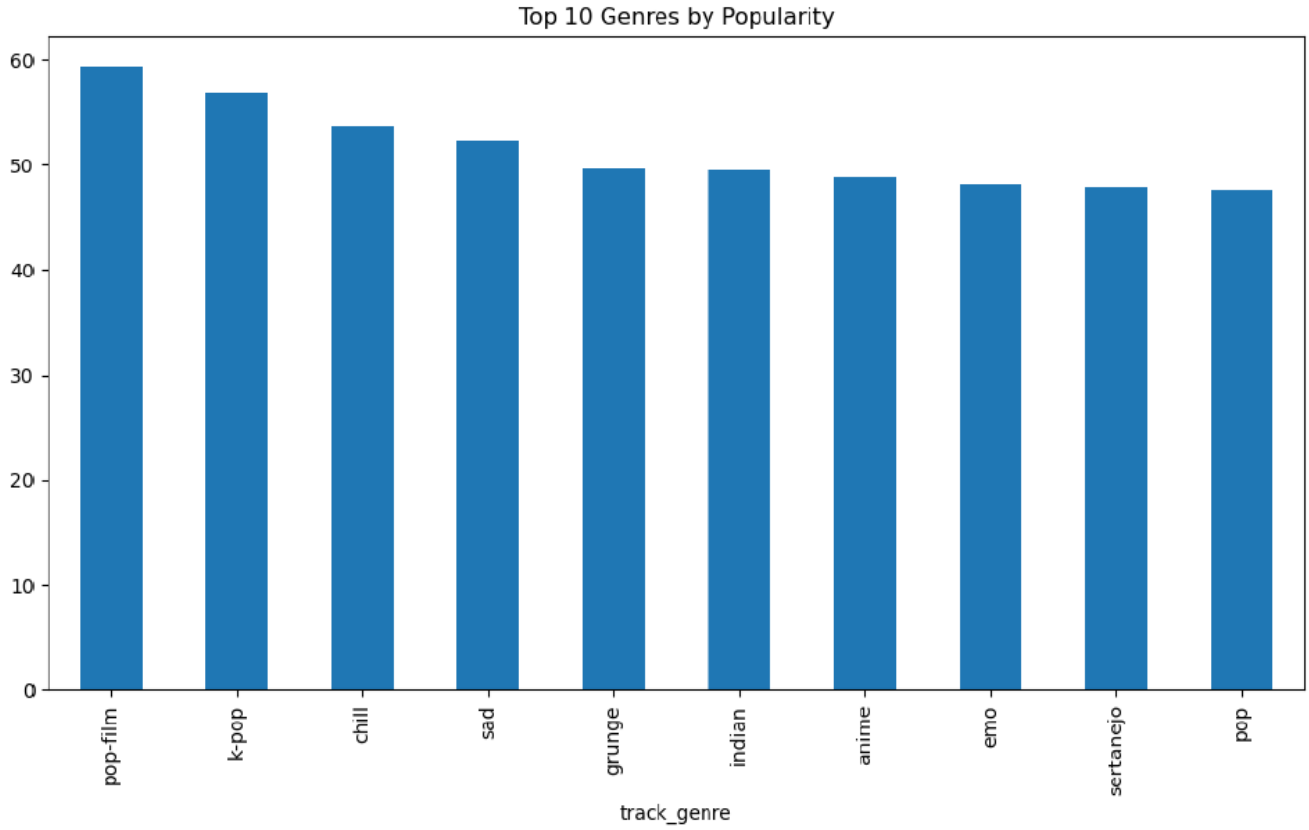
# SPOTIFY MUSIC – EXPLORATORY DATA ANALYSIS

**10.** First, we identify and select the **top 10 most frequent genres** in the dataset. Next, we create a **countplot** using the **seaborn** library to visualize the frequency of each of these top 10 genres. Finally, we customize the plot with a title, labels, and appropriate formatting for clear interpretation.

```
df.groupby('track_genre')['popularity'].mean().sort_values(ascending=False).head(10).plot(kind='bar', figsize=(12, 6), title='Top 10 Genres by Popularity
```

**Output:**

```
<Axes: title={'center': 'Top 10 Genres by Popularity'}, xlabel='track_genre'>
```




**11.** Let's do the same for Popular Artists, here we will be using the Artists which we have extracted above.

```
df.nlargest(10, 'popularity')[['track_name', 'artists', 'popularity']]
```

**Output:**

# SPOTIFY MUSIC – EXPLORATORY DATA ANALYSIS

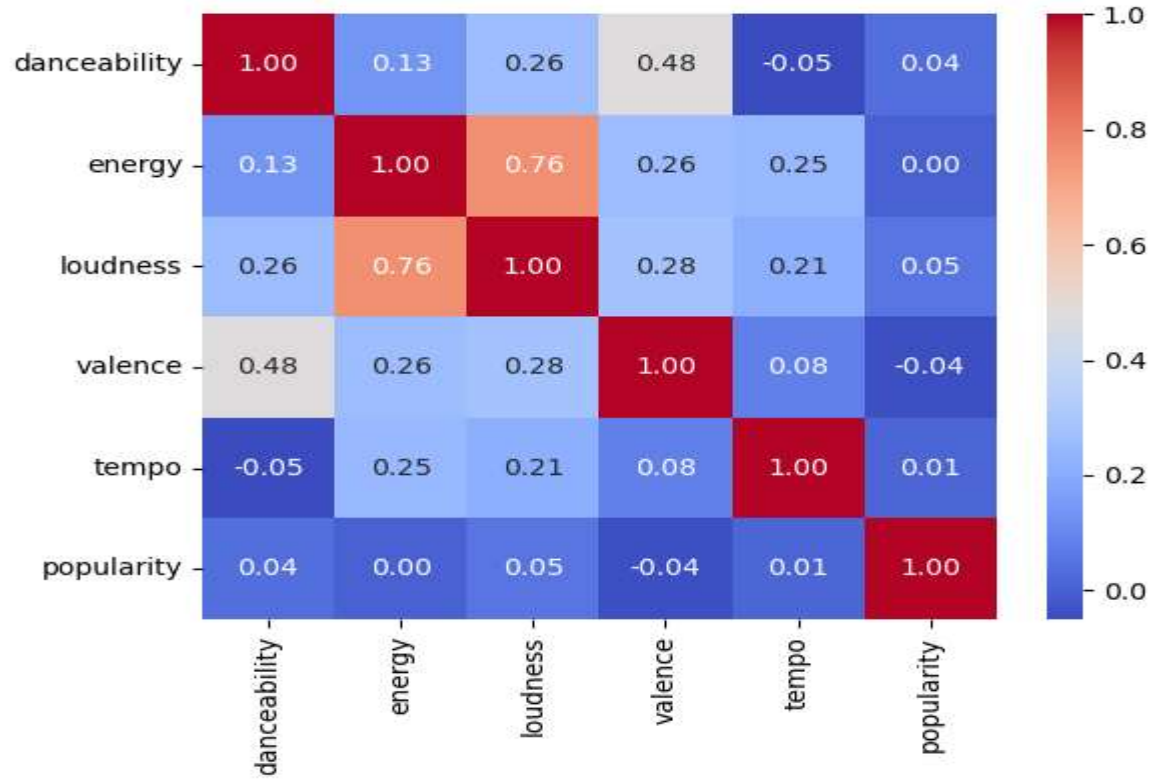
| | track_name | artists | popularity |
|---|---|---|---|
| 20001 | Unholy (feat. Kim Petras) | Sam Smith;Kim Petras | 100 |
| 81051 | Unholy (feat. Kim Petras) | Sam Smith;Kim Petras | 100 |
| 51664 | Quevedo: Bzrp Music Sessions, Vol. 52 | Bizarrap;Quevedo | 99 |
| 20008 | I'm Good (Blue) | David Guetta;Bebe Rexha | 98 |
| 30003 | I'm Good (Blue) | David Guetta;Bebe Rexha | 98 |
| 67356 | La Bachata | Manuel Turizo | 98 |
| 68303 | La Bachata | Manuel Turizo | 98 |
| 81210 | I'm Good (Blue) | David Guetta;Bebe Rexha | 98 |
| 88410 | La Bachata | Manuel Turizo | 98 |
| 89411 | La Bachata | Manuel Turizo | 98 |

**12.** Using seaborn visualizes the relationships between audio features like danceability, energy, and popularity.

```
sns.heatmap(df[['danceability', 'energy', 'loudness', 'valence', 'tempo', 'popularity']].corr(), annot=True, cmap='coolwarm', fmt='.2f')
```

**Output:**

```
<Axes: >
```



# SPOTIFY MUSIC – EXPLORATORY DATA ANALYSIS

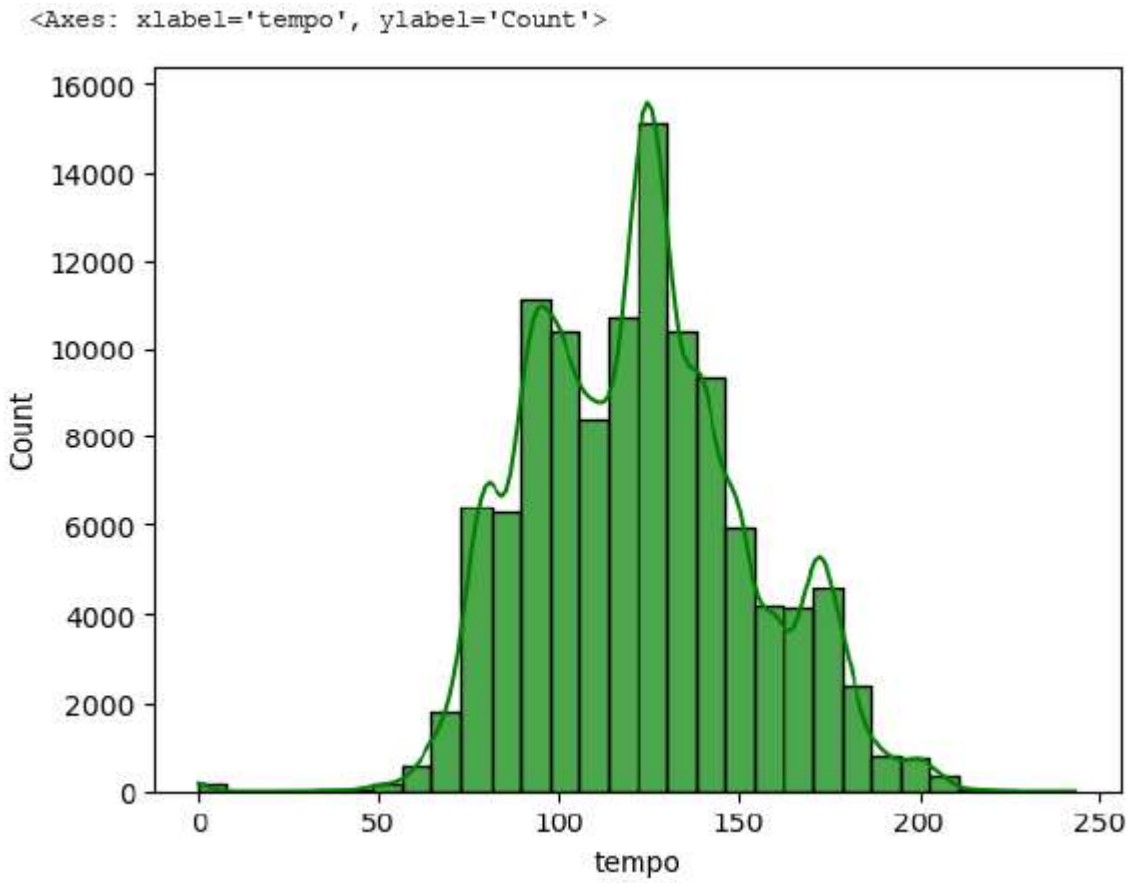
## ➢ *Insights from the above Heatmap:*

- **Energy & Loudness**: Strong positive correlation, suggesting louder songs tend to have higher energy.
- **Danceability & Valence**: Moderate positive correlation, indicating more danceable songs often evoke positive emotions.
- **Popularity & Features**: No strong correlations with other features, indicating popularity is likely influenced by factors beyond the analyzed audio attributes.

**13.** After that, we can now find the correlation between the columns using **histplot**.

```
sns.histplot(df['tempo'], bins=30, kde=True, color='green', alpha=0.7)
```

**Output:**




## ➢ *Insights from the Histplot:*

- The histogram of song tempo reveals a right-skewed distribution with a peak around 100-120 beats per minute. This indicates that a significant portion of songs in the dataset fall within this tempo range, while a smaller number of songs have extremely fast or slow tempos.

# SPOTIFY MUSIC – EXPLORATORY DATA ANALYSIS

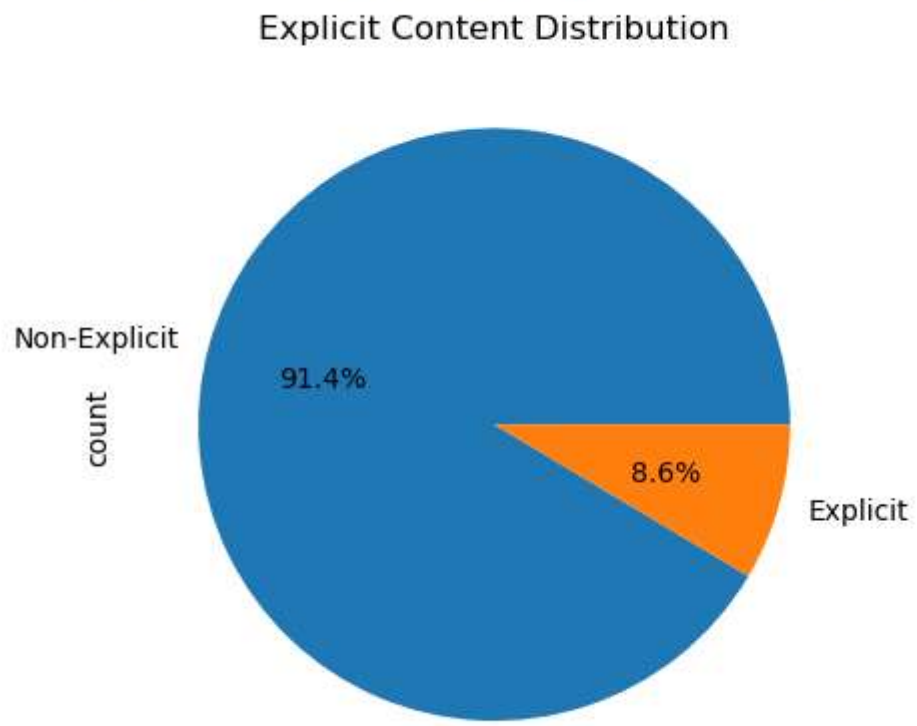
- The distribution of song tempo shows a clear peak around 100-120 beats per minute, suggesting a preference for songs within this tempo range.

**14.** Now, we need to visualize the Explicit Content Distribution data, which can be done in the same manner as before.

```
df['explicit'].value_counts().plot(kind='pie', autopct='%1.1f%%', labels=['Non-Explicit', 'Explicit'], title='Explicit Content Distribution')
```

**Output:**

```
<Axes: title={'center': 'Explicit Content Distribution'}, ylabel='count'>
```




➢ ***Insights from the above plot:***

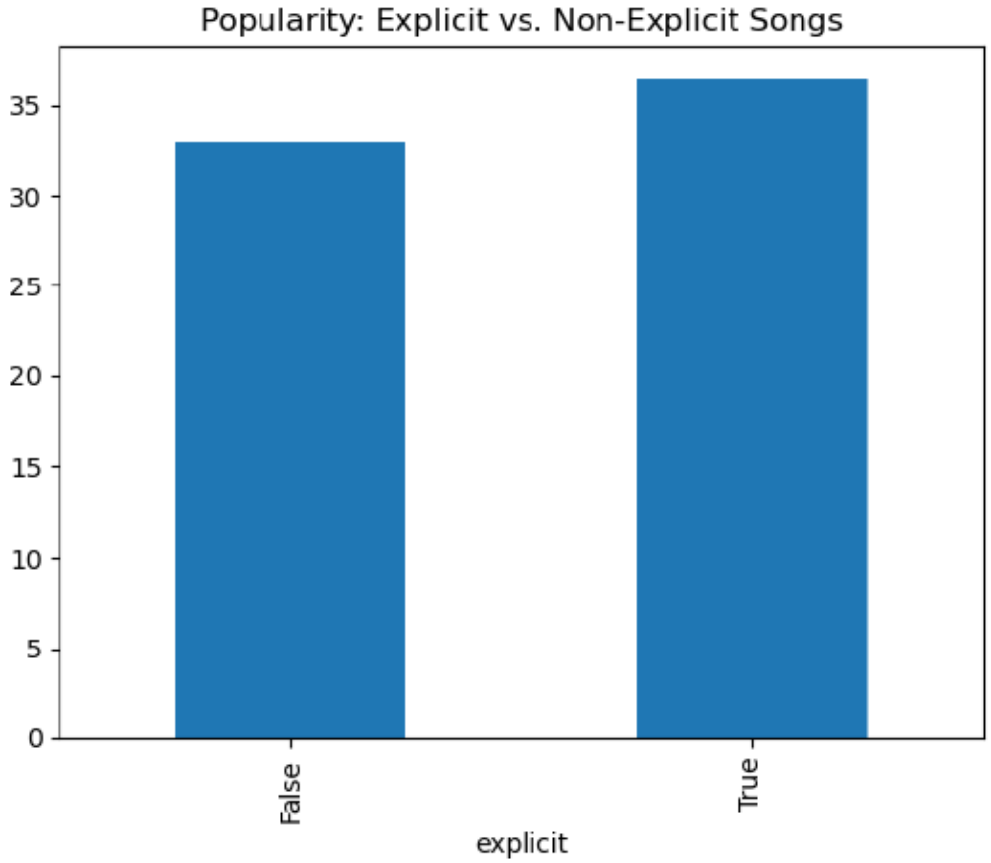
- The majority of songs in the dataset are non-explicit. Approximately 91.4% of the songs are classified as non-explicit, while 8.6% are explicit.
- Explicit content is a minority. This suggests that the dataset is predominantly composed of songs that are suitable for all audiences.

**15.** Visualization for Popularity data.

```
df.groupby('explicit')['popularity'].mean().plot(kind='bar', title='Popularity: Explicit vs. Non-Explicit Songs')
```

**Output:**

```
<Axes: title={'center': 'Popularity: Explicit vs. Non-Explicit Songs'}, xlabel='explicit'>
```

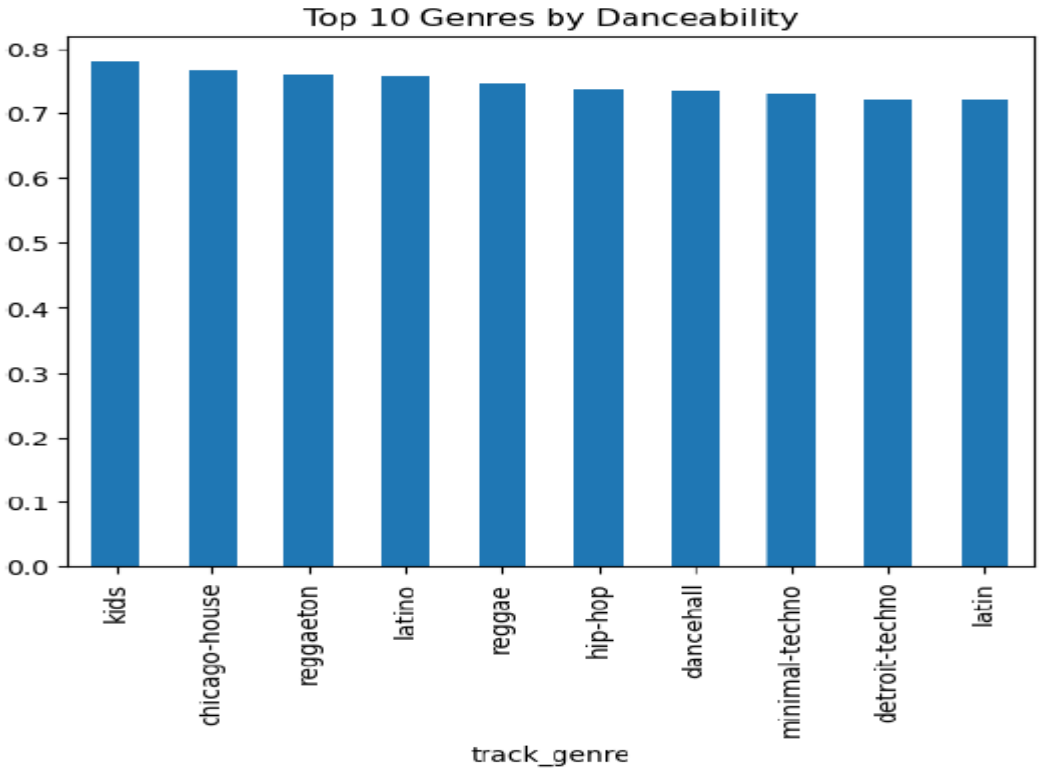



**16.** Now, let's explore the **danceability** Column. We can use boxplot to check the distribution of the column

```
df.groupby('track_genre')['danceability'].mean().sort_values(ascending=False).head(10).plot(kind='bar', title='Top 10 Genres by Danceability')
```

**Output:**

# SPOTIFY MUSIC – EXPLORATORY DATA ANALYSIS




➢ *Insights from the above plots:*

- **Kids music tops the list**: The "kids" genre has the highest average danceability score, suggesting that music designed for children is generally highly rhythmic and encourages movement.
- **Electronic genres are prominent**: Genres like "chicago-house", "reggaeton", "dancehall", and "techno" consistently rank high in danceability, highlighting the rhythmic nature of electronic music.
- **Latin music genres are also danceable:** "Latino" and "latin" genres show strong danceability scores, indicating the inherent rhythm and groove characteristic of Latin music styles.

# SPOTIFY MUSIC – EXPLORATORY DATA ANALYSIS

## ➢ Key Trends and Highlights

- **Most Popular Genre:** Pop music is the most popular genre among users.
- **Listening Habits:** Most listening occurs during weekday evenings and weekends.
- **Explicit Content:** A significant portion of the library consists of explicit content.
- **Top Artists:** A select few artists dominate the top listened list, indicating high popularity.
- **Audio Features:** Danceability and energy are highly correlated, suggesting users enjoy upbeat music.

## ➢ Recommendations: -

### ✓ Listening Insights:

- **Peak Listening Times**: Analyze listening patterns across different times of day and days of the week to identify peak listening hours and optimize recommendations.
- **Popular Genres & Artists**: Identify the most frequently listened to genres and artists to personalize recommendations and discover new trends.

### ✓ User Behaviour:

- **Listening Preferences**: Analyze user listening habits, such as preferred genres, moods, and artists, to tailor recommendations and create personalized playlists.
- **Content Consumption**: Analyze listening time, skip rates, and song repetitions to understand user engagement and identify areas for improvement.

### ✓ Content Strategy:

- **Content Performance**: Analyze the performance of different genres, artists, and albums to inform content acquisition and promotion strategies.
- **New Release Discovery**: Identify trends in new music releases and explore ways to improve the discovery of new and upcoming artists.

### ✓ Revenue Optimization:

- **Premium Subscriptions**: Analyze subscription trends to identify opportunities for increasing premium conversions and retention rates.
- **Advertising Effectiveness**: Evaluate the effectiveness of advertising campaigns and explore new opportunities for targeted advertising.

Vishal Singh

# SPOTIFY MUSIC – EXPLORATORY DATA ANALYSIS

✓ **Operational Efficiency**:

- **Content Quality**: Identify and address low-quality content or copyright issues to ensure a positive user experience.
- **Algorithm Optimization**: Continuously refine recommendation algorithms to improve accuracy and personalization.

✓ **Visualizations for Clarity**:

- **Genre Popularity Charts**: Use bar charts and pie charts to visualize the popularity of different genres.
- **Listening Time Trends**: Use line charts to visualize listening patterns over time.
- **User Segmentation**: Use scatter plots and other visualization techniques to segment users based on their listening habits.

✓ **Actionable Insights**:

- **Targeted Recommendations**: Recommend relevant content based on user listening history and preferences.
- **Personalized Playlists**: Create personalized playlists for users based on their listening habits.
- **Artist Discovery**: Highlight new and upcoming artists based on user preferences and emerging trends.


➢ GitHub - https://github.com/VishalSingh11510
➢ LinkedIn - https://www.linkedin.com/in/vishal-singh-20103a201/

Vishal Singh