Analyzing Music Trends: Insights from top 5000 Albums

Ravivarman Devarajan

24-04-2024

Table of Contents

Project Summary:

The goal of this project is to demonstrate proficiency in data analysis techniques using R Studio and R Markdown by exploring a novel dataset and creating meaningful visualizations. The project will involve selecting a dataset, conducting exploratory data analysis and telling a story with different types of visualizations.

Introduction:

The Data set contains information about 4402 albums, it consists of 19 features from spotify. The data includes the following type of information:

- Basic information such as Artist name, album name, release date etc.
- Genres associated with the album and a textual description of the album.
- Average user rating and potentially the number of ratings/reviews each album has received.
- A range of audio features, these are numerical variables likely extracted from the audio itself, including acousticness, danceability, energy and many more.
- Additional Information such as duration of the album in milliseconds, time signature etc. This data can be used for various music- related analyses and visualizations which includes,
- Exploring music trends across decades, analyse how genres, average ratings, or audio features alike danceability or tempo have changed over time.
- Recommending music based on user preferences and audio feature similarities between albums.

- Compare audio features and user ratings between artists or across different genres.

This data set appears to be a valuable resource for exploring music information, artist styles, user preferences and potentially music trends across time. By analysing the various features, you can gain insights into listener preferences and the sonic characteristics of different genres and artists. However, The objective of my research is,

1. To find the typical ranges and distributions of some of the audio features (numerical variables) and analyse their patterns.
2. To find how the number of songs released per decade changed over time and analyse any noticeable trends or patterns in the distribution of music releases.
3. To select any one audio feature and describe how it has changed over the years and find any discernible trends or patterns in the relationship between that feature and the release year.

Variables of interest and the Techniques used:

The characteristics of variables associated with my research questions and their characteristics are listed below:

1. *Danceability:* A measure of a piece of music's suitability for dancing based on numerous musical components such as rhythm, tempo, and so on. Higher values reflect a song's danceability. Songs with high danceability scores typically include rhythmic patterns and beats that encourage movement and dancing.
2. *Average Rating (avg_rat):* It is the average score or rating assigned to a music by listeners or reviewers. It provides an overall measure of the song's perceived quality or enjoyment by its listeners. It varies according to the rating scale employed and the rating system's context (for example, user-generated ratings on a music system or professional reviews).
3. *Liveness:* It assesses the existence of live performance characteristics in a song, indicating whether it was recorded live or in the studio. It runs from 0 to 1, with higher numbers suggesting a stronger concentration of live performance elements. Songs with greater liveness scores may include audience reactions, crowd noise, or other elements that capture the mood of a live performance.
4. *Energy:* It is the perceptual measure of intensity and activity level in a piece of music. It captures characteristics such as loudness, dynamic range, and tempo that add to a song's overall sense of excitement and energy. Energy levels typically range from 0 to 1, with higher values suggesting music with more energetic or intense features.
5. *Release Date:* This is the date that a song or album was formally released to the public. It enables the investigation of trends and patterns across time. Release dates can be expressed in a variety of ways and divided into years, decades, or particular time periods for analysis.

The various Exploratory Data Analysis(EDA) techniques used in this research are

1. **Loading Packages and Data:**

R packages include a diverse set of functions and tools for data processing and visualization. Before beginning any analysis, you must install and load the relevant packages using the ‘install.packages()’ and ‘library()’ functions. EDA programs commonly used in R include tidyverse (which contains ggplot2, dplyr, tidyr, readr, and so on), lubridate, gridExtra, and rmarkdown.

Data is often kept in files like CSV, Excel, and databases. In R, data can be imported into the workspace using functions such as’read.csv()’ and’read.excel()’.

```r
spotify <- read.csv("D:/Lab/Top 5000 Albums _Spotify features.csv")
View(spotify) #reading the dataset from a csv file
```

2. **Data Inspection:**

Data inspection is used to determine the structure, content, and quality of a dataset. After loading the data, it is frequently useful to examine its structure, summary statistics, and the first few rows to gain a better grasp of its content. Some of the functions useful for dataset examination include ‘glimpse()’,‘summary()’, ‘head()’, ‘is.na()’, and so on. These procedures provide a brief overview of the dataset’s structure, including column names, missing values, data types, and the first few rows of data.

```r
glimpse(spotify)  #displays data types and variable names

## Rows: 4,402
## Columns: 20
## $ X                  <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,
15,…
## $ ars_name           <chr> "Radiohead", "Pink Floyd", "King Crimson",
"Radiohead…
## $ rel_date           <chr> "16-Jun-97", "12-Sep-75", "10-Oct-69", "03-Oct-
00", "…
## $ gens               <chr> "Alternative Rock, Art Rock", "Progressive Rock,
Art …
## $ descs              <chr> "melancholic, anxious, futuristic, alienation,
existe…
## $ avg_rat            <dbl> 4.23, 4.29, 4.30, 4.21, 4.27, 4.24, 4.20, 4.25,
4.23,…
## $ num_rat            <chr> "70,382", "48,662", "44,943", "58,590", "44,206",
"49…
## $ num_revs           <int> 1531, 983, 870, 734, 379, 1223, 1549, 961, 929,
721, …
## $ album              <chr> "OK Computer", "Wish You Were Here", "In the
Court of…
## $ acousticness       <dbl> 0.1357626, 0.6028000, 0.2976862, 0.1232186,
0.3228888…
## $ danceability       <dbl> 0.2880833, 0.3736000, 0.3406250, 0.6012941,
0.5907500…
```

```
## $ energy           <dbl> 0.5659167, 0.4098000, 0.3704750, 0.6767059,
0.7076250…
## $ instrumentalness <dbl> 0.161052767, 0.363040000, 0.327265000,
0.000669017, 0…
## $ liveness         <dbl> 0.15937500, 0.38494000, 0.14913750, 0.31141177,
0.318…
## $ loudness         <dbl> -9.102417, -12.689400, -14.873125, -7.811941, -
5.8020…
## $ speechiness      <dbl> 0.05630833, 0.03974000, 0.04346250, 0.26831765,
0.294…
## $ tempo            <dbl> 115.4507, 130.0188, 118.9206, 116.0451, 103.4446,
119…
## $ valence          <dbl> 0.2917333, 0.2590400, 0.2794000, 0.3982059,
0.4871875…
## $ duration_ms      <dbl> 268435.5, 530512.0, 507644.1, 325379.5, 296225.8,
265…
## $ time_signature   <dbl> 4.000000, 3.600000, 3.875000, 4.058824, 3.812500,
3.6…
```

```
which(is.na(spotify)) #returns the indices for missing values (NA)
```

```
## integer(0)
```

```
summary(spotify) #Provides overview of the structure and distribution
```

```
##        X            ars_name           rel_date             gens          
##  Min.   :   0   Length:4402        Length:4402        Length:4402       
##  1st Qu.:1100   Class :character   Class :character   Class :character  
##  Median :2200   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :2200                                                           
##  3rd Qu.:3301                                                           
##  Max.   :4401                                                           
##     descs              avg_rat        num_rat             num_revs      
##  Length:4402        Min.   :3.520   Length:4402        Min.   :   0.00  
##  Class :character   1st Qu.:3.700   Class :character   1st Qu.:  16.00  
##  Mode  :character   Median :3.750   Mode  :character   Median :  39.00  
##                     Mean   :3.771                      Mean   :  76.54  
##                     3rd Qu.:3.810                      3rd Qu.:  91.00  
##                     Max.   :4.340                      Max.   :1549.00  
##     album            acousticness        danceability        energy       
##  Length:4402        Min.   :0.0000007   Min.   :0.0749   Min.   :0.00236  
##  Class :character   1st Qu.:0.0917321   1st Qu.:0.3695   1st Qu.:0.42287  
##  Mode  :character   Median :0.2613944   Median :0.4739   Median :0.59877  
##                     Mean   :0.3384089   Mean   :0.4773   Mean   :0.57544  
##                     3rd Qu.:0.5439500   3rd Qu.:0.5845   3rd Qu.:0.74623  
##                     Max.   :0.9960000   Max.   :0.9460   Max.   :1.00000  
##  instrumentalness      liveness         loudness         speechiness     
##  Min.   :0.000000   Min.   :0.0321   Min.   :-45.2670   Min.   :0.02620  
##  1st Qu.:0.009679   1st Qu.:0.1384   1st Qu.:-13.0548   1st Qu.:0.04193  
##  Median :0.126991   Median :0.1790   Median : -9.5905   Median :0.05540  
##  Mean   :0.250818   Mean   :0.2011   Mean   :-10.6743   Mean   :0.08715  
##  3rd Qu.:0.423197   3rd Qu.:0.2385   3rd Qu.: -7.0314   3rd Qu.:0.09417
##  Max.   :0.993000   Max.   :0.9370   Max.   :  0.9426   Max.   :0.94400
##      tempo            valence         duration_ms      time_signature
##  Min.   : 60.01   Min.   :0.00001   Min.   :  39765   Min.   :1.000
##  1st Qu.:110.36   1st Qu.:0.27283   1st Qu.: 199852   1st Qu.:3.769
##  Median :119.18   Median :0.42494   Median : 243201   Median :3.917
##  Mean   :119.29   Mean   :0.42410   Mean   : 284950   Mean   :3.855
##  3rd Qu.:128.04   3rd Qu.:0.57749   3rd Qu.: 307199   3rd Qu.:4.000
##  Max.   :197.93   Max.   :0.99000   Max.   :3876277   Max.   :5.000
```

3. **Data Transformation:**

It is used to prepare and manipulate data for analysis and visualization. Data transformation is often carried out via,

Cleaning and filtering the data include detecting and addressing missing values, outliers, and anomalies in the dataset. The release date column contains inconsistencies in the data that are processed to the proper format using a function.

```r
parse_date <- function(date_str) {
  formats <- c("%d-%B-%y","%Y","%B-%y")
  parsed_date <- NA
  for (format in formats) {
    parsed_date <- lubridate::parse_date_time(date_str, orders = format)
    if (!is.na(parsed_date)) {
      break
    }
  }
  return(parsed_date)
}

spotify$release_date <- sapply(spotify$rel_date, parse_date) # Parse dates in
the correct format
```

Creating derived variables involves merging or altering existing variables to obtain more information. To address my findings, I've defined additional variables that separate the year and decade from the rel_date. The'mutate()' method in dplyr is widely used to construct derived variables.

```r
spotify$year <- year(spotify$release_date) # Extract the year
spotify <- spotify %>%
  mutate(decade = floor(year / 10)* 10)
view(spotify)

spotify <- spotify %>%
  filter(decade <= 2020)
view(spotify)  #Removing the decades after 2020
```

Results/Findings (Data Visualization):

This technique is used to visually explore datasets, identify patterns, trends and relationships and communicate insights effectively. The ’ggplot2 package is used to visualize the data using different types of plots as mentioned below:

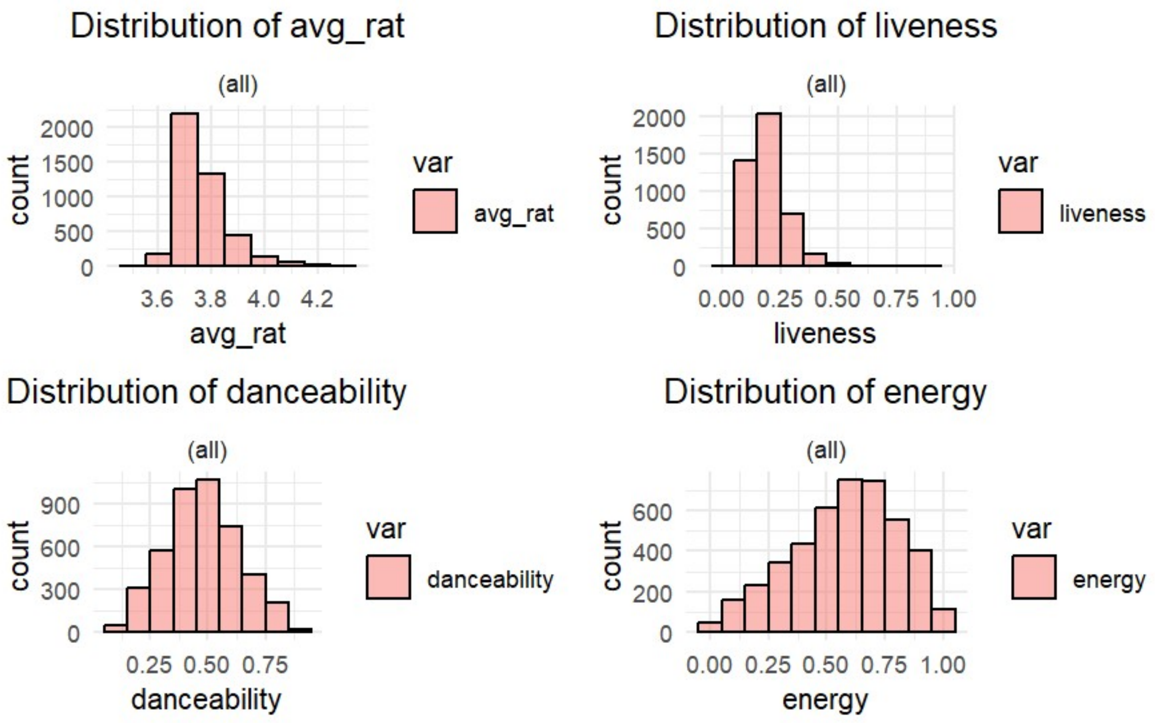
Research Question 1: To Analyse the distribution of some of the audio features

Histograms are used to visually represent the distribution of a single numerical value. They provide information about the central tendency, spread, and skewness of the data. Histograms are used to display the distribution of several audio properties, including danceability, liveness, energy, and average rating.

```r
selected_variables <- c("avg_rat", "liveness", "danceability", "energy")

histograms <- lapply(selected_variables, function(var) {
  ggplot(spotify, aes(x = .data[[var]], fill = var)) +
    geom_histogram(color = "black", alpha = 0.5, binwidth = 0.1) +
    facet_wrap(~ ., scales = "free_y", ncol = 2) +
    labs(title = paste("Distribution of", var)) +
    theme_minimal() +  # Apply a minimal theme for better readability
    theme(plot.title = element_text(hjust = 0.5))  # Center the plot title
})

# Combine all histograms into one plot
multiplot <- do.call(gridExtra::grid.arrange, histograms)
```



```r
multiplot
```

```
## TableGrob (2 x 2) "arrange": 4 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## 3 3 (2-2,1-1) arrange gtable[layout]
## 4 4 (2-2,2-2) arrange gtable[layout]
```

Average Rating:

The distribution has a narrow spread, The peak of the plot is shitfted towards the left side, hence the distribution is negatively skewed (Skewed Right). This indicates that more data points are concentrated on the lower end of the range. The most common rating for all the songs in the data set is found to be 3.7 out of 5 which is given to more that 2000 songs present in the data set.

Liveness:

It is similar to the distribution of the average rating (Skewed Right) and with a narrow spread, the most common ratings ranges from 0.13 to 0.25 in the scale of 0 to 1 which is given to nearly 2000 songs in the data set. The most unusual (less occuring) liveness ratings are 0.5, 0.75 and 1.00.

Danceability:

Danceability has a wider spread and a Bell shaped curve is seen, which shows that the distribution is normal. The most common danceability rating is 0.5 in the scale of 0 to 1 for around 1050 songs in the data set. which ideally will be the mean of the distribution.

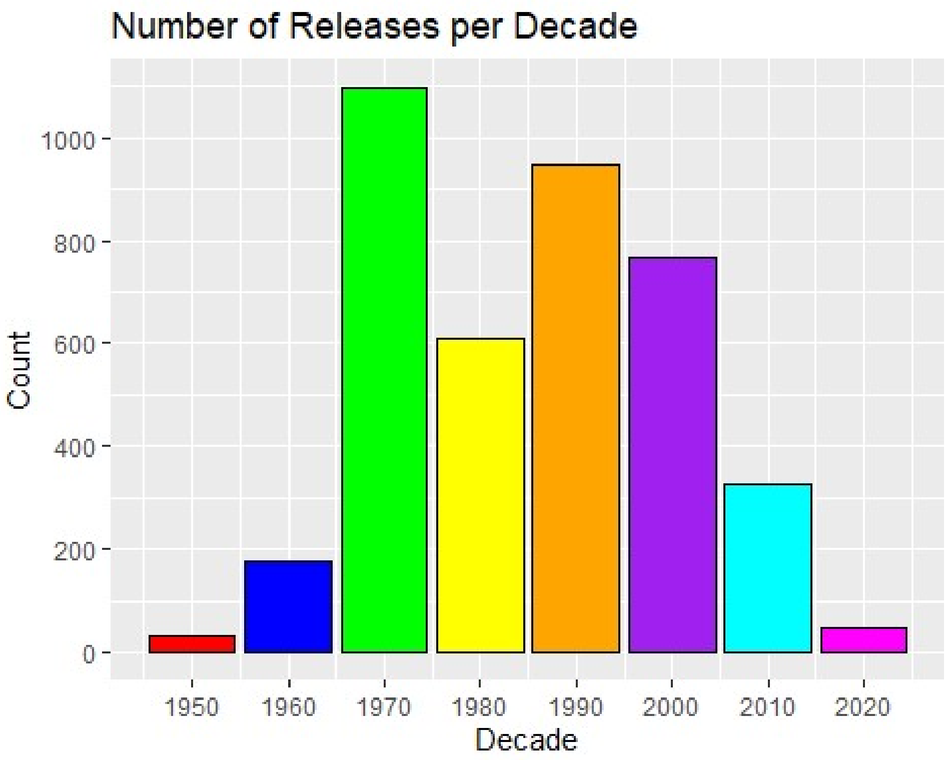
Energy:

It has a wider spread and the peak is towards the right (positively skewed). The most common energy ratings lies within the range of 0.5 to 0.75 in the scale of 0 to 1 for nearly 1400 songs in the data set.

Research Question 2: To Analyze the number of songs released per decade.

Bar plots are commonly used to visualize the frequency or count of categorical variables. They are useful for comparing the distribution of categories or groups. Using rel_date variable, the decade of release for each song is determined and the number of songs released during each decade is plotted.

```
color = c("red", "blue","green","yellow","orange","purple","cyan","magenta")
ggplot(spotify, aes(x = decade)) +
  geom_bar(fill = color, color = "black") +
  scale_y_continuous(breaks = seq(0,1000, by = 200)) +
  scale_x_continuous(breaks = seq(1940,2020, by = 10)) +
  labs(title = "Number of Releases per Decade", x = "Decade", y = "Count")
```



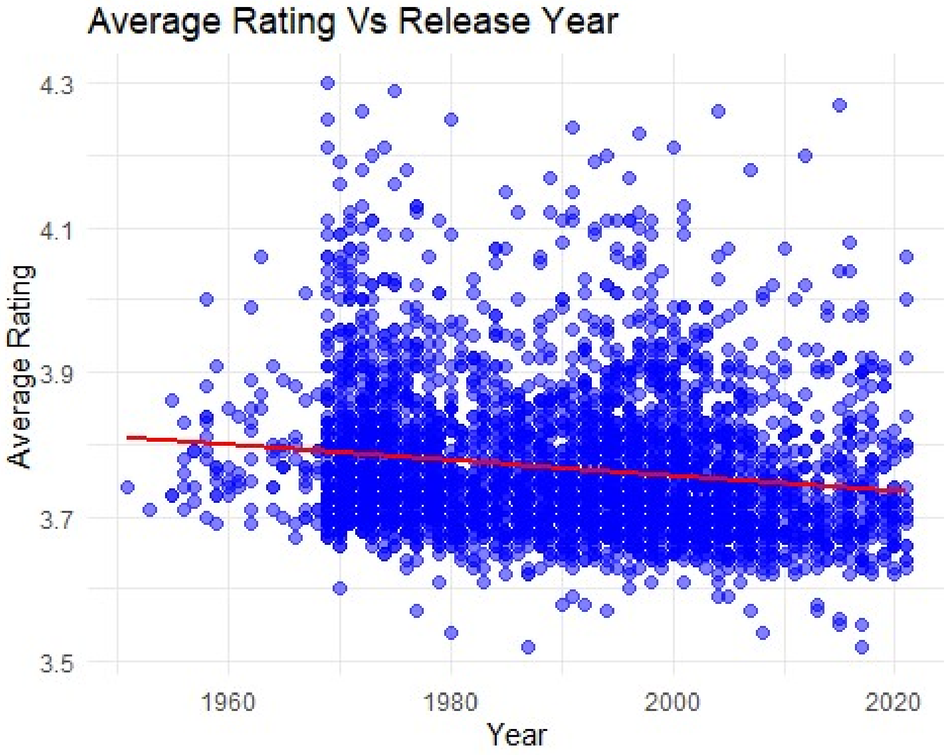
It is seen that the most common decades which holds the maximum number of releases are 1970s, 1980s, 1990s and 2000s. Out of which the 1970s has maximum number of releases of nearly 1100 songs and lowest release decade is the 1950s. There is rapid increase in the number of releases after 1960s and a progressive decline in the count is seen after 1990s. The bar graph thus clearly depicts how the number of songs released changes over each decade.

Research Question 3: Analysing how average rating of the songs has changed over years.

Scatter plots are used to visualize the relationship between two numerical variables. They help in identifying correlations, clusters and outliers in the data. A scatter plot is created comparing the year of release and the danceability audio feature.

```
ggplot(spotify, aes(x = year, y = avg_rat)) +
  geom_point( color = "blue", alpha = 0.5, size = 2.2) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x = "Year",
       y = "Average Rating", title = "Average Rating Vs Release Year") +
  theme_minimal()

## `geom_smooth()` using formula = 'y ~ x'
```



The scatter plot describe how the average rating of the songs distributed over the years. It is clear that most of the songs in the data set are released during the period of 1970 to 2000 and has an average rating between 3.6 to 3.9. There is a spreadout dispersion around the trend line which depicts a weaker relationship between the variables. However, A slightly declining trend in the average rating is noted in the distribution.

Conclusion:

This study conducts a comprehensive analysis of audio features, release trends, and average music track ratings throughout time. The study of distributional characteristics reveals nuanced insights into the diversity and common judgments of key criteria like average rating, liveness, danceability, and energy. Furthermore, studying song releases by decade reveals extraordinary trends in the music industry, with unique peaks and valleys detected throughout time periods. Furthermore, a review of average ratings over time finds a slight decrease in ratings, reflecting possible changes in listener preferences or evolving music production standards. Overall, this study provides significant insights into music trends and consumer behavior, which can help many stakeholders, including artists, record labels, and streaming platforms, make informed decisions about content development, promotion, and curation.

References:

- Wilke, C.O. (no date) Fundamentals of Data Visualization. https://learning.oreilly.com/library/view/fundamentals-of-data/9781492031079/copyright-page01.html.

- Top 5000 Albums of All Time - Spotify features (2022). https://www.kaggle.com/datasets/lucascantu/top-5000-albums-of-all-time-spotify-features/data.

Appendix:

~~~
---
title: "Analyzing Music Trends: Insights from top 5000 Albums (Appendix)"
author: "Ravivarman Devarajan"
date: "2024-04-04"
output:
  word_document: default
  html_document: default
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(warning=FALSE)
library(tidyverse)
library(rmarkdown)
library(lubridate)
library(dplyr)
library(gridExtra)
library(ggplot2)
```
__Project Summary:__

The goal of this project is to demonstrate proficiency in data analysis techniques using R
Studio and R Markdown by exploring a novel dataset and creating meaningful visualizations.
The project will involve selecting a dataset, conducting exploratory data analysis and
telling a story with different types of visualizations.

__Introduction:__

The Data set contains information about 4402 albums, it consists of 19 features from
spotify. The data includes the following type of information:

* Basic information such as Artist name, album name, release date etc.
*       Genres associated with the album and a textual description of the album.
*       Average user rating and potentially the number of ratings/reviews each  album has
received.
*       A range of audio features, these are numerical variables likely extracted from the
audio itself, including acousticness, danceability, energy and many more.
*       Additional Information such as duration of the album in milliseconds, time
signature etc.
This data can be used for various music- related analyses and visualizations which
includes,
*       Exploring music trends across decades, analyse how genres, average ratings, or
audio features alike danceability or tempo have changed over time.
*       Recommending music based on user preferences and audio feature similarities
between albums.
*       Compare audio features and user ratings between artists or across different
genres.

This data set appears to be a valuable resource for exploring music information, artist
styles, user preferences and potentially music trends across time. By analysing the
various features, you can gain insights into listener preferences and the sonic
characteristics of different genres and artists. However, The objective of my research is,

1. To find the typical ranges and distributions of some of the audio features (numerical
variables) and analyse their patterns.

2. To find how the number of songs released per decade changed over time and analyse any
noticeable trends or patterns in the distribution of music releases.

3. To select any one audio feature and describe how it has changed over the years and find
any discernible trends or patterns in the relationship between that feature and the
release year.

__Variables of interest and the Techniques used:__
~~~

The characteristics of variables associated with my research questions and their characteristics are listed below:

1. _Danceability:_ A measure of a piece of music's suitability for dancing based on numerous musical components such as rhythm, tempo, and so on. Higher values reflect a song's danceability. Songs with high danceability scores typically include rhythmic patterns and beats that encourage movement and dancing.

2. _Average Rating (avg_rat):_ It is the average score or rating assigned to a music by listeners or reviewers. It provides an overall measure of the song's perceived quality or enjoyment by its listeners. It varies according to the rating scale employed and the rating system's context (for example, user-generated ratings on a music system or professional reviews).

3. _Liveness:_ It assesses the existence of live performance characteristics in a song, indicating whether it was recorded live or in the studio. It runs from 0 to 1, with higher numbers suggesting a stronger concentration of live performance elements. Songs with greater liveness scores may include audience reactions, crowd noise, or other elements that capture the mood of a live performance.

4. _Energy:_ It is the perceptual measure of intensity and activity level in a piece of music. It captures characteristics such as loudness, dynamic range, and tempo that add to a song's overall sense of excitement and energy. Energy levels typically range from 0 to 1, with higher values suggesting music with more energetic or intense features.

5. _Release Date:_ This is the date that a song or album was formally released to the public. It enables the investigation of trends and patterns across time. Release dates can be expressed in a variety of ways and divided into years, decades, or particular time periods for analysis.

The various Exploratory Data Analysis(EDA) techniques used in this research are

1. __Loading Packages and Data:__

R packages include a diverse set of functions and tools for data processing and visualization. Before beginning any analysis, you must install and load the relevant packages using the 'install.packages()' and 'library()' functions. EDA programs commonly used in R include tidyverse (which contains ggplot2, dplyr, tidyr, readr, and so on), lubridate, gridExtra, and rmarkdown.

Data is often kept in files like CSV, Excel, and databases. In R, data can be imported into the workspace using functions such as'read.csv()' and'read.excel()'.

~~~
```{r Loading_Data}
spotify <- read.csv("D:/Lab/Top 5000 Albums _Spotify features.csv")
View(spotify) #reading the dataset from a csv file
```
~~~

2. __Data Inspection:__

Data inspection is used to determine the structure, content, and quality of a dataset. After loading the data, it is frequently useful to examine its structure, summary statistics, and the first few rows to gain a better grasp of its content. Some of the functions useful for dataset examination include 'glimpse()','summary()', 'head()', 'is.na()', and so on. These procedures provide a brief overview of the dataset's structure, including column names, missing values, data types, and the first few rows of data.

~~~
```{r Data_Inspection}
glimpse(spotify)  #displays data types and variable names
which(is.na(spotify)) #returns the indices for missing values (NA)
summary(spotify) #Provides overview of the structure and distribution
```
~~~

3. __Data Transformation:__

It is used to prepare and manipulate data for analysis and visualization. Data transformation is often carried out via,

Cleaning and filtering the data include detecting and addressing missing values, outliers, and anomalies in the dataset. The release date column contains inconsistencies in the data that are processed to the proper format using a function.

~~~
```{r, Data_Transformation1}
parse_date <- function(date_str) {
  formats <- c("%d-%B-%y","%Y","%B-%y")
  parsed_date <- NA
  for (format in formats) {
    parsed_date <- lubridate::parse_date_time(date_str, orders = format)
    if (!is.na(parsed_date)) {
      break
    }
  }
  return(parsed_date)
}

spotify$release_date <- sapply(spotify$rel_date, parse_date) # Parse dates in the correct
format

```
~~~

Creating derived variables involves merging or altering existing variables to obtain more information. To address my findings, I've defined additional variables that separate the year and decade from the rel_date. The'mutate()' method in dplyr is widely used to construct derived variables.

~~~
```{r Data_Transformation2}
spotify$year <- year(spotify$release_date) # Extract the year
spotify <- spotify %>%
  mutate(decade = floor(year / 10)* 10)
view(spotify)

spotify <- spotify %>%
  filter(decade <= 2020)
view(spotify)  #Removing the decades after 2020
```
~~~

__Results/Findings(Data Visualization):__

This technique is used to visually explore datasets, identify patterns, trends and relationships and communicate insights effectively. The 'ggplot2 package is used to visualize the data using different types of plots as mentioned below:

__Research Question 1: To Analyse the distribution of some of the audio features__

Histograms are used to visually represent the distribution of a single numerical value. They provide information about the central tendency, spread, and skewness of the data. Histograms are used to display the distribution of several audio properties, including danceability, liveness, energy, and average rating.

~~~
```{r histograms}
selected_variables <- c("avg_rat", "liveness", "danceability", "energy")

histograms <- lapply(selected_variables, function(var) {
  ggplot(spotify, aes(x = .data[[var]], fill = var)) +
    geom_histogram(color = "black", alpha = 0.5, binwidth = 0.1) +
    facet_wrap(~ ., scales = "free_y", ncol = 2) +
    labs(title = paste("Distribution of", var)) +
    theme_minimal() +  # Apply a minimal theme for better readability
    theme(plot.title = element_text(hjust = 0.5))  # Center the plot title
})
~~~

~~~
# Combine all histograms into one plot
multiplot <- do.call(gridExtra::grid.arrange, histograms)
multiplot
```
~~~

__Average Rating:__

The distribution has a narrow spread, The peak of the plot is shitfted towards the left side, hence the distribution is negatively skewed (Skewed Right). This indicates that more data points are concentrated on the lower end of the range. The most common rating for all the songs in the data set is found to be 3.7 out of 5 which is given to more that 2000 songs present in the data set.

__Liveness:__

It is similar to the distribution of the average rating (Skewed Right) and with a narrow spread, the most common ratings ranges from 0.13 to 0.25 in the scale of 0 to 1 which is given to nearly 2000 songs in the data set. The most unusual (less occuring) liveness ratings are 0.5, 0.75 and 1.00.

__Danceability:__

Danceability has a wider spread and a Bell shaped curve is seen, which shows that the distribution is normal. The most common danceability rating is 0.5 in the scale of 0 to 1 for around 1050 songs in the data set. which ideally will be the mean of the distribution.

__Energy:__

It has a wider spread and the peak is towards the right (positively skewed). The most common energy ratings lies within the range of 0.5 to 0.75 in the scale of 0 to 1 for nearly 1400 songs in the data set.

__Research Question 2: To Analyze the number of songs released per decade.__

Bar plots are commonly used to visualize the frequency or count of categorical variables. They are useful for comparing the distribution of categories or groups. Using rel_date variable, the decade of release for each song is determined and the number of songs released during each decade is plotted.

~~~
```{r barplot}
color = c("red", "blue","green","yellow","orange","purple","cyan","magenta")
ggplot(spotify, aes(x = decade)) +
  geom_bar(fill = color, color = "black") +
  scale_y_continuous(breaks = seq(0,1000, by = 200)) +
  scale_x_continuous(breaks = seq(1940,2020, by = 10)) +
  labs(title = "Number of Releases per Decade", x = "Decade", y = "Count")
```
~~~

It is seen that the most common decades which holds the maximum number of releases are 1970s, 1980s, 1990s and 2000s. Out of which the 1970s has maximum number of releases of nearly 1100 songs and lowest release decade is the 1950s. There is rapid increase in the number of releases after 1960s and a progressive decline in the count is seen after 1990s. The bar graph thus clearly depicts how the number of songs released changes over each decade.

__Research Question 3: Analysing how average rating of the songs has changed over years.__

Scatter plots are used to visualize the relationship between two numerical variables. They help in identifying correlations, clusters and outliers in the data. A scatter plot is created comparing the year of release and the danceability audio feature.

~~~
```{r scatterplot}
ggplot(spotify, aes(x = year, y = avg_rat)) +
  geom_point( color = "blue", alpha = 0.5, size = 2.2) +
~~~

```
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x = "Year",
       y = "Average Rating", title = "Average Rating Vs Release Year") +
  theme_minimal()

```

The scatter plot describe how the average rating of the songs distributed over the years. It is clear that most of the songs in the data set are released during the period of 1970 to 2000 and has an average rating between 3.6 to 3.9. There is a spreadout dispersion around the trend line which depicts a weaker relationship between the variables. However, A slightly declining trend in the average rating is noted in the distribution.

__Conclusion:__

This study conducts a comprehensive analysis of audio features, release trends, and average music track ratings throughout time. The study of distributional characteristics reveals nuanced insights into the diversity and common judgments of key criteria like average rating, liveness, danceability, and energy. Furthermore, studying song releases by decade reveals extraordinary trends in the music industry, with unique peaks and valleys detected throughout time periods. Furthermore, a review of average ratings over time finds a slight decrease in ratings, reflecting possible changes in listener preferences or evolving music production standards. Overall, this study provides significant insights into music trends and consumer behavior, which can help many stakeholders, including artists, record labels, and streaming platforms, make informed decisions about content development, promotion, and curation.

__References:__

* Wilke, C.O. (no date) Fundamentals of Data Visualization. https://learning.oreilly.com/library/view/fundamentals-of-data/9781492031079/copyright-page01.html.

* Top 5000 Albums of All Time - Spotify features (2022). https://www.kaggle.com/datasets/lucascantu/top-5000-albums-of-all-time-spotify-features/data.