

Synergy between Spotify Song Tracks and their relative Youtube Videos

My Topic

My topic is to investigate the synergy between Spotify songs and YouTube videos through data visualization, uncovering how music performance on both platforms interplays and influences audience engagement and trends. I will be identifying the top trends and characteristics with popular youtube videos and spotify tracks. Data Sources

The following datasets from Kaggle to explore my topics were used:

1. Most Streamed Spotify Songs 2023 : <https://www.kaggle.com/datasets/nelgiriyeewithana/top-spotify-songs-2023>

2. Spotify and Youtube: <https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube/data>

Week 9 Diary:

For week 9, I have planned out the rough layout of my app using some sketches I did on my tablet, however these will be executed in the later weeks when I hopefully firm up my webpage. Here are the following content:

Week 10 Diary:

(1) My Question: What is the synergy between the various elements in a song?

(2) Why do I find this question important to answer?

- As the evolution of music grows, we can learn how the relationships and dynamics of different elements in a song affects how people perceive music to be addictive. Songs with musical groove have become popular as naturalistic stimuli to study interactions between auditory and motor brain regions (O'Connell et al, 2022)
- It is important to understand what makes music popular and what are the specific elements of creating a popular song that is enjoyed by many.
- Different audiences may have different preferences different musical elements, which can be useful in understanding these preferences which can help artists tailor their music to specific groups based on their songs.



Figure 1: Images I sketched to visualise my website

(3) Which rows and columns of the dataset will be used to answer this question?

From Spotify & Youtube Dataset

- Columns:
 - Musical Attributes: Danceability, Energy, Key, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo
 - Duration_ms
 - Views
 - Likes
 - Comments
 - Artist
 - Track
 - Stream
- Rows: All rows

From Most Streamed Spotify Songs 2023 Dataset

- Columns:
 - track_name: Name of the song
 - artist_count: Number of artists contributing to the song
 - in_spotify_playlists: Number of Spotify playlists the song is included in
 - in_spotify_charts: Presence and rank of the song on Spotify charts
 - streams: Total number of streams on Spotify
 - bpm: Beats per minute, a measure of song tempo
 - key: Key of the song

- mode: Mode of the song (major or minor)
 - danceability_%%: Percentage indicating how suitable the song is for dancing
 - valence_%%: Positivity of the song’s musical content
 - energy_%%: Perceived energy level of the song
 - acousticness_%%: Amount of acoustic sound in the song
 - instrumentalness_%%: Amount of instrumental content in the song
 - liveness_%%: Presence of live performance elements
 - speechiness_%%: Amount of spoken words in the song
-

(4) Why did I choose this dataset:

- I have chosen two datasets to answer the questions on the synergy in songs and their performance on YouTube and Spotify separately
- There are a few reasons as to why I chose this datasets
 - Global YouTube Statistics 2023 / Most Streamed Spotify Songs 2023
 - * Both datasets have numerical variables that contains data types such as doubles, integers, complex numbers, logical variables which can be manipulated
 - * Both datasets also have categorical variables that contains nominal variables, string values. This will allow me to understand the qualitative aspects of the dataset as well.

(5) Basic Visualisations to draw:

I will be looking at the following graphs to come up with some possible visualisations, although I will not be creating all of them.

- **Correlation Heatmap:** I will use a heatmap to visualize correlations between musical attributes and popularity metrics.
- **Scatter Plots:** I will plot scatter plots with regression lines to showcase the relationship between two variables. One graph that I can visualise is to plot “Danceability” on the x-axis and “Spotify Streams” on the y-axis to see if more danceable songs get more streams.
- **Bar Chart :** I will create bar charts to compare the number of streams/views/likes for official vs. non-official videos. Also, for categorical variables like “key,” you can visualize the number of songs in each key.
- **Histogram:** I will create histograms to visualize the distribution of continuous variables, such as song tempo or loudness. It’ll help to understand the most common tempo or loudness values among the top songs.
- **Word Cloud:** Based on the “Description” of YouTube videos, I will attempt to create a word cloud to see which words or phrases are most commonly mentioned.

(6) Advanced Graphs:

I hope to enhance my skills to make better and more informed graphs through R.

- **Bivariate Relationship:** I would like to identify those variables that have relationships with each other and to identify whether they can impact each other. I hope to create a pair plot to visualize the bivariate relationships between each pair of variables.

- **Violin Plot for Stream Distributions:** Create a violin plot to visualize the distribution of Spotify Streams across different keys or any other categorical variable, providing more depth than a standard box plot.

Process of cleaning of Data: Excel

This week, I will start by cleaning my data. Here is a snapshot of my variables from “Spotify & Youtube” dataset.

```
## New names:
## Rows: 20718 Columns: 28
## -- Column specification
## ----- Delimiter: "," chr
## (10): Artist, Url_spotify, Track, Album, Album_type, Uri, Url_youtube, T... dbl
## (16): ...1, Danceability, Energy, Key, Loudness, Speechiness, Acousticne... lgl
## (2): Licensed, official_video
## i Use 'spec()' to retrieve the full column specification for this data. i
## Specify the column types or set 'show_col_types = FALSE' to quiet this message.
## * ' -> '...1'

## # A tibble: 20,718 x 28
##   ...1 Artist   Url_spotify   Track Album Album_type Uri   Danceability Energy
##   <dbl> <chr>     <chr>         <chr> <chr> <chr> <chr>         <dbl> <dbl>
## 1      0 Gorillaz https://open~ Feel~ Demo~ album   spot~      0.818 0.705
## 2      1 Gorillaz https://open~ Rhin~ Plas~ album   spot~      0.676 0.703
## 3      2 Gorillaz https://open~ New ~ New ~ single  spot~      0.695 0.923
## 4      3 Gorillaz https://open~ On M~ Plas~ album   spot~      0.689 0.739
## 5      4 Gorillaz https://open~ Clin~ Gori~ album   spot~      0.663 0.694
## 6      5 Gorillaz https://open~ DARE~ Demo~ album   spot~      0.76  0.891
## 7      6 Gorillaz https://open~ New ~ New ~ single  spot~      0.716 0.897
## 8      7 Gorillaz https://open~ She'~ Huma~ album   spot~      0.726 0.815
## 9      8 Gorillaz https://open~ Crac~ Crac~ single  spot~      0.741 0.913
## 10     9 Gorillaz https://open~ Dirt~ Demo~ album   spot~      0.625 0.877
## # i 20,708 more rows
## # i 19 more variables: Key <dbl>, Loudness <dbl>, Speechiness <dbl>,
## #   Acousticness <dbl>, Instrumentalness <dbl>, Liveness <dbl>, Valence <dbl>,
## #   Tempo <dbl>, Duration_ms <dbl>, Url_youtube <chr>, Title <chr>,
## #   Channel <chr>, Views <dbl>, Likes <dbl>, Comments <dbl>, Description <chr>,
## #   Licensed <lgl>, official_video <lgl>, Stream <dbl>
```

Here is a snapshot of my variables from “Most Streamed Spotify Songs 2023” dataset.

```
## Rows: 953 Columns: 24
## -- Column specification -----
## Delimiter: ","
## chr (5): track_name, artist(s)_name, streams, key, mode
## dbl (17): artist_count, released_year, released_month, released_day, in_spot...
## num (2): in_deezer_playlists, in_shazam_charts
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

```
## # A tibble: 953 x 24
##   track_name      'artist(s)_name' artist_count released_year released_month
##   <chr>          <chr>                <dbl>         <dbl>         <dbl>
## 1 Seven (feat. Latt~ Latto, Jung Kook          2          2023           7
## 2 LALA            Myke Towers              1          2023           3
## 3 vampire        Olivia Rodrigo            1          2023           6
## 4 Cruel Summer   Taylor Swift              1          2019           8
## 5 WHERE SHE GOES  Bad Bunny                 1          2023           5
## 6 Sprinter       Dave, Central C~          2          2023           6
## 7 Ella Baila Sola Eslabon Armado,~         2          2023           3
## 8 Columbia       Quevedo                   1          2023           7
## 9 fukumean       Gunna                     1          2023           5
## 10 La Bebe - Remix Peso Pluma, Yng~        2          2023           3
## # i 943 more rows
## # i 19 more variables: released_day <dbl>, in_spotify_playlists <dbl>,
## #   in_spotify_charts <dbl>, streams <chr>, in_apple_playlists <dbl>,
## #   in_apple_charts <dbl>, in_deezer_playlists <dbl>, in_deezer_charts <dbl>,
## #   in_shazam_charts <dbl>, bpm <dbl>, key <chr>, mode <chr>,
## #   'danceability_%' <dbl>, 'valence_%' <dbl>, 'energy_%' <dbl>,
## #   'acousticness_%' <dbl>, 'instrumentalness_%' <dbl>, 'liveness_%' <dbl>, ...
```

Challenges of cleaning of Data:

One of the challenges that I faced while cleaning the data on Excel was the amount of iterations needed. As there were symbols that made it difficult to read, I removed them by using the Find & Replace function on excel. I did this so that it would be easier to call out certain names and make them more readable to the user. **I only did this for the variable “artist” as I will be using this variable to make comparisons**

Old

Antonio Carlos Jobim	https://open.Águas De Marão
Antonio Carlos Jobim	https://open.S Tinha De Ser Com Voc™
Antonio Carlos Jobim	https://open.Pela Luz dos Olhos Teus -
Antonio Carlos Jobim	https://open.The Girl From Ipanema
Antonio Carlos Jobim	https://open.Corcovado
Antonio Carlos Jobim	https://open.Para Machuchar Meu Cor
Antonio Carlos Jobim	https://open.Wave
Antonio Carlos Jobim	https://open.Garota De Ipanema
Antonio Carlos Jobim	https://open.Triste
Antonio Carlos Jobim	https://open.O Morro NÉo Tem Vez

Figure 2: Strings of text with symbols

New

Week 11 Diary:

Antonio Carlos Jobim	https://open.	Águas De Março
Antônio Carlos Jobim	https://open.	S Tinha De Ser Com Voc™
Antônio Carlos Jobim	https://open.	Pela Luz dos Olhos Teus -
Antônio Carlos Jobim	https://open.	The Girl From Ipanema
Antônio Carlos Jobim	https://open.	Corcovado
Antônio Carlos Jobim	https://open.	Para Machuchar Meu Cor
Antônio Carlos Jobim	https://open.	Wave
Antônio Carlos Jobim	https://open.	Garota De Ipanema

Figure 3: After removing the symbol

Antonio Carlos Jobim	https://open.	Águas De Março
Antônio Carlos Jobim	https://open.	S Tinha De Ser Com Voc™
Antônio Carlos Jobim	https://open.	Pela Luz dos Olhos Teus -
Antônio Carlos Jobim	https://open.	The Girl From Ipanema
Antônio Carlos Jobim	https://open.	Corcovado
Antônio Carlos Jobim	https://open.	Para Machuchar Meu Cor
Antônio Carlos Jobim	https://open.	Wave
Antônio Carlos Jobim	https://open.	Garota De Ipanema

Figure 4: After removing the symbol

(1) List the visualizations that you are going to use in your project

Bar Graph: Bivariate Relationship Between Danceability and Energy

(a) I will be using the *Danceability* and *Energy* variables which are of double data type.

(b) This could help to visualize the relationship between how danceable a track is and its energy level. My hypothesis is that there is no bivariate relationship between danceability and energy as there are songs which maybe low in energy but have high danceability in them. (E.g Slow songs that are meant for slow dancing)

Heatmap: Heatmap of correlations

(a) The variables I will be using are danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo.

(b) This will allow me to identify the relationships between the variables and how musicians can use this information to understand their target audience better.

(2) How do you plan to make it interactive?

Bar Graph: Bivariate Relationship Between Danceability and Energy

- Dropdown Menu for Coloring: Allows users to choose whether the points should be colored based on genre or album type.
- Slider for Danceability Range: Lets users filter the data based on a range of danceability values.
- Slider for Energy Range: Lets users filter the data based on a range of energy values.

Heatmap: Heatmap of correlations

- Checkboxes: To select which features to include in the correlation matrix.
- Dropdown: To choose the method of correlation
- Reset Button: to recompute the heatmap after changing the selections.

List of Topics used for my project

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##      combine
```

Topics	Weeks
Bivariate Scatter Plot	Self Learnt
Heat Map	Self Learnt
Shiny features: Dropdown	Week 8
Shiny features: Slider	Week 8
Shiny features: Reset Button	Week 8
ggplot2	Week 2
ggplot	Week 2
Manipulating data: Changing Data types	Week 3
Functions	Week 5
Loops	Week 6

Week 11 Diary:

Error with setting background image

This is a minimal task! But I am still unable to set the correct image and allow it to be displayed on my screen. I will try to get this done before the final submission.

Error with setting video wrap

One Issue that I had was the setting of the video alignment on the About page. I was unable to get a straight video that covers the full section of the screen even after referencing Quarto.org for the guides.

```
26 .fullscreen-video-wrap {  
27   position: relative;  
28   top: 0;  
29   left: 0;  
30   width: 100vw;  
31   height: 100vh;  
32   overflow: hidden;
```

Figure 5: The standardised width was not appropriate for my screen

In the end, I had to tweak around the values for width and stuck with -605vw. I think that there should be some errors in my settings as I understand that it should automatically fit the picture to my screen.

Valence Bargraph

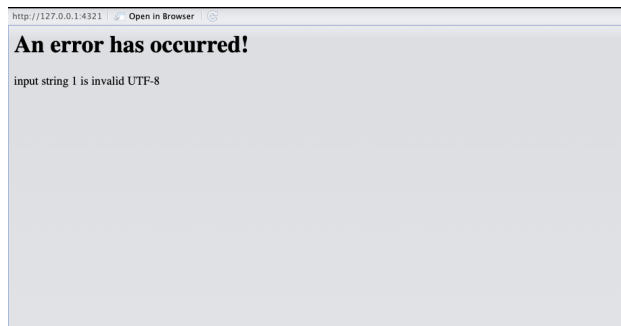


Figure 6: The standardised width was not appropriate for my screen

I had this error when trying to process the column “Stream” as I did not clean my dataset well enough to realise that there was invalid characters in the dataset. This has prevented me from being able to display my bar graph.

As a result, I had referenced some help from Chatgpt & source from this [StackOverflow post] and realised that I had to add the following code:

```
mutate_if(is.character, utf8::utf8_encode)
```

Word Cloud

For this particular error, I had faced this issue of being unable to run my word cloud. One of the reasons was because while searching up on line on how to do a word cloud, I realised that my dataset was too large and the Column “Tracks” had too many rows for the word cloud to show. Hence, rather than setting the min frequency to 1, I set it to 2 instead as seen below.

(1) min.freq : I had increased the limit from 1 to 2 as it showed too little words and I wanted to make it more meaningful by showing more varieties of words.


```

136 spotify_youtube <- read.csv("~/Documents/GitHub/wongweiqihahaha.github.io/spotify_youtube.csv")
137
138
139 word_data <- spotify_youtube %>%
140   unnest_tokens(word, Track) %>%
141   count(word, sort = TRUE)
142
143
144 |
145 wordcloud(words = word_data$word, freq = word_data$n, min.freq = 1)
146
147 ...

```

Warning: seinendan could not be fit on page. It will not be plotted.
 Warning: pancakes could not be fit on page. It will not be plotted.
 Warning: membawamu could not be fit on page. It will not be plotted.
 Warning: broken could not be fit on page. It will not be plotted.
 Warning: essence could not be fit on page. It will not be plotted.
 Warning: sister could not be fit on page. It will not be plotted.
 Warning: veyssel could not be fit on page. It will not be plotted.
 Warning: punishment could not be fit on page. It will not be plotted.
 Warning: deeply could not be fit on page. It will not be plotted.

Figure 7: Word Cloud Adjustment

```

min.freq = 2, max.words = 100)

```

Figure 8: Word Cloud Adjustment

will increase as well and vice versa.

- Acousticness and Energy has a correlation of -0.65, which suggest that there is a negative linear correlation between these 2 variables. This means that whenever Acousticness increases, Energy will decrease. This observation is justified as per an observation Zhou (2023).

Loading required package: RColorBrewer

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

A brief lesson about music

Let's breakdown the dataset!

How are the different musical elements related to one another?

Heatmap: What do these numbers mean?

Histogram: What do these numbers mean?

Valence Bar Graph: What do these numbers mean?

Figure 9: Updated Code

(2) max.word: I had set a limit to 100 as the function was unable to process and fit all the words in my graph.

I plan to enhance this word cloud using this example provided by Rul,2019.This would be really useful as I can make different shapes and sizes of Word Cloud without being limited to HTML's style of word cloud. One issue with the current code that I have is that it is limited to the design I can make for it and I would like to try to be more experimental.

Line graph

I had the error of the wrong data type used in the function sum. I did not realise that streams were all in character type and it should have been in numeric data type in order for me to manipulate them in sum.

As such, I used the as.numeric() function to request R to read them as numeric values instead which worked.

```
# Plotting
ggplot(monthly_streams_2022, aes(x = released_month, y = total_streams, group = 1)) +
  geom_line() + # Line graph
  geom_point() + # Points at each month
  theme_minimal() + # Minimal theme
  labs(title = "Total Streams per Month in 2022",
        x = "Month",
        y = "Total Streams")
...

```

```
Error in `summarise()`:  
! In argument: `total_streams = sum(streams)`.  
! In group 1: `released_month = NA`.  
Caused by error in `sum()`:  
! invalid 'type' (character) of argument  
Backtrace:
```

Figure 10: Updated Code

Pie Chart (In Progress)

I am still in the midst of coming up with the code.