Statistical Analysis Across Different Playlist Genres

Nicole Rodriguez

04/07/2025

Contents

Introduction

This analysis explores the genre of songs by Queen and Selena Gomez utilizing a Spotify dataset from GitHub. Queen, known for blending rock and pop, will be compared to Selena Gomez, a pop icon with influences in dance and Latin music. By examining the genre distribution of both artists, we'll uncover how their musical styles differ and evolve, offering insights into their unique sounds and artistic trajectories.

Data

Load Libraries & Packages

```r
# Load required packages
if (!require("mosaic"))
  install.packages("mosaic")
if (!require("dplyr"))
  install.packages("dplyr")
if (!require("ggplot2"))
  install.packages("ggplot2")
if (!require("tidyverse"))
  install.packages("tidyverse")
if (!require("readr"))
  install.packages("readr")

library(mosaic) # Stats analysis
library(dplyr) # Data manipulation
library(ggplot2) # Data Visualization
library(tidyverse) # Data packages
library(readr) # Read data
```

Data Description

This analysis will use the spotify_songs.csv data set which includes a total of 32,833 observations and 23 variables.

Load & Preview Data

```r
# Load dataset into R
spotify_songs <- read_csv("spotify_songs.csv")

# View data
dim(spotify_songs) # Dimensions
```

```
## [1] 32833    23
```

```r
colnames(spotify_songs) # All variables
```

```
##  [1] "track_id"                "track_name"
##  [3] "track_artist"            "track_popularity"
##  [5] "track_album_id"          "track_album_name"
##  [7] "track_album_release_date" "playlist_name"
##  [9] "playlist_id"             "playlist_genre"
## [11] "playlist_subgenre"       "danceability"
## [13] "energy"                  "key"
## [15] "loudness"                "mode"
## [17] "speechiness"             "acousticness"
## [19] "instrumentalness"        "liveness"
## [21] "valence"                 "tempo"
## [23] "duration_ms"
```

```
head(spotify_songs) # Preview data
```

```
## # A tibble: 6 x 23
##   track_id             track_name track_artist track_popularity track_album_id
##   <chr>                <chr>      <chr>                   <dbl> <chr>
## 1 6f807x0ima9a1j3VPbc7VN I Don't C~ Ed Sheeran               66 2oCs0DGTsRO98~
## 2 0r7CVbZTWZgbTCYdfa2P31 Memories ~ Maroon 5                 67 63rPSO264uRjW~
## 3 1z1Hg7Vb0AhHDiEmnDE79l All the T~ Zara Larsson             70 1HoSmj2eLcsrR~
## 4 75FpbthrwQmzHlBJLuGdC7 Call You ~ The Chainsm~             60 1nqYsOef1yKKu~
## 5 1e8PAfcKUYoKkxPhrHqw4x Someone Y~ Lewis Capal~             69 7m7vv9wlQ4i0L~
## 6 7fvUMiyapMsRRxr07cU8Ef Beautiful~ Ed Sheeran               67 2yiy9cd2QktrN~
## # i 18 more variables: track_album_name <chr>, track_album_release_date <chr>,
## #   playlist_name <chr>, playlist_id <chr>, playlist_genre <chr>,
## #   playlist_subgenre <chr>, danceability <dbl>, energy <dbl>, key <dbl>,
## #   loudness <dbl>, mode <dbl>, speechiness <dbl>, acousticness <dbl>,
## #   instrumentalness <dbl>, liveness <dbl>, valence <dbl>, tempo <dbl>,
## #   duration_ms <dbl>
```

Variables

From the 23 variables, the following will be utilized for this analysis:

- acousticness
- danceability
- duration_ms
- energy
- liveness
- loudness
- speechiness
- tempo
- valence

Data Analysis

Descriptive Statistics

By using `favstats()`, several *descriptive statistics* for `speechiness` separated by `playlist_genre`, can be viewed.

```
# Get descriptive stats for speechiness by playlist_genre
stats <- favstats(speechiness ~ playlist_genre, data = spotify_songs)

# Descriptive stats results
print(stats)
```

```
##   playlist_genre   min      Q1 median     Q3   max       mean         sd    n
## 1            edm 0.0239 0.043800 0.0599 0.0985 0.624 0.08669548 0.07105803 6043
```

```
## 2             latin 0.0232 0.043600 0.0674 0.1270 0.662 0.10265315 0.08772875 5155
## 3               pop 0.0228 0.036900 0.0490 0.0791 0.869 0.07399127 0.06782213 5507
## 4               r&b 0.0224 0.041900 0.0679 0.1580 0.918 0.11679214 0.10718205 5431
## 5               rap 0.0243 0.078825 0.1775 0.2900 0.877 0.19750588 0.13244392 5746
## 6              rock 0.0000 0.032800 0.0419 0.0634 0.488 0.05769568 0.04525378 4951
##    missing
## 1        0
## 2        0
## 3        0
## 4        0
## 5        0
## 6        0
```

In summary, speechiness differs across genres, with rap having the most spoken word elements and rock the least. Latin, R&B, and pop have moderate speechiness, with some songs featuring more spoken or rap-like portions. This variation highlights how different genres incorporate spoken content in varying degrees, influencing their overall sound and style.

Highest Average Speechiness

```r
# Average speechiness by playlist_genre
av_speech <- tapply(spotify_songs$speechiness, spotify_songs$playlist_genre, mean, na.rm = TRUE)
print(av_speech)
```

```
##        edm      latin        pop        r&b        rap       rock
## 0.08669548 0.10265315 0.07399127 0.11679214 0.19750588 0.05769568
```

```r
# Find genre with highest average speechiness
highest_speechiness <- names(av_speech)[which.max(av_speech)]
highest_speechiness_value <- max(av_speech)

# Display highest average level of speechiness results
highest_av_speech <- data.frame(playlist_genre = highest_speechiness,
                                highest_average_speechiness = highest_speechiness_value)
print(highest_av_speech)
```

```
##   playlist_genre highest_average_speechiness
## 1            rap                   0.1975059
```

The genre that has the *highest average* level of `speechiness` is **rap**.

Rap had an average level of 'speechiness' of 0.2.

Lowest Average Speechiness

```r
# Find genre with lowest average speechiness
lowest_speechiness <- names(av_speech)[which.min(av_speech)]
lowest_speechiness_value <- min(av_speech)
```

4

```r
# Display lowest average level of speechiness results
lowest_av_speech <- data.frame(playlist_genre = lowest_speechiness,
                               lowest_average_speechiness = lowest_speechiness_value)
print(lowest_av_speech)
```

```
##   playlist_genre lowest_average_speechiness
## 1           rock                 0.05769568
```

The genre that has the *lowest average* level of `speechiness` is **rock**.

Rock had an average level of 'speechiness' of 0.06.

Queen & Selena Gomez

The dataset was filtered to only observe songs by Queen and Selena Gomez.

```r
# Filter dataset for Queen & Selena Gomez
QueenVsSelenaGomez_songs <- spotify_songs %>%
  filter(track_artist %in% c("Queen", "Selena Gomez"))

# Preview filtered dataset
head(QueenVsSelenaGomez_songs)
```

```
## # A tibble: 6 x 23
##   track_id               track_name track_artist track_popularity track_album_id
##   <chr>                  <chr>      <chr>                   <dbl> <chr>
## 1 0rsRJOlDJxfNPNTEALa2t3 Rare       Selena Gomez               73 0CQPSBCo1cLsE~
## 2 1HfMVBKM75vxSfsQ5VefZ5 Lose You ~ Selena Gomez               93 3tBkjgxDqAwss~
## 3 7HMmFQsKsljwTw8bS7lu19 Rare       Selena Gomez               88 3YPFaTR7WMi1H~
## 4 1HfMVBKM75vxSfsQ5VefZ5 Lose You ~ Selena Gomez               93 3tBkjgxDqAwss~
## 5 4yI3HpbSFSgFZtJP2kDe5m Look At H~ Selena Gomez               86 4OO4QLiBOlGHQ~
## 6 7HMmFQsKsljwTw8bS7lu19 Rare       Selena Gomez               88 3YPFaTR7WMi1H~
## # i 18 more variables: track_album_name <chr>, track_album_release_date <chr>,
## #   playlist_name <chr>, playlist_id <chr>, playlist_genre <chr>,
## #   playlist_subgenre <chr>, danceability <dbl>, energy <dbl>, key <dbl>,
## #   loudness <dbl>, mode <dbl>, speechiness <dbl>, acousticness <dbl>,
## #   instrumentalness <dbl>, liveness <dbl>, valence <dbl>, tempo <dbl>,
## #   duration_ms <dbl>
```

R Analysis

```r
# Variables to be compared
compare_spotify_songs <- QueenVsSelenaGomez_songs %>%
  select(track_artist, danceability, energy, loudness, speechiness,
         acousticness, liveness, valence, tempo, duration_ms)

# Reshaping the data from wide to long format for plotting
long_spotify_songs <- compare_spotify_songs %>%
  pivot_longer(cols = danceability:duration_ms, names_to = "variable", values_to = "value")
```

```r
# Boxplot to compare each variable between Queen and Selena Gomez together
ggplot(long_spotify_songs, aes(x = track_artist, y = value, fill = track_artist)) +
  # Boxplot
  geom_boxplot() +
  # Facet wrap
  facet_wrap(~variable, scales = "free") +
  # Adding specific color to plot
  scale_fill_manual(values = c("Queen" = "#CCCCFF",
                               "Selena Gomez" = "#9ACEEB")) +
  # Labels
  labs(title = "A Comparative Analysis of Musical Attributes
       Between Queen & Selena Gomez",
       x = "Artist", y = "Value", fill = "Artist") +
  theme_minimal()
```


A Comparative Analysis of Musical Attributes Between Queen & Selena Gomez
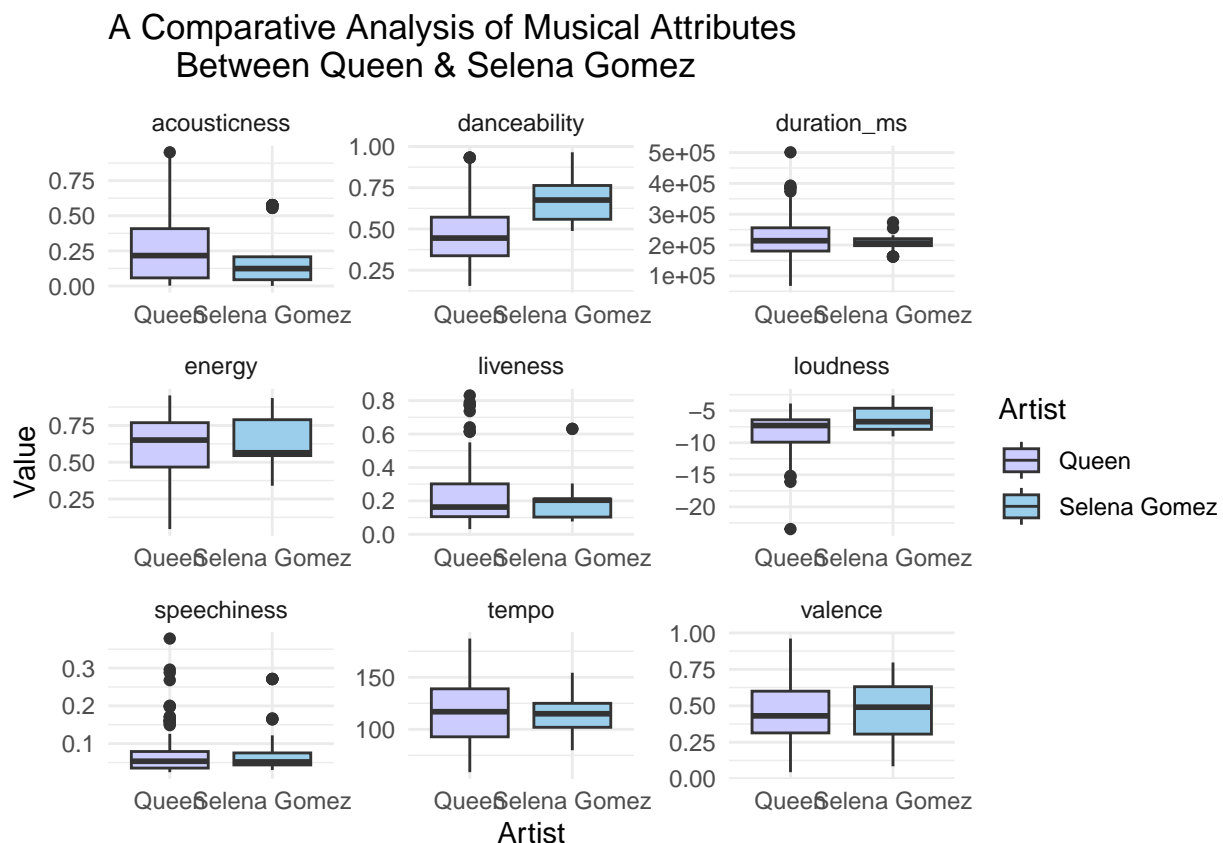
Based on the box plot observations, we can summarize the following:

- Selena Gomez's songs generally have lower acoustics, higher danceability, and more speechiness compared to Queen's, aligning with typical pop characteristics.

- Queen's songs have a wider range in duration, higher liveness (live performance elements), and are louder on average.

- Selena's songs tend to have slightly higher energy and a more upbeat feel, while Queen's tracks have a slightly lower tempo.

- Both artists show similar valence ranges, but Selena's songs have a higher median, suggesting a more positive or upbeat emotional tone.

Canva Presentation

Statistical Analysis Across Different Playlist Genres by Nicole Rodriguez

Canva Presentation Link