Team Members :
Hari Nair Suresh Chandran (UF ID - 24745989)
Arjun Gopalakrishnan Kaliyath (UF ID - 32205237)

Objective of the Project :
We aim to build a music recommendation engine using our datasets that will suggest songs to the end users based on the interests they specify such as genre,mood or artist preferences. Using the datasets that we have, we also intend to explore the evolutionary trends in music across time and explore the characteristics of what makes a hit song and look at correlations between artist popularity and a song becoming a hit.

Type of Tool :
We intend to build an interactive dashboard to visualize our data and generate recommendations and a conversational chatbot to interact with our recommendation engine.

Interactive Dashboard -

1. Will have a component to display the evolution of popular music across decades by genre, characteristics of songs such as acousticness, valence etc.
2. Trends to display popularity of genres across decades.
3. Artist longevity analysis to illustrate which artists have stayed relevant over multiple decades.
4. Create and visualize clusters of genres with similar musical characteristics.
5. Display trends in modern music in the streaming era such as time taken for a song to peak, Visualize popular artist data and correlation of artist popularity with song popularity.
6. Recommendation engine component to recommend similar songs based on genre preference, mood preference, artist preference.

Conversational agent -

1. A chatbot that can be used to interact with the recommendation engine to suggest songs.

Data to be used :

We have obtained the datasets that we would require for our project from the following Kaggle sources -

- Spotify Top 10000 most streamed songs - (Licensing - CC0 Public Domain)
  https://www.kaggle.com/datasets/rakkesharv/spotify-top-10000-streamed-songs

  Size - 644 KB

Columns - 9 Columns including artist name, song name, peak position, total streams.
Rows - Contains 11084
Description - We will be using this dataset to analyze the trends in modern music in the streaming era and analyze the most popular artists in the modern age.

- Large random spotify artists metadata - (Licensing - Open Database License(ODbL))
  https://www.kaggle.com/datasets/sarahjeffreson/large-random-spotify-artist-sample-with-metadata

  Size - 1,297 KB
  Columns - 10 including artist name, genre, followers, monthly listeners, year of first release, year of last release, the total number of releases the artist has had, popularity score
  Rows - 15027
  Description - We will be using this dataset to examine correlation between artist popularity and song success. We will also examine trends in artist longevity and the correlation of artist production output to their streaming popularity.

- Spotify Tracks Genre - (Licensing - CC0 Public Domain)
  https://www.kaggle.com/datasets/thedevastator/spotify-tracks-genre-dataset

  Size - 19647 KB
  Columns - 21 including song name, genre, popularity, duration, time_signature and musical characteristics such as danceability, liveness, valence, acousticness, popularity
  Rows - 114000
  Description - A collection of approx 90000 tracks across 125 genres retrieved from Spotify using their API and with musical characteristics as determined by spotify. This will be used as the main Training set for training our musical recommendation engine and also to visualize data relationships between duration and song success, track genres and popularity metrics and how musical characteristics define song genres.

- Top 10000 songs on Spotify 1950 - Now - (Licensing - CC0 Public Domain)
  https://www.kaggle.com/datasets/joebeachcapital/top-10000-spotify-songs-1960-now

  Size - 7635 KB
  Columns - 35 columns including artist name, song name, genre, date of release.
  Rows - 9997
  Description - A collection of most popular songs across the billboard and ARIA charts retrieved From spotify using their API along with the spotify defined musical characteristic metrics such as Liveness, loudness, valence, popularity.

This dataset will be used to visualize the evolutionary trends in music across 7 decades from the 1950s till present day.

Tech Stack :

We will be using python entirely for the project with additional frameworks.

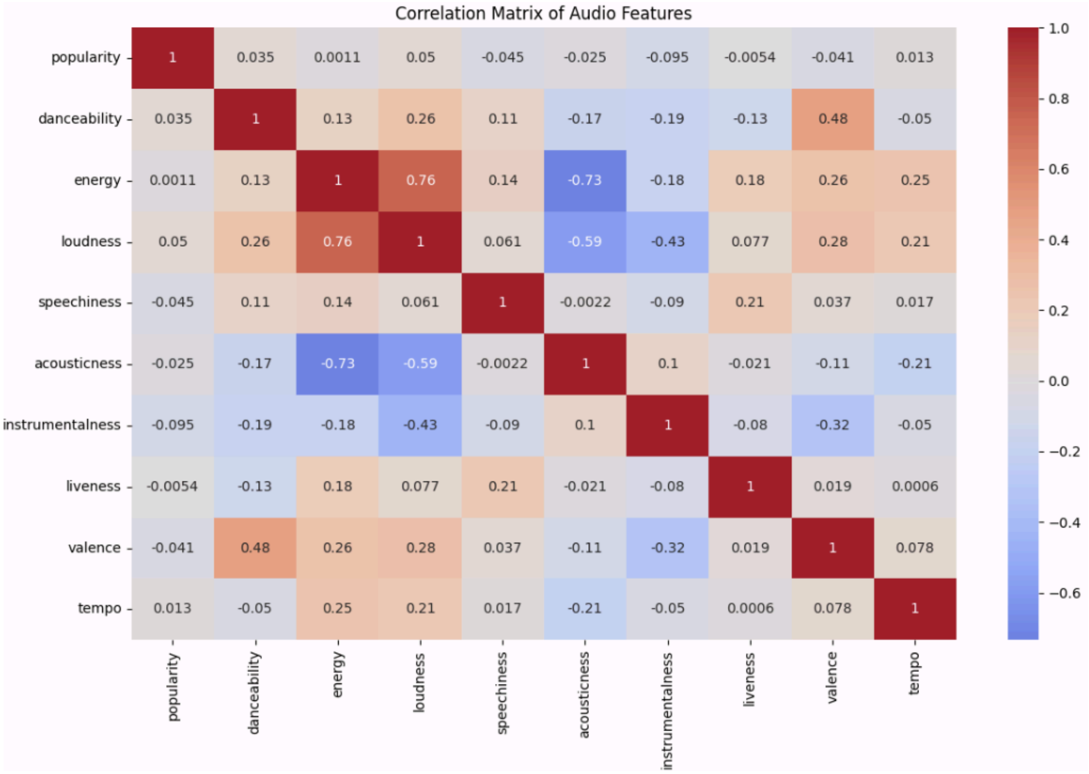
- Interactive Dashboard - Streamlit to build an interactive web app for the dashboard.
- Chatbot - langchain to build our conversational agent.
- ML models - Scikit-learn to train on our datasets for classification and clustering.

Project Timelines :

1. Data collection, Data Preprocessing and Exploratory Data Analysis - 2/23/25
2. Feature engineering - 3/3/25
3. Data modeling and training model - 3/17/25
4. Evaluation of model performance and conclusions - 3/31/25
5. Creation of interactive dashboard and conversational agent - 4/23/25

Exploratory Data Analysis :

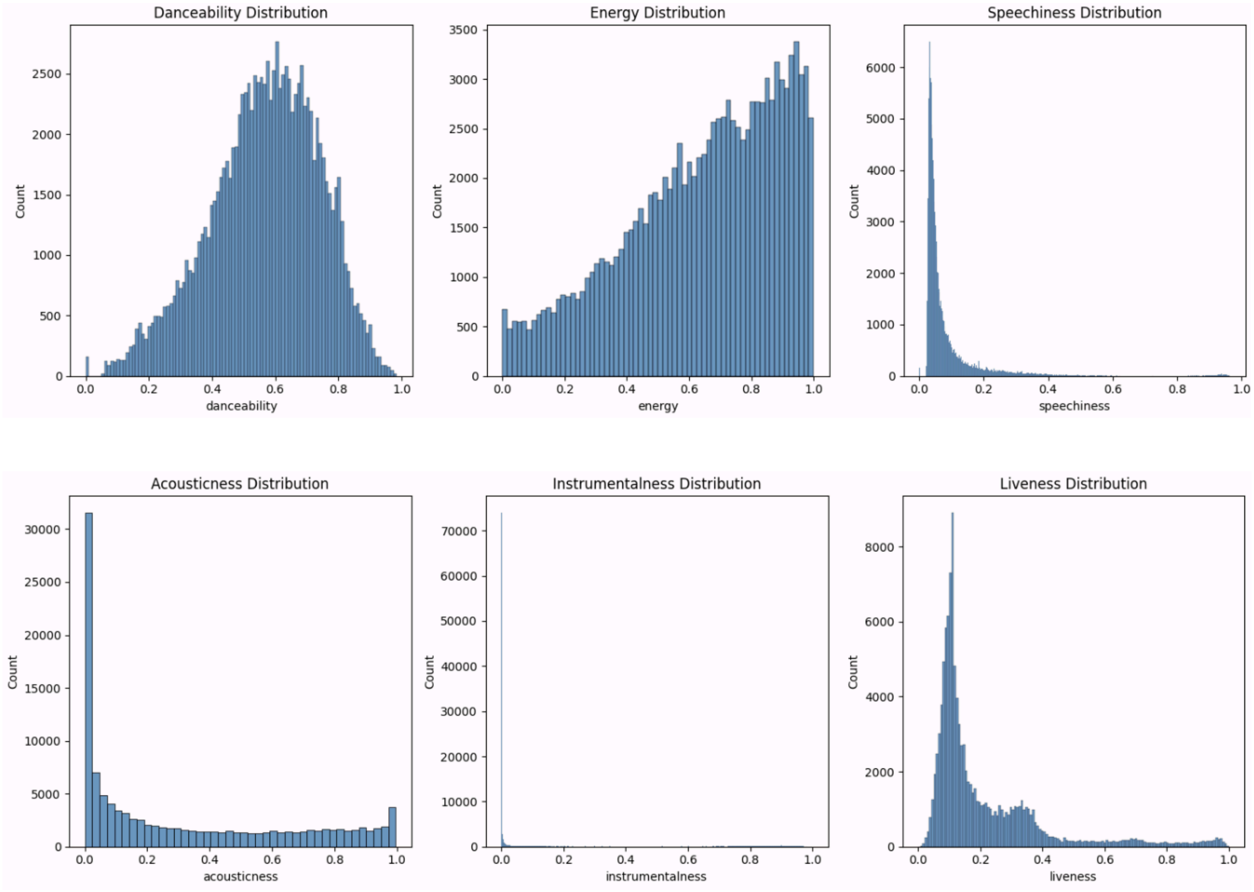
For analysis of our main dataset, we generated the following visualizations :



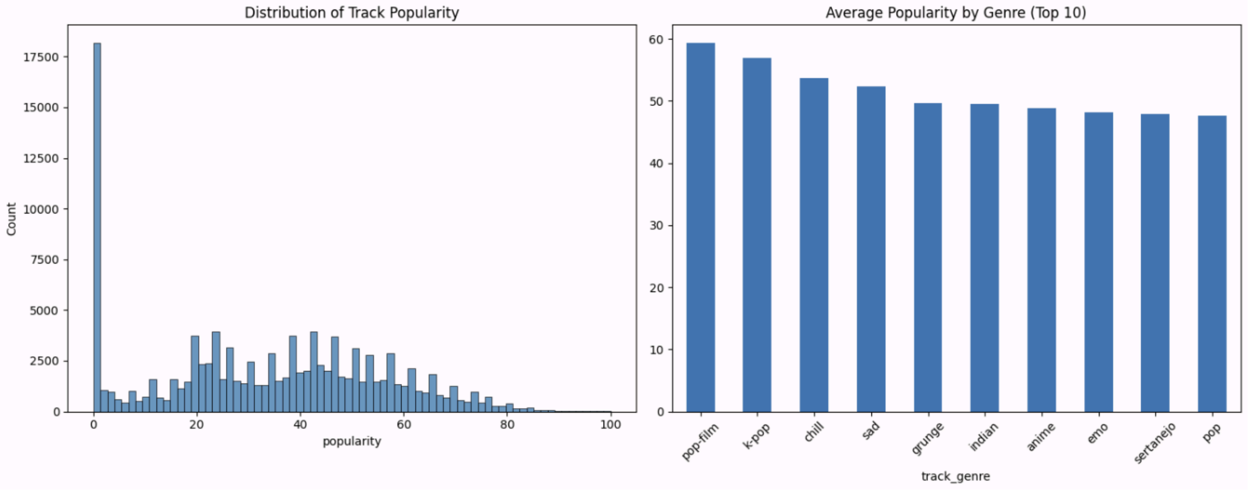
Correlation Matrix of Audio Features

	popularity	danceability	energy	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo
popularity	1	0.035	0.0011	0.05	-0.045	-0.025	-0.095	-0.0054	-0.041	0.013
danceability	0.035	1	0.13	0.26	0.11	-0.17	-0.19	-0.13	0.48	-0.05
energy	0.0011	0.13	1	0.76	0.14	-0.73	-0.18	0.18	0.26	0.25
loudness	0.05	0.26	0.76	1	0.061	-0.59	-0.43	0.077	0.28	0.21
speechiness	-0.045	0.11	0.14	0.061	1	-0.0022	-0.09	0.21	0.037	0.017
acousticness	-0.025	-0.17	-0.73	-0.59	-0.0022	1	0.1	-0.021	-0.11	-0.21
instrumentalness	-0.095	-0.19	-0.18	-0.43	-0.09	0.1	1	-0.08	-0.32	-0.05
liveness	-0.0054	-0.13	0.18	0.077	0.21	-0.021	-0.08	1	0.019	0.0006
valence	-0.041	0.48	0.26	0.28	0.037	-0.11	-0.32	0.019	1	0.078
tempo	0.013	-0.05	0.25	0.21	0.017	-0.21	-0.05	0.0006	0.078	1

The above correlation map indicates some of the intuitive features from our dataset, the music characteristics show that energy is highly correlated with loudness i.e. songs that tend to be loud are more energetic. Another interesting metric is that valence is strongly correlated to danceability which indicates that songs that convey positive emotion generally tend to be the dance songs. Acousticness and instrumentalness are negatively correlated with energy and valence which indicates that songs that feature more acousticness and instrumentalness tend to be melancholic in nature.

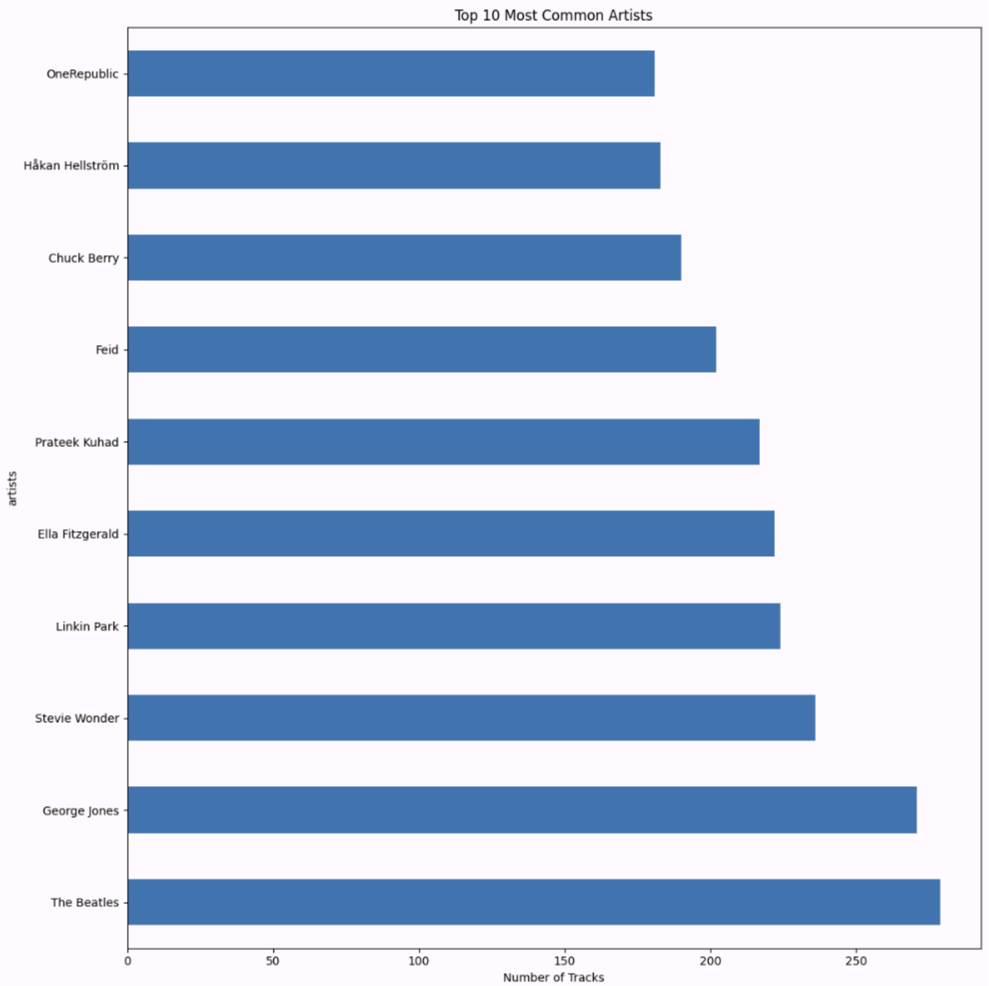


The above histogram plots indicate the occurring frequencies of the various musical characteristics as described by Spotify (normalized between 0 and 1) in our dataset.



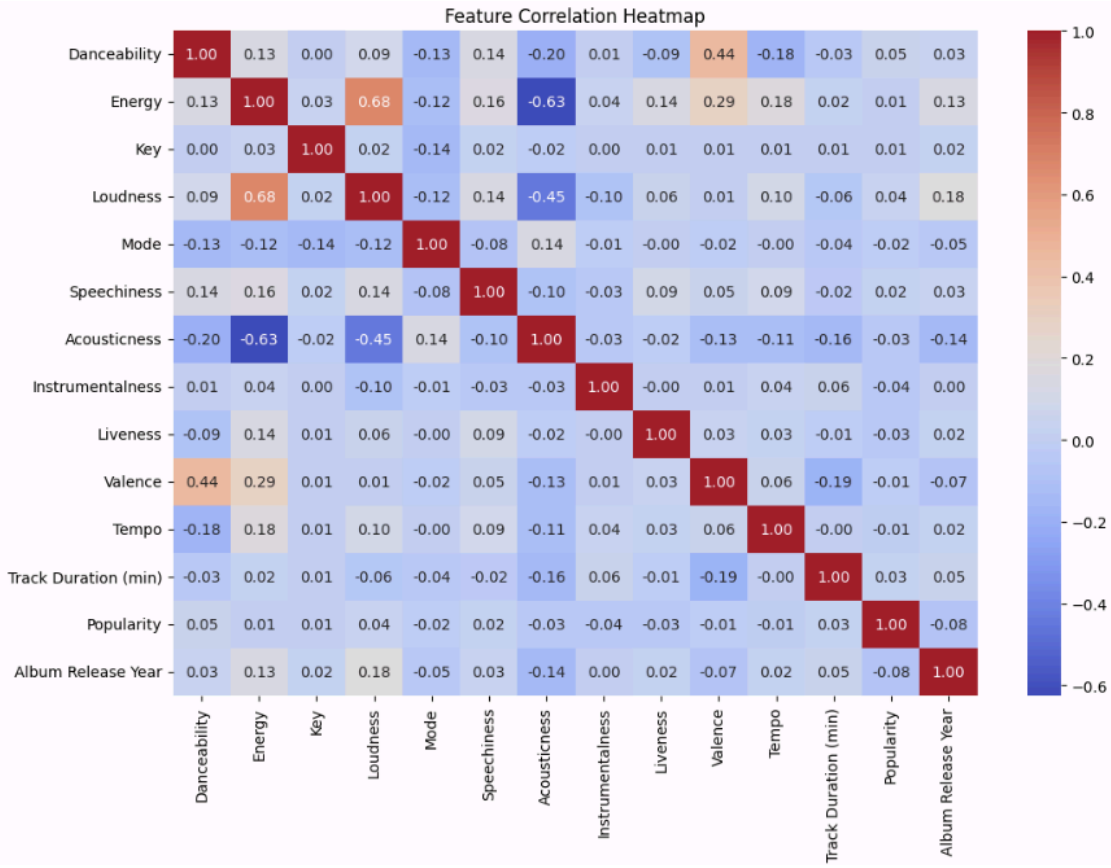
The above plots show the distribution of tracks based on the popularity metric defined by Spotify and the average popularity score for the top genres based on occurrence in the dataset.

The below horizontal bar chart indicates the most commonly occurring artists in our dataset with the Beatles the most commonly occurring artist.



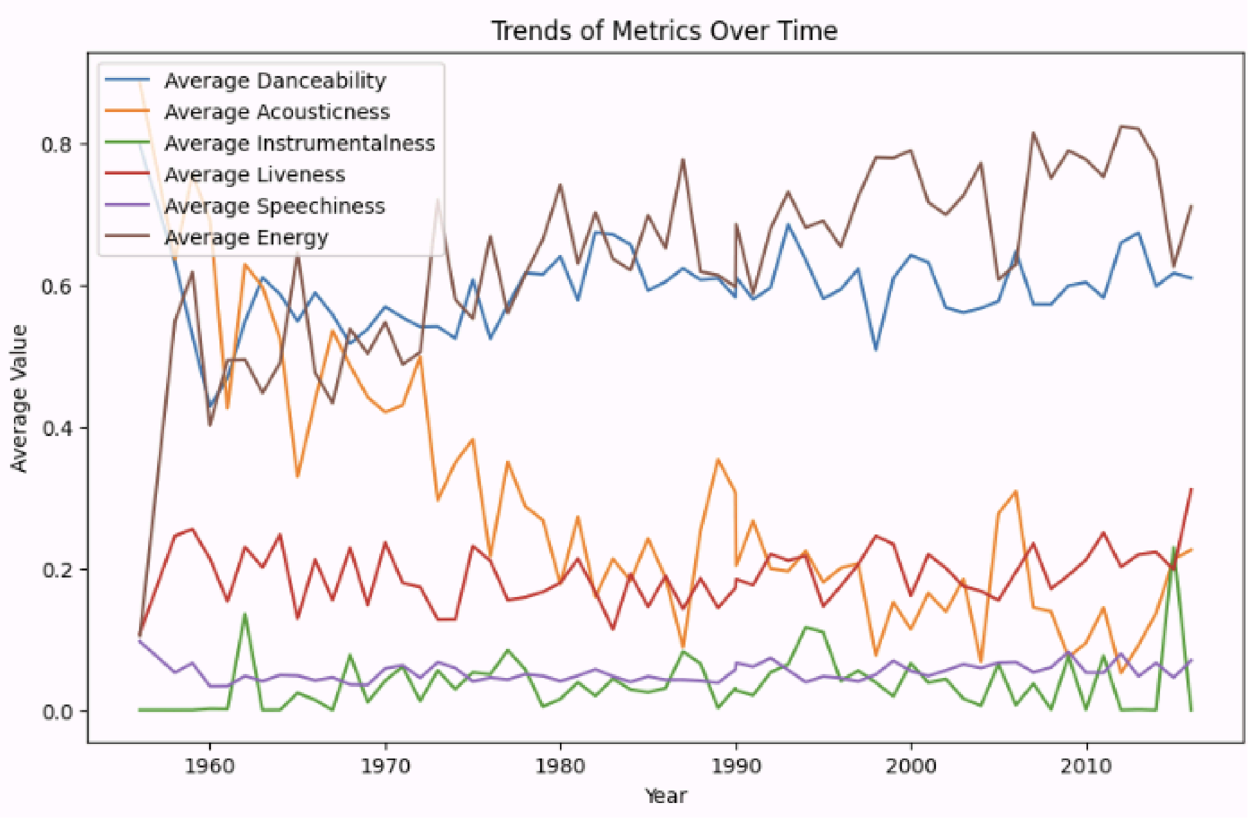
The dataset consists of 114 unique genres and 31347 unique artists.

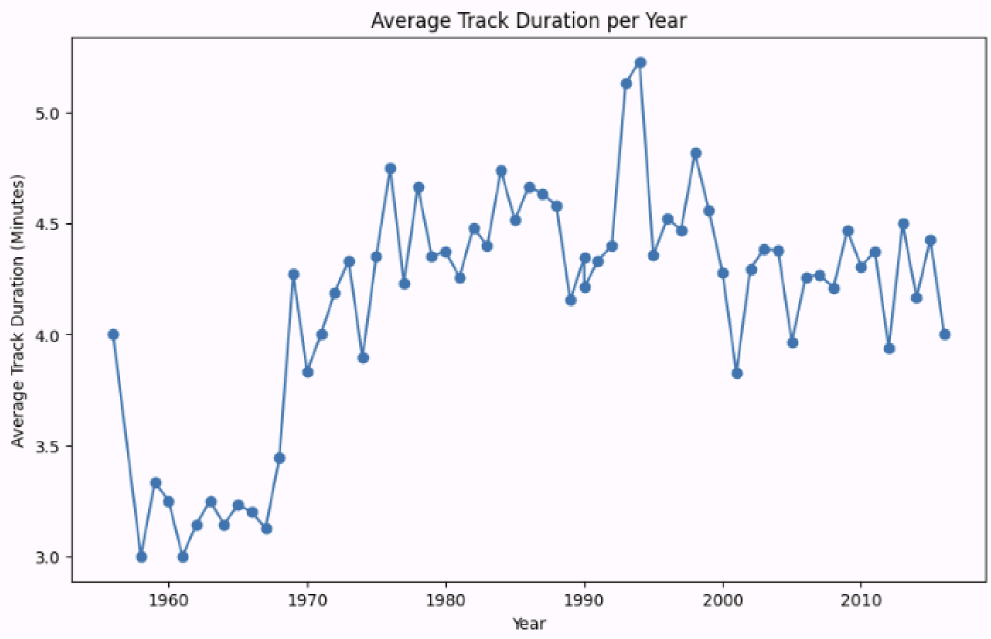
For the analysis of music trend evolution, we performed the following exploratory data analysis :

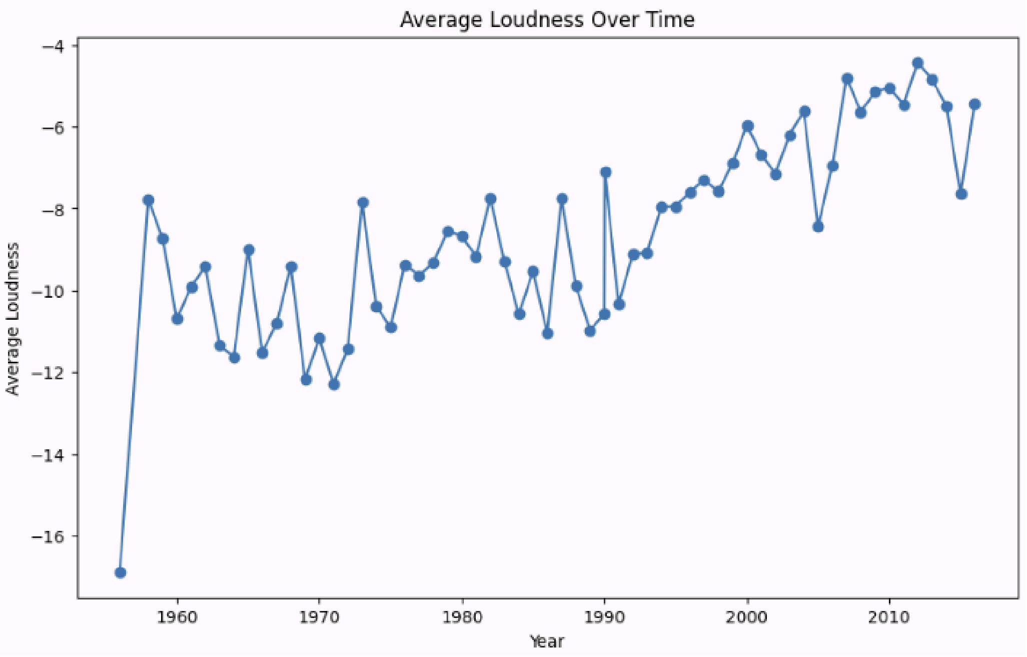


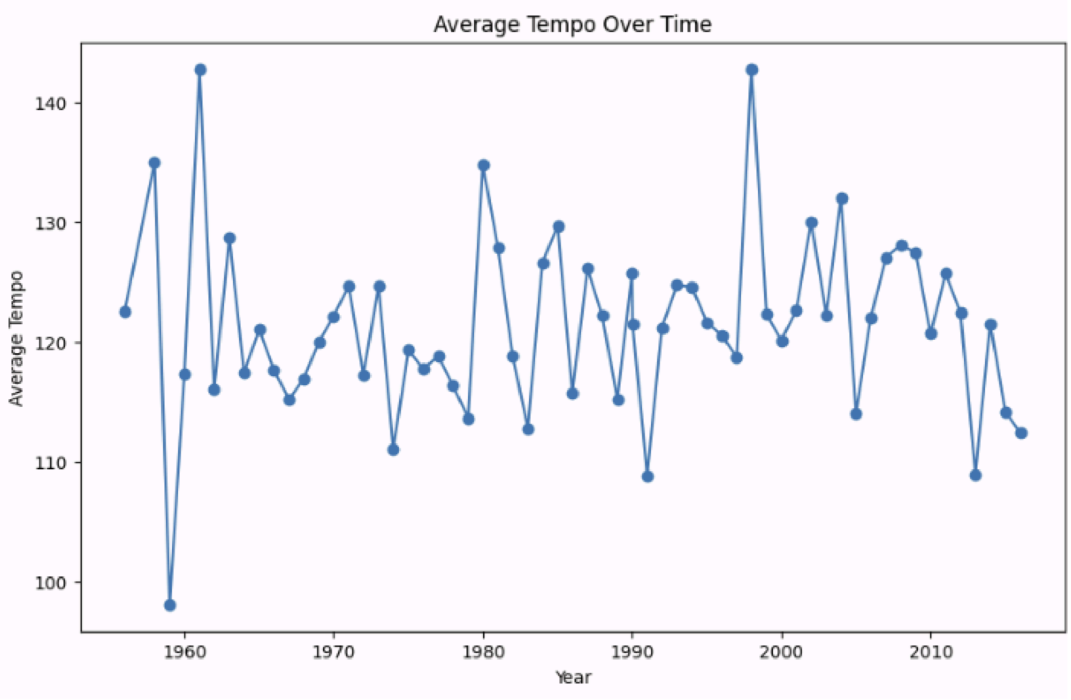
The trends indicated by the correlation heatmap in the case of the dataset of popular songs from 1950 to now indicate the same kind of correlations as seen in the case of our main training dataset which indicates that overall arching theme of musical characteristics is universal across several decades of music.

The below line graphs indicates how the musical characteristics have evolved over the six decades represented in our dataset



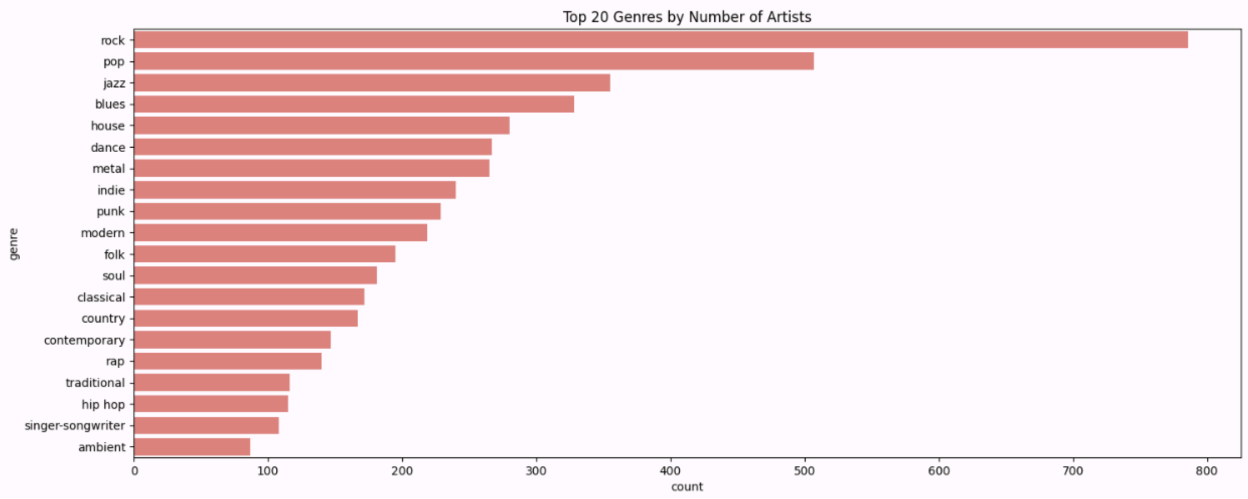


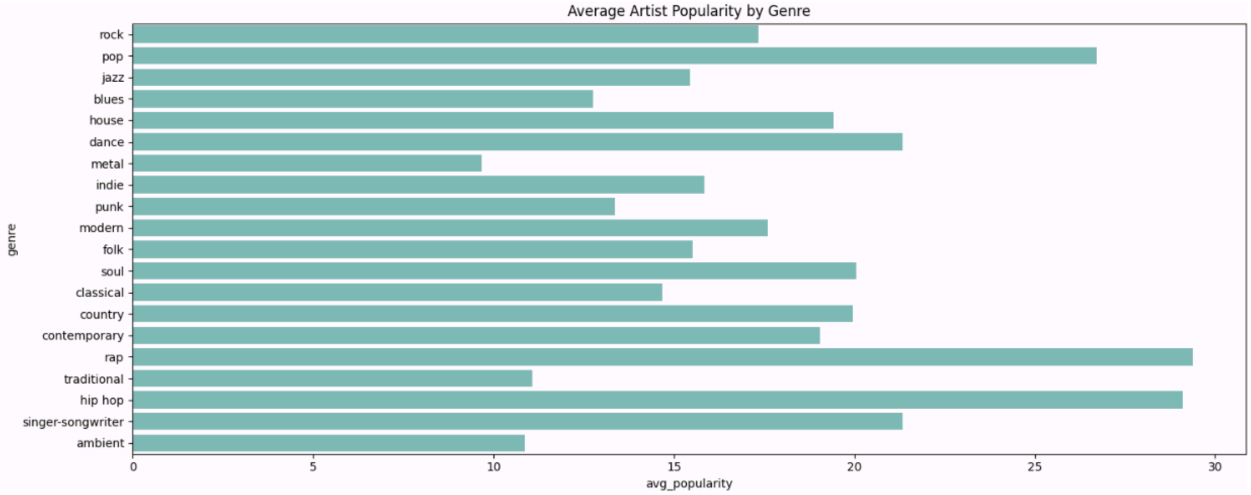




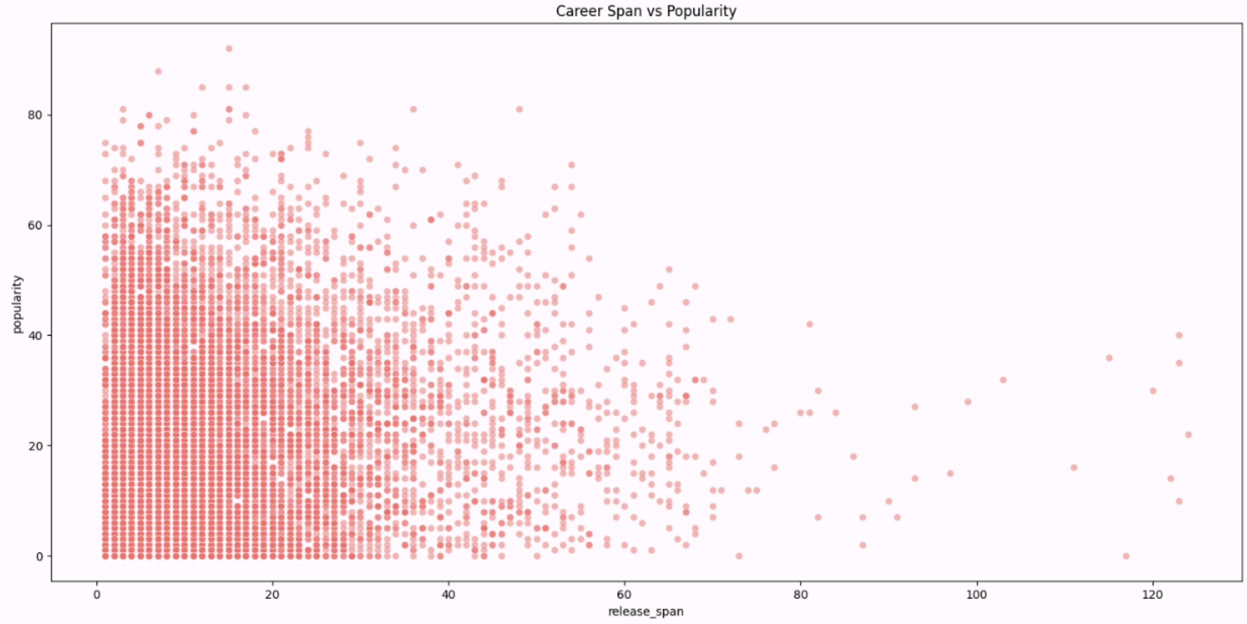
The trends indicate that while the average track duration has consistently stayed between the 3 and a half minute mark and the 5 minute mark apart from a peak in the mid-90s where it was above 5 minutes, the average loudness of tracks has steadily increased over time and the average track tempo of popular tracks has shown no clear discernible pattern.

For exploring artist popularity metrics, we generated the below visuals :

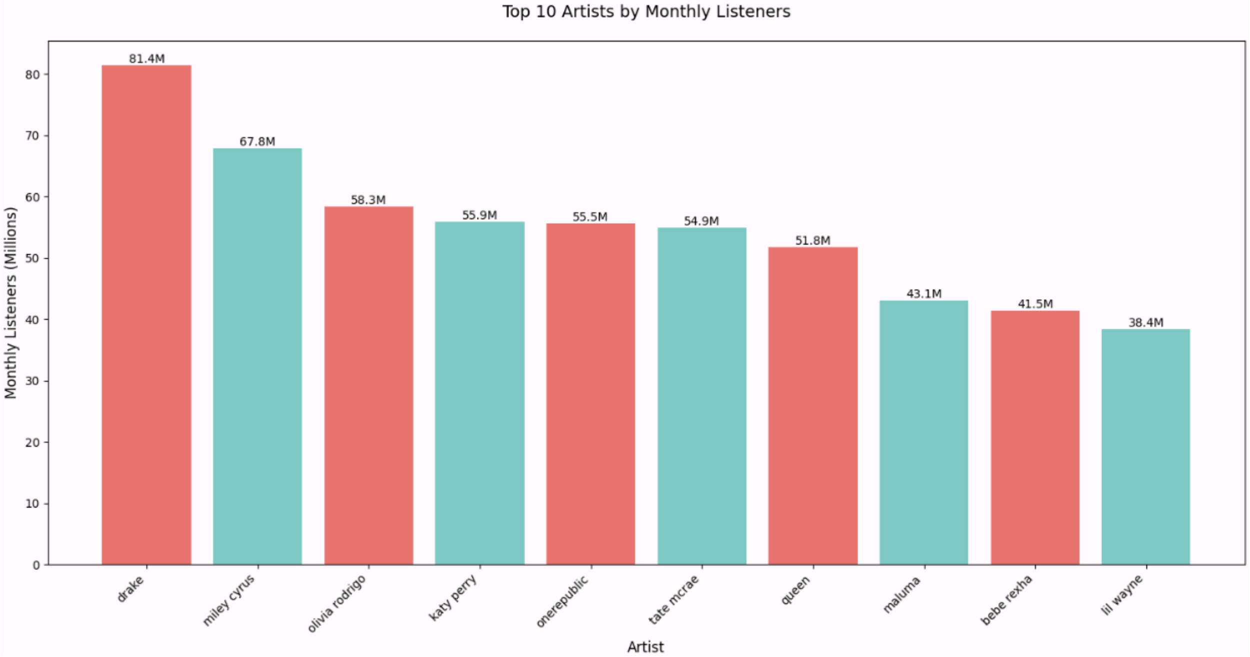




The above two horizontal bar graphs indicate the distribution in our dataset and the genres which are most represented by the artists in our dataset. Also an interesting observation is the average popularity of artists by genre indicating rap and hip hop as the leading genres among the popular artists followed by pop. Niche genres like metal have the least popular artists.



The above scatter plot points to an interesting observation of how some artists have been able to stay relevant over forty years, although they are clearly outliers and the data indicates that artists who are currently in the nascent stages of their careers or within the first twenty years of their careers tend to be more popular in the mainstream.



The above bar graph indicates the most popular artists based on our dataset with Drake having the most number of monthly listeners followed by Miley Cyrus.

Glossary (Important column name definitions as defined in the data) :

Popularity - a score indicating how popular a song or artist is on spotify on a scale of 1 to 100. Popularity score is mostly determined by how many plays a song has and how recent those plays are.

Danceability - a score assigned by spotify which indicates how suitable a song is for dancing on a scale of 0 to 1.

Energy - indicates intensity and fast paced-ness of a song from 0 to 1.

Loudness - indicates how loud or quiet a song is in decibels. Positive values suggest louder songs whereas negative values indicate quieter songs.

Duration_ms - indicates the duration of a track in milliseconds.

Key - represents musical keys such as A,B,C which are represented in integers ranging from 0 to 11.

Valence - measures the positiveness contributed by a track. High valence indicates happy tracks.

Tempo - represents the speed of a song indicated in bpm (beats per minute)

Mode - represents the tonal mode of a track (major or minor indicated by 1 and 0 respectively)

Speechiness - a score from 0 to 1 which represents spoken words in a track.

Acousticness - a score from 0 to 1 which represents the acoustic quality of a track.

Instrumentalness - a score from 0 to 1 which represents the likelihood of a track being instrumental.

Liveness - a score from 0 to 1 which indicates a live audience presence in the track.

Time_signature - the number of beats within a bar of the track.

Monthly_Listeners - The number of listeners that listens to the artists' tracks on a monthly basis. It is usually a strong indicator of an artist's popularity on Spotify.

Followers - The number of people that are subscribed to an artist on Spotify.