Music Platform Analysis

Siddharth Das, Kevin Torres, Sulei Wang

Link to the Google Drive Code

Introduction

The main purpose of this data analysis project is to determine whether songs occupy the top 100 most popular list for a longer duration on Spotify or Youtube. The top 100 most popular songs list is based on the number of streams for that music streaming platform. This investigation will give us meaningful insights regarding platform dominance for sustained song popularity, artist longevity, audience engagement, and marketing effectiveness. Record labels, artists, and the music industry as a whole can utilize the intuition gained from this study to enhance decision-making for optimal profit and popularity. It is logical to assume that there is a positive correlation between the time spent on the top 100 list and the profits reaped from that corresponding platform. Thus, artists and record labels would have increased incentive to distribute their music to the platform that demonstrates increased song longevity. In addition, music awards shows, promoters, and sponsors are more likely to advertise artists on the music platform that indicates a more dedicated audience. We could interpret a music platform as having

increased audience engagement if the song remains in the top 100 list longer on one platform than the other.

Our hypothesis is that songs remain longer on Spotify than Youtube for the top 100 most popular songs list. This theory is based on the assumption that people find streaming music on Spotify to be more practical, rather than viewing videos on YouTube.

To complete this analysis, we will need to overcome various challenges. We will need to import a substantial amount of data, so we will need to decipher a plausible storage method. In addition, we will need to make numerous api queries to receive portions of the information at a time. Thus, we will need to set up our queries for each api to require minimal adjustments for each query call. Lastly, we may need to deal with outliers, such as Mariah Carey's Christmas album, which resurfaces yearly in the top 100 most popular songs list.

Data Acquisition & Processing

We utilized the Python requests package to make HTTP requests to various websites to access the number of streams for the top 100 most popular songs in America. Once we had the data in the JSON format, we transformed it into a Pandas DataFrames for data cleaning, exploration, and visualizations.

Initially, we planned to acquire data by utilizing the Spotify API, Twitter API, Instagram API, Google Trends, and YouTube API. The Spotify API could provide relevant information about song names, artist names, release dates, and the number of streams per song. We could use the iTunes API in a similar fashion to the Spotify API, while also obtaining the number of purchases for any given song.. We intended to utilize both the Twitter and Instagram APIs to acquire data on the number of times any given song was mentioned, and the number of people that posted about the song. This information was going to be accessed through tweets, posts, and hashtags. Google Trends was going to provide insight about the popularity of certain artists by measuring the magnitude of certain artists' online presence around the time period their song was officially released. Lastly, we intended for Youtube to provide information about the popularity of any given song by accessing the number of views and comments for relevant song videos.

However, after attempting to gain access to these data sources, we decided to adjust our approach. This was due to the fact that we encountered difficulties with the Spotify, Itunes, Twitter, and Instagram APIs. The Spotify API did not provide any relevant information about the stream count for songs. The Itunes API costs money. The Instagram API only allowed us to check 30 hashtags per week, which did not align with our goal of working with large datasets. Lastly, the Youtube API does not provide historical data on videos views, so it could not provide anything of substantial value.

As a result, we proceeded with our data collection for songs by harnessing the Music Charts Archive, Youtube Music Charts, and Spotify Music charts. The Music Charts Archive contains data for the Billboard top 50 list. The Music Charts Archive provided us with the song names, artist names, and song popularity ranks for the top 50 most popular songs per week. The YouTube Music Charts supplied us with 2 separate data sets. The first dataset consists of the artist names, artist ranks, and weekly view count for the top 100 ranked artists per week. The second dataset was similarly developed, containing the song names, YouTube channels uploaded on, rank of songs, and weekly streams for the top 100 most popular songs per week. Both the Music Charts Archive and Youtube Music Charts were accessed through Web Scraping methods. Lastly, the Spotify Music Charts contained data for the song names, artist names, number of song streams per week, and song ranks for the top 200 songs per week. Since we are mainly focusing on the top 100 songs/artists in the US, we minimized the Spotify data to solely contain relevant information for the top 100 songs/artists. Initially, we were going to implement Web Scraping to access the Spotify Data. However, since Spotify does not allow Web Scraping, we pursued an alternative Data Acquisition method by directly downloading the CSV files for each week. This was a tedious process as we downloaded a CSV file for every week from 2020 until December 2023. Once we acquired Spotify data for all the desired weeks, we utilized Pandas to merge all the weekly files into one substantial dataframe. This was completed by placing all the saved CSV files into the google drive folder where we could access all the data. We found this to be the most efficient way to process such an extensive dataset in a clear, organized manner. Since a considerable amount of our data was acquired through Web Scraping, we worked with a

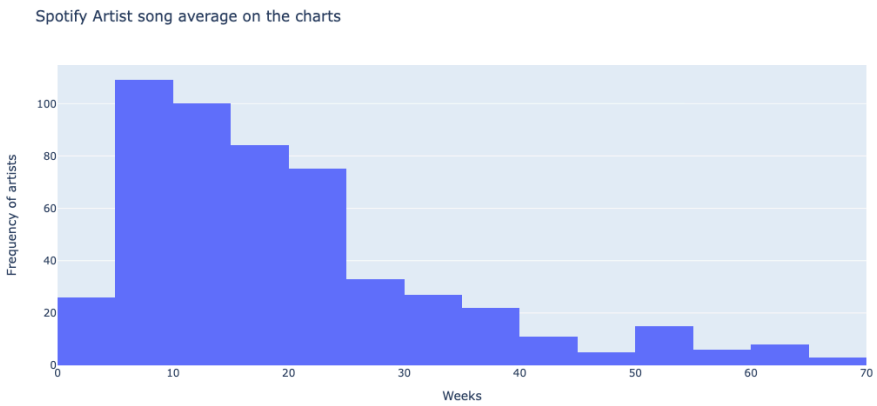
significant amount of unstructured data. This unstructured data was the JSON text that contained dictionaries, lists, etc. The challenge of accessing all these data sources required a consistent group effort split across 4+ weeks.

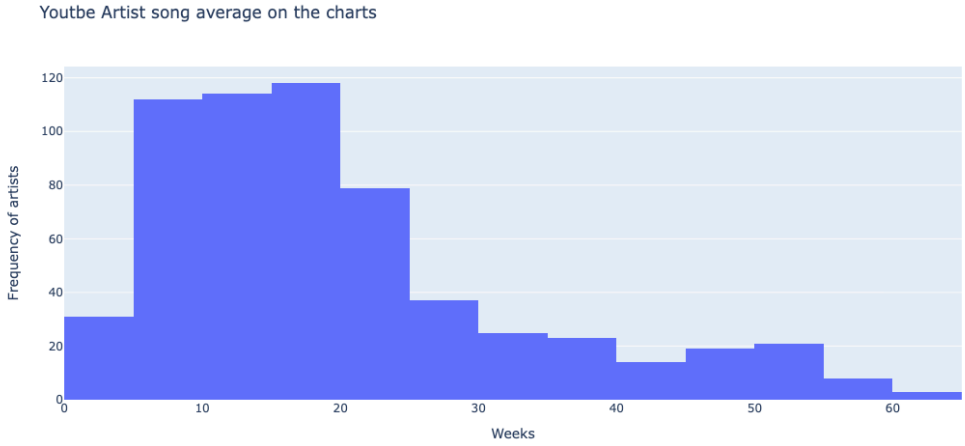
Visualization and Methodology

In the creation of our data visualizations, Plotly served as the primary tool for generating graphical representations of our findings. However, prior to plotting, several critical preprocessing steps were essential to ensure the accuracy and meaningfulness of our visualizations. Notably, the presence of outlier artists, known for seasonal hits like Mariah Carey's "All I Want for Christmas," posed a challenge, as such songs experience peaks in popularity during specific times of the year. Additionally, songs with extremely brief chart appearances, lasting only one or two weeks, had the potential to distort the overall visualization.

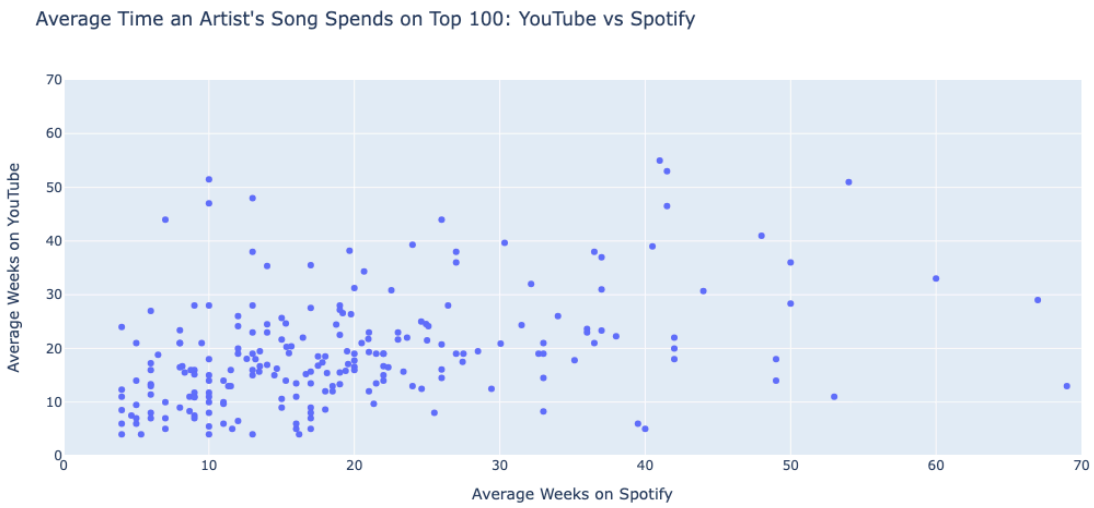
To address these challenges we filtered up the data. Initially, we grouped the data by song and artist, determining the minimum and maximum appearance dates for each song. By calculating the difference between these dates, we obtained the duration in weeks that a song remained on the chart. Subsequently, the Interquartile Range (IQR) was applied to identify and eliminate outliers, particularly songs with unusually extended durations. This step ensured that the visualization focused on more representative data.

Following the outlier removal, the data was once again grouped, this time by artist. Within this grouping, we computed the mean duration for each artist, providing a valuable metric representing the average longevity of their songs on the charts. These refined datasets were then employed in our Plotly visualizations, allowing for a more accurate and insightful depiction of the artists' chart performances.



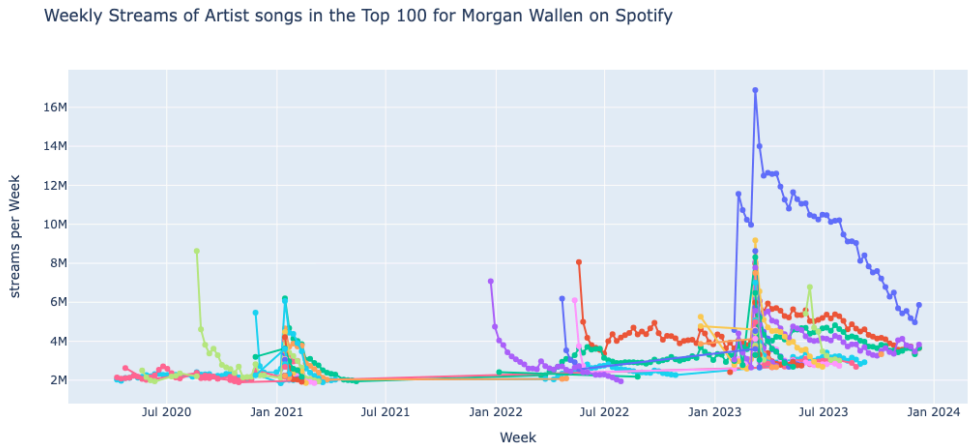


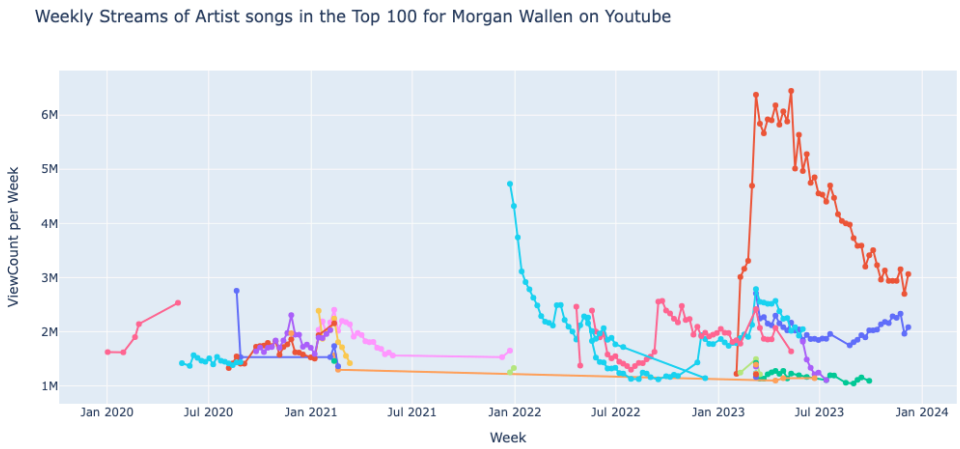
Figures 1 and 2 depict the average number of weeks an artist's song remains on the top 100 charts for their respective platforms, Spotify and YouTube. The distribution appears similar, with both platforms showing a concentration around the 10 - 30 week mark. However, Spotify exhibits a wider spread on its chart compared to YouTube. This implies that Spotify could have more artists with longer average week song time.



In Figure 3, a scatter plot illustrates the amalgamation of average weeks a song by an artist resides on the charts for both YouTube and Spotify, achieved through an inner join. The x-axis denotes the average weeks a song appears on Spotify, while the y-axis represents the corresponding duration on YouTube. The majority of points closely align with the x=y line, indicating a balanced presence on both platforms. However, a notable concentration in the lower right corner suggests that artists on Spotify generally amass significantly more streams compared to YouTube. Conversely, the upper right corner displays fewer points, suggesting a relatively lower frequency of artists garnering more streams on YouTube in comparison to Spotify.

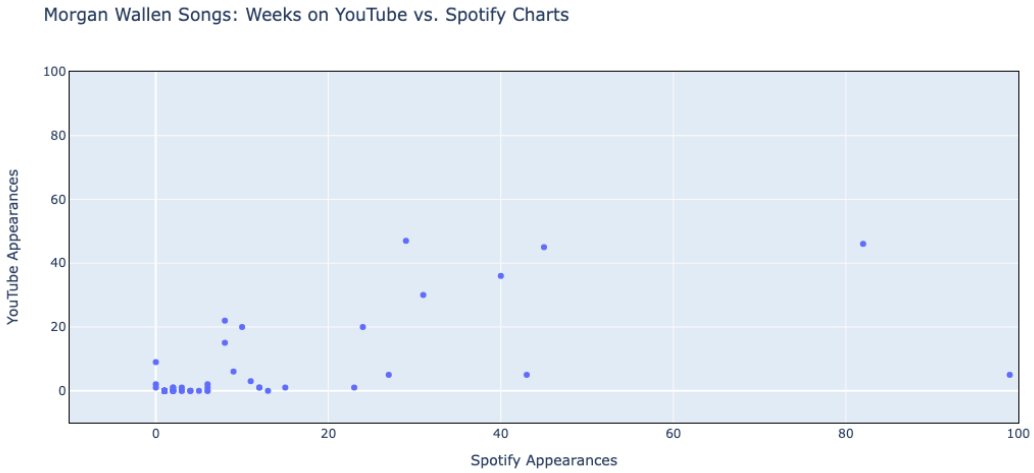
Now we are going to choose artist Morgan Wallen to see if it fits the average.





Figures 4 and 5 showcase Morgan Wallen's songs, distinguished by color, that have secured a position in the top 100 on both Spotify and YouTube. A significant distinction becomes apparent as Morgan Wallen's songs with Spotify streams notably outnumber those with YouTube streams in the top 100. This observation strongly suggests a higher likelihood for Morgan Wallen's songs to attain a position in the top 100 on Spotify compared to YouTube.

Additionally, upon closer examination, it is evident that songs on Spotify tend to have a longer duration on the charts. Towards the left of the charts, a greater number of songs with extended lines is observed, indicating that Morgan Wallen's songs endure for a more prolonged period on the Spotify chart compared to YouTube.



In the presented scatter plot of Morgan Wallen's songs, those that achieved placement in both the Spotify and YouTube top 100 charts generally exhibit comparable view counts on both platforms. However, three distinct points deviate significantly, indicating instances where Morgan Wallen's songs spent nearly 100 weeks on the Spotify chart but only a little over 10 weeks on YouTube.

Interestingly, the absence of a vice versa scenario suggests a potential trend: a hit song by an artist might experience greater success on YouTube than on Spotify. This observation raises intriguing questions about the dynamics of an artist's popularity across different streaming platforms.

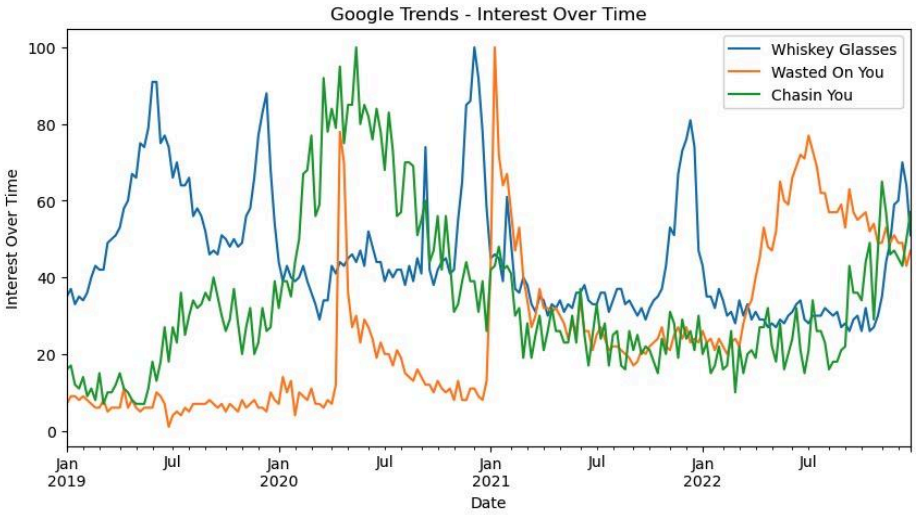


Fig 7. Interest over time from Google trends for Whiskey Glasses, Wasted on You, Chasin You

track_name	Appearances_x	FirstAppearance_x	LastAppearance_x	Appearances_y	FirstAppearance_y	LastAppearance_y	Appearance_Difference
Whiskey Glasses	99.0	2020-04-09	2023-09-07	5.0	2020-01-02	2020-05-07	94.0
Wasted On You	108.0	2021-01-14	2023-12-07	22.0	2021-01-14	2021-12-23	86.0
Chasin' You	43.0	2020-04-09	2023-08-24	5.0	2021-02-11	2023-06-22	38.0

Table 1. Appearances comparison for Whiskey Glasses, Wasted on You, Chasin You.
Appearance_x is Spotify and Apperances_y is YouTube

We examined the initial three song titles by observing the disparity in their appearance on Spotify and YouTube, subsequently analyzing their popularity trends over time. In Figure 7,

'Interest Over Time' illustrates the level of popularity for each song, allowing us to pinpoint the peak periods of popularity. Between 2020 and 2023, 'Whiskey Glasses' registered 99 appearances on Spotify but only 5 on YouTube, a trend consistent with the other two songs where Spotify consistently outperformed YouTube in terms of appearances. As depicted in Figure 7, the peak interest over time for 'Whiskey Glasses' and 'Wasted On You' aligns closely with Spotify's metrics. While 'Chasin You' shows a variance in appearance numbers between Spotify and Google Trends, the preference for searching and listening to pop music on Spotify remains evident, especially when compared to YouTube.

Conclusion

The purpose of this data exploration was to understand whether songs have a longer duration in the top 100 most popular list for Spotify or Youtube. This study intended to elucidate platform superiority for song popularity, artist longevity, audience engagement, and advertisement strategies. Our analysis leads us to believe that our initial hypothesis is true. Songs remain longer on the top 100 list for Spotify than they do for YouTube. This indicates that Spotify might have a more consistent audience. As a result, it would make sense for artists, record labels, and others in the music industry to prioritize marketing through Spotify, instead of YouTube. This could likely lead to optimal profits, popularity, and overall impact of music. Some of the challenges we faced included processing large amounts of data, and processing outliers such as Mariah Carey’s Christmas album. In the future, we intend to explore further by analyzing genres, artists, and time periods. In conclusion, it is exciting to consider how these meaningful insights can impact important business decisions in the music industry.