Analyzing Spotify Trends with R

Isabelle Fortaleza-Tan

2026-05-07

Spotify Dataset

- The Spotify Tracks Dataset contains audio features and popularity information for songs across many genres
- The dataset was obtained from Kaggle

Introduction

- Dataset: “Spotify Tracks Dataset” from Kaggle
- Contains audio features and popularity data for Spotify songs
- Includes over 100000 tracks across various genres
- This project investigates whether certain song characteristics are associated with higher Spotify popularity

Loading the Data

```
##                   track_name track_genre popularity danceability
## 1                     Comedy    acoustic         73        0.676
## 2           Ghost - Acoustic    acoustic         55        0.420
## 3             To Begin Again    acoustic         57        0.438
## 4 Can't Help Falling In Love    acoustic         71        0.266
## 5                    Hold On    acoustic         82        0.618
## 6       Days I Will Remember    acoustic         58        0.688
```

- read.csv() imports the dataset into R
- head() displays the first few rows of the dataset

Summary Statistics

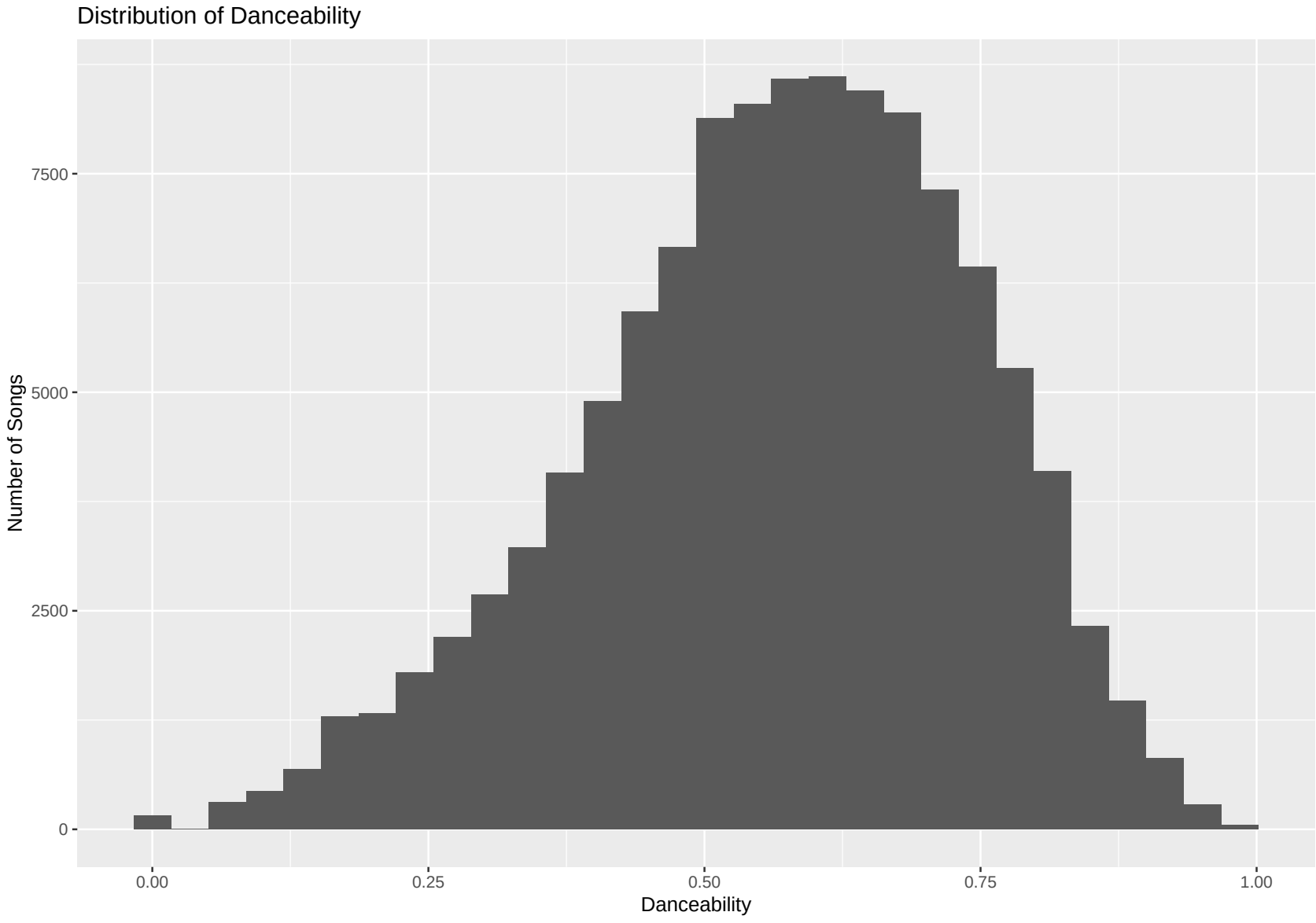
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   17.00   35.00   33.24   50.00  100.00

## [1] 33.23854

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.4560  0.5800  0.5668  0.6950  0.9850
```

- summary() gives descriptive statistics like minimum, median, mean, and maximum
- mean() calculates the average popularity score
- Danceability measures how suitable a song is for dancing
- The average popularity score is about 33, which suggests many songs in the dataset are not extremely mainstream
- The median danceability is around 0.58, meaning many songs are moderately danceable
- Danceability values range widely, showing substantial variation across songs

Distribution of Danceability



Distribution of Danceability

- This histogram shows the distribution of danceability scores across Spotify songs
- Most songs fall between about 0.4 and 0.8 danceability, suggesting moderately danceable songs are most common
- Very low and very high danceability scores appear less frequently

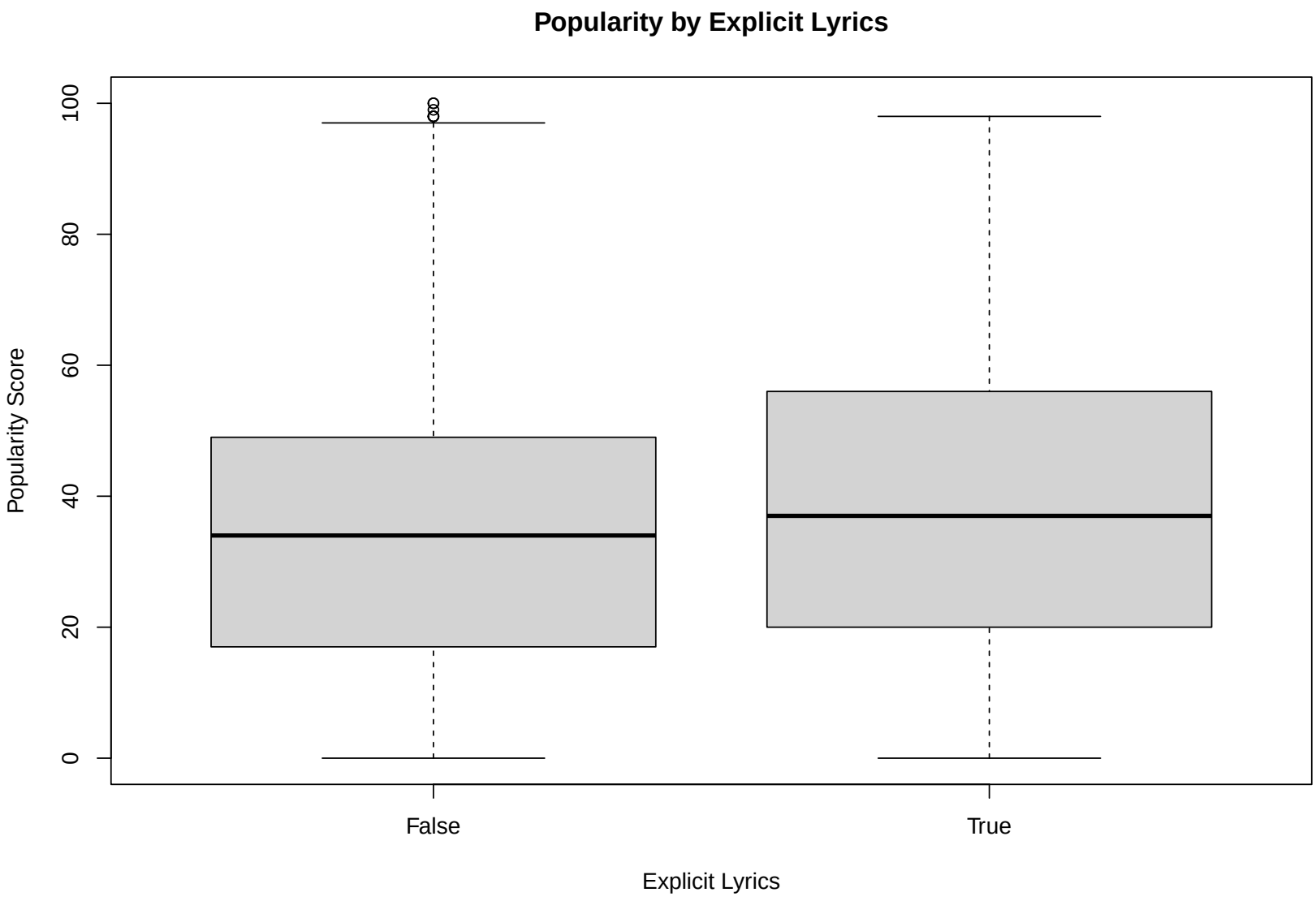
Explicit vs Non-Explicit Songs

```
##
##  False   True   <NA>
## 104253   9747      0
```

```
##
## False  True  <NA>
##  0.91  0.09  0.00
```

- About 91% of songs in the dataset are not explicit
- Only around 9% are labeled explicit
- This suggests non-explicit songs make up the majority of Spotify tracks in this dataset

Popularity by Explicit Lyrics



Popularity by Explicit Lyrics

- This boxplot compares popularity for explicit and non-explicit songs
- The median popularity for explicit songs appears slightly higher
- Explicit songs also show a somewhat wider spread of popularity scores
- However, the distributions overlap substantially, suggesting explicit content alone does not determine popularity

Average Popularity by Genre

```
##    track_genre popularity
## 82    pop-film     59.283
## 66       k-pop     56.896
## 16       chill     53.651
## 95         sad     52.379
## 45      grunge     49.594
## 56      indian     49.539
## 6        anime     48.772
## 34         emo     48.128
## 98   sertanejo     47.866
## 81         pop     47.576
```

- The genres with the highest average popularity include pop-film, k-pop, chill, and sad
- K-pop appears near the top of the rankings, suggesting strong listener engagement on Spotify
- The results suggest popularity on Spotify may vary substantially across musical genres

Conclusion

- The project demonstrated how data visualization and summary statistics can help reveal patterns in large datasets
- Overall, the analysis suggests that song characteristics such as danceability, genre, and explicit content may be related to popularity on Spotify
- Although no single factor fully determines popularity, the analysis shows how R can be used to explore trends and relationships within real-world data