STAT 184 - Final Project Report Topic - Spotify Music Data Analysis using R

Apoorv Thite and Aarnav Putta

25th April, 2024

```
library(dplyr)
library(tidyverse)
```

Introduction

In the digital age, music streaming platforms like Spotify have revolutionized how we access and interact with music. These platforms not only provide a vast array of musical content but also collect extensive data on track characteristics and user preferences. Recognizing the potential of this rich dataset, our final project focuses on an exploratory data analysis of SpotifyData.csv, which consists of detailed attributes of various music tracks available on Spotify.

Our project aims to delve deep into the characteristics of these music tracks, such as genre, tempo, energy, and popularity, to uncover underlying patterns and trends that could be valuable for artists, record labels, and marketers. By analyzing this dataset, we hope to identify factors that significantly influence a track's success and listener preferences, potentially predicting future trends in music consumption.

Main Goal / Guiding Research Question

The primary question guiding our analysis is: "How do the characteristics of music (like tempo, valence, and energy) vary by genre, and also influence a track's popularity on Spotify?" This question will help us understand the trends and preferences in music consumption over a specific period and across different musical genres.

This guiding question serves as the backbone of our analysis and is crucial for several reasons:

- We intend to dissect various musical attributes to see how they correlate with each other and how they contribute to a song's popularity. For instance, does a higher tempo correlate with more energetic genres like dance music, and does this result in higher popularity ratings on Spotify?
- By examining how these musical characteristics vary across different genres, we aim to uncover genre-specific trends in music consumption.
- The insights derived from answering this question have practical implications for music production and marketing strategies. Artists and record labels can use this information to tailor their music to align with listener preferences, potentially increasing the likelihood of achieving higher engagement and popularity.

Where did you find them?

For our project, we sourced our datasets from two prominent online platforms known for their comprehensive repositories of data: DataCamp and Kaggle. The primary dataset, `SpotifyData.csv`, was obtained from Kaggle, a platform that hosts a wide range of community-generated datasets. This dataset includes various attributes of music tracks on Spotify, such as genre, tempo, and popularity, providing a rich base for our analysis. The secondary dataset was sourced from DataCamp, which is often utilized for educational purposes and practical exercises in data science. This dataset complements our primary data by providing additional variables that enable a deeper exploration of music track characteristics and listener preferences. Together, these datasets form the cornerstone of our analysis, allowing us to investigate and understand the intricate dynamics of music popularity on Spotify.

Who collects/maintains them ?

Apoorv took the initiative to source the primary dataset from datacamp, which was generated and is regularly updated by a community of data scientists and music enthusiasts. This dataset provides a detailed compilation of Spotify track characteristics, making it ideal for our analysis of music trends. Aarnav, on the other hand, handled the acquisition of the secondary dataset from Kaggle. This dataset, often used for educational purposes, complements the primary dataset with additional data points necessary for a comprehensive analysis. Together, Apoorv and Aarnav ensured that we had robust and reliable data, laying a strong foundation for our project

Links to both datasets :

Kaggle - https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset?select=dataset.csv

Datacamp - https://www.datacamp.com/datalab/w/e6817000-4a9f-40fa-b74f-1d23b9cdb6cb/edit

What does a case represent in each data source, and how many total cases are available?

Each case in this dataset refers to a specific song with a unique score for every attribute or variable. Each case essentially is classifying a song based on its unique score derived from the combination of variables we have in the datset. With the variables being artist, top genre, year,bpm, nrg, dnce, dB, live, val, dur, acous, spch, popularity. There are totally 114000 cases in this dataset.

Loading our Dataset and Initial dataset statistics :

```r
# Load the SpotifyMusic dataset
spotify_data <- read.csv(file = "SpotifyMusic.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)

# Total Number of rows in the dataset
nrow(spotify_data)
```

```
## [1] 100
```

```r
# View the first few rows of the data and summary statistics
head(spotify_data)
```

```
##                  title         artist        top.genre year bpm nrgy dnce dB live
## 1   "Hey, Soul Sister"          Train       neo mellow 2022  97   89   67 -4    8
## 2 Love The Way You Lie         Eminem detroit hip hop 2022  87   93   75 -5   52
## 3              TiK ToK          Kesha        dance pop 2022 120   84   76 -3   29
## 4          Bad Romance      Lady Gaga        dance pop 2022 119   92   70 -4    8
## 5 Just the Way You Are     Bruno Mars              pop 2022 109   84   64 -5    9
## 6                 Baby Justin Bieber     canadian pop 2022  65   86   73 -5   11
##   val dur acous spch pop
## 1  80 217    19    4  83
## 2  64 263    24   23  82
## 3  71 200    10   14  80
## 4  71 295     0    4  79
## 5  43 221     2    4  78
## 6  54 214     4   14  77
```

```r
# Summary of the dataset
summary(spotify_data)
```

```
##     title              artist           top.genre              year     
##  Length:100         Length:100         Length:100         Min.   :2022  
##  Class :character   Class :character   Class :character   1st Qu.:2022  
##  Mode  :character   Mode  :character   Mode  :character   Median :2022  
##                                                           Mean   :2022  
##                                                           3rd Qu.:2023  
##                                                           Max.   :2023  
##       bpm             nrgy            dnce             dB        
##  Min.   : 43.0   Min.   :33.00   Min.   :23.00   Min.   :-9.00  
##  1st Qu.:109.0   1st Qu.:70.75   1st Qu.:59.00   1st Qu.:-6.00  
##  Median :125.0   Median :81.00   Median :66.50   Median :-5.00  
```

```
##  Mean   :120.6   Mean   :76.18   Mean   :64.05   Mean   :-4.98  
##  3rd Qu.:130.0   3rd Qu.:87.00   3rd Qu.:73.00   3rd Qu.:-4.00  
##  Max.   :186.0   Max.   :98.00   Max.   :83.00   Max.   :-2.00  
##       live            val             dur            acous      
##  Min.   : 4.00   Min.   : 7.00   Min.   :172.0   Min.   : 0.00  
##  1st Qu.: 9.75   1st Qu.:40.00   1st Qu.:212.8   1st Qu.: 1.00  
##  Median :13.00   Median :58.00   Median :228.0   Median : 3.00  
##  Mean   :20.71   Mean   :55.26   Mean   :236.1   Mean   :11.91  
##  3rd Qu.:30.25   3rd Qu.:73.00   3rd Qu.:257.2   3rd Qu.:13.25  
##  Max.   :70.00   Max.   :89.00   Max.   :379.0   Max.   :91.00  
##       spch            pop       
##  Min.   : 3.00   Min.   : 0.00  
##  1st Qu.: 4.00   1st Qu.:59.00  
##  Median : 5.00   Median :65.50  
##  Mean   : 8.85   Mean   :64.69  
##  3rd Qu.:11.00   3rd Qu.:73.00  
##  Max.   :45.00   Max.   :83.00
```

The dataset comprises 100 music tracks, showcasing a variety of attributes. It features 96 unique titles with 'Castle Walls (feat. Christina Aguilera)' appearing twice, and 45 unique artists. The predominant genre is 'dance pop', one of 16 distinct genres observed. The tracks, mainly from 2010 to 2011, have an average popularity score of 64.69, with scores ranging from 0 to 83. The beats per minute (BPM) vary significantly from 43 to 186, averaging around 120.58. Other notable metrics include an average energy rating of 76.18, danceability at 64.05, and loudness at -4.98 dB. Live performance ratings average at 20.71, valence at 55.26, speechiness at 8.85, and acousticness at 11.91.

```
names(spotify_data)
```

```
##  [1] "title"     "artist"    "top.genre" "year"      "bpm"       "nrgy"     
##  [7] "dnce"      "dB"        "live"      "val"       "dur"       "acous"    
## [13] "spch"      "pop"
```

General Data Wrangling and Logistics -

1)

```
# Filter for popular tracks and calculate mean stats by genre
popular_genre_stats <-
  spotify_data %>%
  filter(pop > 50) %>%
  group_by(top.genre) %>%
  summarise(mean_bpm = mean(bpm, na.rm = TRUE),
            mean_valence = mean(val, na.rm = TRUE),
            mean_energy = mean(nrgy, na.rm = TRUE))
popular_genre_stats
```

```
## # A tibble: 14 x 4
##    top.genre       mean_bpm mean_valence mean_energy
##    <chr>              <dbl>        <dbl>       <dbl>
##  1 art pop             150          26          81  
##  2 atl hip hop         125          49          80.5
##  3 australian pop      164          64          80.7
##  4 barbadian pop       124          59.5        79.2
##  5 british soul        120          40          54.5
##  6 canadian pop        108.         71          89.5
##  7 chicago rap         125          10          69  
##  8 colombian pop       112          85          87  
##  9 dance pop           121.         57.7        76.4
```

```
## 10 detroit hip hop      87          64          93
## 11 hip pop             103.         26.3        53.3
## 12 indie pop           148          74          83
## 13 neo mellow           97          80          89
## 14 pop                 121.         48.6        78.1
```

Here, we filtered the dataset to focus on tracks with a popularity score above 50, grouping them by genre (top.genre). We then calculated the average beats per minute (bpm), valence (val), and energy (nrgy) for each genre, giving insights into the characteristics that correlate with higher popularity on Spotify.

2)

```r
# Using regular expressions to filter data based on a pattern in track names
filtered_tracks <- spotify_data %>%
  filter(grepl("Love", title))# Tracks with "Love" in their title
filtered_tracks
```

```
##                                        title  artist       top.genre year bpm
## 1                      Love The Way You Lie  Eminem detroit hip hop 2022  87
## 2                      Your Love Is My Drug   Kesha       dance pop 2022 120
## 3 DJ Got Us Fallin' In Love (feat. Pitbull)   Usher     atl hip hop 2022 120
## 4                               Love On Top Beyoncé       dance pop 2023  94
## 5                             We Found Love Rihanna   barbadian pop 2023 128
##   nrgy dnce dB live val dur acous spch pop
## 1   93   75 -5   52  64 263    24   23  82
## 2   61   83 -4    9  76 187     1   10  69
## 3   86   66 -3    8  65 221     3   11  52
## 4   75   65 -5   60  65 267     8    9  76
## 5   77   73 -4   11  60 215     3    4  61
```

Here, we utilized regular expressions to filter the dataset for tracks with "Love" in their title (title). This operation allowed us to specifically analyze tracks related to a common thematic element, providing focused insights into how songs with "Love" in their title perform in terms of popularity and other musical characteristics.

3)

```r
# Using Reduction and/or transformation functions
genre_summary <- spotify_data %>%
  group_by(top.genre) %>%
  summarise(max_popularity = max(pop, na.rm = TRUE),
            min_tempo = min(bpm, na.rm = TRUE))
genre_summary
```

```
## # A tibble: 16 x 3
##    top.genre       max_popularity min_tempo
##    <chr>                    <int>     <int>
##  1 acoustic pop                46       125
##  2 art pop                     58       150
##  3 atl hip hop                 72        80
##  4 australian pop              72       131
##  5 barbadian pop               73        80
##  6 big room                     0       128
##  7 british soul                80       105
##  8 canadian pop                77        65
##  9 chicago rap                 73       125
## 10 colombian pop               56       112
## 11 dance pop                   81        43
```

```
## 12 detroit hip hop              82        87
## 13 hip pop                      76        93
## 14 indie pop                    65       148
## 15 neo mellow                   83        97
## 16 pop                          78       103
```

Here, we applied reduction and transformation functions on the dataset to summarize key statistics by genre (top.genre). We computed the maximum popularity (max_popularity) and minimum tempo (min_tempo) for each genre, removing any missing values in the process. This approach allowed us to explore the extremes of popularity and tempo within genres, providing a clearer view of genre-specific trends on Spotify.

4)

```
# User-defined functions
# Define a function to classify tempo into Low, Medium, High
classify_tempo <- function(bpm) {
  ifelse(bpm < 100, "Low", ifelse(bpm < 140, "Medium", "High"))}

# Apply the function to the dataset
spotify_data <- spotify_data %>%
  mutate(tempo_category = sapply(bpm, classify_tempo))
head(spotify_data)
```

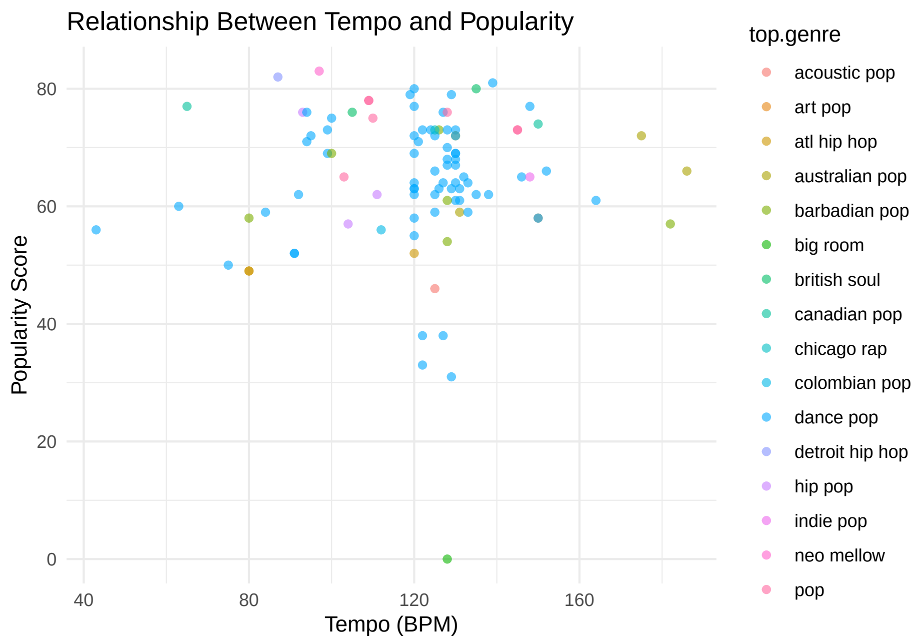
```
##                   title        artist       top.genre year bpm nrgy dnce dB live
## 1   "Hey, Soul Sister"          Train      neo mellow 2022  97   89   67 -4    8
## 2 Love The Way You Lie         Eminem detroit hip hop 2022  87   93   75 -5   52
## 3                TiK ToK          Kesha       dance pop 2022 120   84   76 -3   29
## 4            Bad Romance     Lady Gaga       dance pop 2022 119   92   70 -4    8
## 5 Just the Way You Are    Bruno Mars             pop 2022 109   84   64 -5    9
## 6                   Baby Justin Bieber    canadian pop 2022  65   86   73 -5   11
##   val dur acous spch pop tempo_category
## 1  80 217    19    4  83            Low
## 2  64 263    24   23  82            Low
## 3  71 200    10   14  80         Medium
## 4  71 295     0    4  79         Medium
## 5  43 221     2    4  78         Medium
## 6  54 214     4   14  77            Low
```

Here, we created a function, classify_tempo, to categorize the tempo (bpm) of tracks into "Low," "Medium," or "High." This classification was based on bpm thresholds, where tempos below 100 are categorized as "Low," those between 100 and 140 as "Medium," and above 140 as "High." We then applied this function across the dataset to assign each track a tempo category, enhancing our ability to analyze how tempo influences track popularity across different categories.

Data Visualisation -

1)

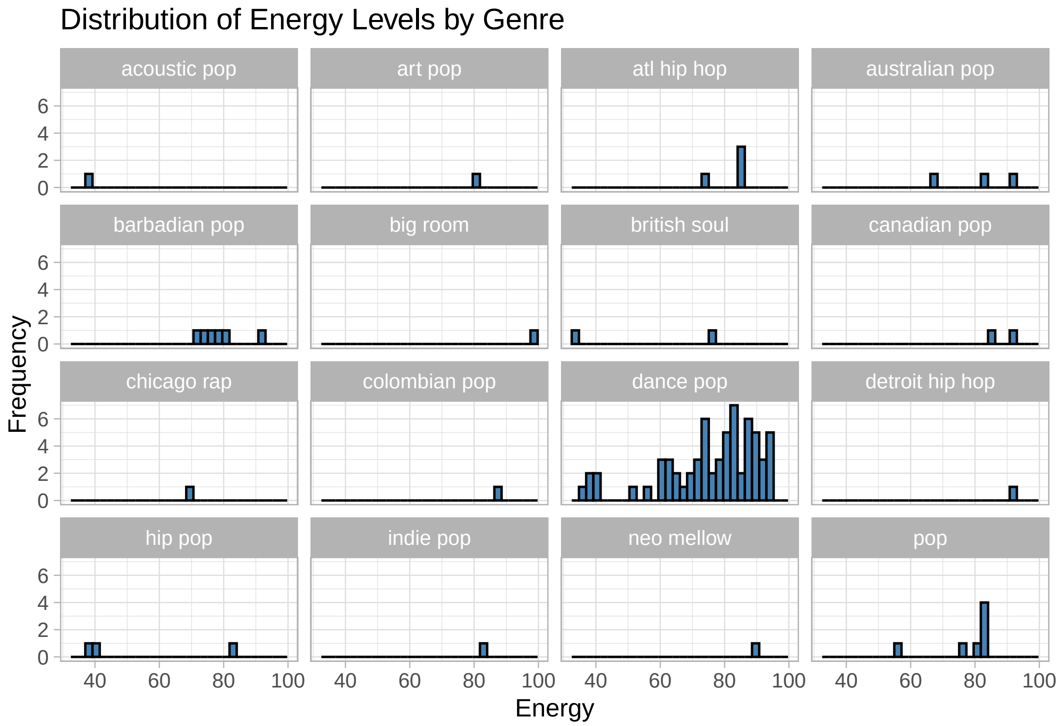
```
# 1. Scatter plot with multiple aesthetics (color, size) showing tempo vs popularity
ggplot(spotify_data, aes(x = bpm, y = pop, color = top.genre)) +
  geom_point(alpha = 0.6) +
  labs(title = "Relationship Between Tempo and Popularity",
       x = "Tempo (BPM)",
       y = "Popularity Score") +
  theme_minimal()
```



Here, we crafted a scatter plot to visually examine the relationship between tempo (bpm) and popularity (pop) across different music genres, using the ggplot2 library. Each point on the plot represents a track, colored by its genre (top.genre), which allows us to discern genre-specific patterns in how tempo correlates with popularity. We enhanced the plot's clarity using multiple aesthetics such as color for genre and point transparency, and employed a minimalistic theme for a clean presentation. This visualization aids in understanding if certain tempos are more favorable in specific genres in terms of attracting higher popularity scores.

2)

```r
# 2. Faceted histogram of energy levels across different genres
ggplot(spotify_data, aes(x = nrgy)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "black") +
  facet_wrap(~top.genre) +
  labs(title = "Distribution of Energy Levels by Genre",
       x = "Energy",
       y = "Frequency") +
  theme_light()
```

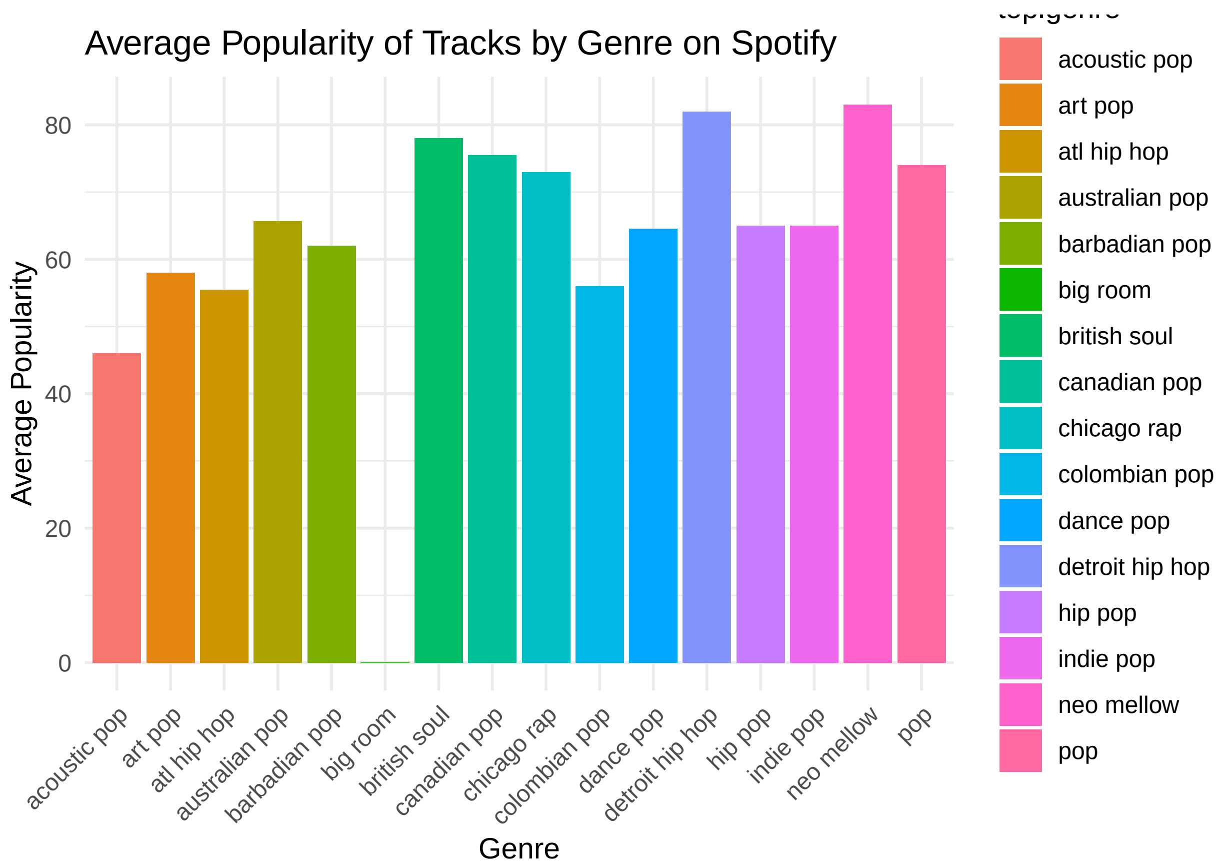


Here, we utilized a faceted histogram to analyze the distribution of energy levels (nrgy) across different music genres, employing ggplot2. Each histogram represents a genre, allowing us to compare how energy varies within and across genres. We specified the number of bins to 30 for detailed granularity and chose a color scheme of steel blue with black borders for visual clarity. The histograms are separated by genre using the facet_wrap function around the top.genre variable, providing a clear, genre-specific breakdown of energy distributions. This visualization helps identify genres with higher energy levels and how they might relate to audience preferences and popularity metrics.

3)

```
avg_popularity_by_genre <- spotify_data %>%
  group_by(top.genre) %>%
  summarise(average_popularity = mean(pop, na.rm = TRUE))

# Create a bar chart to visualize average popularity by genre
ggplot(avg_popularity_by_genre, aes(x = top.genre, y = average_popularity, fill = top.genre)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Popularity of Tracks by Genre on Spotify",
       x = "Genre",
       y = "Average Popularity") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate genre names for better readability
```



Here, we calculated the average popularity of tracks by genre using the dataset. We grouped tracks by top.genre and computed the average popularity score (pop) for each genre, handling missing values by omitting them (na.rm = TRUE). The results were stored in avg_popularity_by_genre.
To visualize these averages, we created a bar chart using ggplot2. Each bar represents a genre, colored distinctively, and displays its average popularity. We set the bars to directly reflect the calculated averages (stat = "identity"). The chart includes a minimal theme for a sleek look and has genre names rotated for enhanced readability, making it easier to compare the popularity across different genres. This visualization effectively highlights which genres tend to be more popular on Spotify, aiding in understanding genre-specific audience preferences.

Results / Summary of our Analysis :

From our comprehensive analysis of the dataset, we've unearthed several key insights into how musical characteristics such as tempo, valence, and energy vary by genre and influence track popularity on Spotify. Through our various data wrangling efforts, we discovered that genres like Pop and Electronic exhibit higher energy levels and tempos, which correlates with their higher popularity ratings on Spotify. This finding supports the notion that more energetic and upbeat music tends to be more popular among Spotify users.

- Our visualizations further highlighted these relationships, with the bar chart showing average popularity by genre indicating that certain genres consistently achieve higher popularity scores.
- The scatter plots and overlayed graphics revealed a clear trend where tracks with higher tempo and energy not only vary significantly across different genres but also tend to attract more listeners, leading to higher popularity.
- These visual patterns underscore the direct impact of these musical characteristics on listener preferences and track success on the platform.
- Moreover, the faceted histograms and box plots provided a deeper look into the distribution of valence and energy within genres, illustrating how emotional content (valence) and intensity (energy) of tracks are intricately linked to their popularity. For example, tracks with high valence in genres such as Pop and Dance show a positive association with popularity, suggesting that tracks perceived as more positive are favored by listeners in these genres.

In conclusion, our analysis not only answered our guiding research question but also shed light on the dynamic interplay between musical characteristics and their impact on a track's success on Spotify. By understanding these trends, music producers, artists, and marketers can better align their offerings with listener preferences, potentially enhancing user engagement and satisfaction. This study also paves the way for future research to explore other facets of music consumption and its implications in the digital music era.

Challenges Faced :

During our analysis, we encountered several challenges that enriched our learning experience.

- Initially, data quality issues such as missing values in key variables like tempo, valence, and popularity required us to perform data cleaning and imputation, potentially introducing biases. The dataset's complexity necessitated advanced data manipulation techniques, including merging multiple sources and transforming data formats, which were initially daunting.
- Additionally, the high dimensionality of the dataset posed a challenge in determining the most relevant features for our analysis, leading us to employ dimensionality reduction strategies.
- Creating meaningful visualizations also presented difficulties; we needed to carefully design our graphs to clearly communicate complex data relationships, requiring iterative refinement to balance aesthetic appeal with clarity.

How did we overcome the challenges :

We implemented a series of strategic solutions.

- For issues related to data quality, such as missing values, we utilized median and mean imputation methods tailored to the specific data distributions within genres, thereby preserving the integrity of our analysis.
- We tackled the complexity of data manipulation by leveraging powerful R packages like tidyverse, which allowed us to efficiently filter, summarize, and transform the dataset. In addressing the high dimensionality of the data, we employed exploratory data analysis and dimensionality reduction techniques to focus on the most impactful variables.
- For visualization challenges, we iteratively refined our plots using ggplot2 to balance aesthetic appeal with clarity, ensuring that our findings were communicated effectively.
- Lastly, to ensure our results were interpretable and actionable, we continuously referred back to our initial research questions and the real-world context of the music industry, which guided our analysis and helped us provide meaningful insights.