STAT 184 - Preliminary EDA Project Topic - Spotify Music Data Analysis using R

Team - Apoorv Thite & Aarnav Putta

04/17/2024

```
library(dplyr)
library(tidyverse)
library(ggplot2)
```

Introduction

As part of our coursework, We decided to work on an exploratory data analysis project using a dataset titled SpotifyData.csv, which comprises information about various music tracks on Spotify. Our aim is to delve into the intricacies of this dataset to extract meaningful insights and patterns that can inform further research.

Main Goal / Guiding Research Question

The primary question guiding our analysis is: "How do the characteristics of music (like tempo, valence, and energy) vary by genre, and also influence a track's popularity on Spotify?" This question will help us understand the trends and preferences in music consumption over a specific period and across different musical genres.

Where did you find them?

We found our Data sources on Datacamp and on Kaggle. These are the two platforms where we found our datasets.

Who collected/maintains them?

Apoorv found the dataset, Aarnav is responsible for maintaining the dataset and keeping a track of the variables and cases in the dataset

When & Why were they originally collected?

The data was collected by the motive of finding meaningful insights about the spotify music that's trending these days and historically, the genre of those songs and additional details about it. Apoorv collected the data from the above mentioned sources and we started analyzing its features.

What does a case represent in each data source, and how many total cases are available?

Each case in this dataset refers to a specific song with a unique score for every attribute or variable.Each case essentially is classifying a song based on its unique score derived from the combination of variables we have in the datset. With the variables being artist, top genre, year,bpm, nrg, dnce,dB,live,val,dur,acous,spch,popularity. There are totally 114000 cases in this dataset.

```
spotify_data <- read.csv(file = "SpotifyData.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)
nrow(spotify_data)
```

```
## [1] 114000
```

What are some of the variables that you plan to use ?

We plan to use variables such as energy, danceability, tempo, speechiness to identify correlation of these various genres to the popularity during a specific timespan , in order to identify the trends and changes in people's music taste of a desired time period. In addition to that variables like loudness emphasize the importance of factors such as volume in people's music preference.

```
names(spotify_data)
```

```
##  [1] "X"                "track_id"         "artists"          "album_name"
##  [5] "track_name"       "popularity"       "duration_ms"      "explicit"
##  [9] "danceability"     "energy"           "key"              "loudness"
## [13] "mode"             "speechiness"      "acousticness"     "instrumentalness"
## [17] "liveness"         "valence"          "tempo"            "time_signature"
## [21] "track_genre"
```

Explore intuition related to the research question

In this part we perform couple simple plots to derive a correlation between the various variables. This is just a small sample size of the analysis to be done for the final project.

Loading our data:

```
spotify_data <- read.csv(file = "SpotifyData.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)
nrow(spotify_data)
```

```
## [1] 114000
```

Short summary and logistics:

```
summary(spotify_data)
```

```
##        X            track_id           artists           album_name
##  Min.   :     0   Length:114000      Length:114000      Length:114000
##  1st Qu.: 28500   Class :character   Class :character   Class :character
##  Median : 57000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 57000
##  3rd Qu.: 85499
##  Max.   :113999
##   track_name          popularity      duration_ms        explicit
##  Length:114000      Min.   :  0.00   Min.   :      0   Length:114000
##  Class :character   1st Qu.: 17.00   1st Qu.: 174066   Class :character
##  Mode  :character   Median : 35.00   Median : 212906   Mode  :character
##                     Mean   : 33.24   Mean   : 228029
##                     3rd Qu.: 50.00   3rd Qu.: 261506
##                     Max.   :100.00   Max.   :5237295
##   danceability        energy            key            loudness
##  Min.   :0.0000   Min.   :0.0000   Min.   : 0.000   Min.   :-49.531
##  1st Qu.:0.4560   1st Qu.:0.4720   1st Qu.: 2.000   1st Qu.:-10.013
##  Median :0.5800   Median :0.6850   Median : 5.000   Median : -7.004
##  Mean   :0.5668   Mean   :0.6414   Mean   : 5.309   Mean   : -8.259
##  3rd Qu.:0.6950   3rd Qu.:0.8540   3rd Qu.: 8.000   3rd Qu.: -5.003
##  Max.   :0.9850   Max.   :1.0000   Max.   :11.000   Max.   :  4.532
```

```
##       mode         speechiness       acousticness    instrumentalness
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00e+00
##  1st Qu.:0.0000   1st Qu.:0.03590   1st Qu.:0.0169   1st Qu.:0.00e+00
##  Median :1.0000   Median :0.04890   Median :0.1690   Median :4.16e-05
##  Mean   :0.6376   Mean   :0.08465   Mean   :0.3149   Mean   :1.56e-01
##  3rd Qu.:1.0000   3rd Qu.:0.08450   3rd Qu.:0.5980   3rd Qu.:4.90e-02
##  Max.   :1.0000   Max.   :0.96500   Max.   :0.9960   Max.   :1.00e+00
##     liveness         valence           tempo        time_signature
##  Min.   :0.0000   Min.   :0.0000   Min.   :  0.00   Min.   :0.000
##  1st Qu.:0.0980   1st Qu.:0.2600   1st Qu.: 99.22   1st Qu.:4.000
##  Median :0.1320   Median :0.4640   Median :122.02   Median :4.000
##  Mean   :0.2136   Mean   :0.4741   Mean   :122.15   Mean   :3.904
##  3rd Qu.:0.2730   3rd Qu.:0.6830   3rd Qu.:140.07   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :0.9950   Max.   :243.37   Max.   :5.000
##  track_genre
##  Length:114000
##  Class :character
##  Mode  :character
##
##
##
```

Number of cases

```
num_cases <- nrow(spotify_data)
num_cases
```

```
## [1] 114000
```

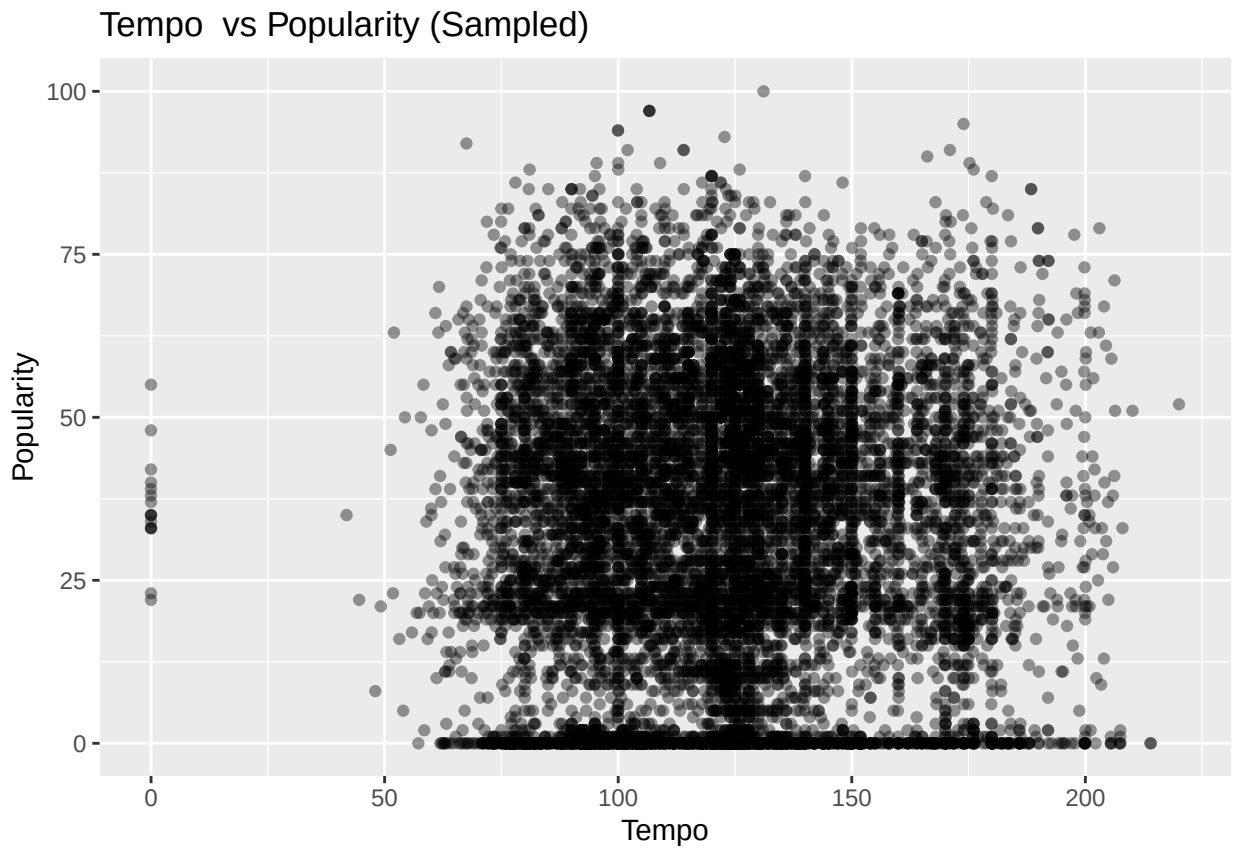
(A) informative plot (3 infomrative plots)

Noticing that our dataset has so many cases to establish a relationship between the variables we decided to use a method of sampling to give us an approximate that we can map onto the entire dataset.

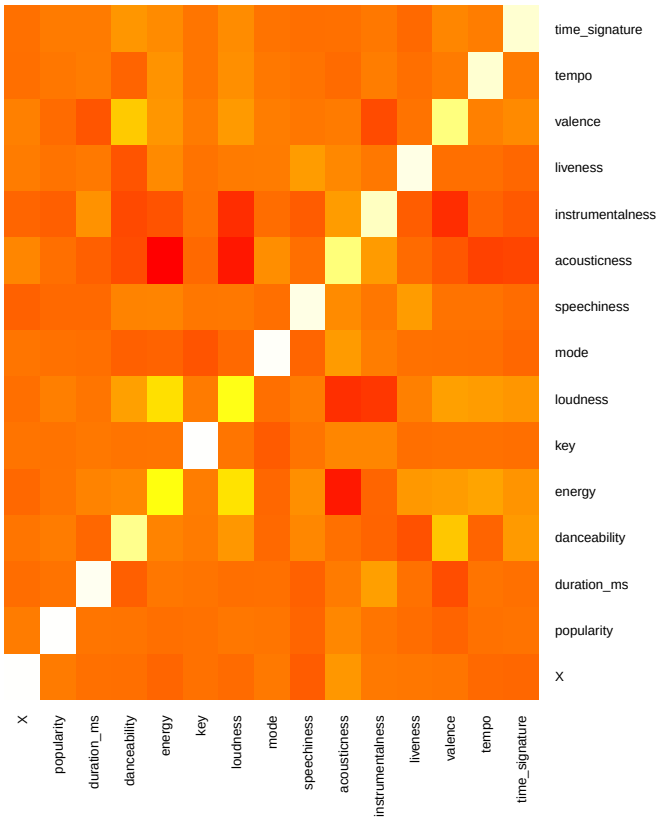
```
spotify_data <- read.csv(file = "SpotifyData.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)
nrow(spotify_data)
```

```
## [1] 114000
```

```
# sample 10% of the data as an approximate
sampled_data <- sample_frac(spotify_data, 0.1)

ggplot(sampled_data, aes(x = tempo, y = popularity)) +
  geom_point(alpha=0.4) +
  labs(title = "Tempo  vs Popularity (Sampled)", x = "Tempo", y = "Popularity")
```

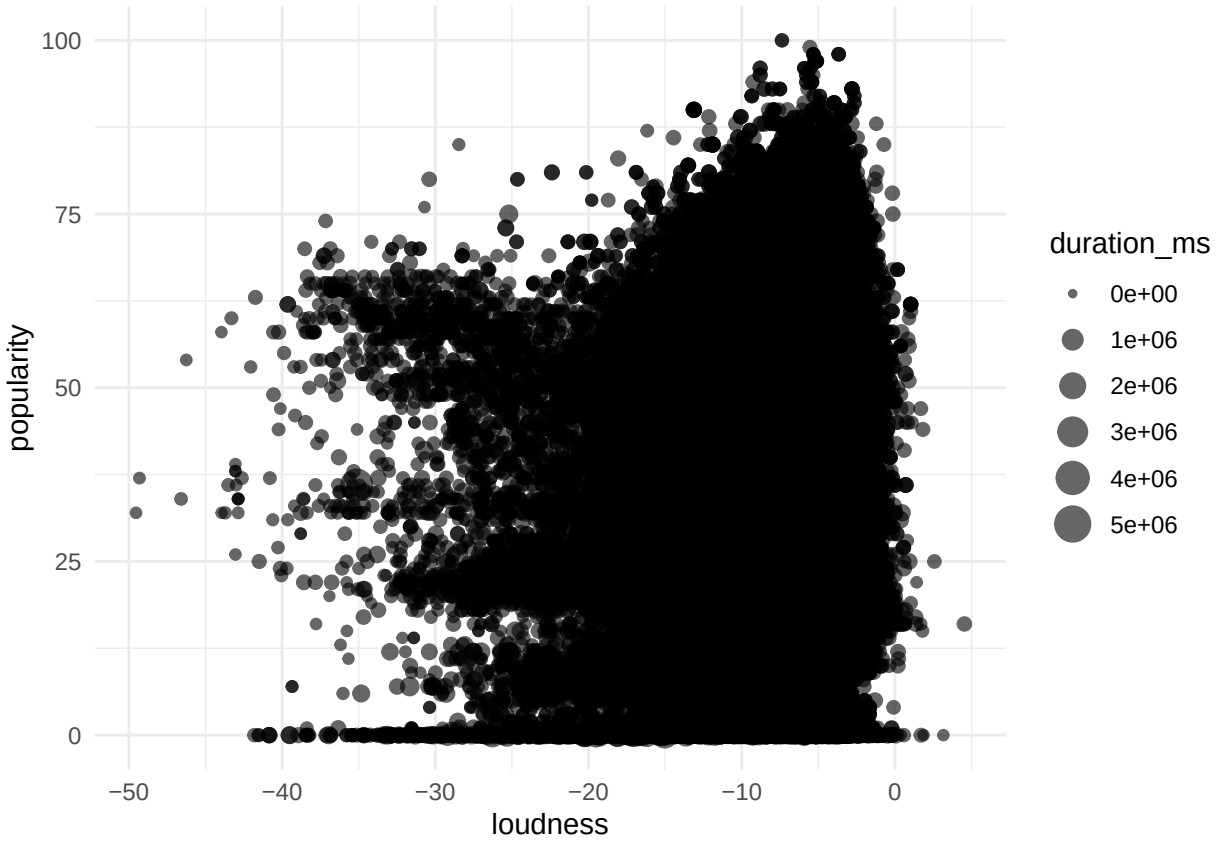



```r
spotify_data <- read.csv("SpotifyData.csv", stringsAsFactors = FALSE)

correlation_matrix <- cor(spotify_data[sapply(spotify_data, is.numeric)])


heatmap(correlation_matrix, Rowv = NA, Colv = NA, col = heat.colors(256), scale="column",
        margins = c(5,10), cexRow = 0.5, cexCol = 0.5)
```



We try to correlate popularity to loudness

```
ggplot(spotify_data, aes(x = loudness, y = popularity, size = duration_ms)) +
  geom_point(alpha = 0.6) +
  theme_minimal()
```



(B) preliminary observations/intuition about RQ

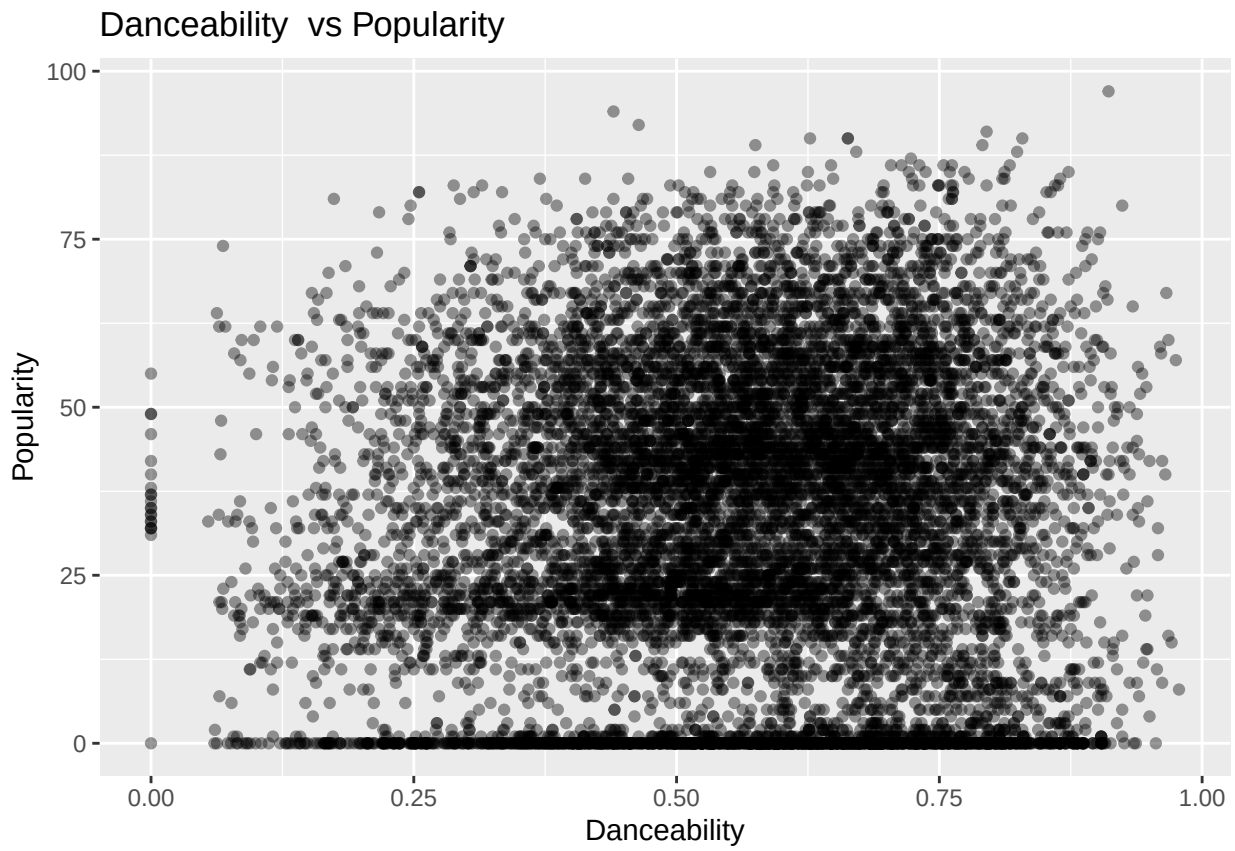
Preliminary Observations:

A couple of preliminary observations I have observed were the correlation of popularity vs danceability showing how the most popular songs on spotify seem to have the highest danceability score.This shows the evolution of the current taste of music b, requiring the music to be mopre danceable to be liked.

```
spotify_data <- read.csv(file = "SpotifyData.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)
nrow(spotify_data)
```

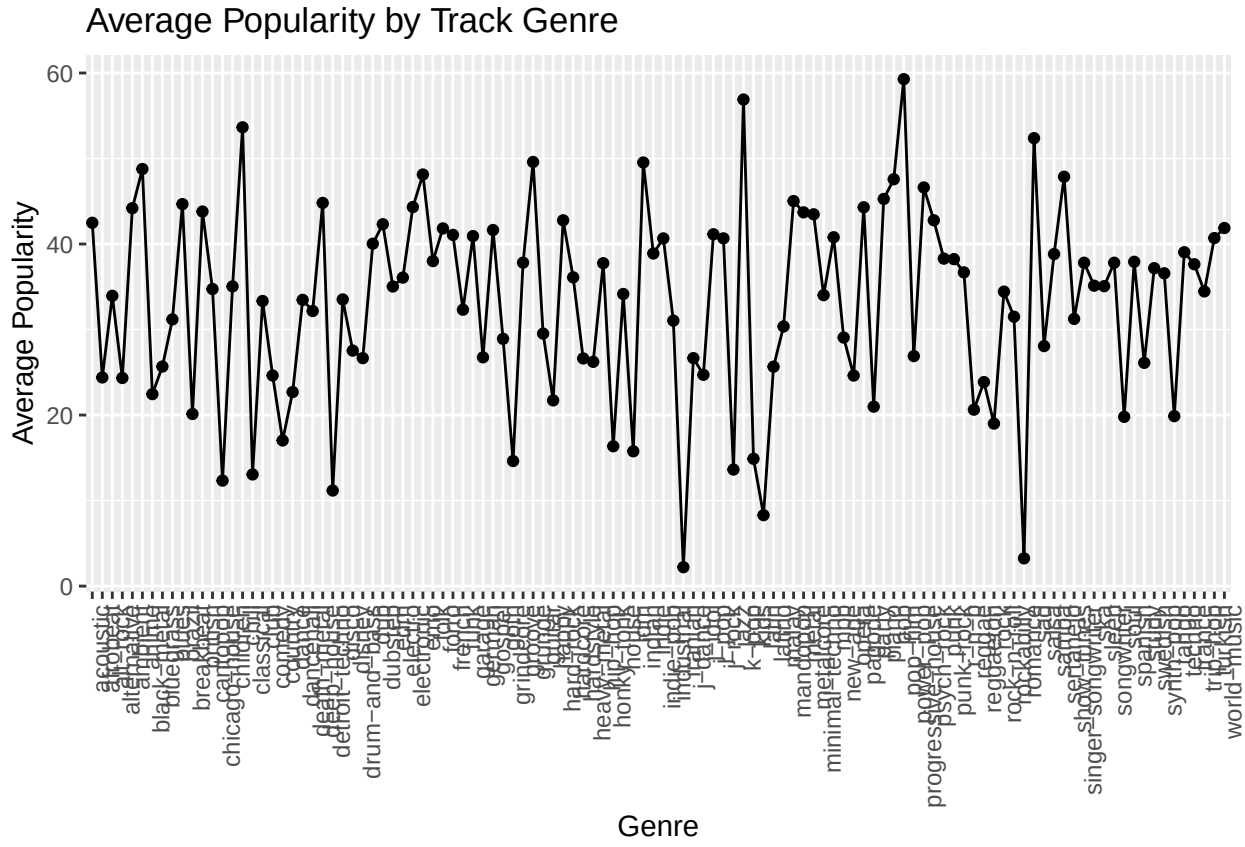
```
## [1] 114000
```

```
sampled_data <- sample_frac(spotify_data, 0.1)

ggplot(sampled_data, aes(x = danceability, y = popularity)) +
  geom_point(alpha=0.4) +
  labs(title = "Danceability  vs Popularity", x = "Danceability", y = "Popularity")
```



Another Preliminary Observation I noticed was high average popularity for jazz.

```
average_popularity_genre <- spotify_data %>%
  group_by(track_genre) %>%
  summarize(average_popularity = mean(popularity, na.rm = TRUE)) %>%
  arrange(desc(average_popularity))

ggplot(average_popularity_genre, aes(x = track_genre, y = average_popularity, group = 1)) +
  geom_line() +
  geom_point() +  # Adding points for clarity
  labs(title = "Average Popularity by Track Genre",
       x = "Genre",
       y = "Average Popularity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate genre labels for readability
```



My intuition is that we can start of by seeing initially that old genres such as jazz, and ragtime would be prevalent. These genres have low loundness, energy, danceability , however we would notice a pickup as we approach a time period closer to now.

(C) Plan for Addressing Your Research Question

Data Cleaning and Preparation:

We want to handle missing values, outliers, and normalize data if necessary to ensure quality analysis.

Exploratory Data Analysis (EDA):

We also want to implement the plots we used in part A to understand data distributions and preliminary correlations.

Statistical Analysis:

Depending on the data distribution and initial findings, we want to apply appropriate statistical tests or regression analysis to quantify relationships between features and popularity.

Model Building:

If predicting popularity is of interest, consider building a predictive model using regression or or other packages we learnt in class based on identified significant features.

Interpretation and Conclusion:

Draw conclusions from the analyses and visualizations, interpreting how music characteristics influence popularity and vary by genre. We also consider the implications for music producers and marketers.

TWO Data Sources (at least)

(A) Primary data must not be loaded from an R package & includes mix of variable types & has enough cases

link of the primary dataset: https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

```
# Load the SpotifyMusic dataset
spotify_data <- read.csv(file = "SpotifyData.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)
head(spotify_data)
```

```
##   X               track_id                artists
## 1 0 5SuOikwiRyPMVoIQDJUgSV             Gen Hoshino
## 2 1 4qPNDBW1i3p13qLCt0Ki3A            Ben Woodward
## 3 2 1iJBSr7s7jYXzM8EGcbK5b Ingrid Michaelson;ZAYN
## 4 3 6lfxq3CG4xtTiEg7opyCyx            Kina Grannis
## 5 4 5vjLSffimiIP26QG5WcN2K       Chord Overstreet
## 6 5 01MVOl9KtVTNfFiBU9I7dc            Tyrone Wells
##                                           album_name
## 1                                             Comedy
## 2                                  Ghost (Acoustic)
## 3                                    To Begin Again
## 4 Crazy Rich Asians (Original Motion Picture Soundtrack)
## 5                                             Hold On
## 6                                Days I Will Remember
##                   track_name popularity duration_ms explicit danceability
## 1                     Comedy         73      230666    False        0.676
## 2          Ghost - Acoustic         55      149610    False        0.420
## 3            To Begin Again         57      210826    False        0.438
## 4 Can't Help Falling In Love         71      201933    False        0.266
## 5                   Hold On         82      198853    False        0.618
## 6       Days I Will Remember         58      214240    False        0.688
##   energy key loudness mode speechiness acousticness instrumentalness liveness
## 1 0.4610   1   -6.746    0      0.1430       0.0322         1.01e-06   0.3580
## 2 0.1660   1  -17.235    1      0.0763       0.9240         5.56e-06   0.1010
## 3 0.3590   0   -9.734    1      0.0557       0.2100         0.00e+00   0.1170
## 4 0.0596   0  -18.515    1      0.0363       0.9050         7.07e-05   0.1320
## 5 0.4430   2   -9.681    1      0.0526       0.4690         0.00e+00   0.0829
## 6 0.4810   6   -8.807    1      0.1050       0.2890         0.00e+00   0.1890
##   valence   tempo time_signature track_genre
## 1   0.715  87.917              4    acoustic
## 2   0.267  77.489              4    acoustic
## 3   0.120  76.332              4    acoustic
## 4   0.143 181.740              3    acoustic
## 5   0.167 119.949              4    acoustic
## 6   0.666  98.017              4    acoustic
```

```
nrow(spotify_data)
```

```
## [1] 114000
```

```
names(spotify_data)
```

```
##  [1] "X"                "track_id"         "artists"          "album_name"
##  [5] "track_name"       "popularity"       "duration_ms"      "explicit"
##  [9] "danceability"     "energy"           "key"              "loudness"
## [13] "mode"             "speechiness"      "acousticness"     "instrumentalness"
## [17] "liveness"         "valence"          "tempo"            "time_signature"
## [21] "track_genre"
```

(B) at least one other data source has been identified and examined

This additional dataset aids us in comparing with the main dataset on aspects like track popularity or audio features. This might reveal trends or discrepancies between the datasets, like differences in popularity.By combining data from both datasets, we can enhance the analysis, by observing the popularity or characteristics of tracks that appear in both datasets.This datasetalso gives us the much needed flexibility to establish the correlation of these characteristics with popularity over time.

```
# Load the SpotifyMusic dataset
spotify_data <- read.csv(file = "top50.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)
head(spotify_data)
```

```
##                  title        artist the.genre.of.the.track year
## 1     Hey, Soul Sister         Train             neo mellow 2010
## 2 Love The Way You Lie        Eminem        detroit hip hop 2010
## 3              TiK ToK         Kesha              dance pop 2010
## 4          Bad Romance     Lady Gaga              dance pop 2010
## 5 Just the Way You Are    Bruno Mars                   pop 2010
## 6                 Baby Justin Bieber           canadian pop 2010
##   Beats.Per.Minute..The.tempo.of.the.song
## 1                                      97
## 2                                      87
## 3                                     120
## 4                                     119
## 5                                     109
## 6                                      65
##   Energy..The.energy.of.a.song...the.higher.the.value..the.more.energtic
## 1                                                                     89
## 2                                                                     93
## 3                                                                     84
## 4                                                                     92
## 5                                                                     84
## 6                                                                     86
##   Danceability...The.higher.the.value..the.easier.it.is.to.dance.to.this.song
## 1                                                                          67
## 2                                                                          75
## 3                                                                          76
## 4                                                                          70
## 5                                                                          64
## 6                                                                          73
##   Loudness.dB...The.higher.the.value..the.louder.the.song
## 1                                                      -4
## 2                                                      -5
## 3                                                      -3
## 4                                                      -4
## 5                                                      -5
## 6                                                      -5
##   Liveness...The.higher.the.value..the.more.likely.the.song.is.a.live.recording
## 1                                                                             8
## 2                                                                            52
## 3                                                                            29
## 4                                                                             8
## 5                                                                             9
## 6                                                                            11
##   Valence...The.higher.the.value..the.more.positive.mood.for.the.song
## 1                                                                  80
## 2                                                                  64
## 3                                                                  71
## 4                                                                  71
## 5                                                                  43
## 6                                                                  54
```

```
##   Length...The.duration.of.the.song
## 1                               217
## 2                               263
## 3                               200
## 4                               295
## 5                               221
## 6                               214
##   Acousticness...The.higher.the.value.the.more.acoustic.the.song.is
## 1                                                                19
## 2                                                                24
## 3                                                                10
## 4                                                                 0
## 5                                                                 2
## 6                                                                 4
##   Speechiness...The.higher.the.value.the.more.spoken.word.the.song.contains
## 1                                                                         4
## 2                                                                        23
## 3                                                                        14
## 4                                                                         4
## 5                                                                         4
## 6                                                                        14
##   Popularity..The.higher.the.value.the.more.popular.the.song.is
## 1                                                            83
## 2                                                            82
## 3                                                            80
## 4                                                            79
## 5                                                            78
## 6                                                            77
```

Code Appendix

```
# This template file is based off of a template created by Alex Hayes
# https://github.com/alexpghayes/rmarkdown_homework_template

# Setting Document Options
knitr::opts_chunk$set(
  echo = TRUE,
  warning = FALSE,
  message = FALSE,
  fig.align = "center"
)
library(dplyr)
library(tidyverse)
library(ggplot2)


spotify_data <- read.csv(file = "SpotifyData.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)
nrow(spotify_data)
names(spotify_data)
spotify_data <- read.csv(file = "SpotifyData.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)
nrow(spotify_data)

summary(spotify_data)
num_cases <- nrow(spotify_data)
num_cases
spotify_data <- read.csv(file = "SpotifyData.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)
nrow(spotify_data)


# sample 10% of the data as an approximate
sampled_data <- sample_frac(spotify_data, 0.1)

ggplot(sampled_data, aes(x = tempo, y = popularity)) +
  geom_point(alpha=0.4) +
  labs(title = "Tempo  vs Popularity (Sampled)", x = "Tempo", y = "Popularity")




spotify_data <- read.csv("SpotifyData.csv", stringsAsFactors = FALSE)

correlation_matrix <- cor(spotify_data[sapply(spotify_data, is.numeric)])


heatmap(correlation_matrix, Rowv = NA, Colv = NA, col = heat.colors(256), scale="column",
        margins = c(5,10), cexRow = 0.5, cexCol = 0.5)


ggplot(spotify_data, aes(x = loudness, y = popularity, size = duration_ms)) +
  geom_point(alpha = 0.6) +
  theme_minimal()

spotify_data <- read.csv(file = "SpotifyData.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)
nrow(spotify_data)
sampled_data <- sample_frac(spotify_data, 0.1)

ggplot(sampled_data, aes(x = danceability, y = popularity)) +
  geom_point(alpha=0.4) +
```

```r
  labs(title = "Danceability  vs Popularity", x = "Danceability", y = "Popularity")

average_popularity_genre <- spotify_data %>%
  group_by(track_genre) %>%
  summarize(average_popularity = mean(popularity, na.rm = TRUE)) %>%
  arrange(desc(average_popularity))

ggplot(average_popularity_genre, aes(x = track_genre, y = average_popularity, group = 1)) +
  geom_line() +
  geom_point() +  # Adding points for clarity
  labs(title = "Average Popularity by Track Genre",
       x = "Genre",
       y = "Average Popularity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate genre labels for readability

# Load the SpotifyMusic dataset
spotify_data <- read.csv(file = "SpotifyData.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)
head(spotify_data)

nrow(spotify_data)
names(spotify_data)
# Load the SpotifyMusic dataset
spotify_data <- read.csv(file = "top50.csv", header = TRUE, sep = ",", stringsAsFactors = FALSE)
head(spotify_data)
```