Exploratory Analysis of Spotify Tracks

Rishita Shukla, Khush Chutani, Xiang Lin

2025-12-17

Introduction

Music connects people all across the world and hence music streaming platforms generate a lot of data. Our dataset consists of information regarding 32,833 Spotify tracks authored by Charlie Thompson, Josiah Parry, Donal Phipps, and Tom Wolff. It includes 23 attributes such as track popularity, loudness, liveliness, energy, etc. This project follows an exploratory data analysis done on these audio features and the main goal is to find various relationships among them.

This exploratory data analysis answers these three research questions:

1. Which artists produce the highest-energy music on average, and how does energy relate to loudness?
2. Is there a relationship between a song's emotional positivity (valence) and its danceability?
3. How does tempo vary across different Spotify playlist genres?

Data Provenance

This data was released by Charlie Thompson, Josiah Parry, Donal Phipps, and Tom Wolff on 21st January 2020 and was collected via the spotifyr package. It originates from Spotify's Web API and contains information of about 32,833 songs ranging over 23 attributes. It was released on the weekly social data project, Tidy Tuesday, organized by the Data Science Learning Community with the purpose of providing real-world datasets for people to work with.

FAIR and CARE Principles

FAIR Principles:

1. Findable: The dataset is easily discoverable through public repositories and it contains documentation that well described its contents and structure.
2. Accessible: The dataset is easily accessible and downloadable without any restrictive permissions, which allows it to be used for analysis.
3. Interoperable:The dataset is provided in common formats such as CSV, which makes it easy to use since its compatible with tools like R, Python, etc. All the variable names follow the convention standards.
4. Reusable: The dataset is extremely reuseable and can be used for many different analytical goals purposes.

Hence, the dataset satisfies the FAIR principles.

CARE Principles:

1. Collective Benefit: The main purpose of this dataset is educational and research use, which in turn leads us to better understanding various audio features and their patterns.
2. Authority to Control:Spotify is responsible for the original data collection, so users do not have authority over it. This principle is partially satisfied.
3. Responsibility: The dataset avoids any personal information and focuses on audio features which minimizes ethical risk.
4. Ethics: Since the dataset describes music tracks and audio features and does not include sensitive information, it is ethical.

Hence, the dataset partially satisfies CARE principles.

Attributes

The attributes used throughout this project include:

1. track_artist: This is a categorical attribute and it lists out the name of the artist of any particular track.
2. energy: This is a numeric attribute and it measures how intense the track is.

3. loudness: This is a numeric attribute and it measures the loudness of a track in decibels(dB).
4. valence: This is a numeric attribute and it measures the positivity of a track.
5. danceability: This is a numeric attribute and it measures how danceable a track is.
6. playlist_genre: This is a categorical attribute and it lists out the genre of any particular track. Genres include: rock, rap, pop, latin, edm, r&b.
7. tempo: This is a numeric attribute and it measures the speed of a track in beats per minute(BPM).

Top Artists by Energy and Loudness

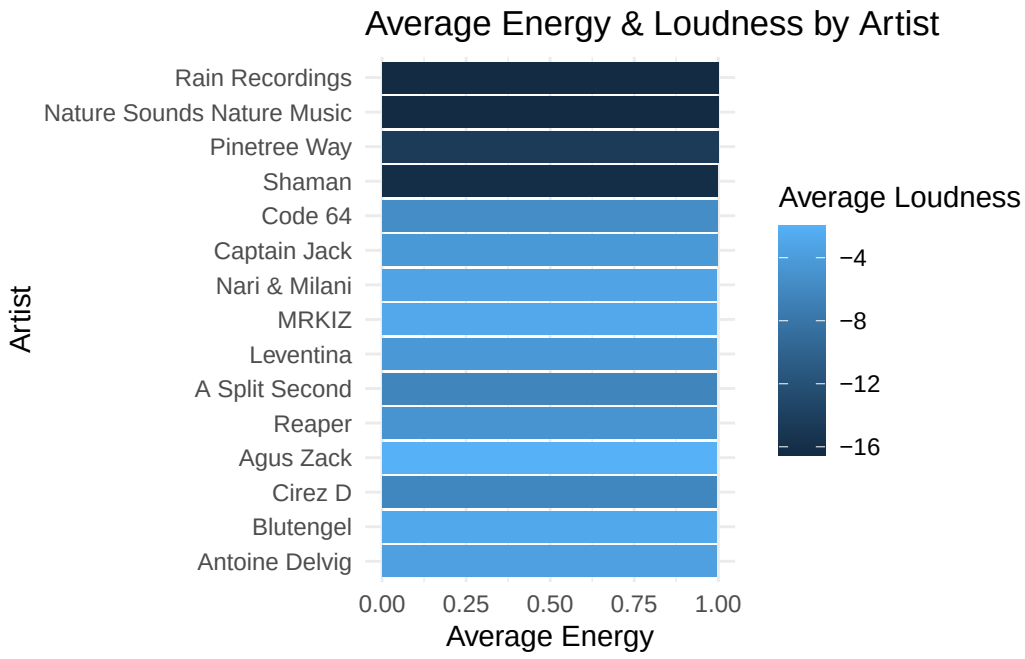
Table

Table 1: Top Artists by Average Energy and Loudness

Artist	Average Energy	Average Loudness (dB)
Nature Sounds Nature Music	1.00	-16.4
Rain Recordings	1.00	-16.5
Pinetree Way	1.00	-14.6
Captain Jack	1.00	-4.4
Code 64	1.00	-5.6
Shaman	1.00	-16.1
A Split Second	1.00	-6.4
Leventina	1.00	-4.5
MRKIZ	1.00	-2.9
Nari & Milani	1.00	-3.2
Reaper	1.00	-4.8
Agus Zack	0.99	-2.0
Antoine Delvig	0.99	-3.6
Blutengel	0.99	-2.9
Cirez D	0.99	-6.2

The image is a table titled “Top Artists by Average Energy and Loudness.” It consists of three columns: “Artist,” “Average Energy,” and “Average Loudness (dB).” The table contains information for fourteen artists, with their respective average energy and loudness readings. All entries have an average energy of either 1.00 or 0.99. The average loudness varies, with values ranging from -16.5 dB to -2.0 dB. The table is neatly organized and formatted with alternating row shading for clarity.

Bar Chart



The image is a horizontal bar chart titled "Average Energy & Loudness by Artist." It displays the average energy and loudness levels for various artists. The artists are listed along the vertical axis on the left side, while the horizontal axis at the bottom indicates average energy, ranging from 0.00 to 1.00. The bars are color-coded by average loudness, with a gradient scale to the right indicating values from -4 to -16. The top of the chart shows darker bars representing artists like Rain Recordings and Nature Sounds Nature Music, indicating lower loudness levels. It progresses to lighter shades with artists like Antoine Delvig at the bottom, suggesting higher loudness levels.

Relationship between Valence and Danceability

Table

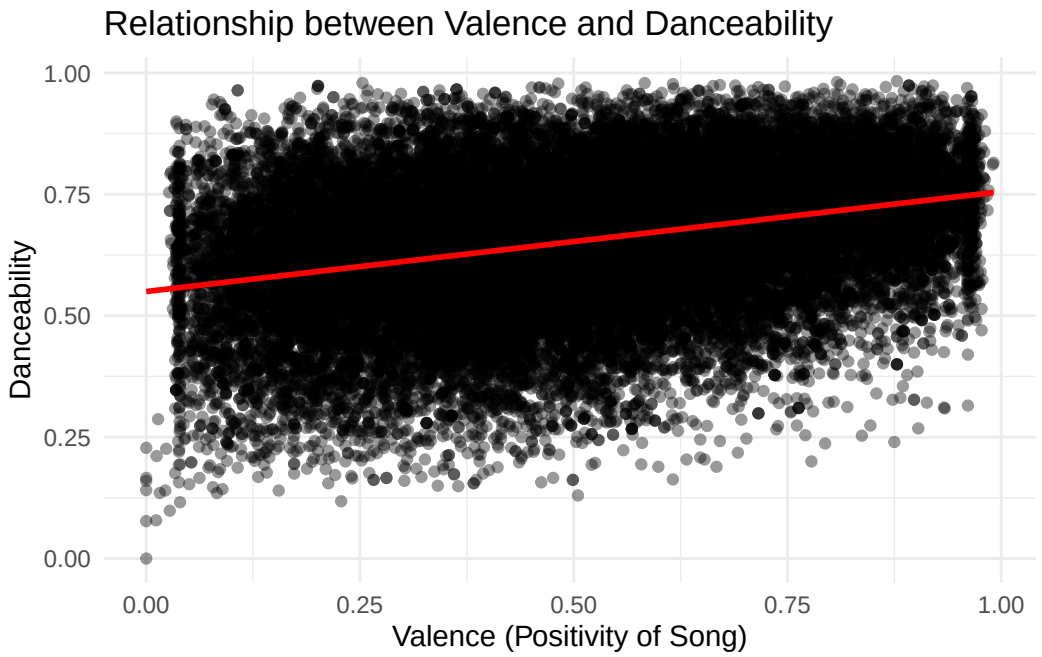
The image is a table titled "Average Danceability by Valence Bin." It consists of three columns labeled: "Valence Bin," "Average Danceability," and "Number of Songs." The table displays several rows, each corresponding to a range of valence values, their associated average danceability scores, and the count of songs within those ranges. The valence bins are presented in intervals of 0.1, starting from (0,0.1] up to (0.9,1], with an additional row labeled "NA." The data shows a progressive increase in average danceability with the increase in valence bin

Table 2: Average Danceability by Valence

Valence Bin	Average Danceability	Number of Songs
(0,0.1]	0.5616663	991
(0.1,0.2]	0.5883724	2610
(0.2,0.3]	0.6029460	3466
(0.3,0.4]	0.6187275	4337
(0.4,0.5]	0.6355416	4424
(0.5,0.6]	0.6621385	4810
(0.6,0.7]	0.6914027	4281
(0.7,0.8]	0.7034048	3743
(0.8,0.9]	0.7275738	2705
(0.9,1]	0.7360171	1460
NA	0.0000000	1

intervals, except for the row labeled "NA," which displays a danceability score of 0 and a count of 1 song.

Scatter Plot



The image is a scatter plot illustrating the relationship between "Valence (Positivity of Song)"

on the x-axis and "Danceability" on the y-axis. Both axes range from 0.00 to 1.00 in increments of 0.25. The plot contains a dense collection of black dots representing data points scattered across the graph. A red trend line spans diagonally from the lower left to the upper right, indicating a positive correlation between valence and danceability. The majority of dots are concentrated towards the center, forming a thick band around the trend line, with fewer dots dispersed at the extremities.

Tempo Distribution by Genre

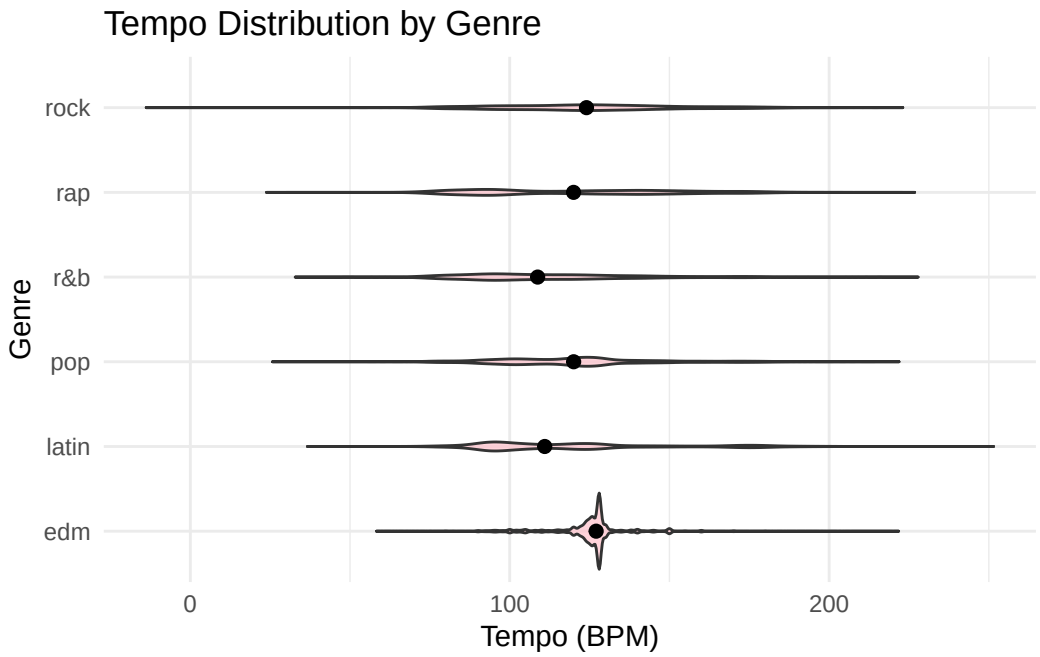
Table

Table 3: Tempo Summary by Genre

Genre	Median Tempo	Mean Tempo	Min Tempo	Max Tempo	Number of Songs
edm	127.056	125.7680	60.045	219.991	6043
latin	110.962	118.6309	48.718	239.440	5153
pop	120.017	120.7432	35.477	212.137	5507
r&b	108.744	114.2222	46.169	214.516	5431
rap	119.988	120.6606	38.985	211.644	5743
rock	124.035	124.9888	0.000	209.257	4951

The image is a table titled "Tempo Summary by Genre." It consists of six rows and six columns. Each row is dedicated to a specific music genre: edm, latin, pop, r&b, rap, and rock. The columns represent different tempo-related statistics and the number of songs, with headings indicating "Genre," "Median Tempo," "Mean Tempo," "Min Tempo," "Max Tempo," and "Number of Songs." The numerical values under each heading provide tempo statistics for each genre. The entire table has a light gray background with bold column headings.

Violin Plot



The image is a violin plot depicting the tempo distribution by music genre. The horizontal axis represents tempo measured in beats per minute (BPM), ranging from 0 to 200. The vertical axis lists music genres: rock, rap, r&b, pop, latin, and edm. Each genre has a corresponding, symmetrically shaped pink violin plot indicating the distribution and density of tempo values. A black dot marks the median tempo for each genre. The edm genre has a notably more complex and multi-modal distribution compared to the other genres, which have more elongated, narrow shapes with central bulges.

Author Contribution

Rishita Shukla led the data cleaning, exploratory analysis, visualization design, and made the final report and delegated tasks to the other team members.

Khush Chutani contributed to the narrative development, FAIR and CARE principles, and interpretation of results.

Xiang Lin contributed to research on the attributes, data provenance, and assisted in table formatting.

All authors contributed to atleast one visualization and reviewed and approved the final report.

Code Appendix

```
#------STEP 1: Load necessary packages and dataset------
library(tidyverse)
library(kableExtra)

spotify_songs <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytues

#------STEP 2: Clean dataset------

spotify_clean <- spotify_songs %>%

  distinct() %>%
  filter(!is.na(track_popularity),
         !is.na(track_artist),
         !is.na(track_name)) %>%
  mutate(
    playlist_genre = tolower(playlist_genre),
    playlist_genre = case_when(
      playlist_genre %in% c("", "na", "n/a", "none") ~ NA,
      TRUE ~ playlist_genre
    )
  ) %>%
  mutate(
    across(c(danceability, energy, valence, acousticness,
             instrumentalness, speechiness),
           ~ ifelse(. < 0 | . > 1, NA, .))
  ) %>%
  mutate(
    track_artist = as.factor(track_artist),
    playlist_genre = as.factor(playlist_genre),
    playlist_subgenre = as.factor(playlist_subgenre)
  )

#------STEP 3: Visualizations------

#TABLE AND PLOT 1
artist_energy <- spotify_clean %>%
  group_by(track_artist) %>%
  summarise(
    mean_energy = mean(energy, na.rm = TRUE),
```

```
    mean_loudness = mean(loudness, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  arrange(desc(mean_energy)) %>%
  slice(1:15)  # pick top 15 artists

artist_energy %>%
  mutate(
    mean_energy = round(mean_energy, 2),
    mean_loudness = round(mean_loudness, 1)
  ) %>%
  kable(
    caption = "Top Artists by Average Energy and Loudness",
    col.names = c("Artist", "Average Energy", "Average Loudness (dB)"),
    align = c("l", "c", "c")
  ) %>%
  kable_styling(
    full_width = FALSE
  )

artist_energy <- spotify_clean %>%
  group_by(track_artist) %>%
  summarise(mean_energy = mean(energy, na.rm = TRUE),
            mean_loudness = mean(loudness, na.rm = TRUE)) %>%
  arrange(desc(mean_energy))

ggplot(artist_energy, aes(x = reorder(track_artist, mean_energy),
                          y = mean_energy, fill = mean_loudness)) +
  geom_col() +
  coord_flip() +
  labs(title = "Average Energy & Loudness by Artist",
       x = "Artist", y = "Average Energy",
       fill = "Average Loudness") +
  theme_minimal()

#TABLE AND PLOT 2
valence_table <- spotify_clean %>%
  mutate(valence_bin = cut(valence, breaks = seq(0, 1, 0.1))) %>%
  group_by(valence_bin) %>%
  summarise(
    avg_danceability = mean(danceability, na.rm = TRUE),
    n_songs = n(),
```

```
    .groups = "drop"
  )

valence_table %>%
  kable(
    caption = "Average Danceability by Valence Bin",
    col.names = c("Valence Bin", "Average Danceability", "Number of Songs"),
    align = c("l", "c", "c")
  ) %>%
  kable_styling(full_width = FALSE)

ggplot(spotify_clean, aes(x = valence, y = danceability)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Relationship between Valence and Danceability",
       x = "Valence (Positivity of Song)",
       y = "Danceability") +
  theme_minimal()

#TABLE AND PLOT 3
tempo_table <- spotify_clean %>%
  group_by(playlist_genre) %>%
  summarise(
    median_tempo = median(tempo, na.rm = TRUE),
    mean_tempo = mean(tempo, na.rm = TRUE),
    min_tempo = min(tempo, na.rm = TRUE),
    max_tempo = max(tempo, na.rm = TRUE),
    n_songs = n(),
    .groups = "drop"
  )

tempo_table %>%
  kable(
    caption = "Tempo Summary by Genre",
    col.names = c("Genre", "Median Tempo", "Mean Tempo", "Min Tempo", "Max Tempo", "Number o
    align = c("l", "c", "c", "c", "c", "c")
  ) %>%
  kable_styling(full_width = FALSE)

ggplot(spotify_clean, aes(x = playlist_genre, y = tempo)) +
  geom_violin(trim = FALSE, fill = "pink", alpha = 0.7) +
  stat_summary(fun = median, geom = "point", color = "black", size = 2) +
```

```
  coord_flip() +
  labs(title = "Tempo Distribution by Genre",
       x = "Genre", y = "Tempo (BPM)") +
  theme_minimal()
```