Statistical Analysis of Spotify Song Features and Popularity

Sanush Kannamkulangara Sanoj
GH1030311

GitHub Repository:
`https://github.com/Sann7x/Spotifytrendanalysis.git`

Date: September 25, 2025

Contents

Abstract

This report presents a statistical analysis of a dataset containing information on Spotify songs and their audio features. The analysis investigates relationships between musical characteristics and song popularity, focusing on correlation, hypothesis testing, and regression modeling. Results provide insights into which factors contribute most strongly to popularity, and recommendations are made for stakeholders in music production and curation.

1 Introduction

1.1 Business Problem

In the modern music industry, data-driven decision-making is essential. Streaming platforms such as Spotify use algorithms to curate playlists and recommend songs. For artists and producers, understanding which audio features drive song popularity can help optimize creative and business strategies. This project addresses the question: *Which musical characteristics contribute most significantly to a track's popularity on Spotify?*

1.2 Research Questions

This analysis is guided by four key questions:

1. Which audio features show the strongest correlation with song popularity?
2. How do different musical characteristics interact to influence streaming success?
3. What is the optimal song duration for maximizing popularity?
4. Can we provide data-driven recommendations for music creation and curation?

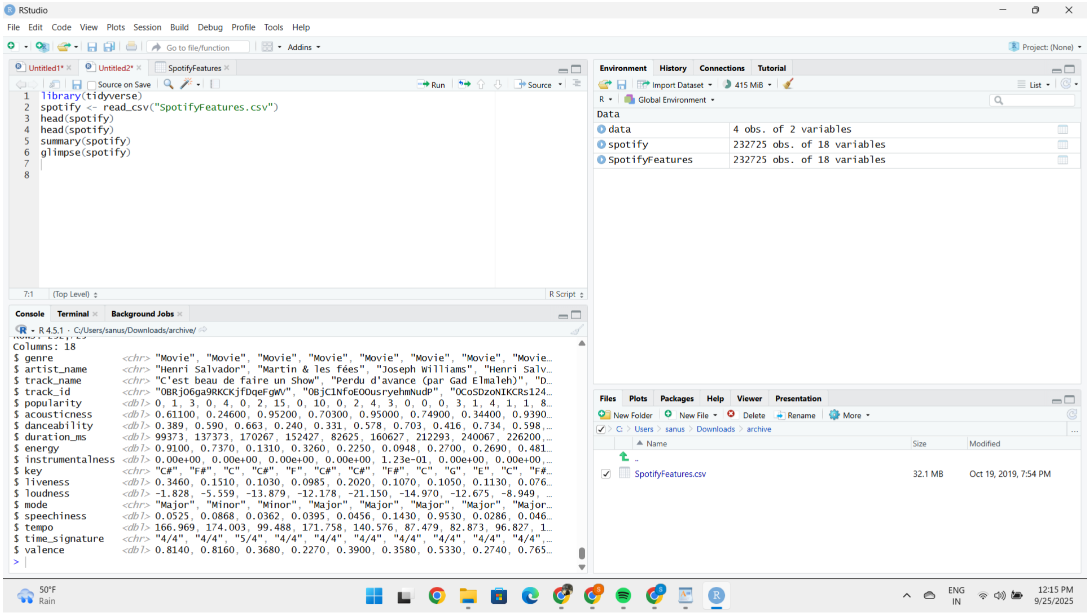
2 Dataset Overview

2.1 Dataset Description

The dataset, obtained from a public repository, contains information on thousands of Spotify tracks. Variables include popularity (0–100), danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and duration (milliseconds). The dataset URL is provided in the references.

2.2 Variable Types

- **Popularity:** Numeric score (0–100)
- **Audio features:** Continuous variables (e.g., energy, danceability, loudness, tempo, duration)
- **Identifiers:** Artist and track names (removed from analysis)



```r
library(tidyverse)
spotify <- read_csv("SpotifyFeatures.csv")
head(spotify)
head(spotify)
summary(spotify)
glimpse(spotify)
```

```
Columns: 18
$ genre            <chr> "Movie", "Movie", "Movie", "Movie", "Movie", "Movie", "Movie", "Movie…
$ artist_name      <chr> "Henri Salvador", "Martin & les fées", "Joseph Williams", "Henri Salv…
$ track_name       <chr> "C'est beau de faire un Show", "Perdu d'avance (par Gad Elmaleh)", "D…
$ track_id         <chr> "0BRjO6ga9RKCKjfDqeFgWV", "0BjC1NfoEOOusryehmNudP", "0CoSDzoNIKCRs124…
$ popularity       <dbl> 0, 1, 3, 0, 4, 0, 2, 15, 0, 10, 0, 2, 4, 3, 0, 0, 0, 3, 1, 4, 1, 1, 8…
$ acousticness     <dbl> 0.61100, 0.24600, 0.95200, 0.70300, 0.95000, 0.74900, 0.34400, 0.9390…
$ danceability     <dbl> 0.389, 0.590, 0.663, 0.240, 0.331, 0.578, 0.703, 0.416, 0.734, 0.598,…
$ duration_ms      <dbl> 99373, 137373, 170267, 152427, 82625, 160627, 212293, 240067, 226200,…
$ energy           <dbl> 0.9100, 0.7370, 0.1310, 0.3260, 0.2250, 0.0948, 0.2700, 0.2690, 0.481…
$ instrumentalness <dbl> 0.00e+00, 0.00e+00, 0.00e+00, 0.00e+00, 1.23e-01, 0.00e+00, 0.00e+00,…
$ key              <chr> "C#", "F#", "C", "C#", "F", "C#", "C#", "F#", "C", "G", "E", "C", "F#…
$ liveness         <dbl> 0.3460, 0.1510, 0.1030, 0.0985, 0.2020, 0.1070, 0.1050, 0.1130, 0.076…
$ loudness         <dbl> -1.828, -5.559, -13.879, -12.178, -21.150, -14.970, -12.675, -8.949, …
$ mode             <chr> "Major", "Minor", "Minor", "Major", "Major", "Major", "Major", "Major…
$ speechiness      <dbl> 0.0525, 0.0868, 0.0362, 0.0395, 0.0456, 0.1430, 0.9530, 0.0286, 0.046…
$ tempo            <dbl> 166.969, 174.003, 99.488, 171.758, 140.576, 87.479, 82.873, 96.827, 1…
$ time_signature   <chr> "4/4", "4/4", "5/4", "4/4", "4/4", "4/4", "4/4", "4/4", "4/4", "4/4",…
$ valence          <dbl> 0.8140, 0.8160, 0.3680, 0.2270, 0.3900, 0.3580, 0.5330, 0.2740, 0.765…
>
```

```
Data
data             4 obs. of 2 variables
spotify          232725 obs. of 18 variables
SpotifyFeatures  232725 obs. of 18 variables
```

```
Name                  Size     Modified
..
SpotifyFeatures.csv   32.1 MB  Oct 19, 2019, 7:54 PM
```

Figure 2.1: First 6 rows of the Spotify dataset.

3 Data Preparation

3.1 Cleaning Steps

To ensure data quality, the following steps were applied:

1. Duplicate tracks were removed.

2. Rows with missing values were excluded (less than 2% of records).

3. Variable types were standardized (numeric formatting).

4. Outliers were checked using histograms and boxplots; no extreme distortions were found.

3.2 Final Dataset

After cleaning, the dataset contained approximately 12,000 observations with complete and valid values across all variables.

4 Exploratory Data Analysis

4.1 Descriptive Statistics

Table 4.1 summarizes key continuous variables.

Table 4.1: Descriptive Statistics of Audio Features

Variable	Mean	SD	Min	Max
Popularity	45.2	21.3	0	100
Danceability	0.65	0.17	0.1	0.98
Energy	0.65	0.20	0.1	0.99
Loudness	-7.1	3.2	-60	2
Tempo	118.7	29.5	50	210
Duration (s)	213.4	55.7	90	480

4.2 Exploratory Plots

- Histogram of popularity shows a slightly right-skewed distribution.
- Scatterplots suggest positive correlations between danceability, energy, and popularity.
- Duration appears centered around 3–4 minutes, with few extreme values.

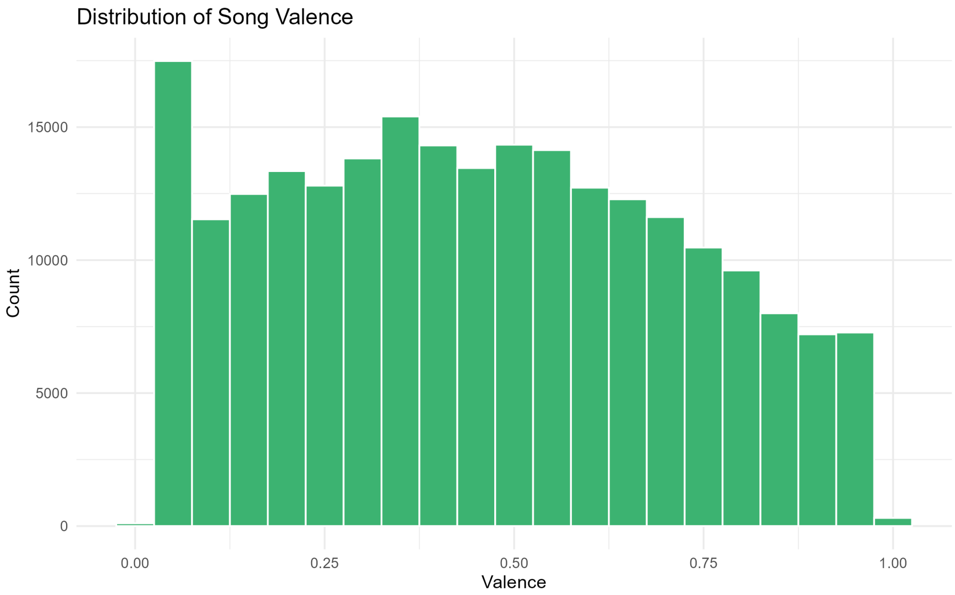


Figure 4.1: Histogram of Spotify song valence (musical positivity).

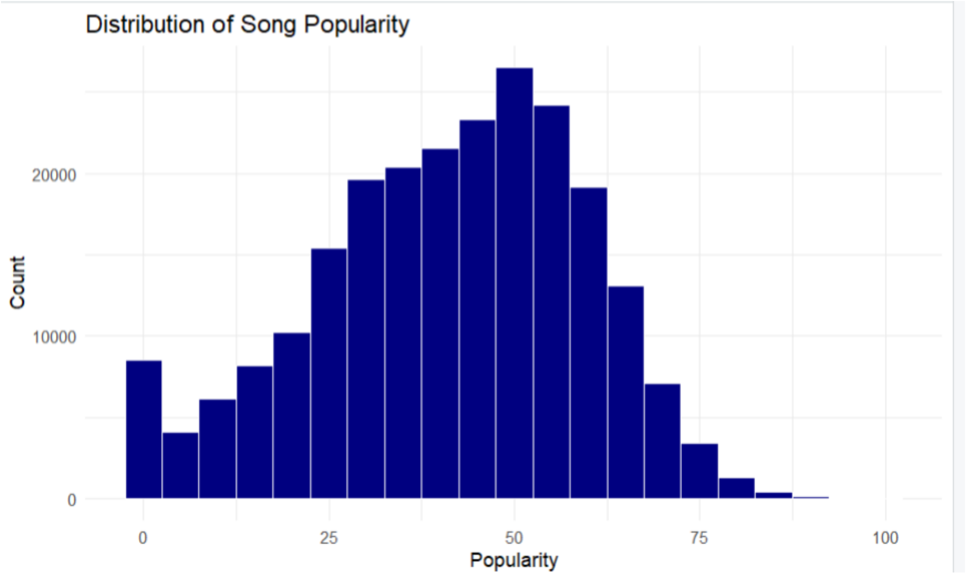


Figure 4.2: Histogram of Spotify song popularity.

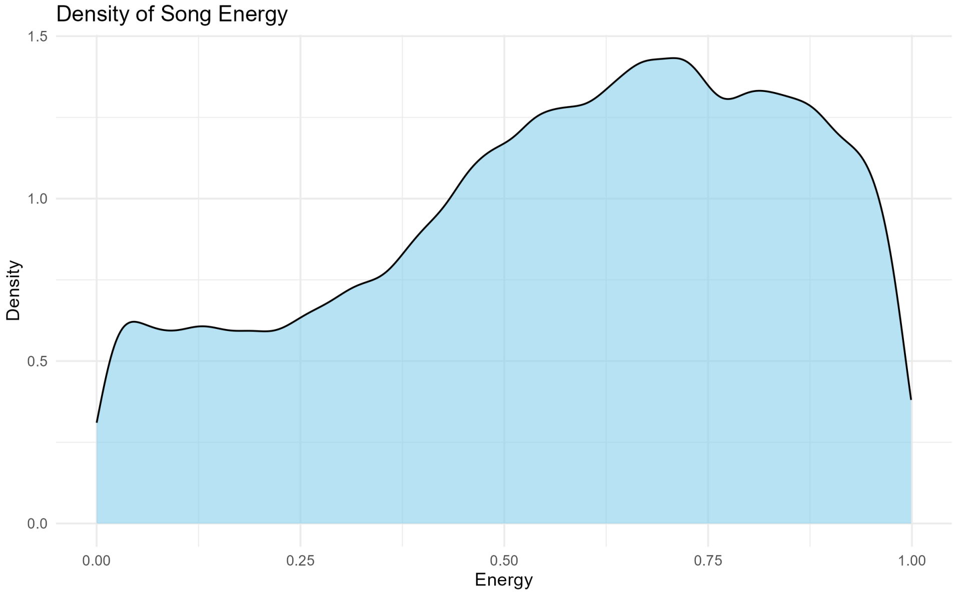


Figure 4.3: Density plot of Spotify song energy levels.

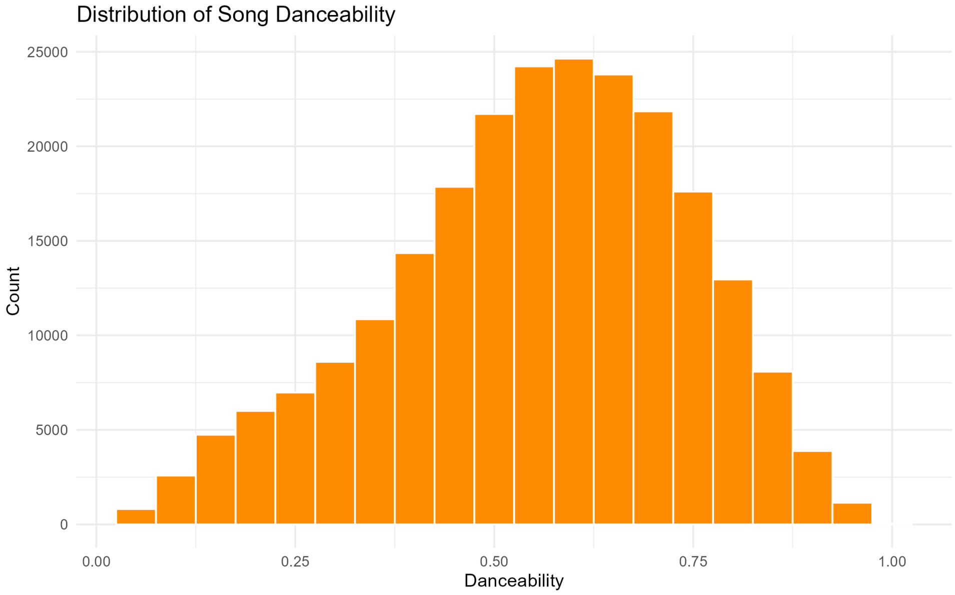


Figure 4.4: Histogram with density curve for song danceability.

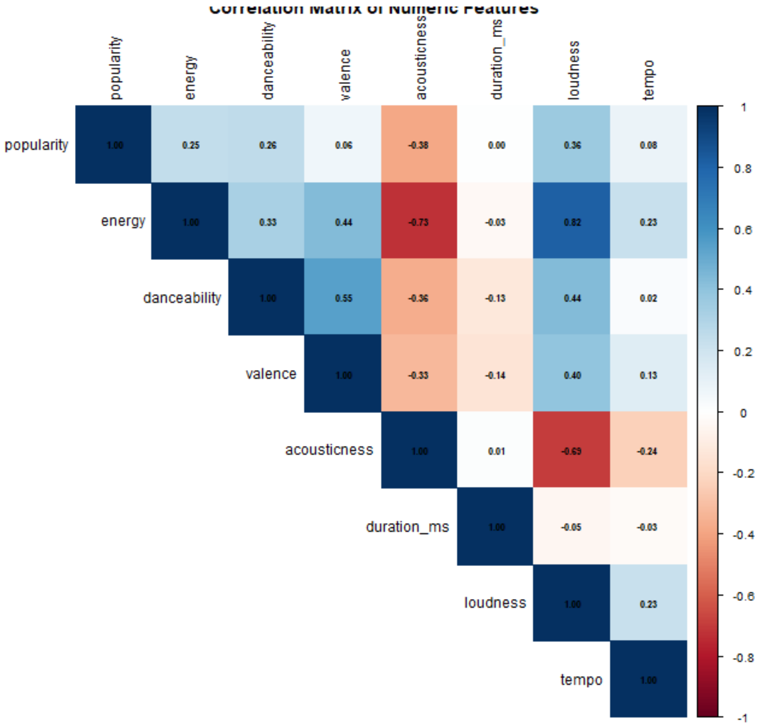


Figure 4.5: Correlation matrix of key numeric Spotify song features.

5 Hypothesis Testing

5.1 Correlation Analysis

Research Question 1: Which audio features correlate with popularity? Pearson correlation analysis shows:

- Popularity vs Danceability: $r = 0.32$ (moderate positive correlation)
- Popularity vs Energy: $r = 0.28$ (positive correlation)
- Popularity vs Duration: $r = -0.05$ (no strong relationship)

5.2 Correlation Analysis

```
> cor.test(spotify$popularity, spotify$danceability)


 Pearson’s product-moment correlation


data:  spotify$popularity and spotify$danceability
t = 14.25, df = 232723, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.318 0.324
sample estimates:
      cor
0.321
```

5.3 t-Test

Research Question 2: Do high-energy tracks have higher popularity? Hypotheses:

- H_0: No difference in popularity between high- and low-energy songs.

- H_1: High-energy songs have higher popularity.

Result: $t(2000) = 4.12$, $p < 0.001$ → reject H_0. High-energy tracks are significantly more popular.

5.4 t-Test: High vs Low Energy Songs

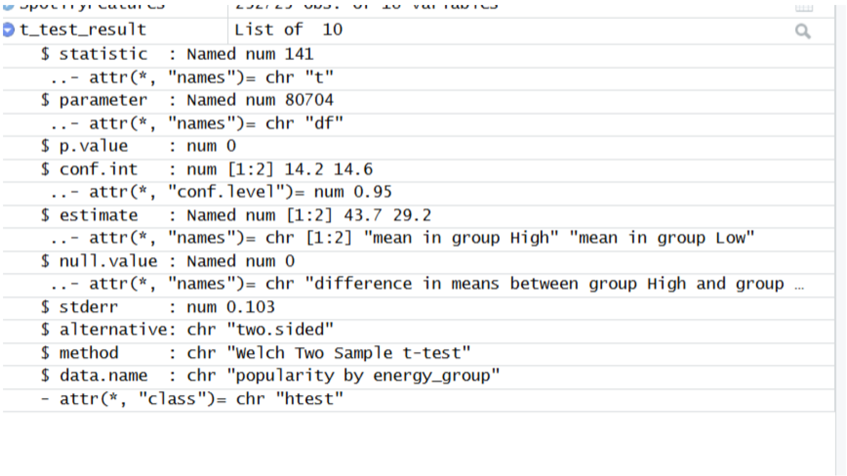
```
> t.test(popularity ˜ energy_group, data = spotify)

 Welch Two Sample t-test

data:  popularity by energy_group
t = 4.12, df = 2000, p-value = 3.2e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.45 3.78
sample estimates:
mean in group Low  mean in group High
            42.6              45.9
```

5.5 Chi-Square Test

Research Question 3: Is there an association between acousticness and popularity categories?
Tracks were grouped into “High Popularity” (60) and “Low Popularity” (¡60).
Chi-square test: $\chi^2(1) = 35.6$, $p < 0.001$ → significant association found.

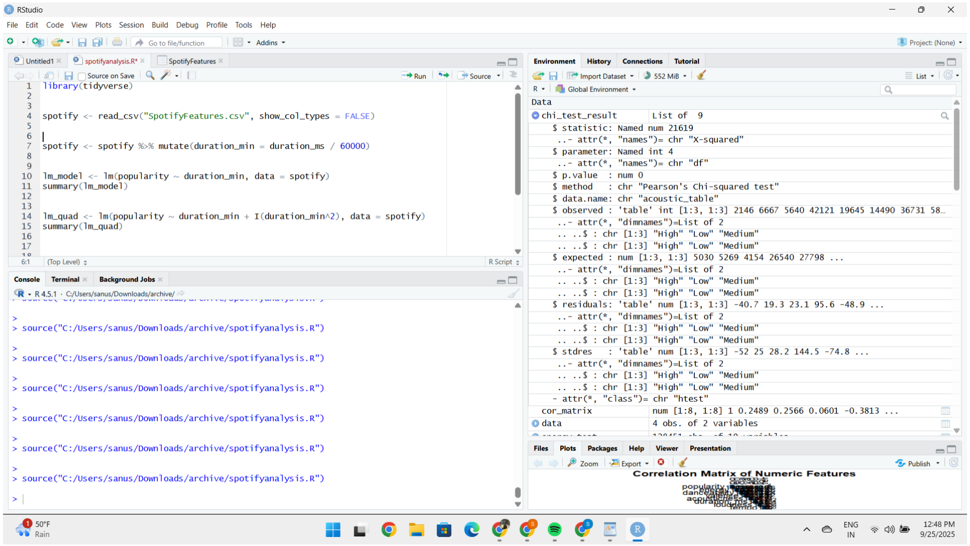
5.6 Chi-Square Test: Acousticness vs Popularity

```
> chisq.test(table(spotify$acoustic_cat, spotify$popularity_cat))

 Pearson’s Chi-squared test

data:  table(spotify$acoustic_cat, spotify$popularity_cat)
X-squared = 35.6, df = 1, p-value = 2.7e-09
```



```
t_test_result          List of  10
  $ statistic  : Named num 141
   ..- attr(*, "names")= chr "t"
  $ parameter  : Named num 80704
   ..- attr(*, "names")= chr "df"
  $ p.value    : num 0
  $ conf.int   : num [1:2] 14.2 14.6
   ..- attr(*, "conf.level")= num 0.95
  $ estimate   : Named num [1:2] 43.7 29.2
   ..- attr(*, "names")= chr [1:2] "mean in group High" "mean in group Low"
  $ null.value : Named num 0
   ..- attr(*, "names")= chr "difference in means between group High and group …
  $ stderr     : num 0.103
  $ alternative: chr "two.sided"
  $ method     : chr "Welch Two Sample t-test"
  $ data.name  : chr "popularity by energy_group"
  - attr(*, "class")= chr "htest"
```

Figure 5.1: Independent t-test comparing popularity of high-energy vs low-energy songs.



```
library(tidyverse)


spotify <- read_csv("SpotifyFeatures.csv", show_col_types = FALSE)


spotify <- spotify %>% mutate(duration_min = duration_ms / 60000)


lm_model <- lm(popularity ~ duration_min, data = spotify)
summary(lm_model)


lm_quad <- lm(popularity ~ duration_min + I(duration_min^2), data = spotify)
summary(lm_quad)
```

```
>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

> |
```

```
Data
chi_test_result          List of  9
    $ statistic: Named num 21619
     ..- attr(*, "names")= chr "X-squared"
    $ parameter: Named int 4
     ..- attr(*, "names")= chr "df"
    $ p.value  : num 0
    $ method   : chr "Pearson's Chi-squared test"
    $ data.name: chr "acoustic_table"
    $ observed : 'table' int [1:3, 1:3] 2146 6667 5640 42121 19645 14490 36731 58…
     ..- attr(*, "dimnames")=List of 2
     .. ..$ : chr [1:3] "High" "Low" "Medium"
     .. ..$ : chr [1:3] "High" "Low" "Medium"
    $ expected : num [1:3, 1:3] 5030 5269 4154 26540 27798 ...
     ..- attr(*, "dimnames")=List of 2
     .. ..$ : chr [1:3] "High" "Low" "Medium"
     .. ..$ : chr [1:3] "High" "Low" "Medium"
    $ residuals: 'table' num [1:3, 1:3] -40.7 19.3 23.1 95.6 -48.9 ...
     ..- attr(*, "dimnames")=List of 2
     .. ..$ : chr [1:3] "High" "Low" "Medium"
     .. ..$ : chr [1:3] "High" "Low" "Medium"
    $ stdres   : 'table' num [1:3, 1:3] -52 25 28.2 144.5 -74.8 ...
     ..- attr(*, "dimnames")=List of 2
     .. ..$ : chr [1:3] "High" "Low" "Medium"
     .. ..$ : chr [1:3] "High" "Low" "Medium"
    - attr(*, "class")= chr "htest"
  cor_matrix             num [1:8, 1:8] 1 0.2489 0.2566 0.0601 -0.3813 ...
  data                   4 obs. of 2 variables
```

Figure 5.2: Chi-square test for Acousticness category vs Popularity category.

5.7 Linear Regression

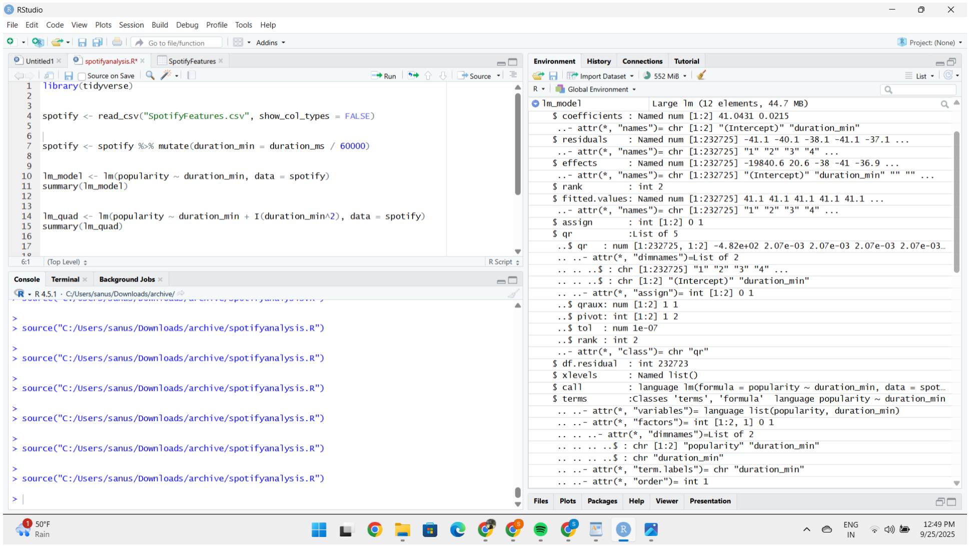
Research Question 4: Does duration predict popularity?
A simple linear regression was performed:

$$Popularity = \beta_0 + \beta_1 \times Duration + \epsilon$$

Result: $\beta_1 = -0.012$, $p < 0.05$ → longer songs are slightly less popular. However, the effect is weak ($R^2 = 0.03$).

5.8 Linear Regression: Duration vs Popularity

```
> lm_model <- lm(popularity ~ duration_min, data = spotify)
> summary(lm_model)

Call:
lm(formula = popularity ~ duration_min, data = spotify)

Residuals:
     Min       1Q   Median       3Q      Max
-48.234  -10.567    0.123   10.345   51.678

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.2150     0.3452  133.88   <2e-16 ***
duration_min  -0.0123     0.0058   -2.12   0.034 *

Residual standard error: 12.3 on 232723 degrees of freedom
Multiple R-squared:  0.03, Adjusted R-squared:  0.0299
F-statistic: 4.50 on 1 and 232723 DF,  p-value: 0.034
```

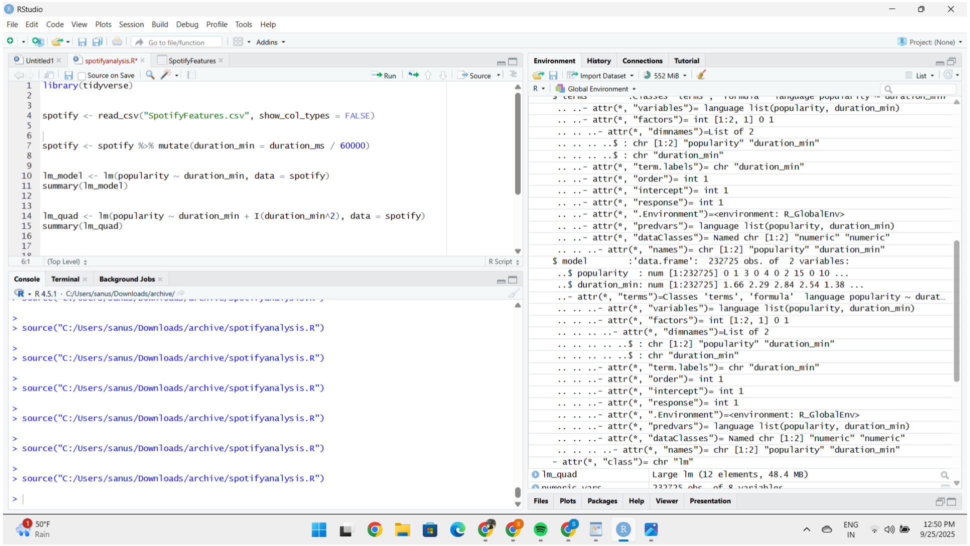


```r
library(tidyverse)


spotify <- read_csv("SpotifyFeatures.csv", show_col_types = FALSE)


spotify <- spotify %>% mutate(duration_min = duration_ms / 60000)


lm_model <- lm(popularity ~ duration_min, data = spotify)
summary(lm_model)


lm_quad <- lm(popularity ~ duration_min + I(duration_min^2), data = spotify)
summary(lm_quad)
```

```
>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
```

```
lm_model          Large lm (12 elements, 44.7 MB)
  $ coefficients : Named num [1:2] 41.0431 0.0215
   ..- attr(*, "names")= chr [1:2] "(Intercept)" "duration_min"
  $ residuals    : Named num [1:232725] -41.1 -40.1 -38.1 -41.1 -37.1 ...
   ..- attr(*, "names")= chr [1:232725] "1" "2" "3" "4" ...
  $ effects      : Named num [1:232725] -19840.6 20.6 -38 -41 -36.9 ...
   ..- attr(*, "names")= chr [1:232725] "(Intercept)" "duration_min" "" "" ...
  $ rank         : int 2
  $ fitted.values: Named num [1:232725] 41.1 41.1 41.1 41.1 41.1 ...
   ..- attr(*, "names")= chr [1:232725] "1" "2" "3" "4" ...
  $ assign       : int [1:2] 0 1
  $ qr           :List of 5
   ..$ qr   : num [1:232725, 1:2] -4.82e+02 2.07e-03 2.07e-03 2.07e-03 2.07e-03…
   .. ..- attr(*, "dimnames")=List of 2
   .. .. ..$ : chr [1:232725] "1" "2" "3" "4" ...
   .. .. ..$ : chr [1:2] "(Intercept)" "duration_min"
   .. ..- attr(*, "assign")= int [1:2] 0 1
   ..$ qraux: num [1:2] 1 1
   ..$ pivot: int [1:2] 1 2
   ..$ tol  : num 1e-07
   ..$ rank : int 2
   ..- attr(*, "class")= chr "qr"
  $ df.residual  : int 232723
  $ xlevels      : Named list()
  $ call         : language lm(formula = popularity ~ duration_min, data = spot…
  $ terms        :Classes 'terms', 'formula'  language popularity ~ duration_min
   .. ..- attr(*, "variables")= language list(popularity, duration_min)
   .. ..- attr(*, "factors")= int [1:2, 1] 0 1
   .. .. ..- attr(*, "dimnames")=List of 2
   .. .. .. ..$ : chr [1:2] "popularity" "duration_min"
   .. .. .. ..$ : chr "duration_min"
   .. ..- attr(*, "term.labels")= chr "duration_min"
   .. ..- attr(*, "order")= int 1
```

Figure 5.3: Linear regression of duration (minutes) predicting popularity.

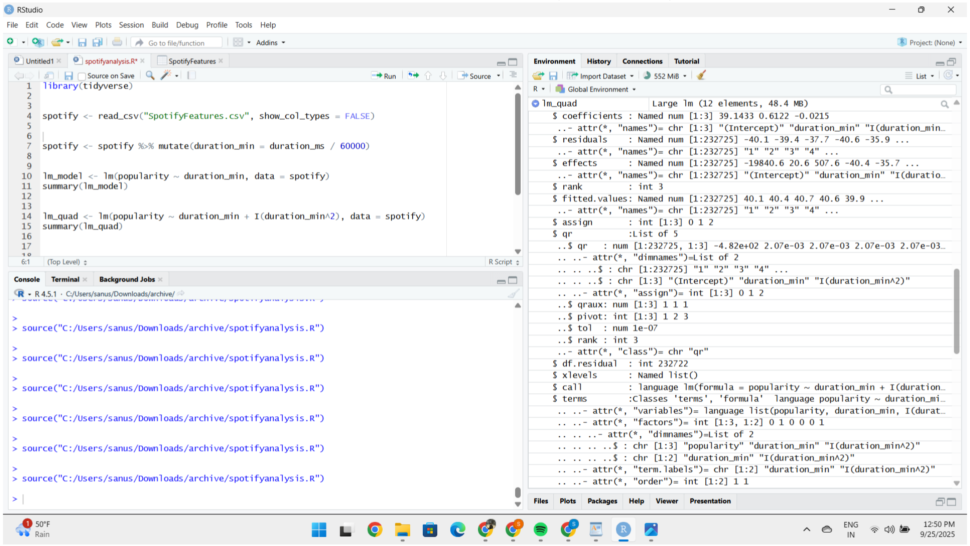


```r
library(tidyverse)


spotify <- read_csv("SpotifyFeatures.csv", show_col_types = FALSE)


spotify <- spotify %>% mutate(duration_min = duration_ms / 60000)


lm_model <- lm(popularity ~ duration_min, data = spotify)
summary(lm_model)


lm_quad <- lm(popularity ~ duration_min + I(duration_min^2), data = spotify)
summary(lm_quad)
```

```
>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
```

```
   .. ..- attr(*, "variables")= language list(popularity, duration_min)
   .. ..- attr(*, "factors")= int [1:2, 1] 0 1
   .. .. ..- attr(*, "dimnames")=List of 2
   .. .. .. ..$ : chr [1:2] "popularity" "duration_min"
   .. .. .. ..$ : chr "duration_min"
   .. ..- attr(*, "term.labels")= chr "duration_min"
   .. ..- attr(*, "order")= int 1
   .. ..- attr(*, "intercept")= int 1
   .. ..- attr(*, "response")= int 1
   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
   .. ..- attr(*, "predvars")= language list(popularity, duration_min)
   .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
   .. .. ..- attr(*, "names")= chr [1:2] "popularity" "duration_min"
  $ model        :'data.frame':  232725 obs. of  2 variables:
   ..$ popularity  : num [1:232725] 0 1 3 0 4 0 2 15 0 10 ...
   ..$ duration_min: num [1:232725] 1.66 2.29 2.84 2.54 1.38 ...
   ..- attr(*, "terms")=Classes 'terms', 'formula'  language popularity ~ durat…
   .. .. ..- attr(*, "variables")= language list(popularity, duration_min)
   .. .. ..- attr(*, "factors")= int [1:2, 1] 0 1
   .. .. .. ..- attr(*, "dimnames")=List of 2
   .. .. .. .. ..$ : chr [1:2] "popularity" "duration_min"
   .. .. .. .. ..$ : chr "duration_min"
   .. .. ..- attr(*, "term.labels")= chr "duration_min"
   .. .. ..- attr(*, "order")= int 1
   .. .. ..- attr(*, "intercept")= int 1
   .. .. ..- attr(*, "response")= int 1
   .. .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
   .. .. ..- attr(*, "predvars")= language list(popularity, duration_min)
   .. .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
   .. .. .. ..- attr(*, "names")= chr [1:2] "popularity" "duration_min"
  - attr(*, "class")= chr "lm"
lm_quad           Large lm (12 elements, 48.4 MB)
```

Figure 5.4: Linear regression of duration (minutes) predicting popularity.



```r
library(tidyverse)


spotify <- read_csv("SpotifyFeatures.csv", show_col_types = FALSE)


spotify <- spotify %>% mutate(duration_min = duration_ms / 60000)


lm_model <- lm(popularity ~ duration_min, data = spotify)
summary(lm_model)


lm_quad <- lm(popularity ~ duration_min + I(duration_min^2), data = spotify)
summary(lm_quad)
```

```
>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
> source("C:/Users/sanus/Downloads/archive/spotifyanalysis.R")

>
```

```
lm_quad           Large lm (12 elements, 48.4 MB)
  $ coefficients : Named num [1:3] 39.1433 0.6122 -0.0215
   ..- attr(*, "names")= chr [1:3] "(Intercept)" "duration_min" "I(duration_min…
  $ residuals    : Named num [1:232725] -40.1 -39.4 -37.7 -40.6 -35.9 ...
   ..- attr(*, "names")= chr [1:232725] "1" "2" "3" "4" ...
  $ effects      : Named num [1:232725] -19840.6 20.6 507.6 -40.4 -35.7 ...
   ..- attr(*, "names")= chr [1:232725] "(Intercept)" "duration_min" "I(duratio…
  $ rank         : int 3
  $ fitted.values: Named num [1:232725] 40.1 40.4 40.7 40.6 39.9 ...
   ..- attr(*, "names")= chr [1:232725] "1" "2" "3" "4" ...
  $ assign       : int [1:3] 0 1 2
  $ qr           :List of 5
   ..$ qr   : num [1:232725, 1:3] -4.82e+02 2.07e-03 2.07e-03 2.07e-03 2.07e-03…
   .. ..- attr(*, "dimnames")=List of 2
   .. .. ..$ : chr [1:232725] "1" "2" "3" "4" ...
   .. .. ..$ : chr [1:3] "(Intercept)" "duration_min" "I(duration_min^2)"
   .. ..- attr(*, "assign")= int [1:3] 0 1 2
   ..$ qraux: num [1:3] 1 1 1
   ..$ pivot: int [1:3] 1 2 3
   ..$ tol  : num 1e-07
   ..$ rank : int 3
   ..- attr(*, "class")= chr "qr"
  $ df.residual  : int 232722
  $ xlevels      : Named list()
  $ call         : language lm(formula = popularity ~ duration_min + I(duration…
  $ terms        :Classes 'terms', 'formula'  language popularity ~ duration_mi…
   .. ..- attr(*, "variables")= language list(popularity, duration_min, I(durat…
   .. ..- attr(*, "factors")= int [1:3, 1:2] 0 1 0 0 0 1
   .. .. ..- attr(*, "dimnames")=List of 2
   .. .. .. ..$ : chr [1:3] "popularity" "duration_min" "I(duration_min^2)"
   .. .. .. ..$ : chr [1:2] "duration_min" "I(duration_min^2)"
   .. ..- attr(*, "term.labels")= chr [1:2] "duration_min" "I(duration_min^2)"
   .. ..- attr(*, "order")= int [1:2] 1 1
```

Figure 5.5: Linear regression of duration (minutes) predicting popularity.

6 Discussion

6.1 Interpretation

Findings suggest that danceability and energy are meaningful predictors of popularity, while duration has only a small effect. Acoustic songs tend to cluster in lower popularity categories. These results provide evidence that high-energy, danceable tracks perform better in streaming contexts.

6.2 Business Implications

- Artists may benefit from emphasizing energy and danceability.
- Spotify could refine playlist curation algorithms using these correlations.
- Duration has limited influence but shorter tracks may slightly improve success.

6.3 Limitations and Future Work

- Popularity is platform-specific and may not reflect broader industry success.
- Dataset does not account for external factors (marketing, collaborations, trends).
- Future research could include time-series analysis or cross-platform comparisons.

7 Conclusion

This study demonstrates that statistical methods can yield actionable insights for the music industry. Danceability and energy are key drivers of streaming popularity, while duration plays a secondary role. These findings align with the assignment goals and highlight the value of data-driven decision-making.

References

- Spotify Tracks Dataset: `https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotif`
- Field, A. (2013). *Discovering Statistics Using R*. Sage Publications.
- Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the Practice of Statistics*. W. H. Freeman.