Exploratory Analysis of Music Characteristics and Popularity Using SQL

Project Overview

This project analyzes a dataset of songs from Spotify and YouTube, obtained from Kaggle. Its purpose is to explore the relationship between various musical characteristics and popularity. The analysis covers a wide range of topics, including the distribution of musical characteristics, the relationship between song duration and popularity, the most common musical keys and tempos, the correlations between musical characteristics like loudness, energy, and valence, and the correlation of official videos and licensed content to performance metrics.

This project provides valuable insights for musicians and music enthusiasts. It includes key questions and answers, as well as the SQL code used to arrive at those insights.

▼ Popularity

Most Streamed Song

The most streamed song in the dataset is "Blinding Lights" by The Weeknd with 3.39 billion streams. This suggests that the song was extremely popular and widely listened to during the time period covered by the dataset.

```
SELECT
  track
  ,artist
  ,stream
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
ORDER BY
  stream DESC
```

Are there any particular artists that stand out in terms of the number of streams, views, likes, or comments?

- Post Malone has the highest total streams on Spotify, followed by Ed Sheeran and Dua Lipa.
- Ed Sheeran has the highest total views on YouTube, followed by CoComelon and Katy Perry.
- On YouTube, BTS has the most total likes and comments, followed by BLACKPINK and Charlie Puth for likes, and BLACKPINK and Stray Kids for comments.

```
SELECT
  artist,
  SUM(stream) as total_streams,
  SUM(views) as total_views,
  SUM(likes) as total_likes,
  SUM(comments) as total_comments
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
GROUP BY
  artist
ORDER BY
  total_likes DESC -- replace this with total_views, total_comments or total_streams
```

Average Number of Likes for Videos with Official Music Videos

Videos with official music videos have a significantly higher average number of likes compared to those without. Videos with official music videos have an average of 802,605 likes, while those without have an average of 175,808 likes. This suggests that having an official music video can increase the popularity and engagement of a song on the platform. It's important to note that other factors, such as the popularity of the artist and the quality of the song, may also contribute to the number of likes a video receives.

```
SELECT
  official_video
  ,AVG(likes) as avg_likes
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
GROUP BY
  official_video
ORDER BY
  avg_likes desc
```

How does the presence of an official video affect the popularity of a song on Spotify and YouTube?

Having an official music video significantly impacts the popularity of songs on both Spotify and YouTube. On Spotify, songs with an official music video have almost double the average number of streams compared to songs without an official video. Similarly, on YouTube, songs with an official video have an average of five times more views than songs without an official video. These findings suggest that creating an official music video can be an effective strategy for artists to boost their visibility and increase their reach on both platforms.

```
SELECT
  official_video,
  AVG(stream) AS avg_spotify_streams,
  AVG(views) AS avg_youtube_views
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
GROUP BY
  official_video
ORDER BY
  official_video desc
```

Are there any particular channels that are more likely to publish official videos or licensed content?

The percentage of channels that are more inclined to publish official videos or licensed content is approximately 17.7% (3,590 out of 20,248).

```
-- Selects the channel and percentage of songs with official videos or licensed content
SELECT
  channel,
  COUNT(CASE WHEN official_video = TRUE OR licensed = TRUE THEN 1 ELSE NULL END) / COUNT(*) * 100 AS percentage
FROM
  portfolio-projects-382014.spotify_youtube_songs.songs
-- Groups the results by channel
GROUP BY
  channel
-- Filters out channels with 0 official videos or licensed content
HAVING
  COUNT(CASE WHEN official_video = TRUE OR licensed = TRUE THEN 1 ELSE NULL END) > 0
ORDER BY
  percentage DESC
```

How does the likelihood of channels to publish official videos or licensed content correlate with their performance metrics such as streams, views, likes, or comments?

Channels with 50% or more official videos or licensed content have significantly higher metrics compared to those with less than 50%. Specifically, these channels have 5.7 times more average views, 5.6 times more average likes, 5.2 times more average comments, and 1.26 times more streams on Spotify. This suggests that having a higher proportion of official videos or licensed content can positively impact a channel's performance on both Spotify and YouTube.

It is worth noting that the channels with 50% or more official videos or licensed content belong to well-known artists. These artists likely have a large and loyal fan base, which could contribute to the higher performance metrics of their channels. Additionally, well-known artists often have more resources and support to produce high-quality content and promote their channels, which could also contribute to their success. This suggests that while having a higher proportion of official videos or licensed content can positively impact a channel's performance, the influence of the artist's popularity and resources cannot be ignored.

```
-- Selects the average metrics for all channels with a likelihood of 50% or greater
SELECT
  '>= 50%' as likelihood,
  AVG(avg_views) as avg_views,
  AVG(avg_likes) as avg_likes,
  AVG(avg_comments) as avg_comments,
  AVG(avg_streams) as avg_streams
FROM
  (
    -- Subquery that calculates various metrics for each channel
    SELECT
      channel,
      COUNT(CASE WHEN official_video = TRUE OR licensed = TRUE THEN 1 ELSE NULL END) / COUNT(*) * 100 AS likelihood,
      AVG(views) AS avg_views,
      AVG(stream) AS avg_streams,
      AVG(likes) AS avg_likes,
      AVG(comments) AS avg_comments
    FROM
      `portfolio-projects-382014.spotify_youtube_songs.songs`
    GROUP BY
      channel
  ) AS sub
```

```
WHERE
  sub.likelihood >= 50

-- Combines the previous results with the average metrics for all channels with a likelihood less than 50%
UNION ALL

-- Selects the average metrics for all channels with a likelihood less than 50%
SELECT
  '< 50%' as likelihood,
  AVG(avg_views) as avg_views,
  AVG(avg_likes) as avg_likes,
  AVG(avg_comments) as avg_comments,
  AVG(avg_streams) as avg_streams
FROM
  (
    -- Subquery that calculates various metrics for each channel
    SELECT
      channel,
      COUNT(CASE WHEN official_video = TRUE OR licensed = TRUE THEN 1 ELSE NULL END) / COUNT(*) * 100 AS likelihood,
      AVG(views) AS avg_views,
      AVG(stream) AS avg_streams,
      AVG(likes) AS avg_likes,
      AVG(comments) AS avg_comments
    FROM
      `portfolio-projects-382014.spotify_youtube_songs.songs`
    GROUP BY
      channel
  ) AS sub
WHERE
  sub.likelihood < 50
```

Is there a relationship between the duration of a song and its popularity on Spotify and YouTube?

Songs lasting 3-5 minutes have the highest Spotify streams and YouTube views, followed by 0-3 minute songs on Spotify and 5-7 minute songs on YouTube. This suggests a preference for songs within the 3-5 minute range. Other factors such as genre, artist, and marketing may also impact popularity. Further analysis is recommended to fully understand the relationship between song duration and popularity.

```
SELECT
  -- The CASE statement categorizes the songs into duration ranges, from 0-3 min to 9+ min
  CASE
    WHEN duration_ms <= 180000 THEN '0-3 min'
    WHEN duration_ms > 180000 AND duration_ms <= 300000 THEN '3-5 min'
    WHEN duration_ms > 300000 AND duration_ms <= 420000 THEN '5-7 min'
    WHEN duration_ms > 420000 AND duration_ms <= 540000 THEN '7-9 min'
    ELSE '9+ min'
  END AS duration_range,
  -- Calculate the average number of streams (avg_spotify_popularity) and average number of views (avg_youtube_views) for each duration range
  AVG(stream) AS avg_spotify_popularity,
  AVG(views) AS avg_youtube_views

FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`

-- Group the results by the duration_range column, so we can see the average Spotify popularity and YouTube views for each duration range
GROUP BY
  duration_range

ORDER BY
  avg_youtube_views DESC
```

▼ Characteristics

What is the distribution of the danceability, energy, and valence variables across the dataset?

The danceability variable ranges from 0.0 to 0.975 with an average value of 0.619, the energy variable ranges from 2.03e-05 to 1.0 with an average value of 0.635, and the valence variable ranges from 0.0 to 0.993 with an average value of 0.529.

```
SELECT
  -- Select column 'danceability' and its corresponding min, max and avg values
  'danceability' as variable,
  MIN(danceability) as min_value,
  MAX(danceability) as max_value,
  AVG(danceability) as avg_value
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
-- Union the previous result set with the following query, and show the results together
UNION ALL
SELECT
  -- Select column 'energy' and its corresponding min, max and avg values
  'energy' as variable,
```

```
  MIN(energy) as min_value,
  MAX(energy) as max_value,
  AVG(energy) as avg_value
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
-- Union the previous result set with the following query, and show the results together
UNION ALL
SELECT
  -- Select column 'valence' and its corresponding min, max and avg values
  'valence' as variable,
  MIN(valence) as min_value,
  MAX(valence) as max_value,
  AVG(valence) as avg_value
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
```

Are there any patterns in the key variable? Are certain keys more commonly used in popular music?

Songs composed in key 11 have the highest average views and streams

```
-- Select the musical key, average views, and average stream from the 'songs' table in the 'spotify_youtube_songs' dataset
SELECT
  key,
  avg(views) as avg_views,
  avg(stream) as avg_stream
FROM
-- Select the key, views, and stream columns from the 'songs' table in the 'spotify_youtube_songs' dataset
  (
  SELECT
    key,
    views,
    stream
  FROM
    portfolio-projects-382014.spotify_youtube_songs.songs
  -- Filter the results to include only songs with views or streams greater than the overall average
  WHERE
  views > (SELECT avg(views) FROM portfolio-projects-382014.spotify_youtube_songs.songs)
  OR stream > (SELECT avg(stream) FROM portfolio-projects-382014.spotify_youtube_songs.songs)
  ) as sub
GROUP BY
  key
-- Order the results by the average views in descending order
ORDER BY
  avg_views DESC
```

Most Common Key

The most common key for the songs in the dataset is key 0 with a count of 2305. This suggests that many of the songs in the dataset are likely to be in the key of C or A minor. The choice of key can have an impact on the overall sound and mood of a song, so this information could be useful for musicians or music enthusiasts who want to analyze or understand the musical characteristics of the songs in the dataset.

```
SELECT
  key
  ,COUNT(key) as count
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
GROUP BY
  key
ORDER BY
  count desc
```

What is the range of tempos found in the dataset, and is there a relationship between tempo and danceability or energy?

The range of tempos found in the dataset ranges from 0.0 to 243.372. The average tempo is 120.638 beats per minute, with a standard deviation of 29.579 beats per minute.

This dataset reveals that there is a slight positive correlation between tempo and energy. However, there is no significant correlation observed between tempo and danceability or valence.

```
-- What is the range of tempos found in the dataset?
SELECT
  MIN(tempo) as min_tempo,
  MAX(tempo) as max_tempo,
  AVG(tempo) as avg_tempo,
  STDDEV(tempo) as tempo_stddev,
FROM
```

```
  `portfolio-projects-382014.spotify_youtube_songs.songs`
-----------------------------------------------------------------------------------------------------------------------------
-- is there a relationship between tempo and danceability or energy or valence?
SELECT
  CORR(tempo, energy) AS tempo_energy_corr,
  CORR(tempo, danceability) AS tempo_danceability_corr,
  CORR(tempo, valence) AS tempo_valence_corr,
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
```

Is there a correlation between the loudness and energy variables, and if so, how strong is it?

There is a strong positive correlation between the loudness and energy of songs, with a correlation coefficient of 0.7448. This indicates that as the loudness of a song increases, so does its energy, and vice versa.

Note: Correlation coefficient ranges from -1 to +1: -1 indicates perfectly inverse correlation, +1 indicates perfectly positive correlation, and close to 0 suggests little to no correlation.

```
SELECT
  CORR(loudness, energy) as correlation_coefficient
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
```

Correlation Between Tempo and Energy

The correlation between tempo and energy is positive, but relatively weak with a value of 0.16. This suggests that there may be a slight tendency for songs with higher tempos to also have higher energy levels, but it's not a strong relationship. Other factors besides tempo likely play a bigger role in determining a song's energy level.

```
SELECT
  CORR(tempo, energy) AS tempo_energy_corr
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
```

Correlation Between Danceability and Energy

The correlation between danceability and energy is 0.2365959, but it is not significant. This means that there is a weak positive relationship between danceability and energy, but the relationship is not strong enough to be statistically significant.

```
SELECT
  corr(danceability, energy) as correlation
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
```

Correlation Between Valence and Energy

The correlation between valence and energy is 0.3891578, indicating a slightly positive relationship between the two variables. This means that there is a tendency for songs with higher energy levels to also have higher valence (more positive mood), but the relationship is not very strong. Other factors may also play a role in determining valence and energy levels of a song.

```
SELECT
  corr(valence, energy) as correlation
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
```

Correlation Between Danceability and Valence

There is a moderate positive correlation (0.47) between danceability and valence. This suggests that songs with higher danceability scores tend to have higher valence scores as well, meaning they are generally more positive or upbeat in nature.

```
SELECT
  corr(danceability, valence) as correlation
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
```

Is there a correlation between the speechiness and instrumentalness variables.

The correlation between speechiness and instrumentalness is negative but very small, suggesting that there is little relationship between the presence of vocals and the use of instruments in the music in the dataset.

```
SELECT
  CORR(speechiness, instrumentalness) AS correlation
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
```

Album Type with Highest Average Danceability

The "single" album type has the highest average danceability with a value of 0.665. This suggests that singles may be more upbeat and danceable than full albums or compilations. However, it's important to note that this is a general trend based on the data and individual songs or albums may vary.

```
SELECT
  album_type
  ,AVG(danceability) as avg_danceability
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
GROUP BY
  album_type
ORDER BY
  avg_danceability desc
```

Most Common Song Duration

The most common duration of a song in the dataset is between 3 to 5 minutes, with 13,473 songs falling in this range. This suggests that this duration range is a popular choice among songwriters and listeners alike. However, it's important to note that song duration can vary depending on the genre and style of music, and there may be other factors that influence the length of a song.

```
SELECT
  duration_range
  ,count(duration_range) as count
FROM
  (
  SELECT
    CASE
      WHEN duration_ms <= 180000 THEN '0-3 min'
      WHEN duration_ms > 180000 AND duration_ms <= 300000 THEN '3-5 min'
      WHEN duration_ms > 300000 AND duration_ms <= 420000 THEN '5-7 min'
      WHEN duration_ms > 420000 AND duration_ms <= 540000 THEN '7-9 min'
      ELSE '9+ min'
    END AS duration_range
  FROM
    `portfolio-projects-382014.spotify_youtube_songs.songs`
  ) AS sub
GROUP BY
  duration_range
ORDER BY
  count DESC
```

Artist with Highest Average Loudness

Kordhell has the highest average loudness with a value of -1.4701. This means their songs may be louder and more energetic than other artists in the dataset, making them great for activities like workouts or parties. However, loudness is subjective, so what one person finds loud may not be the same for another.

```
SELECT
  artist
  ,AVG(loudness) as loudness
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
GROUP BY
  artist
ORDER BY
  loudness desc
LIMIT 10
```

Song with Highest Speechiness Score

The high speechiness score (0.675) of "How High (Remix)" suggests that the song has a significant amount of spoken word elements, such as rap verses or spoken word poetry. This may appeal to listeners who enjoy a focus on spoken word in their music. However, speechiness is subjective, and what one person considers speech-like may differ from another.

```
SELECT
  track
  ,speechiness
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
WHERE
  speechiness < 0.7 -- anything greater than 0.7 are audiobooks, not songs.
ORDER BY
  speechiness desc
```

Average Acousticness of Songs in Albums versus Singles

The average acousticness of songs released as singles is lower (0.2646) compared to those released in albums (0.2978). Songs in albums may be more diverse in terms of instrument use and may include more acoustic instruments, leading to a higher average acousticness. However, it's important to note that the differences in acousticness between singles and albums may also be influenced by other factors such as genre, etc.

```
SELECT
  album_type
  ,AVG(acousticness) as avg_acousticness
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
GROUP BY
  album_type
```

What is the most common album type in the dataset?

Albums are the most popular way of releasing music. It's more common for songs to be part of an album than to be released as singles.

```
SELECT
  album_type,
  COUNT(*) as count
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
GROUP BY
  album_type
ORDER BY
  count DESC
```

Comparing musical characteristics between singles and albums

There is no any substantial differences in the musical characteristics between singles and albums in the dataset.

```
SELECT
  CASE
    WHEN album_type = 'single' THEN 'single'
    ELSE 'album'
  END as release_type,
  AVG(danceability) as avg_danceability,
  AVG(energy) as avg_energy,
  AVG(loudness) as avg_loudness,
  AVG(tempo) as avg_tempo,
  AVG(valence) as avg_valence
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
GROUP BY
  release_type
```

Top 25 Songs with the Highest Danceability

The query results display the top 25 songs with the highest danceability scores, showcasing a diverse range of artists and genres. High danceability scores imply that these songs have strong rhythms and tempos suitable for dancing or physical movement. These insights may aid individuals seeking energetic or upbeat songs for activities that require high energy levels. Notably, several songs feature collaborations between various artists, indicating that combining different styles and voices can create particularly danceable music.

```
SELECT
  DISTINCT(track)
  ,artist
  ,danceability
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
```

```
ORDER BY
  danceability desc
LIMIT 25
```

Top 25 Songs with the Highest Valence (Most Positive Songs)

The query results show the top 25 most positive songs, as measured by their valence scores. The songs come from a variety of genres and artists, ranging from children's music to rock and roll. The high valence scores suggest that these songs have upbeat and happy themes, making them ideal for boosting mood and energy. These insights may be useful for people looking for positive and uplifting music for workouts, parties, or other activities. It's interesting to note that some of the most positive songs come from traditional Mexican music, suggesting that cultural context can play a role in shaping the emotional impact of music.

```
SELECT
  DISTINCT(track)
  ,artist
  ,valence
FROM
  `portfolio-projects-382014.spotify_youtube_songs.songs`
ORDER BY
  valence desc
LIMIT 25
```

Conclusion

This project analyzed musical characteristics and their relationship to popularity using SQL. The findings suggest a preference for songs in the 3-5 minute range and certain musical keys, such as key 11, being more commonly used in popular music. Strong positive correlations were found between loudness and energy, and moderate positive correlations between danceability and valence. Additionally, songs with official music videos or licensed content were found to have significantly higher numbers of streams, views, likes, and comments compared to those without. This project showcases the power of using SQL to extract insights and has helped me improve my data analysis skills.

What Could Be Done to Improve This Project with More Time

If given more time, there are several areas in which this project could be expanded or improved:

- **Analyze the musical characteristics of the most popular songs:** We will gather insights about the danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo of the most popular songs. By analyzing these characteristics, we hope to gain a deeper understanding of what makes these songs popular and identify any commonalities among them.
- **Create visualizations:** Although the project used SQL to extract insights, incorporating visualizations could make the findings more accessible and easier to understand for non-technical people without a background in data.
- **Analyze other factors that impact popularity:** While this project focused on musical characteristics, there are many other factors that can impact a song's popularity, such as marketing, promotion, and the artist's reputation. Analyzing these factors in conjunction with musical characteristics could provide a more complete understanding of what drives popularity.

References

- **Spotify and Youtube (Dataset)**. https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube