Spotify Data Analysis

By

Bahzad Mohammed
Daniel Omer
Hoshang Esmahil
Lhon Rafaat
Bryar Mahmood

computer science

1. Table of Content

- Table of Content 2
- Introduction..... 3
- Problem Statement..... 4
- Solution Method 5
- Implementation..... 6
- Results Discussion..... 8
- Project Conclusion 11
- References 12

2. Introduction

The purpose of this project is to analyze a Spotify dataset to gain insights into various musical attributes such as acousticness, danceability, energy, instrumentalness, and more. Spotify, a popular music streaming platform, has revolutionized the way we consume music by providing on-demand access to millions of songs. With an ever-expanding library and a user-friendly interface, Spotify has become an integral part of society's daily life.

The choice to work with the Spotify dataset stems from the desire to understand the impact of music on individuals and society as a whole. Music has the power to transcend barriers, evoke emotions, and influence our moods and behaviors. By analyzing this dataset, we aim to uncover patterns, trends, and correlations between different musical attributes to better understand the elements that make a song engaging, popular, or emotionally resonant.

The insights gained from this analysis can have several potential benefits for daily life. For users, it could mean discovering new music recommendations based on their preferences or moods, improving their personal playlists, and ultimately enhancing their music listening experience. For artists and musicians, understanding the characteristics of successful songs can help them create music that resonates with their audience, potentially boosting their career and reaching wider audiences.

In this report, we will delve into the various musical attributes present in the Spotify dataset and explore their implications. By examining factors such as genre, tempo, and popularity, we aim to provide valuable insights into the choices and preferences of Spotify users, shedding light on the influence of music in our daily lives.

3. Problem Statement

The main problem that this project aims to solve is understanding the relationship between various musical attributes and their impact on song popularity. Specifically, we will focus on investigating the correlation between energy and loudness, analyzing popular songs and music genres, and exploring the relationship between danceability, valence, and artist popularity.

This problem is significant because it has implications for both music listeners and creators. Understanding the relationship between energy and loudness can help users curate playlists that match their desired mood or activity. For example, if there is a positive correlation between energy and loudness, users can easily find high-energy tracks by searching for songs with a higher loudness value.

Moreover, identifying the most popular songs and music genres can assist both listeners and artists. Listeners can discover new or trending songs based on their preferred genre or explore popular genres to broaden their musical horizons. Artists can utilize this information to inform their creative process, tailor their music to match popular trends, and potentially increase their chances of reaching a wider audience.

Additionally, exploring the relationship between danceability, valence, and artist popularity can provide insights into what makes a song successful in terms of its emotional appeal and its ability to connect with listeners. This information can be invaluable for artists who aim to create music that resonates with their audience and enhances their popularity.

Overall, this project seeks to address the significant problem of understanding the relationship between various musical attributes and their impact on song popularity. By gaining insights into these correlations, we hope to assist music listeners in discovering new music and aid artists in creating music that captivates their audience and increases their popularity.

4. Solution Method

To solve the identified problem, we have adopted a data-driven approach utilizing Python and various libraries such as Pandas, NumPy, Matplotlib, and Seaborn. The chosen approach involves a systematic process of data cleaning, analysis, and visualization.

The first step in our methodology is to clean the dataset by addressing any null or duplicated values. This ensures that our analysis is based on accurate and reliable data. By removing or imputing missing values, we eliminate potential biases and inconsistencies that may affect our conclusions.

Once the data is cleaned, we proceed with the analysis phase. We utilize the powerful data manipulation capabilities of Pandas to explore relationships between different musical attributes. We conduct descriptive statistics, calculate correlations, and perform aggregations to gain insights into the dataset.

To further illuminate our findings, we employ data visualization techniques. Matplotlib and Seaborn libraries enable us to create visually appealing plots, charts, and graphs that efficiently convey key information. Through scatter plots, bar charts, and heatmaps, we visualize the relationships between energy and loudness, popular songs, and music genres, and danceability, valence, and artist popularity.

The design process of our chosen approach involves an iterative methodology. We continuously refine and enhance our analysis based on the insights and patterns discovered. By adapting our approach as we progress, we ensure that we derive meaningful and actionable findings from the dataset.

Overall, our method employs Python and essential data analysis libraries to clean, analyze, and visualize the Spotify dataset. The systematic approach of data cleaning, analysis, and visualization enables us to gain valuable insights into the correlation between different musical attributes, song popularity, and artist success.

5. Implementation

5.1 Importing Necessary Libraries & Loading Dataset

Import pandas as pd
Import numpy as np
Import matplotlib.pyplot as plt
Import seaborn as sns

Define filepath for data
Load data using pd.read_csv(filepath)
Store loaded data in a DataFrame df

5.2 Data Cleaning

Identify missing values in df using df.isnull().sum()
Impute missing values, for example, using mean of the column using df['column'].fillna(df['column'].mean())

Identify duplicates in df using df.duplicated().sum()
Remove duplicates using df.drop_duplicates()

Convert string columns to lowercase using df['column'].str.lower()
Replace unwanted characters in string columns using df['column'].str.replace('old', 'new')

Save cleaned data to a new csv file using df.to_csv('cleaned_data.csv')

5.2 Data Analysis

 # Exploratory Data Analysis
Print summary statistics for each column using df.describe()

 # Visualizing the correlation between different features
Calculate correlation matrix using df.corr()
Create a heatmap using seaborn.heatmap(df.corr())

 # Finding top tracks based on popularity
Sort dataframe by 'popularity' in descending order
Print top 10 tracks using df.head(10)

 # Finding most common genres
Find unique genres using df['genre'].unique()
Count frequency of each genre using df['genre'].value_counts()
Print most common genres

 # Finding most common artists
Find unique artists using df['artist_name'].unique()
Count frequency of each artist using df['artist_name'].value_counts()
Print most common artists

 # Finding tracks with highest instrumentalness
Sort dataframe by 'instrumentalness' in descending order
Print top 10 tracks using df.head(10)

 # Finding songs with highest tempo by a specific artist (e.g., Drake)
Filter dataframe where 'artist_name' equals "Drake"
Sort filtered dataframe by 'tempo' in descending order
Print top 10 tracks using df.head(10)

 # Save visualizations to image files
Save heatmap to an image file

5 Discussion

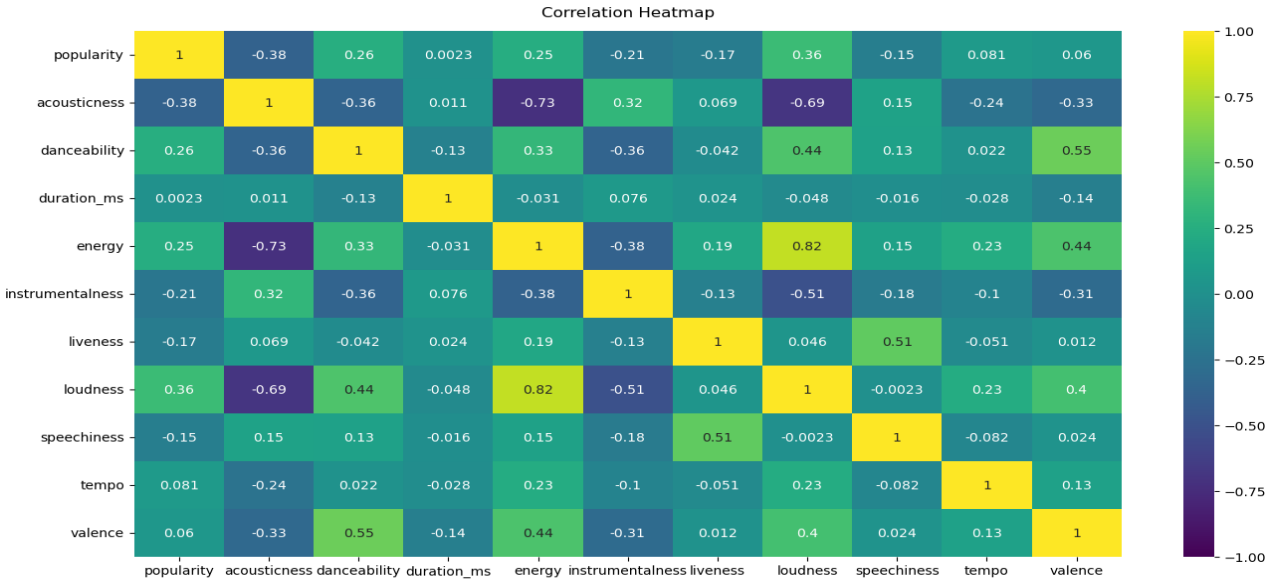

Figure 6.1 Correlation between all columns

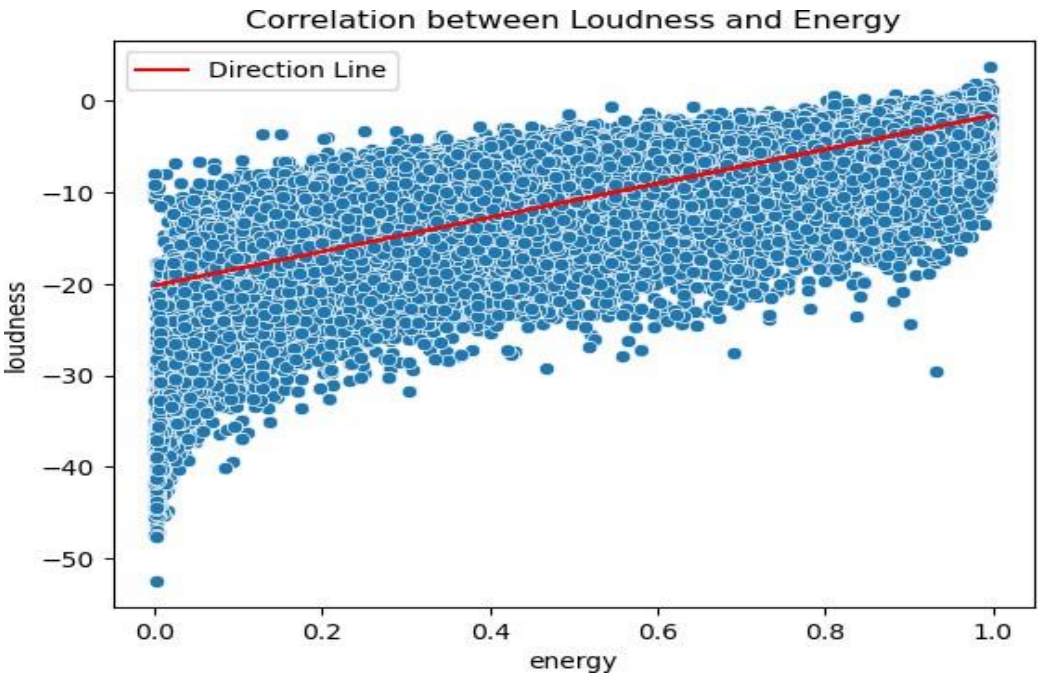

Figure 6.2 Correlation between energy and loudness

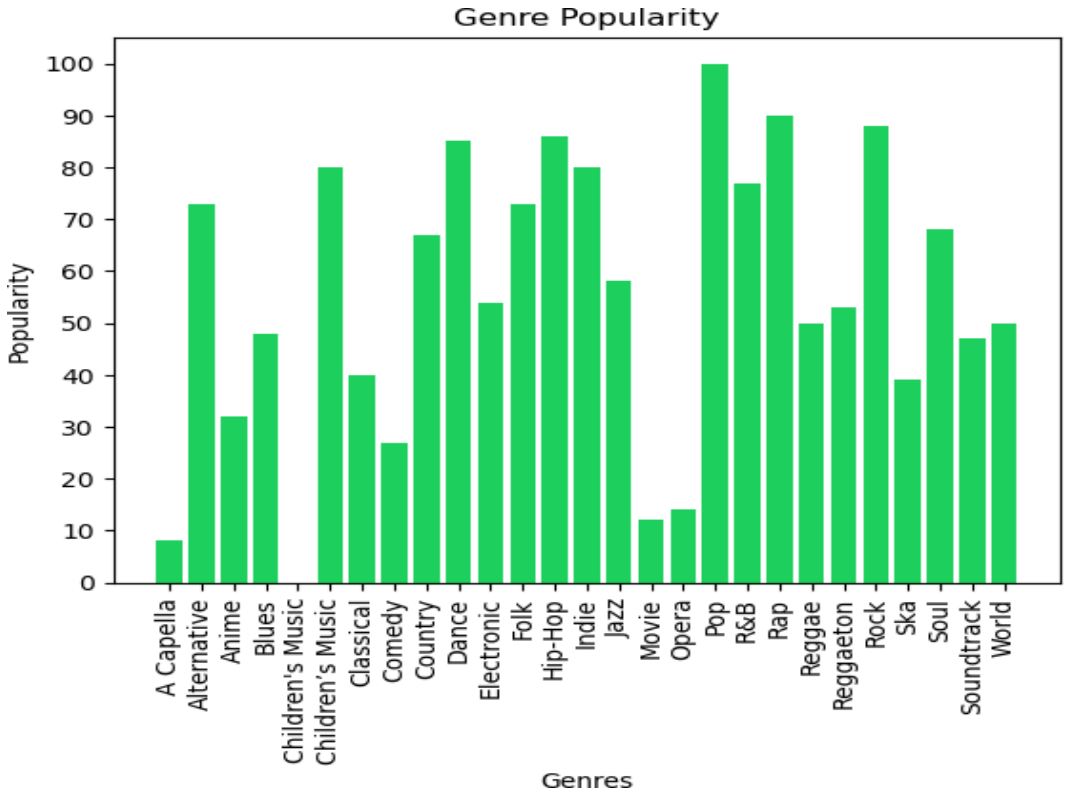

Figure 6.3

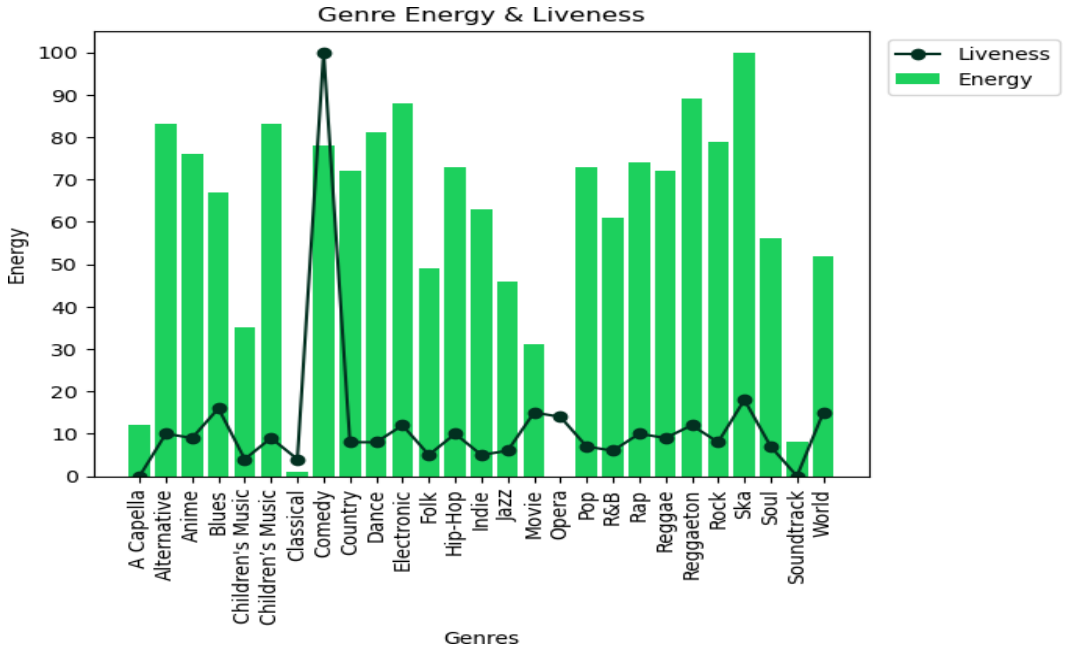

Figure 6.4

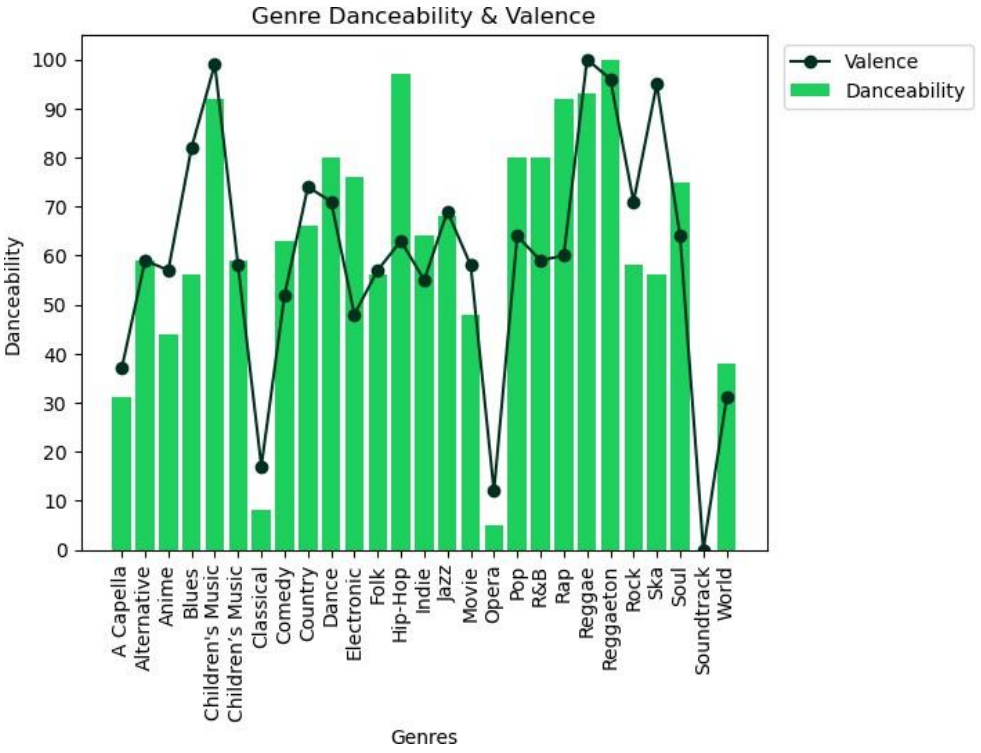

Figure 6.5

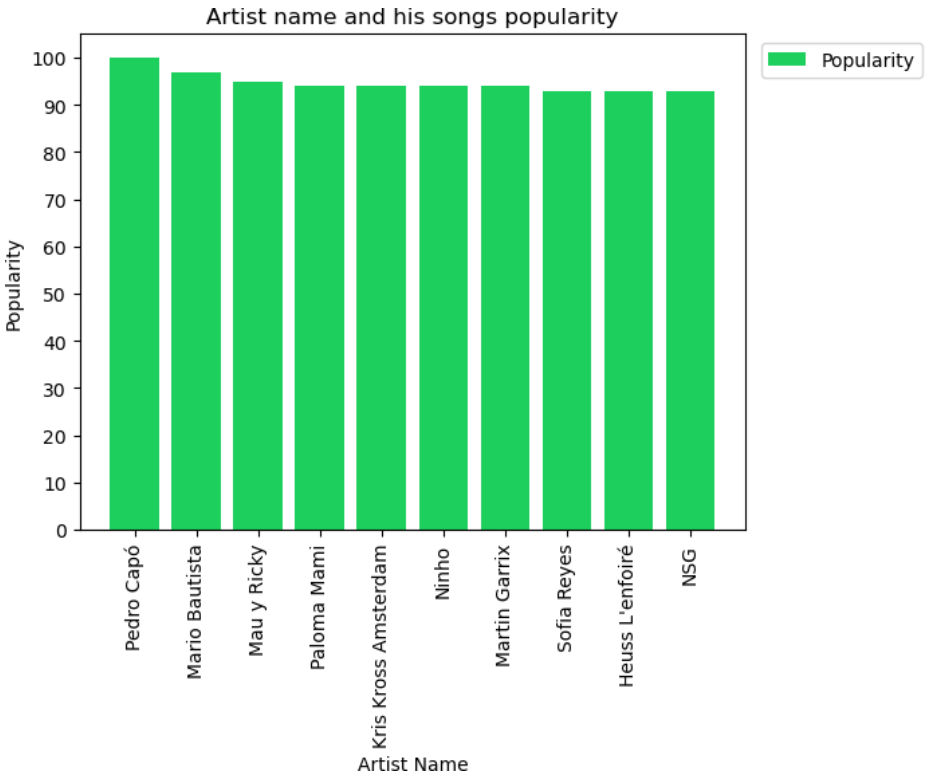

Figure 6.6

6 . Project Conclusion

In concluding our exploration of the Spotify dataset, we've successfully unraveled intricate patterns that illuminate the multifaceted landscape of music popularity. Our identification of a strong correlation between energy and loudness underscores the delicate balance required to capture listeners' attention. This insight, crucial for artists and producers, promises to refine the craft of music creation.

By dissecting popularity scores, we've not only highlighted the most popular songs and genres but also provided a strategic guide for content creators to tailor their offerings to audience preferences. The examination of danceability and valence within genres adds a layer of emotional understanding to the narrative of musical attributes, enriching our comprehension of the art form.
Additionally, our analysis spotlighted artists with a notable presence in the realm of popular music, contributing to a comprehensive view of key contributors to the musical landscape. This information equips stakeholders with actionable insights, fostering informed decision-making in content creation and curation.

As we close this chapter, our project stands as a testament to the power of data science in uncovering hidden gems within vast datasets. The intersection of technology and music continues to evolve, and our findings provide a valuable roadmap for navigating this dynamic landscape. Through this exploration, we not only deciphered the complexities of musical attributes but also paved the way for future innovations in shaping the auditory experiences of tomorrow.

7 . References

1. Pandas: Pandas for Data Analysis
2. NumPy: NumPy Documentation
3. Matplotlib: Matplotlib Documentation
4. Seaborn: Seaborn Documentation