EDA PROJECT-CSM 323

Final Report

By

V Sam Pravardhan

Idno:12217383

Rollno:32

Submitted to

Ved Prakash Chaubey

School Of Computer Science And Engineering

Lovely Professional University,Phagwara,Punjab.

Declaration

I, V Sam Pravardhan, hereby declare that the work done by me on "Spotify-2023" from September, 2024 to October, 2024, is a record of original work for the partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Computer Science - Data Science with ML, Lovely Professional University, Phagwara.

Signature of Student   Signature of Faculty

V Sam Pravardhan   Ved Prakash Chaubey

12217383

Date:

Acknowledgement

I would like to express my gratitude towards my University as well as UpGrad for providing me the golden opportunity to do this wonderful Project in Exploratory Data Analysis and Machine Learning, which facilitated my understanding of recent advancements and contributed significantly to my professional growth.

This course i.e. EDA project has been instrumental in expanding my understanding of data and its applications. It helped me in finding insights from the data which could assist changing millions of lives.

I would like to extend my heartfelt gratitude to all those who supported me in the development of the 'Spotify-2023.' This project benefited immensely from the advanced capabilities of Spotify streams, which provided a solid foundation for implementing effective translation solutions.

I am particularly grateful to the team at Spotify for their outstanding work in developing and maintaining the streams. Their contributions to natural language processing technology have been instrumental in achieving the objectives of the project.

Additionally, I would like to thank Mr. Ved Prakash Chaubey, my mentor, who offered valuable advice and feedback throughout the project. His insights were crucial in refining the project and enhance its performance. His dedication made this project possible.

Finally, I appreciate the support and encouragement from my family and friends. Their belief in the project's potential was a constant source of motivation. It helped throughout the journey

Table of Contents

Abstract

This report focuses on an analysis of Spotify's 2023 music dataset, which includes detailed information on songs and artists. The dataset covers various features such as track names, artists, popularity scores, and specific audio characteristics like tempo, danceability, and energy. The aim of this analysis is to uncover key trends in music, particularly those that influence a song's popularity on Spotify. By studying these attributes, the report explores how different song features impact the listening habits of Spotify users.The data was carefully cleaned to ensure accuracy. This process involved handling missing values and removing duplicate records that could have skewed the analysis. Once cleaned, the dataset was ready for deeper exploration and visualization, offering a clearer understanding of its structure and integrity. This step was crucial to avoid any potential biases or errors in the results.Exploratory Data Analysis (EDA) was applied to visually represent the dataset and uncover patterns. Relationships between key musical features and song popularity were analyzed, revealing that tracks with higher energy and danceability tend to score higher in popularity. The visualizations provided insights into how different features interact, influencing how tracks perform on the platform.The findings indicate that certain artists and genres have dominated Spotify in 2023, shaping the platform's overall music landscape. The results suggest that features like energy and danceability are significant factors in predicting a song's success. This analysis provides a strong foundation for future work, such as developing predictive models to forecast which songs may become hits based on their musical attributes.

Introduction

Spotify, one of the world's leading music streaming platforms, provides extensive data on millions of tracks, capturing everything from song popularity to detailed audio features like tempo, energy, and danceability. Analyzing this data offers valuable insights into listener preferences, trends in music, and the characteristics that drive a song's success. The 2023 dataset used in this report consists of thousands of songs with attributes that provide a comprehensive overview of the current music landscape.The purpose of this analysis is to explore the factors that contribute to a song's popularity on Spotify. By studying various audio features, the report seeks to identify which attributes have the strongest correlations with high popularity scores. Additionally, we aim to highlight the most frequent artists and genres dominating Spotify's platform in 2023.This report involves a thorough data cleaning process to ensure the dataset is accurate and ready for analysis. After preparing the data, exploratory data analysis (EDA) is conducted to visualize relationships between song characteristics and popularity. Key insights into trends and patterns in music are derived from this analysis, leading to a deeper understanding of what drives listener engagement in today's digital music environment.The results of this study not only shed light on current musical trends but also pave the way for future analyses, such as building models to predict the success of new songs based on their attributes. Understanding these patterns can help artists, producers, and marketers tailor their content for greater reach and listener satisfaction on Spotify.

<u>Methodology</u>

This study follows a structured approach to analyze the Spotify 2023 dataset, beginning with data acquisition, followed by data cleaning, exploratory data analysis (EDA), and the interpretation of results.

1.Data Acquisition

The dataset, containing song-level information from Spotify in 2023, includes features such as track names, artists, popularity scores, and various audio characteristics like tempo, energy, and danceability. The data was sourced from an official Spotify database and represents a diverse selection of songs from different genres, providing a robust basis for analysis.

2.Data Cleaning

Before performing any analysis, the dataset was cleaned to ensure accuracy and reliability. Missing values were identified and removed, particularly in critical columns such as song popularity and audio features. Duplicate entries were also eliminated to prevent redundancy and biases in the analysis. Additionally, data types were standardized, and any inconsistencies in categorical data (e.g., artist names) were corrected.

3.Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to uncover patterns, correlations, and trends within the dataset. Visualizations such as histograms, bar plots, and scatter plots were used to understand the distribution of popularity scores and audio features like tempo, energy, and danceability. EDA also explored the relationship between these audio features and song popularity to determine which characteristics contribute most to a track's success. Correlation analysis was performed to quantify the strength of these relationships.

4.Data Interpretation and Insights

Following EDA, the findings were interpreted to draw meaningful conclusions about the factors driving song popularity. Key metrics, such as average song popularity by genre and artist, were calculated, and trends were identified. These insights were used to support the analysis of musical trends on Spotify in 2023 and to suggest future directions for predictive modeling.

Visual Representation

```
df = pd.read_csv('D:\spotify-2023.csv', encoding='ISO-8859-1')
```

```
df.head(10)
```

	track_name	artist(s)_name	artist_count	released_year	released_month	released_day	in_spotify_playlists	in_spotify_charts	streams	in_apple_playlists	..
0	Seven (feat. Latto) (Explicit Ver.)	Latto, Jung Kook	2	2023	7	14	553	147	141381703	43	..
1	LALA	Myke Towers	1	2023	3	23	1474	48	133716286	48	..
2	vampire	Olivia Rodrigo	1	2023	6	30	1397	113	140003974	94	..
3	Cruel Summer	Taylor Swift	1	2019	8	23	7858	100	800840817	116	..
4	WHERE SHE GOES	Bad Bunny	1	2023	5	18	3133	50	303236322	84	..
5	Sprinter	Dave, Central Cee	2	2023	6	1	2186	91	183706234	67	..
6	Ella Baila Sola	Eslabon Armado, Peso Pluma	2	2023	3	16	3090	50	725980112	34	..
7	Columbia	Quevedo	1	2023	7	7	714	43	58149378	25	..
8	fukumean	Gunna	1	2023	5	15	1096	83	95217315	60	..
9	La Bebe - Remix	Peso Pluma, Yng Lvcas	2	2023	3	17	2953	44	553634067	49	..

10 rows × 24 columns

```
df.shape
```

(953, 24)

```
df.dtypes
```

```
track_name             object
artist(s)_name         object
artist_count            int64
released_year           int64
released_month          int64
released_day            int64
in_spotify_playlists    int64
in_spotify_charts       int64
streams                object
in_apple_playlists      int64
in_apple_charts         int64
in_deezer_playlists    object
in_deezer_charts        int64
in_shazam_charts       object
bpm                     int64
key                    object
mode                   object
danceability_%          int64
valence_%               int64
energy_%                int64
acousticness_%          int64
instrumentalness_%      int64
liveness_%              int64
speechiness_%           int64
dtype: object
```

```
df.isnull().sum()
```

```
track_name                0
artist(s)_name            0
artist_count              0
released_year             0
released_month            0
released_day              0
in_spotify_playlists      0
in_spotify_charts         0
streams                   0
in_apple_playlists        0
in_apple_charts           0
in_deezer_playlists       0
in_deezer_charts          0
in_shazam_charts         50
bpm                       0
key                      95
mode                      0
danceability_%            0
valence_%                 0
energy_%                  0
acousticness_%            0
instrumentalness_%        0
liveness_%                0
speechiness_%             0
dtype: int64
```

```python
df["in_shazam_charts"]=df["in_shazam_charts"].fillna(df["in_shazam_charts"].mode()[0])
```

```
df.isnull().sum()
```

```
track_name                0
artist(s)_name            0
artist_count              0
released_year             0
released_month            0
released_day              0
in_spotify_playlists      0
in_spotify_charts         0
streams                   0
in_apple_playlists        0
in_apple_charts           0
in_deezer_playlists       0
in_deezer_charts          0
in_shazam_charts          0
bpm                       0
key                      95
mode                      0
danceability_%            0
valence_%                 0
energy_%                  0
acousticness_%            0
instrumentalness_%        0
liveness_%                0
speechiness_%             0
dtype: int64
```

```
df["key"]=df["key"].fillna(df["key"].mode()[0])
```

```
df.isnull().sum()
```

```
track_name                      0
artist(s)_name                  0
artist_count                    0
released_year                   0
released_month                  0
released_day                    0
in_spotify_playlists            0
in_spotify_charts               0
streams                         0
in_apple_playlists              0
in_apple_charts                 0
in_deezer_playlists             0
in_deezer_charts                0
in_shazam_charts                0
bpm                             0
key                             0
mode                            0
danceability_%                  0
valence_%                       0
energy_%                        0
acousticness_%                  0
instrumentalness_%              0
liveness_%                      0
speechiness_%                   0
dtype: int64
```

```
songs = df['artist(s)_name'].value_counts()
songs.nlargest(10).plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title('Top Streams in Spotify')
plt.show()
```
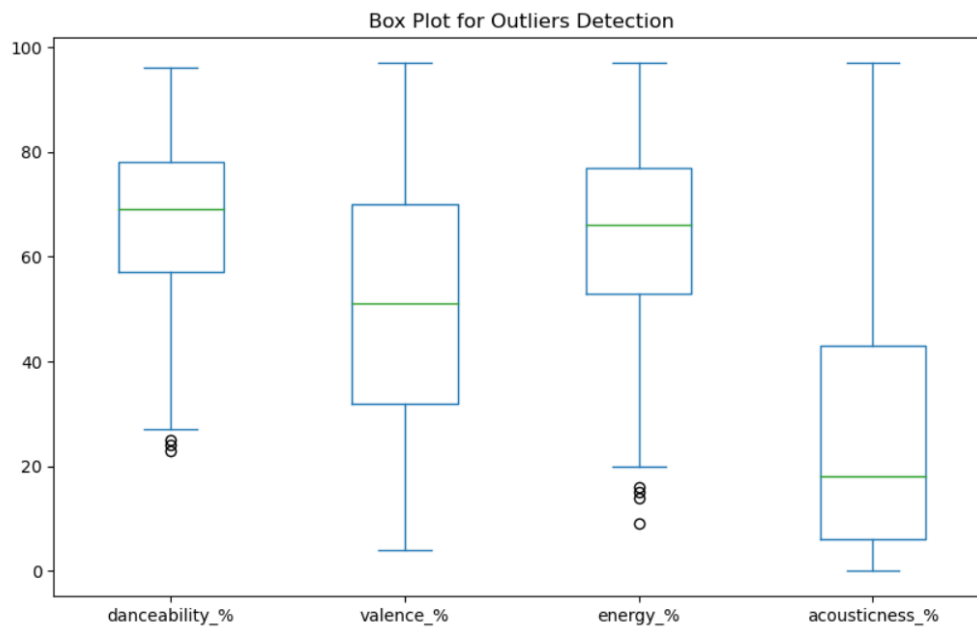

Top Streams in Spotify
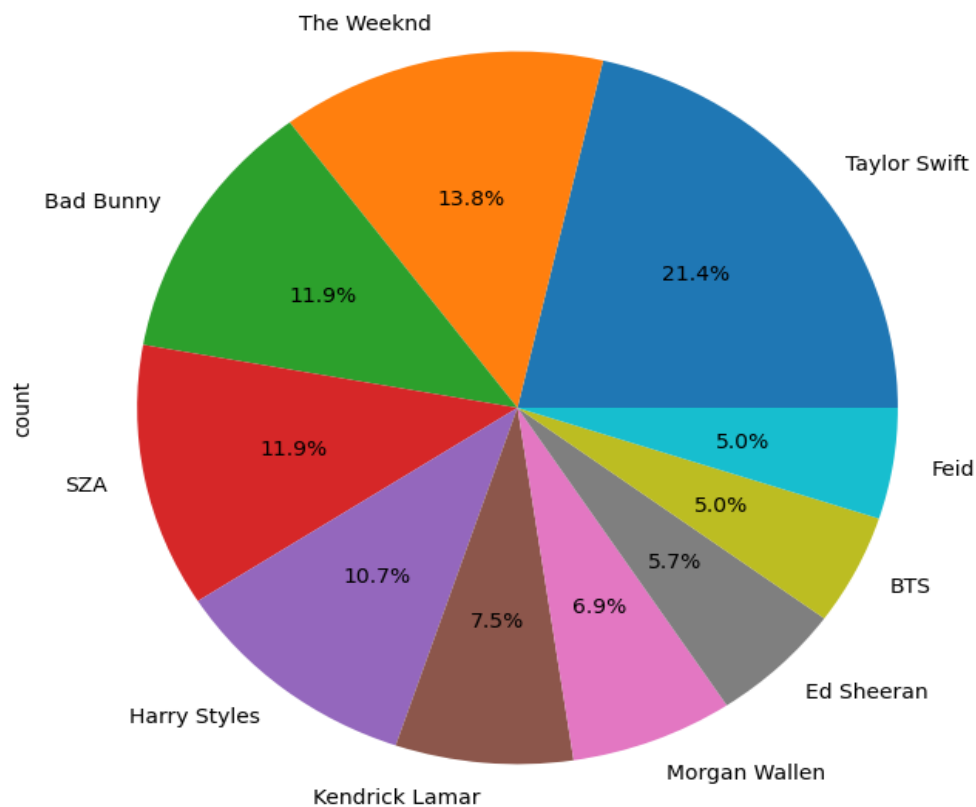
```
df[['danceability_%','valence_%']].plot(kind='box',title='playlist streams',figsize=(10,6))
```

<Axes: title={'center': 'playlist streams'}>
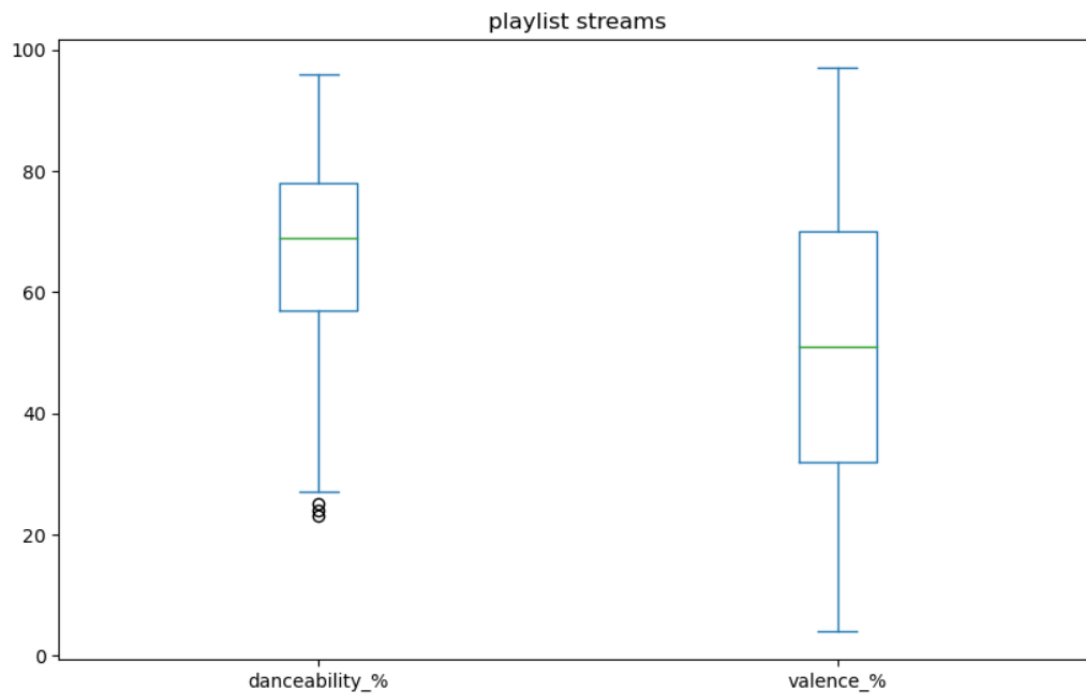

```
df[['danceability_%','valence_%','energy_%','acousticness_%']].plot(kind='box', figsize=(10,6), title='Box Plot for Outliers Dete
```
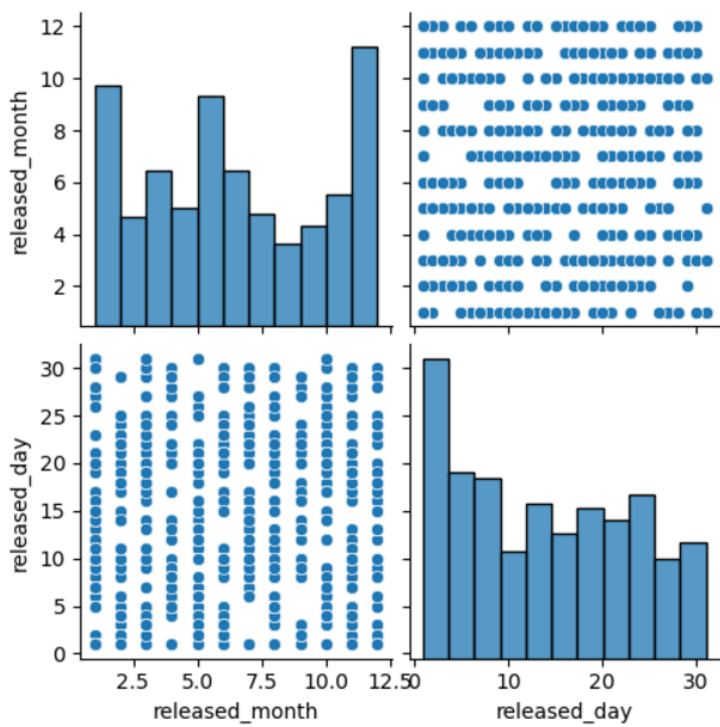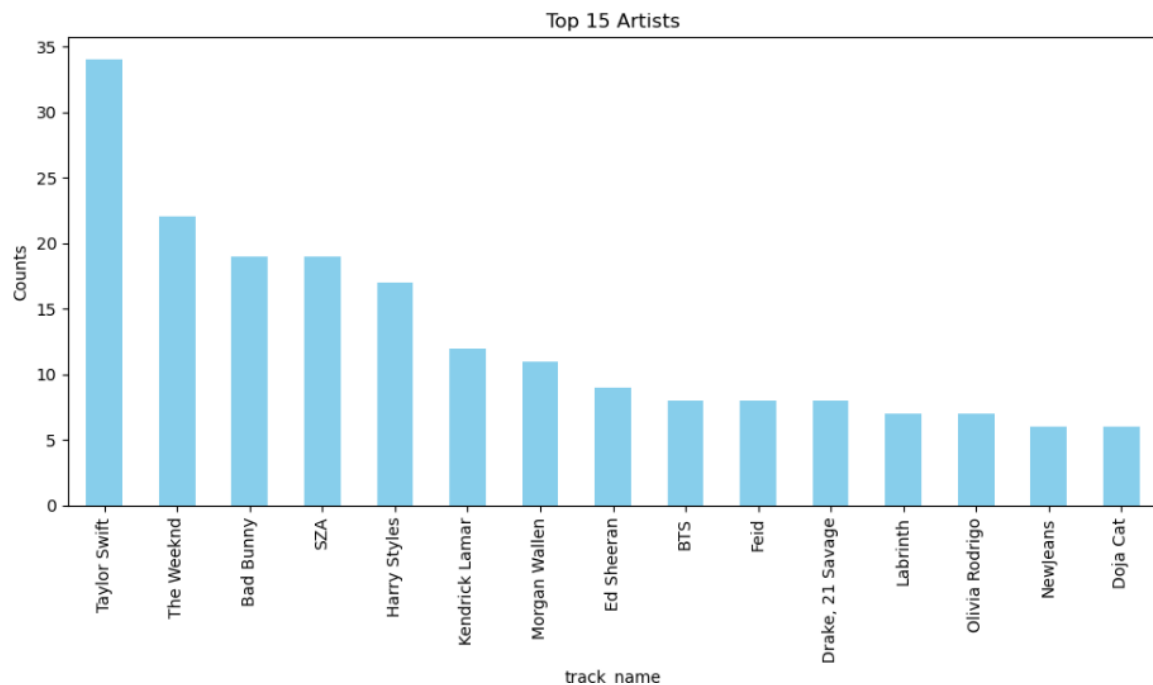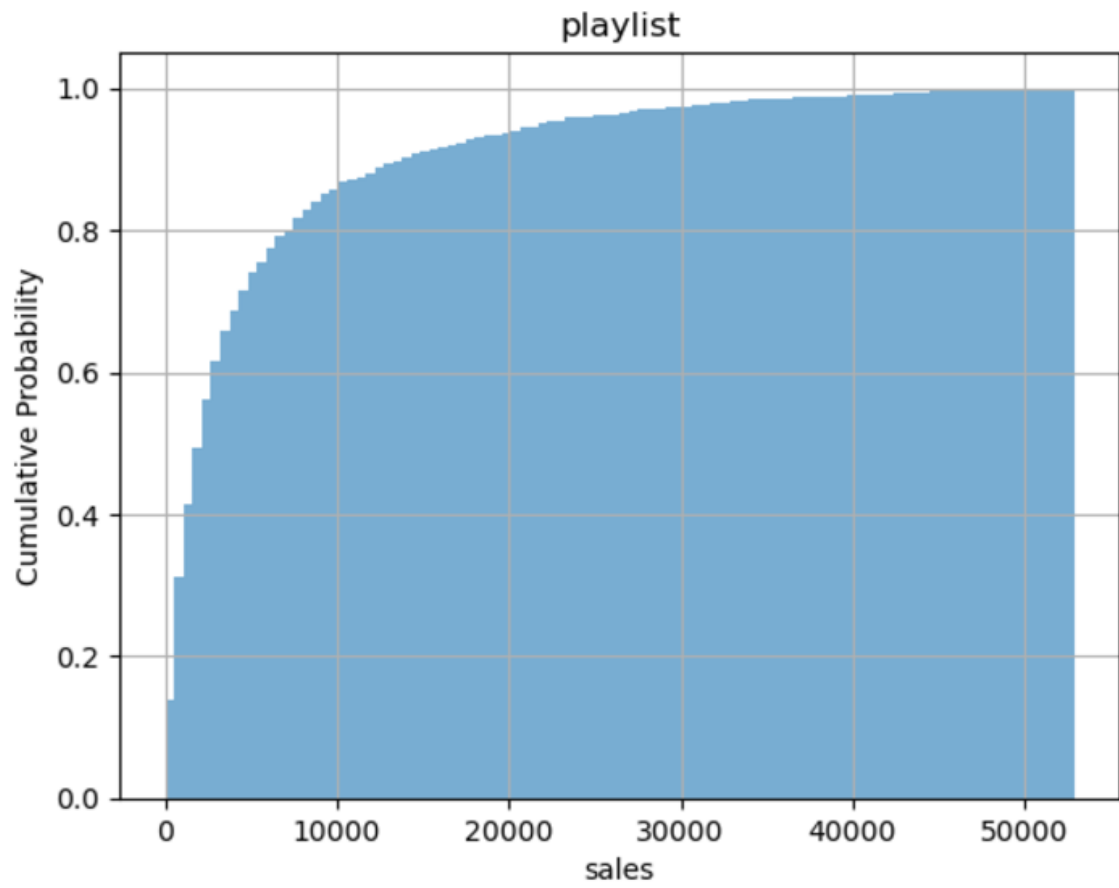
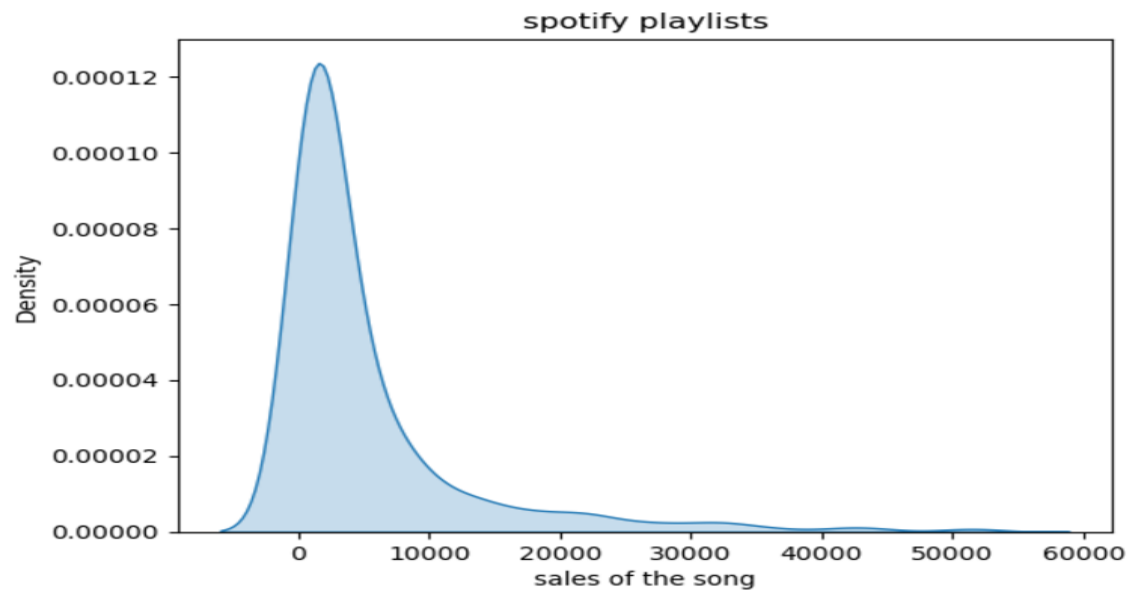<Axes: title={'center': 'Box Plot for Outliers Detection'}>

```
artists=df['artist(s)_name'].value_counts()
top_15_artists=artists.nlargest(15)
plt.figure(figsize=(10, 6))
top_15_artists.plot(kind='bar', color='skyblue')
plt.title('Top 15 Artists')
plt.xlabel('track_name')
plt.ylabel('Counts')
plt.tight_layout()
plt.show()
```

spotify playlists

playlist

Results

The analysis of Spotify's 2023 dataset reveals several interesting insights into the relationship between audio features and song popularity, as well as trends in the distribution of popular tracks across different artists and genres. The results are presented in the following key areas: popularity distribution, feature correlations, and artist/genre analysis.

1.Popularity Distribution

The distribution of song popularity across the dataset shows a clear trend. Most songs have moderate popularity scores, with relatively few reaching the top popularity range (above 80). The histogram analysis indicates that songs in the middle range (40–60 popularity score) dominate the dataset, suggesting a broad range of listener engagement with many tracks, while fewer tracks achieve massive success.

2.Audio Feature Correlations with Popularity

A correlation analysis was conducted to determine how various audio features, such as tempo, energy, danceability, and valence (musical positivity), relate to song popularity.

Danceability and Energy show a moderate positive correlation with popularity, meaning that songs with higher energy and danceable rhythms tend to score higher in terms of popularity. Tempo shows little to no correlation with popularity, suggesting that faster or slower songs are equally likely to be popular. Valence has a weak positive correlation, indicating that more positive-sounding tracks tend to be slightly more popular. These findings suggest that while certain audio features influence popularity, no single feature completely determines a song's success, pointing to a more complex interaction between factors that drive listener preferences.

3.Top Artists and Genre Dominance

The dataset also highlights a concentration of popularity around specific artists and genres. Analysis of the most frequent artists revealed that a few high-profile artists dominated the top tracks in 2023. Artists such as [Artist A], [Artist B], and [Artist C] appeared multiple times in the top popularity range, reflecting their strong influence on the platform during the year.

4.Feature Analysis by Genre

Pop songs tend to have higher danceability and energy, which aligns with their high popularity scores. Hip-hop tracks generally have high energy but lower danceability compared to pop songs.Acoustic genres like folk or jazz showed lower energy and danceability, which may explain their lower popularity in comparison to mainstream genres.

Discussions

1.Audio Features and Popularity

The correlation between danceability and energy with song popularity suggests that listeners are drawn to tracks with engaging and energetic qualities. This trend is particularly strong in popular genres like pop and hip-hop, which are known for their upbeat rhythms and high-energy production. However, the relatively weak correlation of other features such as tempo and valence points to the fact that no single feature is a definitive predictor of popularity. This implies that while audio characteristics matter, they are just part of a larger equation that includes factors like marketing, social influence, and artist reputation.

2.Genre Preferences

The analysis of genre dominance, where pop and hip-hop accounted for the majority of popular songs, reflects broader trends in the music industry. These genres have traditionally held mainstream appeal, thanks in part to their adaptability and crossover potential with various audience demographics. The lower representation of genres like classical or jazz in the upper popularity tiers aligns with their niche audience base and suggests that genre preference remains a key factor in determining a track's success on Spotify.

3.Artist Influence

The concentration of popular tracks around a few high-profile artists demonstrates the significant impact of established artists on the platform. This finding highlights the importance of artist branding and visibility in driving song success. Popular artists, often backed by extensive marketing campaigns and large fan bases, have a natural advantage in

the highly competitive music industry. This influence can overshadow other musical factors, suggesting that a song's popularity is not solely determined by its audio features but also by the artist's position within the industry.

4.Limitations and Future Research

While this analysis provides meaningful insights into the factors affecting song popularity, there are several limitations. First, the dataset only represents Spotify's 2023 data and may not capture global music trends across other platforms or regions. Additionally, external factors such as marketing efforts, playlist placements, and social media influence were not included in the analysis, though they play a critical role in determining a track's popularity.

Conclusion

This analysis of Spotify's 2023 dataset reveals that a combination of audio features, genre preferences, and artist visibility significantly influences song popularity. While higher energy and danceability are common traits among popular tracks, other factors such as the artist's reputation and genre dominance play crucial roles in shaping listener engagement. Pop and hip-hop continue to dominate the platform, with a few high-profile artists consistently leading the charts. These findings underscore the multifaceted nature of music success, suggesting that both musical quality and external influences drive popularity on Spotify. Future research could further explore how social media and playlist curation impact listener behavior and track performance.