SPOTIFY DATA ANALYSIS

1.PROJECT OVERVIEW

As a person has a keen interested in listening to music, someone who spends around 3 hours a day immerse in music on Spotify. I often find myself curious about the trends behind the song I love. What genres dominate the charts? Which artists consistently deliver the hits? And how does Spotify curate those personalized playlist that feel like they know me so well?

This project dives into exploratory data analysis (EDA) to uncover patterns in music preferences. By analyzing top hits, we will explore which genre resonate the most and highlight the artists who have captured the most attention. The insight gained will also shed light on the algorithms that power Spotify's recommendations, offering a glimpse into how music streaming platform shape out listening experiences.

2.DATA SET

- **genre:** Type of song
- **artist_name:** Artist name
- **track_name:** Name of the song
- **popularity:** Popularity percentage of the song
- **danceability**: Describes how suitable a track is for dancing based on a combination of music elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is the best.
- **energy**: Describes how energetics the song is [0.0,1.0]
- **key:** key track is in Integer map to pitches using standard Pitch Class notation
  eg: 0 = C, 2 = D,..
- **loudness:** The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Values typical range between [-60,0] dB.
- **mode:** Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented 1 and minor is 0.
- **valence**: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (eg: happy, cheerful, euphoric) while track with low valence sound more negative (eg: sad, depressed)
- **tempo**: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- **speechiness**: Detects the presence of spoken words in a track. The more exclusively speech-like the recording (eg: talk show, audio book,..), the closer to 1.0 of value attribute. Value above 0.66 describe tracks that are probably made entirely of spoken words. Value between 0.33 and 0.66 describe tracks that may contain both music and speech. Value below 0.33 most likely represent music and other non-speech like tracks.

- **time_signature**: The time signature is a notional convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7, indicating time signature of ¾ to 7/4.

3.EXPLORATORY DATA ANALYSIS

In this part, I will analyze and visualize the dataset to discover patterns, spot anomalies, summary statistics to understand the underlying structure of the data, data cleaning, and prepare for further analysis or model.

```
df.shape
# The data set include 20 columns and 586672 rows (observations)
```
✓ 0.0s

```
(586672, 20)
```

- Overall, the data set include 20 columns and 586672 rows (observations)

Missing data:

```
# Count if is there any null value
df.isnull().sum()
```
✓ 0.1s

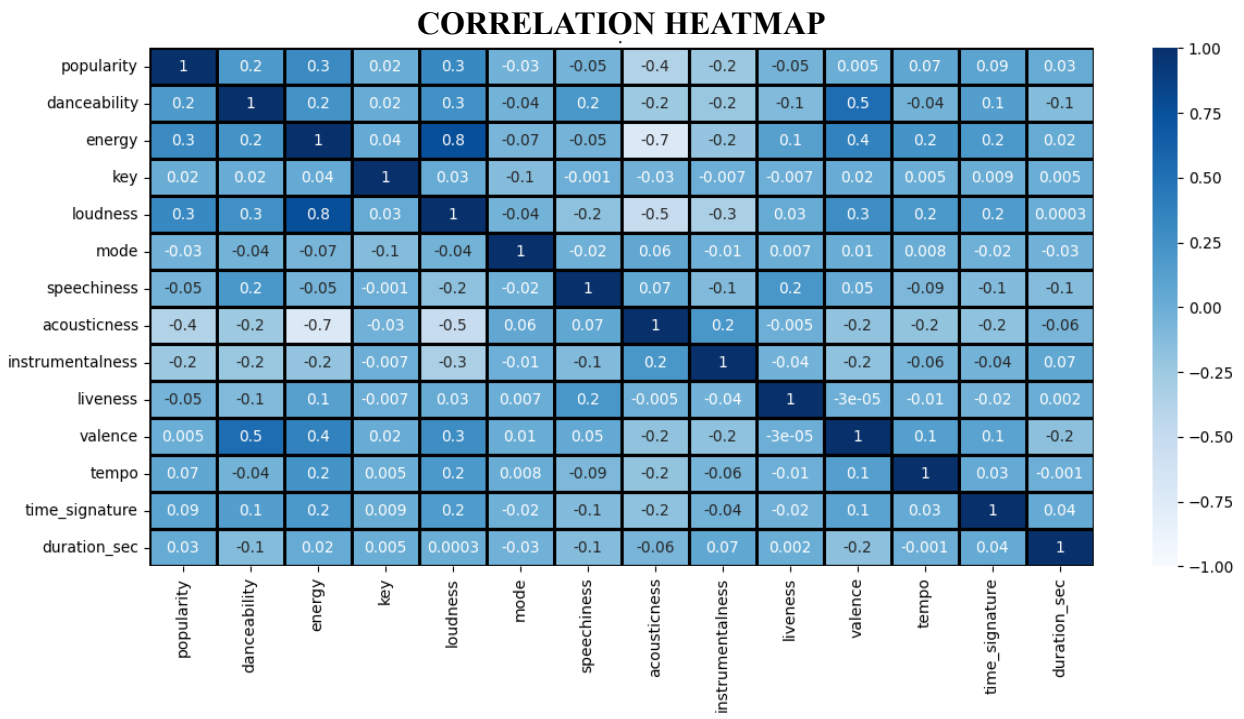
```
id                  0
name               71
popularity          0
duration_ms         0
explicit            0
artists             0
id_artists          0
release_date        0
danceability        0
energy              0
key                 0
loudness            0
mode                0
speechiness         0
acousticness        0
instrumentalness    0
liveness            0
valence             0
tempo               0
time_signature      0
dtype: int64
```

- There are just 71 null values in 'name' column, which is not a big matter and does not affect too much to our data set.
- In general, the data set is quite clean.

Identifying pattern: I convert the duration_ms (from millisecond) to second.

Correlation between variables:

- Correlation refers to the statistical relationships between variables in a dataset. It determines how variables are related to each other – whether they move together (positive correlation), or in opposite direct (negative correlation), or unrelated (no correlation).
  + Positive correlation (1): when two variables increase or decrease together.
  + Negative correlation (-1): when one variable increases, the other decreases.
  + No correlation (0): when two variables show no linear relationship.



Positive Correlation:

- **Energy and Loudness (0.8):** song with higher energy tend to be louder
- **Valence and Danceability (0.5):** valence measure the musical positivity, is often moderately positively correlated with danceability, meaning happier songs are often more danceable.
- **Energy and Valence (0.4):** energetic songs tend to have positive feel.

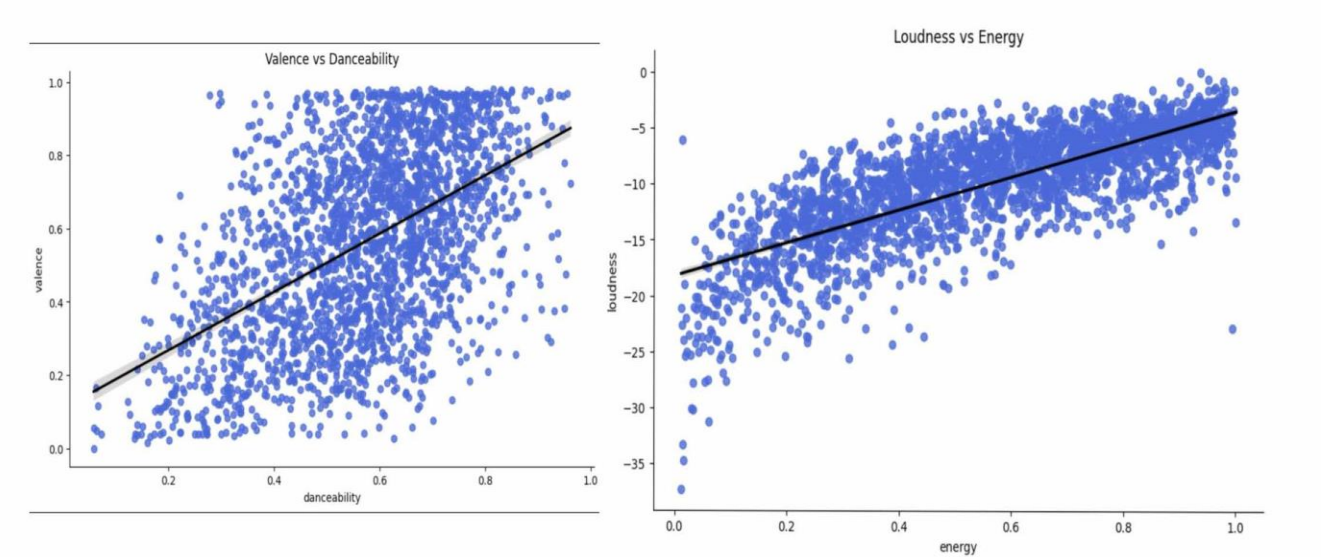
Negative Correlation:

- **Acousticness and Energy (-0.7):** more acoustic tracks tend to be less energetic.
- **Acousticness and Loudness (-0.5):** more acoustic tracks tend to be quieter.
- **Acousticness and Popularity (-0.4):** acoustic songs might be less popular in the data set.

Weak Correlation:

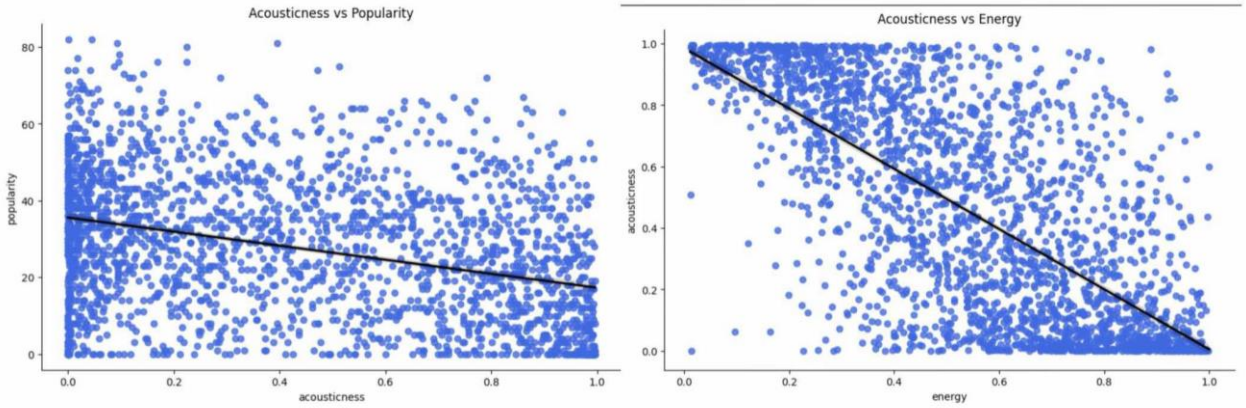
- Features like **liveness, spechiness, time_signature, and duration_sec** show little to no strong correlation with other features, meaning these variables may not play a vital role in determining song characteristics or popularity

Overall, the heatmap gives a comprehensive look at how different musical attributes interact with each other, which is valuable when exploring and interpreting the data set.



The left plot shows the relationship between Valence and Danceability, we can see that the higher the valence is, the higher danceability portion account for. Which means that happier songs are often danceable, in opposite, songs have the sad rhythms have lower danceability.

The right plot shows the relationship between Energy and Loudness, which show a strong positive linear relationship between 2 variables. Energetic songs tend to be louder, and the less energetic tend to be quieter.

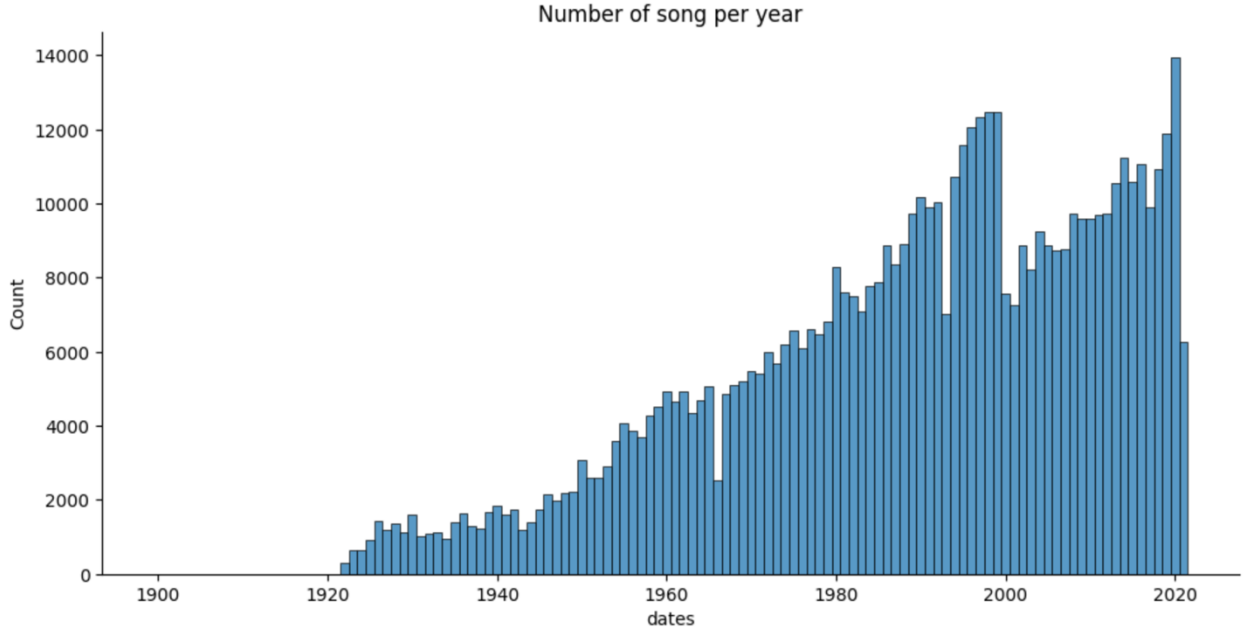


The left plot shows the relationship between Acousticness and Popularity, which indicates that Acousticness genre is not popular based on the given data set. The higher proportion of acoustic rhythms, the less likely to become popular.

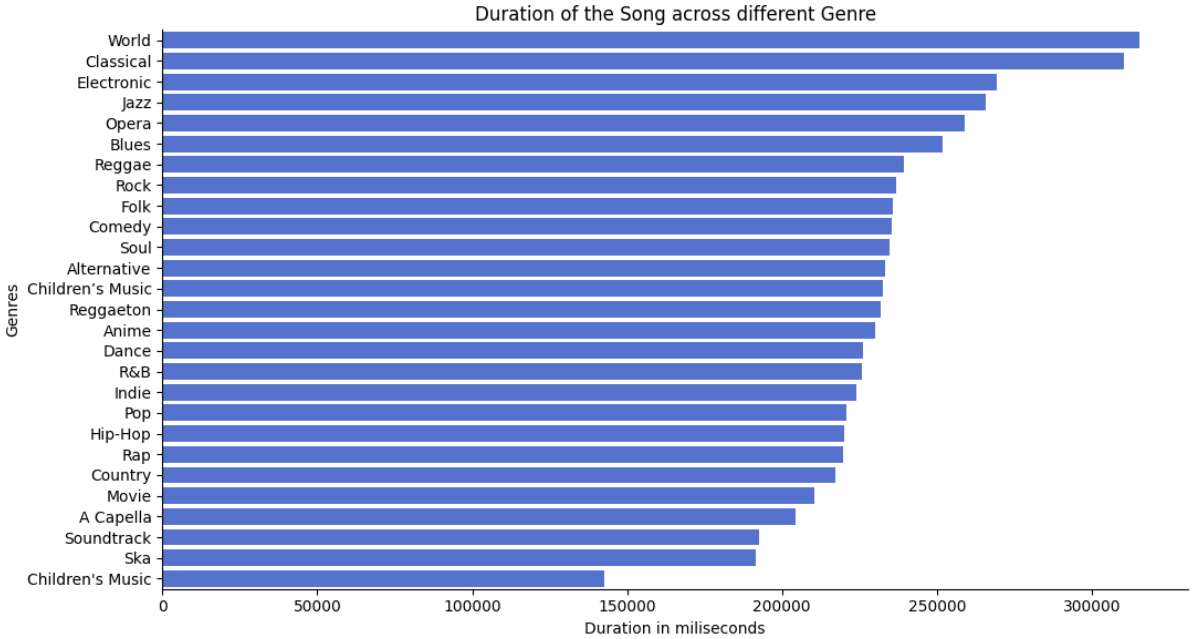
The right plot shows the relationship between Acousticness and Energy, indicating a strong negative relationship between these 2 variables. The more acoustic songs tend to be less energetic, in opposite, the less acoustic rhythms can make the song become more popular.

4.DATA VISUALIZATION

NUMBER OF SONGS WAS RELEASED YEARLY DURING THE GIVEN PERIOD



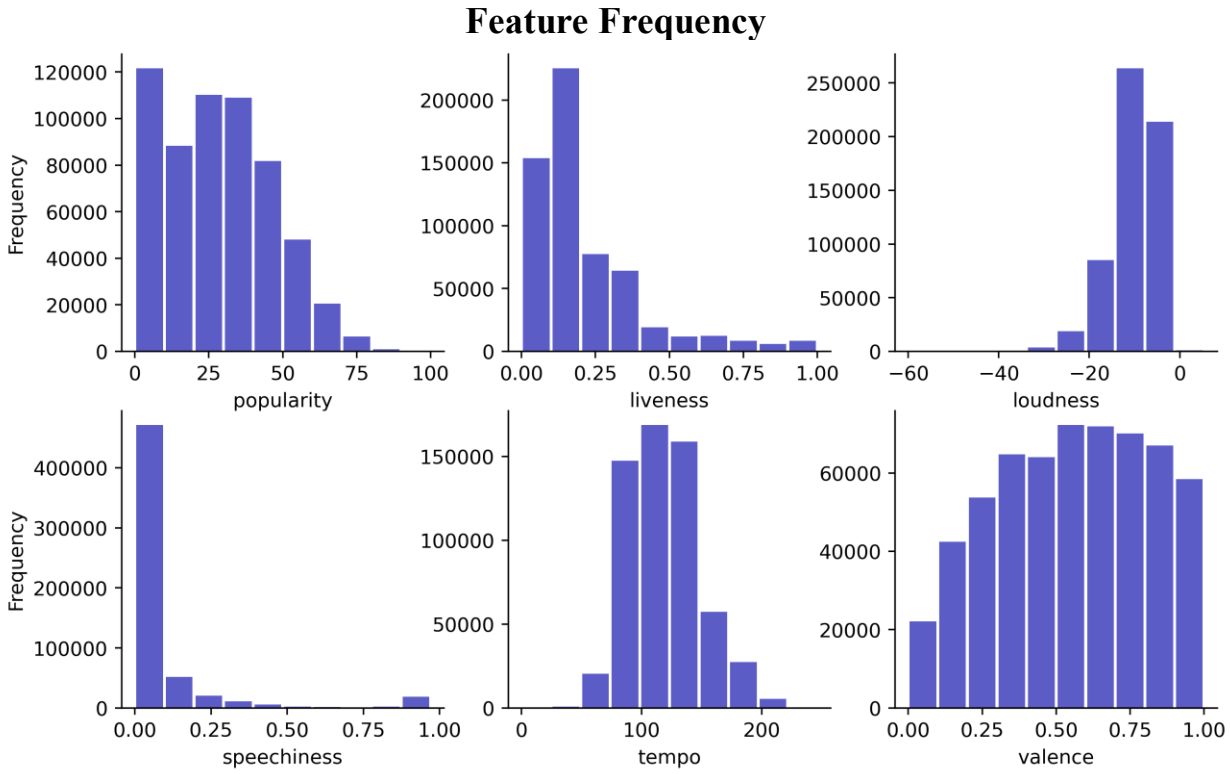
- Overall, the number of songs released per year steadily increased over the period, particularly from the mid 20th century onwards.
- There are noticeable dips in the number of songs during some periods, but overall, the trend is upward.
- The number of songs released picked around 2020, with around 14000 songs was released in this year.



The bar chart illustrates the average track's duration in different genre. This visualization offers several insightful observations about the nature and structure of music within different genre:

- **Longest Average Duration**: Genre like 'World', 'Electronic', and 'Classical' have the longest average duration. With 'World' music has longest duration at around 300000 milliseconds (approximately 5 minutes)
- **Shortest Average Duration:** 'Soundtrack' and 'Ska' are the two genre that have the shortest average track duration (approximately 3 minutes)
- The other rest genres seem to be approximately same at around 225000 millisecond (nearly 4 minutes)

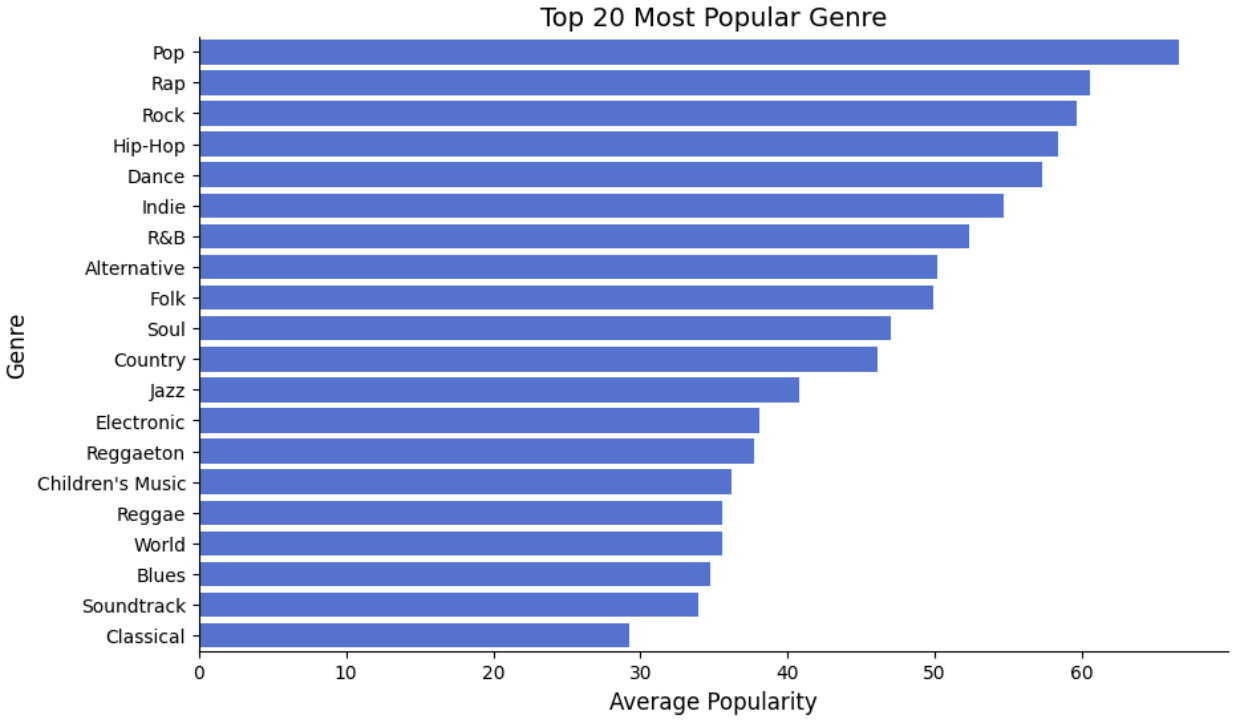
The chart highlights how different genres cater to their audiences with varying song structures, from the elaborate and extended tracks in World and Classical music to the shorter like Soundtracks. In general, most audiences prefer listening to tracks that are around 3.5 to 4 minutes long, which is why many genres have an average duration in that range.



The histogram provides a visual representation of the distribution of various audio features in the data set, including:

- **Popularity**: The distribution popularity is heavily right skewed, with a high frequency of songs having low popularity score (0-25). This suggest that most of songs in data set are not popular, only few of them achieved high popularity score.
- **Liveness**: The liveness features, which detects the presence of a live audience, show a strong right skewed. Most songs have a low liveness value (close to 0), indicating that the majority of tracks are studio recording rather than live performances. Just a small proportion of songs have higher liveness value, suggesting a live concert setting.

- **Loudness**: Loudness is normally distributed with a peak around -10 to -20 decibels. This indicates that most songs have a moderate loudness level, typically of professional music productions, where tracks are normalized to avoid distortion.
- **Spechiness**: The spechiness histogram shows a significant right skewed, with most songs have low spechiness values (close to 0). This indicates that most tracks are more musical rather than speech based. A few tracks with higher spechiness values may represent for podcast, rap, or spoken words.
- **Tempo**: Tempo measures the speed of a song, is distributed normally with a peak around 100-120 beats per minutes (BPM). This range aligns with common tempo ranges for popular music genre like pop, rock, and electronic music.
- **Valence**: Valence represents for the musical positiveness conveyed by a track, has a uniform distribution, with a slight peak around 0.5 to 0.7. This suggests that the data set contains a balance mix of happy and sad song (representing for high valence and low valence song), reflecting a variety of moods depends on audiences' preference.

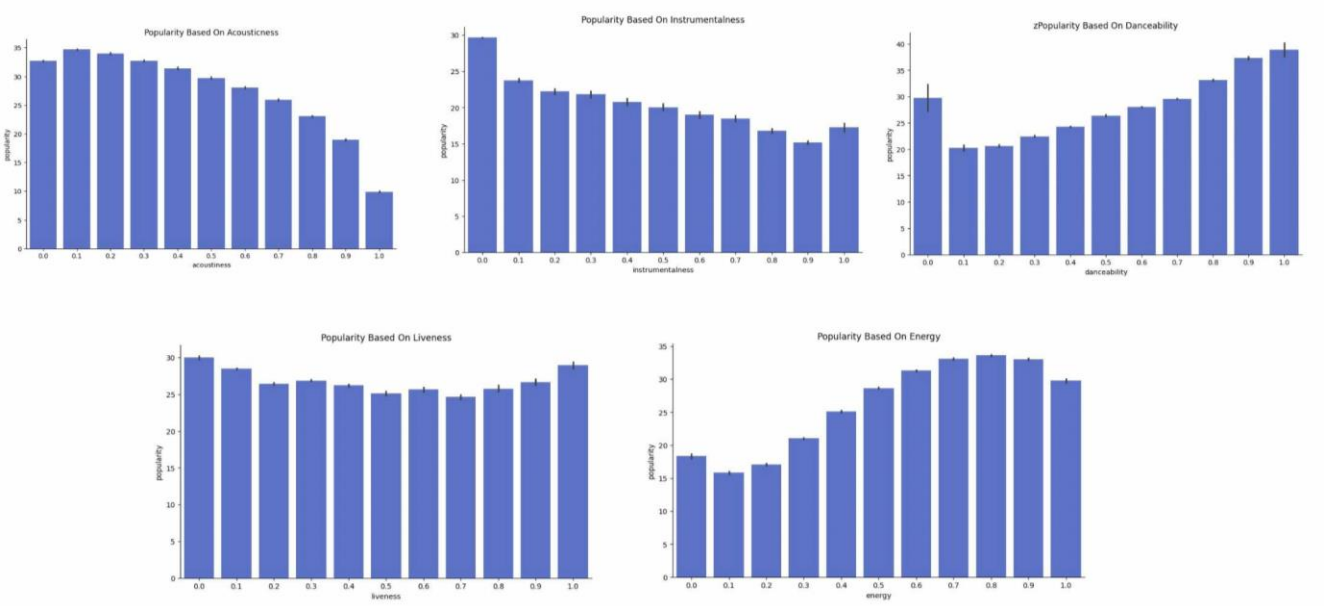


The chart shows the top 20 most popular genres based on their average popularity score:

- **Pop** leads as the most popular genre with the highest average popularity score, indicating that songs classified under the pop genre generally have wide appeal and receive a significant amount of attention.
- **Rap and Rock** follow closely behind Pop genre, suggesting these genres also substantial popularity among listeners.

- **The least popular genre in top 20:** Classical, Soundtracks, and Blues are the lower end of the top 20 in terms of average popularity. This could indicate a more niche audience or less mainstream appeal compared to other on the list.

POPULARITY VIA FEATURES



The collection of bar charts provides an overview of how the average popularity of songs varies with different audio such as acousticness, instrumentalness, danceability, liveness, and energy.

- Overall, the popularity across different audio features reveals that **songs with their lower acousticness, lower instrumentalness, higher danceability, higher energy level tends to be more popular.** These finding suggests that audiences generally prefer songs that are more upbeat, danceable, and engaging, with less focus on acoustic and instrumental-only tracks.
- The relationship between popularity and liveness is relatively flat, with a slight peak at lower liveness level. Liveness detects the presence of audience in the recording, songs with low liveness value (studio recording) are slightly more popular, but the effect is not as pronounced with other features.

Recommendation: Artists can base on these features to understand the audiences' trend, helping in create targeted playlist, music recommendation, or marketing strategies that align with listeners preference.

TOP 3 ARTISTS AND THEIR MOST POPULAR SONG

Artist Name	Track Name	Popularity Score
Arianna Grande	7 Rings	100
Post Malone	WOW.	99
Daddy Yankee	Con Calma	98

5. CREATING PLAYLIST AND RECOMMENDATION ANALYSIS

- **Insight:**

Platforms like Spotify utilize complex algorithms to curate playlist based on song attributes and listeners preference. By understanding the common attributes of songs in different types of playlists (eg: workout, chill, study), we can gain insights into how these platforms personalize recommendations and cater to different listening moods or activities.
This analysis can provide a glimpse into the features that make a song more likely to be recommended for specific context or moods, enhancing our understanding of recommendation algorithms.

- **Analysis:**

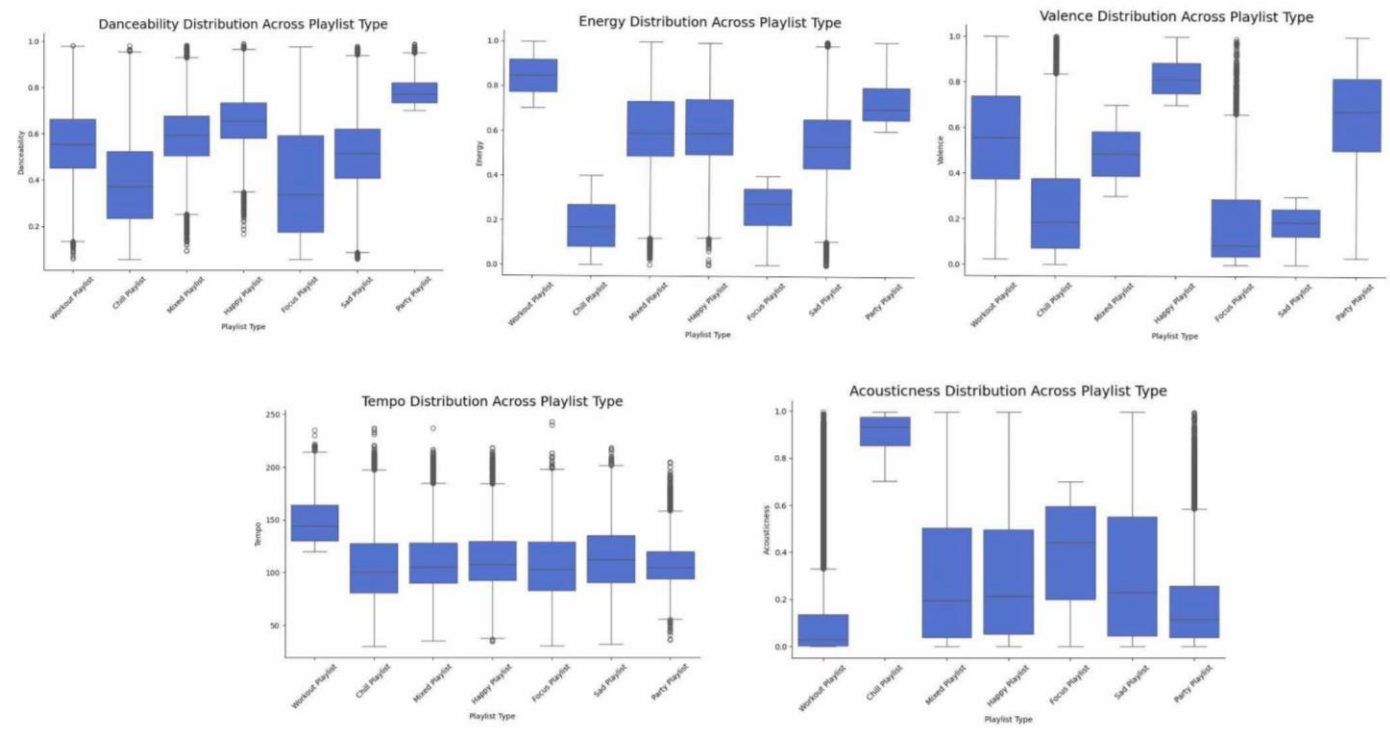
To analyze how Spotify or similar platforms might curate playlists, I investigate the average value of various audio features (danceability, acousticness, instrumentalness, and so on) in different types of playlists.

- **Playlist Created:**

Firstly, I created a function named **categorize_playlist()** to create a specific playlist based on features of the track. There are 7 playlists were created:

1) **Workout Playlist:** This is suitable for listeners during their workout or high cardio activities. This playlist includes songs that are high energy and high tempo, keeping listener motivated and fast-paced tracks that match the rhythm of intense activities.
2) **Chill Playlist**: With high acoustiness and low energy score to help audience feel relax and unwind. Tracks that are more organic, acousticness, calm, and soothing will be included.
3) **Party Playlist**: Perfect for parties with energetic and danceable tracks. This playlist includes high danceability, high energy, and high loudness songs to amplify the party environment.
4) **Focus Playlist**: Helps listeners maintain focus during study or work. High instrumentalness and low energy tracks that aid concentration and minimize distraction from lyrics.
5) **Happy Playlist:** A collection of upbeat, positive tracks to uplift the listener's mood. Tracks in this list will be convey a happy and cheerful mood.
6) **Sad Playlist:** Aimed at evoking or resonating with sad emotions. Tracks that are more melancholic or introspective in nature with low valence.
7) **Mixed Playlist**: A diverse playlist that includes a variety of music types and moods.

FEATURE DISTRIBUTION ACROSS PLAYLIST TYPE



The box plots provide a comprehensive visualization of how different playlists are composed based on various audio features:

- **Workout and Party Playlists** are characterized by high energy, high danceability, and moderate to high tempo, making them suitable for active, high-energy environments.
- **Chill and Focus Playlists** are more acoustic and low in energy, suitable for relaxation and concentration.
- **Happy and Sad Playlists** are distinguished primarily by their valence, reflecting the emotional tone of the songs.
- **Mixed Playlist** shows a wide range of values across all features, representing a diverse mix of songs that don't conform to a single mood or activity type.

These insights are valuable for understanding how different audio features contribute to the creation of playlists tailored for specific moods or activities, which aligns with the way music streaming platforms curate content for diverse listening experiences.

6. OVERVIEW

After conducting a thorough analysis of the dataset, I gained valuable insights into how Spotify's recommendation system operates and what makes certain types of music more likely to become popular. The analysis highlighted that songs with specific audio features—such as high energy and danceability for workout and party playlists, or high acousticness and low energy for chill and focus playlists, have a greater chance of resonating with listeners.

By understanding these patterns, I was able to categorize tracks into various playlist types based on their attributes, shedding light on how Spotify curates playlists to align with different moods and activities. This project deepened my understanding of the role that song characteristics play in shaping user preferences and how streaming platforms leverage these features to enhance the listening experience, enhancing the user experience through personalized music recommendations.