EXPLORATORY DATA ANALYSIS OF MY SPOTIFY STREAMING HISTORY USING SQL AND TABLEAU

Mgbecheta Paschal Chukwudubem

January 2024

TABLE OF CONTENT

CONTENTS

INTRODUCTION

Music offers a unique perspective into my personality, with song choices reflecting who I am and listening patterns revealing its role in my daily life. To explore whether music is a source of motivation, a pastime, or something deeper, I decided to analyze my streaming data for clearer connections.

This project highlights my approach to extracting data, over a 9-month period, between Feb 13th 2023 to Oct 26th 2023, from my Spotify music streaming history, analyzing it with MySQL and visualizing insights in Tableau to gain a deeper understanding of my musical preferences and overall listening habits.

OBJECTIVES

Key questions I want answered are outlined below

- How much time I've spent steaming on spotify overall.
- Discover my favorite artists, tracks, and albums over the past 9 months.
- Discover peak listening times by both time of day and day of the week.
- Time changes in my music consumption over the 9-month period.
- Determine my preferred device for consuming musical content.
- Uncover patterns in how my top music preferences evolved month by month.
- Identify my music-related habits, such as reasons for ending tracks, shuffling preference

DATA COLLECTION AND PREPARATION

DATA SOURCE AND EXTRACTION

One feature I appreciate about Spotify is the ability to request a download of your streaming history via their developer platform. The data was sent to my linked email address in JSON format, which I converted to CSV using Excel. I utilized the *Extended Streaming History dataset*, which contains 21 columns, including timestamps, user details, track information, and streaming metadata, as outlined below.

- Ts - Date and time of when the stream ended in UTC format (Coordinated Universal Time zone).
- Username – My Spotify username.
- Platform - Platform used when streaming the track (e.g. Android OS, Google Chromecast)
- Ms_played - For how many milliseconds the track was played.
- Conn_country - Country code of the country where the stream was played.
- Ip_addr_decrypted - IP address used when streaming the track.
- User_agent_decrypted - User agent used when streaming the track (e.g. a browser, like Mozilla Firefox, or Safari).

- Master_metadata_track_name - Name of the track.
- Master_metadata_album_artist_name - Name of the artist, band or podcast.
- Master_metadata_album_album_name - Name of the album of the track.
- Spotify_track_uri - A Spotify Track URI, that is identifying the unique music track.
- Episode_name - Name of the episode of the podcast.
- Episode_show_name - Name of the show of the podcast.
- Spotify_episode_uri - A Spotify Episode URI, that is identifying the unique podcast episode.
- Reason_start - Reason why the track started (e.g. previous track finished or you picked it from the playlist).
- Reason_end - Reason why the track ended (e.g. the track finished playing or you hit the next button).
- Shuffle: null/true/false - Whether shuffle mode was used when playing the track.
- Skipped: null/true/false - Information whether the user skipped to the next song.
- Offline: null/true/false - Information whether the track was played in offline mode.
- Offline_timestamp - Timestamp of when offline mode was used, if it was used.
- Incognito_mode: null/true/false - Information whether the track was played during a private session.

At first glance, the data was organized into a folder with multiple distinct datasets, each representing different aspects of my Spotify personal data. My primary focus was on the dataset containing my complete streaming history.

As I reviewed the dataset, I identified several key columns that would be essential for answering my predefined questions and uncovering deeper insights. However, I also noticed a few columns that were irrelevant to my objectives, making them redundant for this analysis.

I noticed that a crucial column, "music genre", was missing from the data. This column is essential as it categorizes the genre of every track I've listened to. While resolving this issue wasn't impossible, it required sourcing the missing information from external resources.

IMPORTING DATA INTO MySQL

Given the large size of the dataset, I chose MySQL as my primary tool for data cleaning and querying to extract insights. I created a schema named "streaming history" within MySQL workspace, imported the CSV file of the "my_spotify_streaming_audio_2023" dataset as a table using the Table data import wizard option for this purpose.

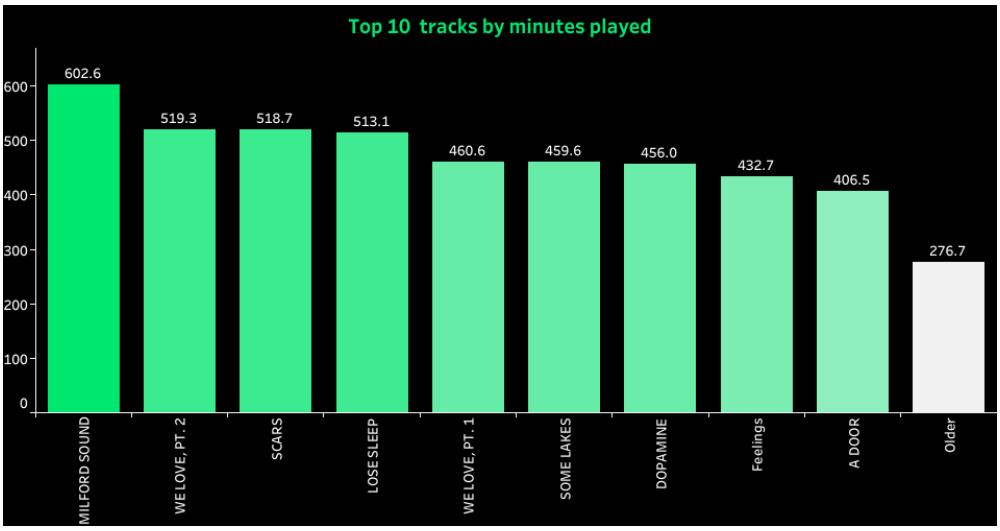
I used SQL for cleaning, formatting and a bit of analysis of the data. The code used can be found on my GitHub page, CLICK HERE

DATA VISUALIZATION USING TABLEAU

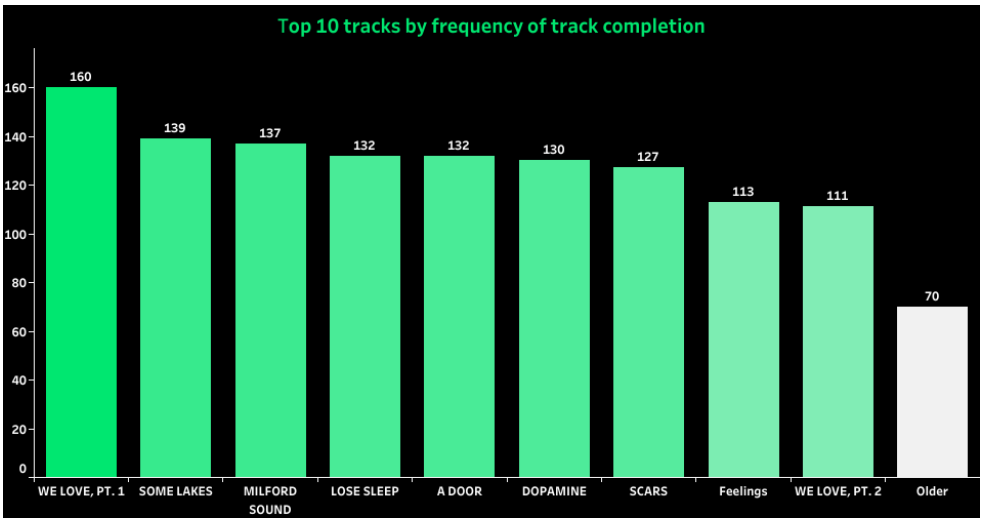
Data visualizations tell compelling stories, are visually appealing, and most importantly, reveal insights that SQL queries alone may miss. This is why I chose Tableau—it highlights trends and patterns not immediately obvious in raw data or tables.

In my visualizations, I employed various chart types, each selected for their ability to effectively communicate the insights I was aiming to uncover. Below are some key chart types and the insights they reveal:

- **TOP 10 TRACKS:** This bar chart represents my top 10 tracks by two metrics
    - total minutes played and

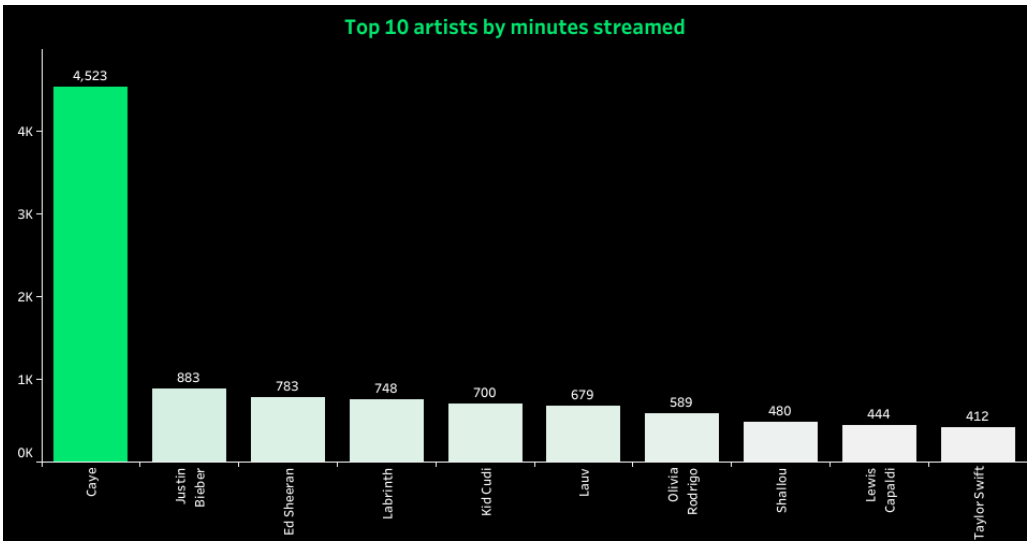


    - frequency of completion.

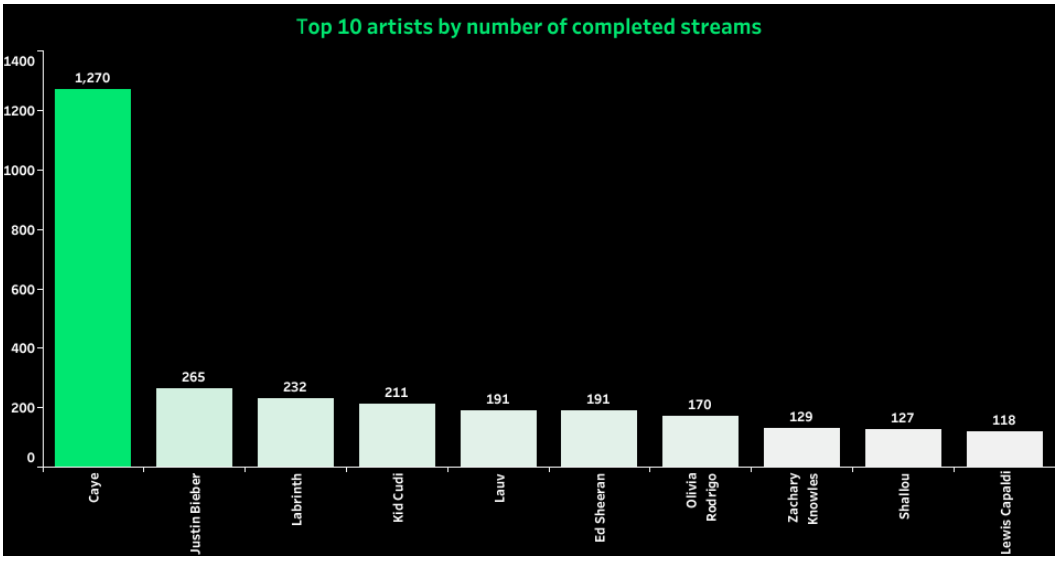


MILFORD SOUND was my top track with 602.6 minutes streamed, followed closely by *WE LOVE PT2* WITH 519.3 minutes by the total minutes played metric. *WE LOVE PT1* takes the top spot under the frequency of completion metric. Most of these top tracks were from the artist "*Caye,*" whose default genre is *POP* suggesting that pop was my dominant genre during the period.

- **TOP 10 ARTISTS:** Similarly, I visualized my favorite artists by
    - minutes played and

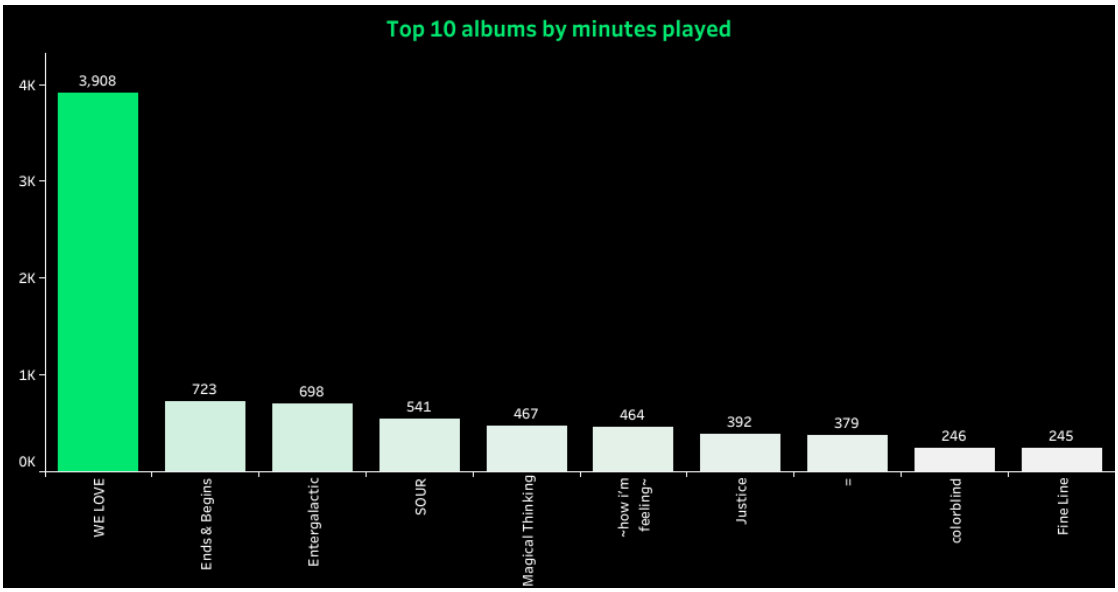


    - track completion frequency.



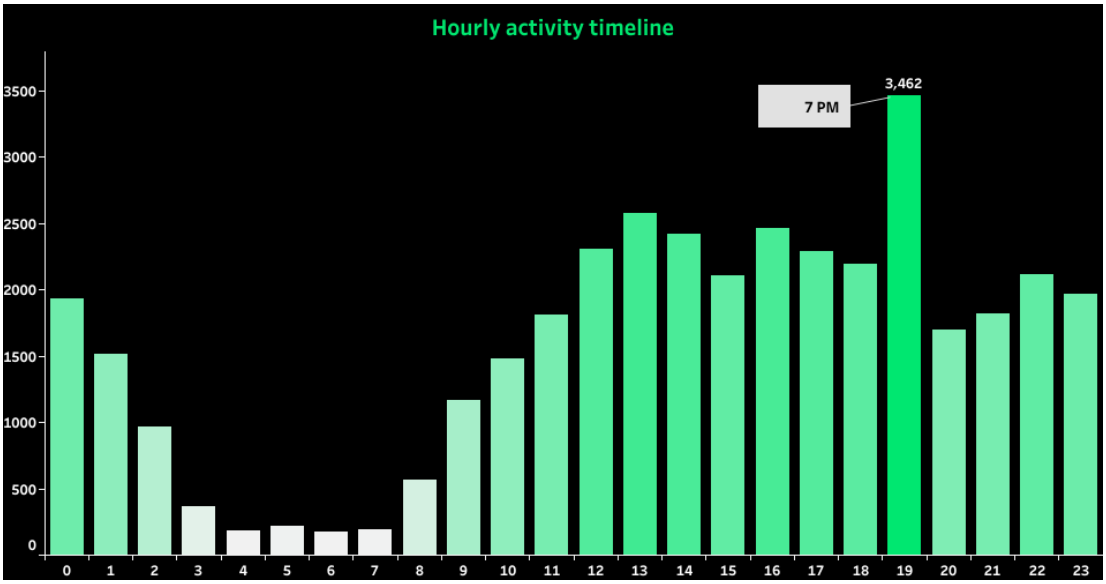
Caye again ranked highest, with 4,523 minutes played and 1270 respectively; *Justin Bieber* coming in second with 883 minutes played and 265 respectively with a majority of artists in the top 10 being pop artists, reinforcing my preference for *POP* music.

- **MOST STREAMED ALBUMS**: This is a bar chart representing my top 10 albums by total minutes played. The bars represent the albums and its corresponding height represents the total minutes spent playing the tracks it contains.

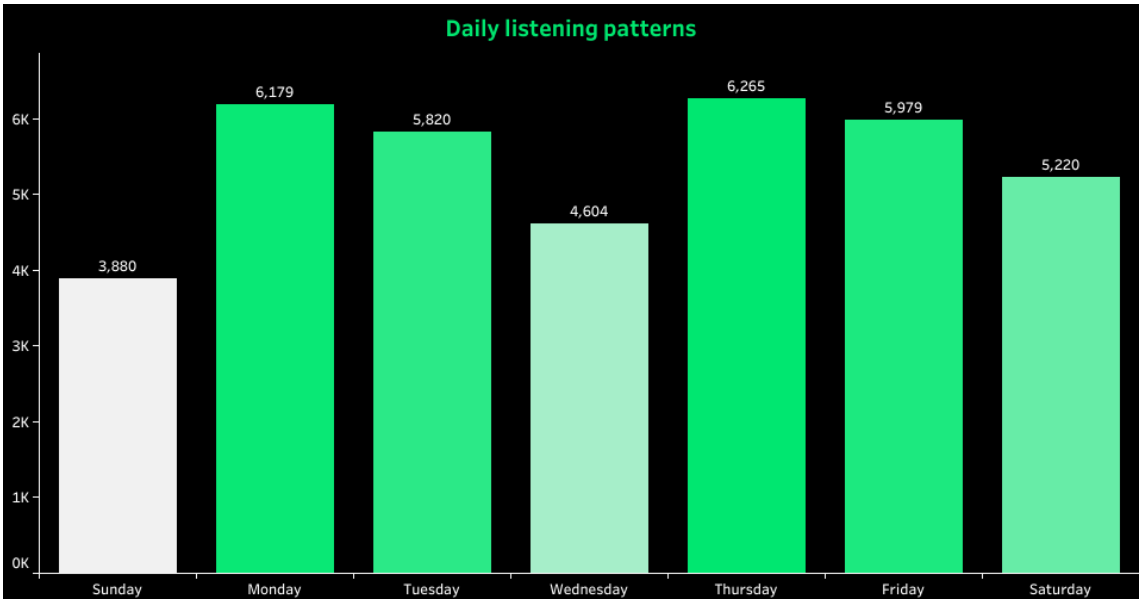


WE LOVE album by *Caye* ranks the highest in this metric, with 3,908 minutes played, solidifying my earlier hypothesis made during my "top tracks" analysis, followed by *ENDS & BEGINS* by *Labrinth* with 723 minutes played. This implies that *POP* is indeed my favorite music genre with a mix of *HipHop* largely due to the *ENTERGALACTIC* album by *Kid Kuli* coming in 3rd with 698 minutes played.

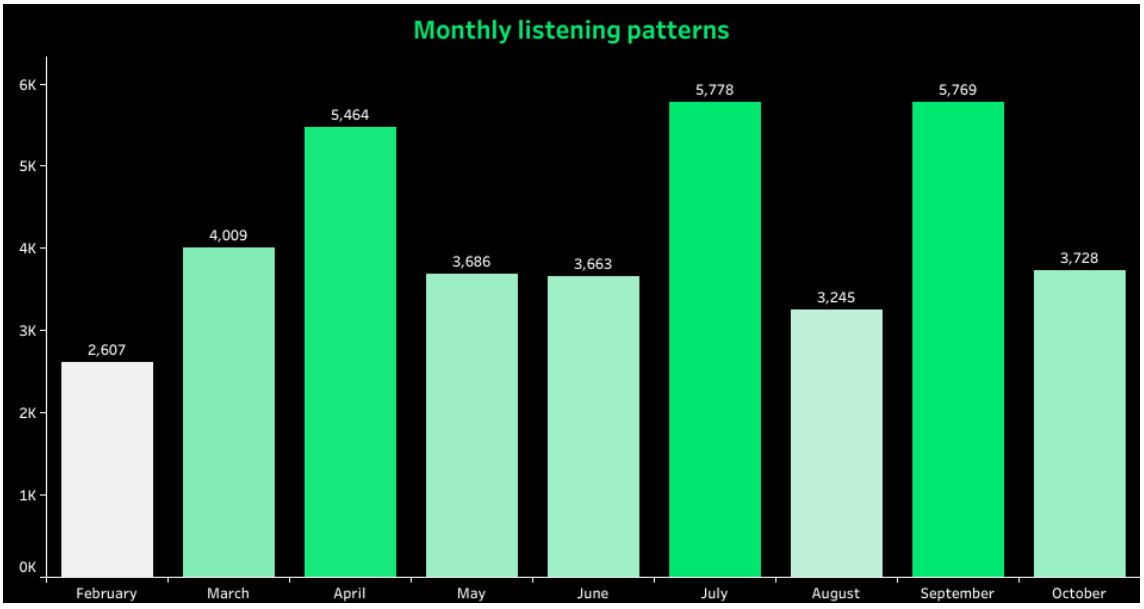
- **HOURLY ACTIVITY TIMELINE:** A bar chart showing my streaming activity across different hours of the day. It reveals that my listening peaks during the day, ranging from around 9 am in the mornings, then peaks at 1pm, 4pm before hitting its highest point at 7pm reflecting that I stream music mostly while working or during daily routines.



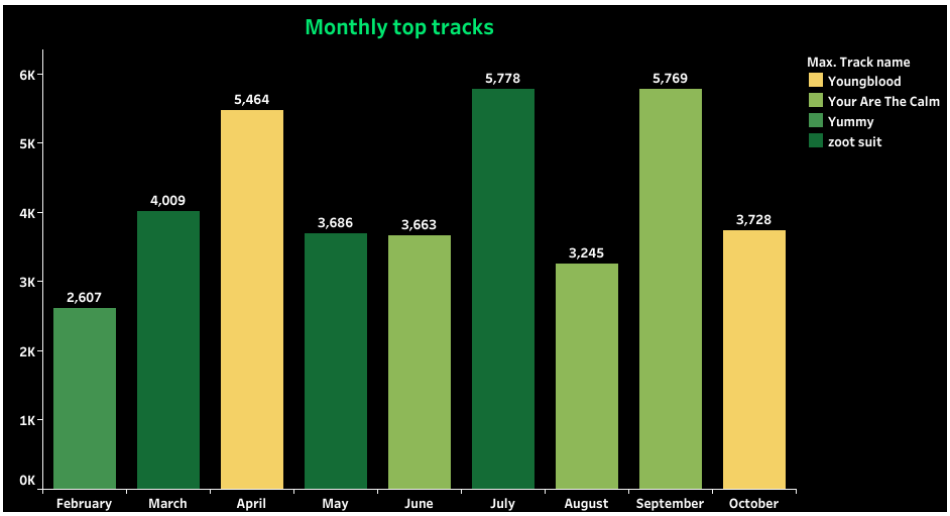
- **DAILY LISTENING PATTERNS:** Another bar chart that highlights streaming activity by day of the week. Monday and Thursday saw the highest streaming activity with 6265 and 6179 minutes streamed respectively as I listen to music while I work, while Sunday had the lowest, likely because I take time off from work on that day.



- **MONTHLY LISTENING PATTERNS:** This bar chart illustrates the variance in my monthly listening habits. Peaks were observed in April, July, and September, suggesting a quarterly trend in increased activity and February being my least active most probably because I was new to spotify as a platform at the time and needed to adapt

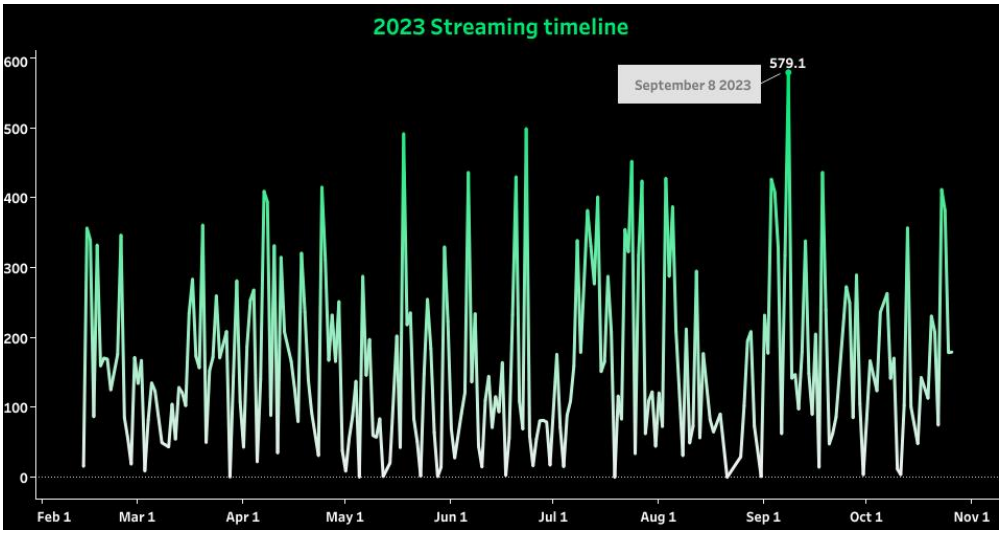


- MONTHLY TOP TRACKS: This visualization displays my most streamed artist and track respectively on a month-by-month basis based on their total minutes streamed. Each bar represents an artist or track respectively and its height is the total minute played

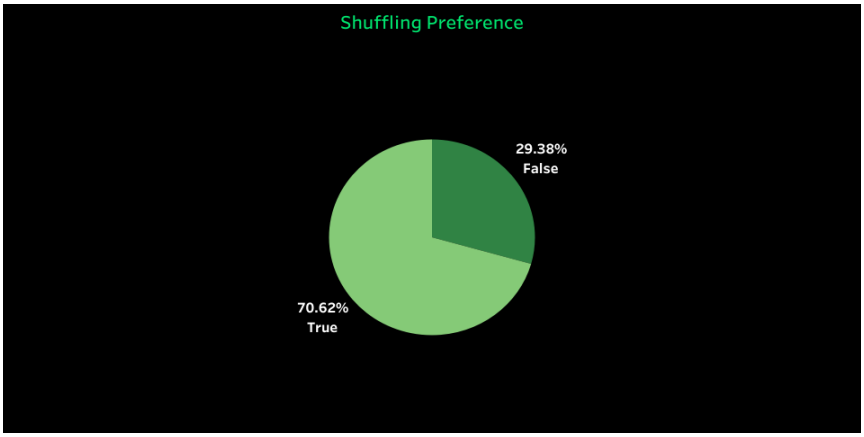


My analysis for both visuals are not the same even as their charts are identical, at a glance Z*OOT SUIT* and *YOU ARE THE CALM* are tied as the most streamed songs month over month. They are both "sleep music", songs used to induce sleep, I used to listen to soothing songs before bed on most days, this means that on these peak months is where I listened to these types of songs the most.

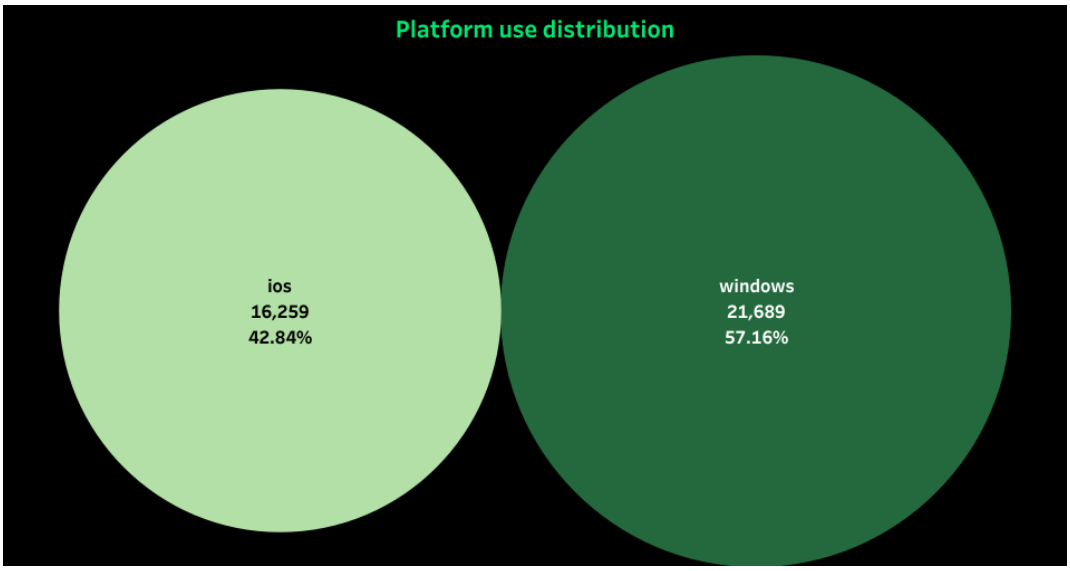
- 2023 STREAMING TIMELINE: I used a line graph to depict my streaming timeline for 2023. The chart shows a spike in activity on September 8th with 579.1 minutes streamed. When zoomed out, the data reveals an overall increase in streaming activity throughout the year, with Q3 2023 (14,793 minutes played) being the most active.



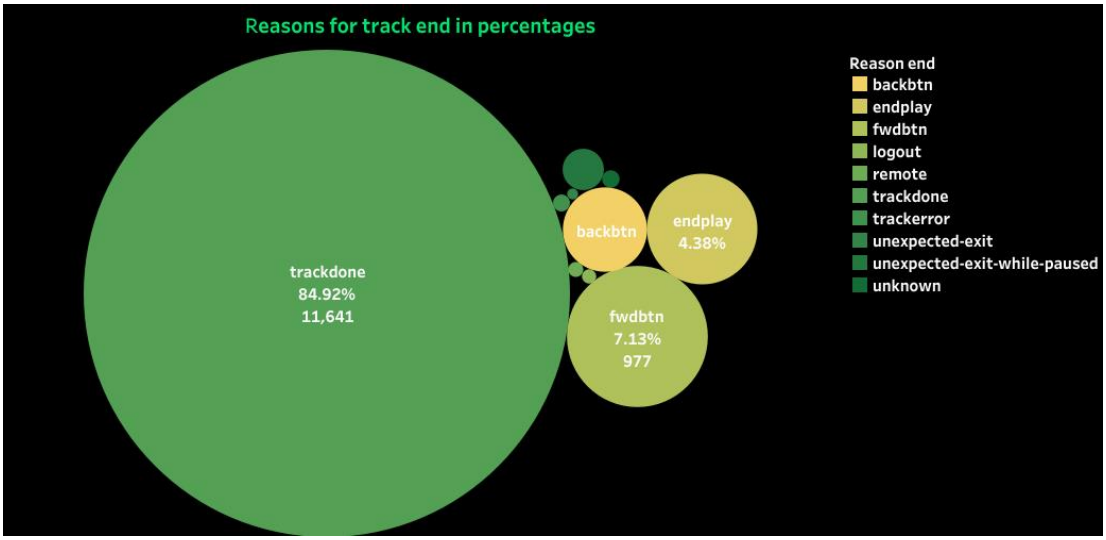
- **SHUFFLING PREFERENCES:** I used a pie chart to visualize my shuffling preferences. The data shows that I prefer listening to songs in shuffle mode, likely because I enjoy the unpredictability.



- PLATFOM USE DISTRIBUTION: I used packed bubbles to compare the devices I use for streaming. The visualization shows that I stream more on my laptop than on my phone, although the gap between the two is closing as I adapt to using the mobile app more frequently.



- REASONS FOR TRACK END IN PERCENTAGES: The packed bubble displayed the different reasons spotify documented as to why a track would end represented by the different sized and colored bubbles and the frequencies that they occurred represented by the sizes of the different bubbles.



I let songs play till completion 84.92% of the time, which means that I rarely interfere with the app once I start playing music or I put my songs on repeat a lot.

Here is a link to my Tableau Interactive Charts and Dashboards CLICK HERE

Link to the Github repository CLICK HERE

CONCLUSIONS

Analyzing my Spotify streaming data has been an eye-opening experience, challenging my assumptions and revealing new insights. While I've always known that I enjoy listening to music, I was unaware of just how much time I spend streaming. Tracking my habits for the first time has provided valuable clarity.

The most significant discovery was that pop music is my default favorite genre. I previously paid little attention to specific genres, focusing only on whether I liked a song or not. This newfound knowledge has been empowering. Additionally, I always thought of myself as an Ed Sheeran fan, but the data revealed that I listen to Caye far more often. This insight has made me rethink how I identify my favorite artists and songs.

Another surprise was learning that I primarily stream music during the day, despite considering myself more nocturnal. The data shows that I tend to listen to music while working, building projects, or browsing the web, as it helps improve my concentration.

Other interesting revelations include my tendency to listen to tracks in full, my preference for shuffling songs, and the fact that I mostly stream music from my laptop.

RECOMMENDATIONS

- While Spotify's AI already tailors my experience based on my music choices, as a data analyst, I recommend introducing more pop music from sub-genres like Afro Pop and Art Pop, as well as related genres such as rock and urban. This would add variety while staying aligned with my current preferences. Public playlists featuring a blend of these genres could also enhance the user experience

- it would be efficient to aggregate my most played or most repeated songs into a dedicated playlist for quick access.

- Using the "smart shuffle" feature rather than regular shuffle could enhance my music experience as spotify's algorithm would play only the most relevant and related tracks, introducing me to even more artists within my preferred genre.

- Being able to listen to music offline would boost retention further because streaming music would no longer be solely determined by an internet connection.

- Finally, further analysis should explore the influence of external factors like mood or exercise on my music choices, providing deeper insights into my listening habits.