Diary Entry: Final Submission

Chua Jieh Yih Audra

2023-11-24

Week 9

Finalised topic (Changed)

I originally wanted to analyse the ratings for Animal Crossing by critics and users, and its sales in relation to that. However, upon receiving feedback on my dataset I have decided to change my topic.

My new topic is: what factors affect the number of streams a song gets on Spotify?

Data sources I have curated so far

Here is my new data source: https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube

Week 10

What is the question that I am going to answer?

What factors affect the number of streams a song gets on Spotify?

Why is this an important question?

According to the Bulgarian Comparative Education Society (BCES), 2021, music is present everywhere around us, and acts as a medium which shapes our environment, self-perception, and interpersonal interactions, thus leaving a large impact on society. Source: https://files.eric.ed.gov/fulltext/ED614071.pdf

According to Business of Apps statistics for 2023, Spotify is the most popular music streaming platform in the world, and it has over 350 million users and 150 million subscribers. Source: https://www.businessofapps.com/data/music-streaming-market/

Finally, according to Statista, in 2022, total revenue generated by the recorded music industry was 26.2 billion U.S. dollars, of which the majority of revenue came from streaming services, showing the economic significance they have. Source: https://www.statista.com/chart/4713/global-recorded-music-industry-revenues/

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

Rows: all Columns: Stream, Danceability, Energy, Speechiness, Instrumentalness, Liveness, Valence, Tempo, Duration_ms, Views, Likes, Title, Description, Licensed, official_video

Upon opening my dataset in Excel, it gave me a message letting me know that some numbers had gotten stored as text. Excel helped me to automate the process of fixing these and converting them back to numbers.

Week 11

Visualisations I will use in my project

I will use ggplot line graphs, whereby no. of Spotify streams would be the x axis, and my other variables would be the y axis.

How do I plan to make it interactive

I would like to let the users be able to view each specific data by hovering over the area they want to know the information of. This will allow them to view the peaks in data to learn about what they want to know. I would also like to create a plot that allows users to select more than one variable to view at once (So the x axis, no. of Spotify streams, would remain the same, but the y axis would be determined by the user). This is a challenge since I do not if this will work out yet. Most likely it seems I'll have to use a reactive() function and replace the variable name in ggplot's 'aes' section for y with a variable that changes depending on user input.

What concepts incorporated in your project were taught in the course and which ones were self-learnt?

Table of Concepts Used

Topic/ Concept	Weeks
Variables (numeric and categorical)	3
Manipulating data.	4
Functions (most likely for input)	5
Visualization with ggplot2	7
Exploratory data analyses	9
If/ Else	
Plotly	
Reactive() Expression/ Function	
Text analysis (str_detect() probably)	

Week 12

Challenges and Errors I faced and How I'll Overcome Them

I realised my original plan to use plotly does not work with Quarto for some reason (it works in R, but faces an error when I try to render it, and when I searched online in forums it seems to need fixing on Quarto's end). To recover from this I will change to use Shiny instead.

Week 13: Final Submission

The theme of my data story is variables of music and how they interact, especially in determining the number of streams a track gets on Spotify.

According to the Bulgarian Comparative Education Society (BCES), 2021, music is present everywhere around us, and acts as a medium which shapes our environment, self-perception, and interpersonal interactions, thus leaving a large impact on society (Petrušić, 2021).

According to Business of Apps statistics for 2023, Spotify is the most popular music streaming platform in the world, and it has over 350 million users and 150 million subscribers (Curry, 2023).

Finally, according to Statista, in 2022, total revenue generated by the recorded music industry was 26.2 billion U.S. dollars, of which the majority of revenue came from streaming services, showing the economic significance they have (Richter, 2023).

I chose the data source “Spotify and Youtube” on Kaggle by Salvatore Rastelli for my project (Rastelli, 2023).

The data provided in this data set is both recent and extensive, with over 20,000 entries. It also breaks down aspects of music into many numerical variables that can be used to objectively compare the tracks based on different criteria. For example, a song like “Feel Good Inc.” by Gorillaz that would typically be subjectively described as “funky”, or placed in a nominal category such as “alternative rock”, is instead rated on factors such as “danceability” and “energy”, which help to capture its funk, and “musicalness” or “speechiness” to capture its melodic nature. This rating system is essential to drawing conclusions about the tracks.

Furthermore, the dataset includes information about Spotify streams, YouTube, as well as the music itself, providing a rich array of data for each song.

For my plots, I decided to give the user more freedom by letting them freely pick whatever x and y variables they would like to use to generate the chart. Beyond just exploring the relationship between Spotify streams and musical or YouTube-based variables, users are also able to look at relationships between musical variables and each other, or between musical and YouTube-based variables. I believe that this would improve users’ appreciation and understanding of music, thus deepening their understanding of how these variables interact with each other to affect the performance of the track on Spotify.

The insights from my data are listed on my website, but can be summarised.

There is an overall preference for more danceable and energetic songs, meaning that faster tempos, loud, and noisy music gains more streams on Spotify. This aligns with the insights provided by the Stream vs Tempo graph, which showed a preference for faster songs at around 100-130bpm or 170bpm, rather than ballads of below 100bpm.

Less speechy tracks, involving less rap or speaking, perform better on Spotify, and a near minimal amount of speechiness at a value of 0.03 is preferred. Coupled with a preference towards less instrumentallness based on streams, we can conclude that songs involving a lot more singing tend to be more popular on Spotify.

Songs that do not sound like they involve an audience or are being performed live perform better on Spotify.

Songs conveying both positive and negative emotions or sentiments perform relatively equally well on Spotify, but songs on the extreme ends of valence perform a little less well.

Songs of moderate length perform best, with the best track length being around 3 and a half minutes.

Shorter titles on the tracks’ videos on YouTube of about 50 characters long correlate to more streams on Spotify, and having up to 9-10 links in one’s YouTube description also showed a positive relationship with Spotify streams, but surprisingly the number of views, likes, and comments on YouTube did not strongly correlate to number of Spotify streams.

Finally, there are positive relationships between the video on YouTube representing licensed content, and being the official video for the song, and the number of streams on Spotify.

To implement this project, I used data transformation, and finally plotly for the graph.

For data transformation, I used mutate to add new columns including "Length_of_Title" which uses nchar() to count the number of characters in each song's YouTube title, "Links_in_Description" which uses str_count() to count the number of times the string "https://" appears in the description, "Licensed_Binary" and "Official_Video_Binary" which use as.integer() and as.logical() to convert the True and False values in the original "Licensed" and "official video" variables to 1s and 0s respectively. Finally, I used select to select the final relevant columns for my cleaned dataset.

I used plotly and enclosed it within a function that allowed for user input to determine variables x and y using a new concept, selectInput(). I then used layout() to have the title of the graph and axis labels change alongside the x and y variables to ensure readability. Plotly allows users to see the coordinates of a point upon hovering over it. This was also a new concept to me as I had not learnt about Plotly before.

References

Curry, D. (2023, July 20). *Music streaming app revenue and Usage Statistics (2023).* Business of Apps. https://www.businessofapps.com/data/music-streaming-market/

Petrušić, D. (2021). The Cultural Impact of Music on Society with a Special Emphasis on Consumerism. In *BCES conference books* (Vol. 19, p. 138). essay, Bulgarian Comparative Education Society.

Rastelli, S. (2023, March 20). *Spotify and YouTube.* Kaggle. https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube/data

Richter, F. (2023, October 6). *Infographic: Streaming Drives Global Music Industry resurgence.* Statista Daily Data. https://www.statista.com/chart/4713/global-recorded-music-industry-revenues/