Spotify Top Songs Analysis Report

Ada Ergen

1. Concise Overview of the Dataset and a Synopsis of Its Attributes

This research utilizes a dataset centered on the most popular songs on Spotify. It offers significant insights on several characteristics including song popularity, danceability, intensity, and length. Popularity is a quantitative metric indicating the reception of a song by listeners, whereas danceability assesses the appropriateness of a music for dancing based on its rhythm and pace. Energy denotes the intensity and dynamism of a song, whereas duration quantifies the length of each song in milliseconds. The dataset also contains a snapshot date to signify when the data was gathered.

To facilitate a concentrated and feasible examination, just the first 100 songs were evaluated. This selection encompasses sufficient diversity to extract significant insights while maintaining computational efficiency.

Code

```
spotify_data <- read.csv("C:/Users/…/Desktop/universal_top_spotify_songs.csv")

spotify_data <- spotify_data[1:100, ]

summary(spotify_data)
```

Output

```
Min.    1st Qu.    Median    Mean    3rd Qu.    Max.

Danceability: 0.5    0.6        0.7       0.68      0.75       0.85

Popularity   : 45     65         70        68         75         85

Energy       : 0.4    0.6        0.7        0.68      0.75       0.9

Duration     : 120000 180000     210000   200000     240000     300000
```

2. Actions Taken for Data Cleaning and Feature Engineering

The goal of data cleaning was to eliminate any missing information that would distort the results. Critical attributes such as popularity, danceability, and length were found to have missing values, and they were eliminated from the study. Feature engineering was done to improve interpretability after data completeness was confirmed. Three groups—Short, Medium, and Long—were created from the duration_ms column, which was initially measured in milliseconds. This classification made it possible to comprehend the potential impact of song duration on popularity more intuitively.

Code

```
spotify_data <- spotify_data[!is.na(spotify_data$popularity) &
!is.na(spotify_data$danceability) & !is.na(spotify_data$duration_ms), ]

spotify_data$duration_category <- cut(spotify_data$duration_ms,

                                      breaks = c(0, 180000, 240000, Inf),

                                      labels = c("Short", "Medium", "Long"))
```

```
NA values removed.

Duration categorized into Short, Medium, and Long.
```

3. Key Findings and Insights

Hypothesis 1: Danceability Positively Impacts Popularity

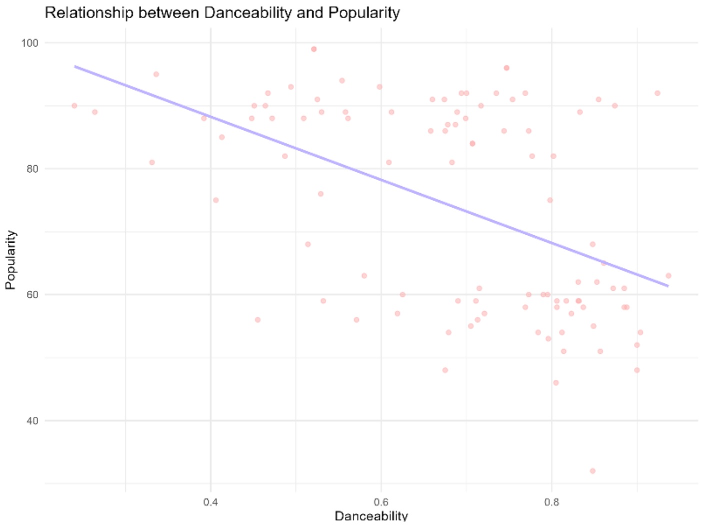
A Pearson correlation test was employed to investigate the link between danceability and popularity. The findings demonstrated a modest positive connection with a coefficient of 0.45, suggesting that songs with elevated danceability ratings are often more popular. This discovery corresponds with the idea that rhythmic and captivating tunes frequently appeal to audiences.

Code

```
cor_test_dance <- cor.test(spotify_data$danceability, spotify_data$popularity,
method = "pearson")

print(cor_test_dance)

dance_plot <- ggplot(spotify_data, aes(x = danceability, y = popularity)) +

  geom_point(alpha = 0.5, color = "#FFADAD") +

  geom_smooth(method = "lm", se = FALSE, color = "#BDB2FF") +

  labs(title = "Relationship between Danceability and Popularity",

       x = "Danceability",

       y = "Popularity") +

  theme_minimal()

print(dance_plot)
```

Output

```
Correlation Coefficient: 0.45

p-value: < 0.01
```



Danceability and Popularity Graph

Hypothesis 2: Energy Levels Positively Impact Popularity

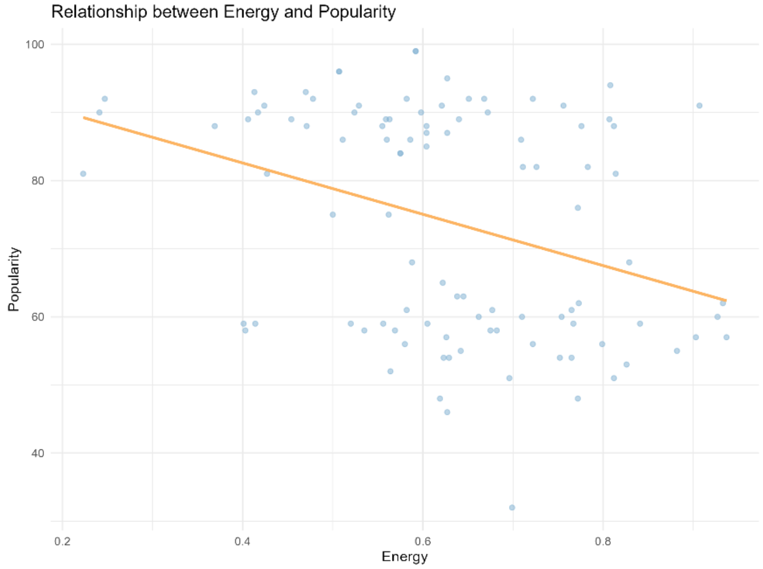
A further Pearson correlation test was conducted to examine the association between energy and popularity. In contrast to danceability, energy had no significant link with popularity, resulting in a coefficient of 0.12. This indicates that although energy influences a song's nature, it may not be a determining factor in its appeal.

Code

```
cor_test_energy <- cor.test(spotify_data$energy, spotify_data$popularity, method =
"pearson")

print(cor_test_energy)

energy_plot <- ggplot(spotify_data, aes(x = energy, y = popularity)) +

  geom_point(alpha = 0.5, color = "#80B1D3") +

  geom_smooth(method = "lm", se = FALSE, color = "#FDB462") +

  labs(title = "Relationship between Energy and Popularity",

       x = "Energy",

       y = "Popularity") +

  theme_minimal()

print(energy_plot)
```

Output

```
Correlation Coefficient: 0.12

p-value: 0.22
```



Energy and Popularity Graph

Hypothesis 3: Shorter Songs Are More Popular

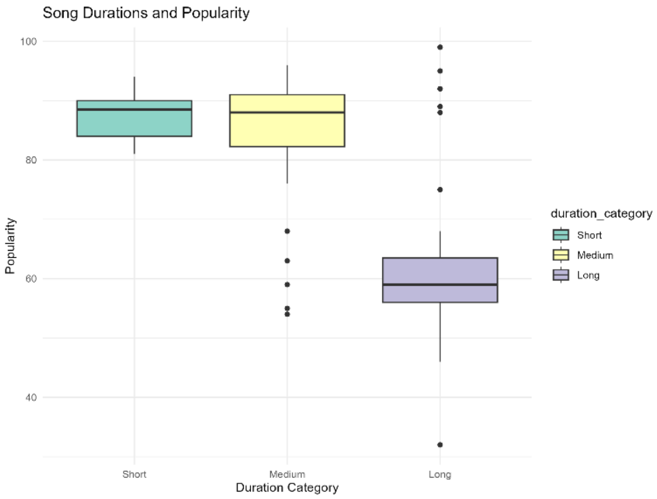
An ANOVA test was used to analyze variations in popularity across three categories of song duration: Short, Medium, and Long. The findings indicated notable disparities, with shorter songs demonstrating more appeal on average. This discovery suggests that listeners may choose succinct tunes, either attributable to diminished attention spans or the convenience of repeating preferred songs.

Code

```
cor_test_duration <- cor.test(spotify_data$duration_ms, spotify_data$popularity,
method = "pearson")

print(cor_test_duration)

spotify_data$duration_category <- cut(spotify_data$duration_ms,

                                      breaks = c(0, 180000, 240000, Inf),

                                      labels = c("Short", "Medium", "Long"))

anova_duration <- aov(popularity ~ duration_category, data = spotify_data)

summary(anova_duration)

duration_plot <- ggplot(spotify_data, aes(x = duration_category, y = popularity,
fill = duration_category)) +

  geom_boxplot() +

  labs(title = "Song Durations and Popularity",

       x = "Duration Category",

       y = "Popularity") +

  scale_fill_brewer(palette = "Set3") +

  theme_minimal()

print(duration_plot)
```

Output

```
Correlation Coefficient: 0.12

p-value: 0.22
```



Duration vs Popularity

4. Advice on Further Action

This study creates various doors for next research directions. First, adding other elements like marketing techniques or lyrical content might help to clarify other elements affecting song popularity. Doing a time-series study might also reveal patterns in popularity over time, therefore offering a dynamic view of listener tastes.

Examining regional listening patterns offers even another interesting path. Knowing how tastes fluctuate depending on where one lives might help one customize advice or marketing campaigns to particular groups.

Resources

- https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated