Data Analysis of Songs That Have Reached Over a Billion Streams on Spotify

Petra Habjanec[1]

[1]student of Computing at the University of Zagreb, Faculty of Electrical Engineering and Computing

Abstract

Purpose – The purpose of this research paper is to conduct an in-depth data analysis of songs that have achieved over a billion streams on Spotify. The analysis aims to uncover trends and insights in the music industry, including the time it took for songs to reach a billion streams, the distribution of popularity, the prevalence of collaborations, and the characteristics of highly streamed songs.

Design/Methodology/Approach – The research paper utilises a quantitative approach, analysing a dataset of songs that have reached over a billion streams on Spotify. The analysis includes statistical methods such as linear regression, polynomial regression, and descriptive statistics to examine various attributes of the songs, including duration, energy, release year, popularity, explicitness, genres, and artist popularity. The paper also includes visual representations, such as box plots, trend lines, histograms, and histograms, to enhance the presentation of findings.

Findings – The analysis reveals several key findings. It shows that songs with billion streams on Spotify span a wide range of genres and time periods, with certain genres like pop, rap, and hip hop being prevalent. The research highlights the link between collaborations and different song characteristics. The analysis also uncovers trends in song duration, energy, and explicitness. Furthermore, it identifies the artists with the most billion-stream songs and the prevalence of specific words in song titles.

Originality/Value – This research paper contributes to the understanding of the factors contributing to the success of highly streamed songs on Spotify. The analysis provides insights into the dynamics of popular music, the evolving trends in the music industry, and the preferences of listeners. The findings have implications for artists, record labels, and music streaming platforms in terms of marketing strategies, song creation, and playlist curation.

Keywords – Music streaming, Spotify, billion streams, popularity, collaborations, genres

Paper Type – Research paper.

1. Introduction

The popularisation of music streaming services has changed the way people consume music, giving people access to a vast library of songs at their fingertips. Among the platforms, Spotify has established itself as a global leader, providing a platform where both mainstream and smaller independent artists can show off their music. In recent years, the streaming industry has reached such heights that the most popular songs are songs getting over a billion streams on Spotify.

The purpose of this research paper is to conduct an in-depth data analysis of songs that have achieved this significant milestone on Spotify. By examining various attributes and characteristics of these songs, we aim to gain insight into factors contributing to their popularity.

1.1. Research focus

In the analysis of highly streamed songs, we examined the time it takes for songs to reach a billion streams, studied popularity distribution, and investigated collaborations among artists. Additionally, we explored the sequencing of songs on albums, trends in song duration and energy, the ongoing popularity of associated artists, and the change in the explicitness of songs through time. The study also includes a comparison of singles vs. album tracks and an examination of common words in song titles. With this analysis, we aim to contribute to a better understanding of the characteristics and dynamics of songs that have achieved tremendous streaming success on Spotify.

RQ1. Which songs through time got onto the list and how long it took them to get there?

RQ2. Which record labels are the most present?

RQ3. How do the songs with multiple artists compare to those with only one?

RQ4. How has the average duration and energy of popular songs changed over time?

RQ5. How popular are the artists currently?

RQ6. How do the "one-hit wonders" compare to artists with multiple popular songs?

RQ7. How has the explicitness of songs changed over time?

RQ8. How many of the songs are singles compared to those that are on albums?

RQ9. Which genres are the most prevalent?

RQ10. Which artists are the most present and how many songs do they have on the list?

RQ11. Which words are the most prevalent in song titles?

From this, we got the following hypotheses:

- ***Ha0****:* There is no significant difference in the distribution of the duration of songs with multiple artists and songs with only one artist.
- ***Ha1****:* There is a significant difference in distribution of the duration of songs with multiple artists and songs with only one artist.
- ***Hb0***: There is no significant difference in the distribution of popularity of songs with multiple artists and songs with only one artist.
- ***Hb1****:* There is a significant difference in the distribution of popularity of songs with multiple artists and songs with only one artist.
- ***Hc0****:* There is no significant difference in the distribution of danceability of songs with multiple artists and songs with only one artist.
- ***Hc1****:* There is a significant difference in the distribution of danceability of songs with multiple artists and songs with only one artist.
- ***Hd0***: There is no significant difference in the distribution of energy of songs with multiple artists and songs with only one artist.
- ***Hd1***: There is a significant difference in the distribution of energy of songs with multiple artists and songs with only one artist.
- ***He0****:* There is no significant difference in the distribution of loudness of songs with multiple artists and songs with only one artist.

- ***He1***: There is a significant difference in the distribution of loudness of songs with multiple artists and songs with only one artist.
- ***Hf0***: There is no significant difference in the distribution of speechiness of songs with multiple artists and songs with only one artist.
- ***Hf1***: There is a significant difference in the distribution of speechiness of songs with multiple artists and songs with only one artist.
- ***Hg0***: There is no significant difference in the distribution of valence of songs with multiple artists and songs with only one artist.
- ***Hg1***: There is a significant difference in the distribution of valence of songs with multiple artists and songs with only one artist.
- ***Hh0:*** The average track ratio is equal or or greater than 1/3.
- ***Hh1:*** The average track ratio is less than 1/3.

In this analysis, an array of different metrics about individual songs was extracted from Spotify API. Popularity, with the value from 0 to 100, which is mostly determined by algorithm based on the total number of plays and how recent the plays are. Acousticess, spanning from 0.0 to 1.0, is a confidence measure, offering insights into whether the song has electrical amplification. Danceability, in contrast, explains the rhythmical realm, accessing how danceable the track is depending on the tempo, rhythm stability, and beat strength, where 0.0 represents the least danceable and 1.0 the most danceable. Energy, measured on a scale of 0.0(least energetic) to 1.0(most energetic), provides a measure intensity and activity of the track. It considers the dynamic range and perceived loudness, whereas energetic tracks are fast-paced, and loud and have dynamic musical elements. Instrumentalness indicates the likelihood of vocal content on the track. Values close to 1.0 signify a greater probability of instrumental dominance, distinguishing the track as predominantly devoid of vocals. The key is represented by integers following *Pitch Class* notation, while mode, indicates major(1) or minor(0). Liveness describes whether the track is a live performance, with values above 0.8, suggesting a strong likelihood of a live recording. Loudness, expressed in dB, provides an averaged perspective across the entire track. Speechiness indices the presence of spoken words, with values of 0.66 signaling predominantly spoken content. Tempo reflects the estimated beats per minute, defining the pace and rhythm of the track. Time signature, ranging from 3 to 7, estimates, the number of beats in each bar. Finally, valence, ranging from 0.0 and 1.0, paints how emotional the track is, conveying its musical positiveness. Higher valence indicates a more positive emotional tone. Together, these metrics will provide us with a good understanding of the auditory journey within a track.

2. Related work

The analysis of highly streamed songs on music streaming platforms has been a topic of interest for researchers and music industry professionals. Several studies have explored the musical characteristics of highly streamed songs.

Nijkamp (2018) analysed the 1000 songs from Spotify API from different genres. They used regression, to build a prediction model, and concluded that audio features from Spotify have little to moderate explanatory power for higher stream count. Also, Luo (2018) analysed the audio features and genres of top-ranking songs on Spotify in 2017. They used machine learning that predicts genres for songs based on the song's audio features. Lastly, Nugroho, Manongga, and Purnomo (2023) analysed the trend of Spotify API's song analysis of the most popular songs over the years of the COVID-19 pandemic, compared to before it.

3. Methodology

Our data was retrieved in November of 2023 from the Spotify playlist BILLION CLUB which contains all of the songs that have more than one million. At the time the playlist contained 509 songs. Our data set contains tracks with the timestamp of when they got added to the playlist, the name of the album the track is on, the release date of the album, total number of tracks on the album, genres, a list of the artists on the track, list of country codes where the track is available, duration of the track in ms, name of the song, the popularity of the song, track number, a measure of danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and time signature. Based on this data set we aim to answer questions(*RQ1-RQ11*) and test the hypotheses(*Ha-Hh*).

3.1. Summary statists of our data set

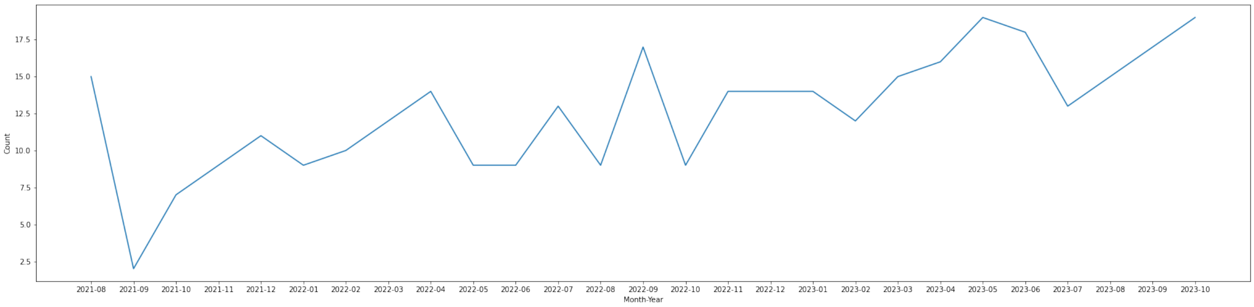
The summary statistics of our data set is shown in Table 1. In the table, we can see that most of the song durations lie between 3mins 15s and 4mins. Where the 3 shortest ones are *Everybody Dies In Their Nightmares*, *Hope* and *Jocelyn Flores* all by XXXTENTACION, and the three longest ones are *Te Boté* by Nio Gacia, Casper Magico, Bad Bunny, Darell, Ozuna and Nicky Jam, *Hotel California - 2013 Remastered* by Eagles and *Nothing Else Matters(Remastered)* by Metallica. We can also see that the popularity of most of the songs lies between 78 and 86, with *Cruel Summer* by Taylor Swift, *Seven(feat. Latto) (Explicit Ver.)* by Jung Kook and Latto, and *I Wanna Be Yours* by Arctic Monkeys as the most popular, and *Bohemian Rhapsody - 2011 Remastered* by Queen, *Roar* by Katy Perry, and *Bad Romance* by Lady Gaga as least popular. For danceability majority of the tracks are situated between 0,578 and 0,755. The most danceable ones are *The Real Slim Shady* by Eminem, *WAP* by Cardi B and Mega Thee Stallion, *Another One Bites The Dust - Remastered 2011* by Queen, and with least danceable ones being *All I Want* by Kodaline, *Fix You* by Coldplay and *Dusk Till Dawn - Radio Edit* by ZAYN and Sia. The energy of the most popular songs falls between 0,533 and 0,773. Where *Welcome To The Jungle* by Guns N' Roses, *Hey Ya!* By Outkast, and *Promiscuous* by Nelly and Timbaland are the most energetic, and *when the party's over* by Billie Eilish, *Say Something* by A Great Big World and Christina Aguilera, and *White Christmas -1947 Version* by Bing Crosby, Ken Darby Singers, and John Scott Trotter & His Orchestra as least energetic. For key most of the songs are between D and G#. Loudness of most streamed songs falls between -7,463 and -6,283. *Africa* by Toto, *White Christmas - 1947 Version* by Bing Crosby, Ken Darby Singers, and John Scott Trotter & His Orchestra, and *Revenge* by XXXTENTACION stand out as the quietest, and *Hey Ya!* by Outkast, *FRIENDS* by Marshmellow and Anne-Marie, and *What Makes You Beautiful* by One Direction as loudest. For speechiness majority of the songs are situated between 0,038 and 0,090. Some of the most speechy songs are *Life Is Good* by Future and Drake, *Youngblood* by 5 Seconds of Summer, and *Panda* by Desiigner, and some of the least speechy songs are *Perfect* by Ed Sheeran, *The Scientist* by Coldplay, and *Set Fire to the Rain* by Adele. In the table, we can see that the acousticness mostly falls between 0,035 and 0,310. *Smells Like Teen Spirit* by Nirvana, *Thunderstruck* by AC/DC, *Come As You Are* by Nirvana being the least acoustic,*when the party's over* by Billie Eilish, *The Night We Met* by Lord Huron, and *Bruises* by Lewis Capaldi being most acoustic. For instrumentalness, most of the values fall into 0,000, which means not instrumental at all. The ones with the greatest values of instrumentalness are *everything i wanted* by Billie Eilish, *White Christmas - 1947 Version* by Bing Crosby, Ken Darby Singers, John Scott Trotter & His Orchestra, and *Better* by Khalid. Values of liveness lie between 0,093 and 0,193. The tracks with the lowest values of liveness are *Flowers* by Miley Cyrus, *Uptown Funk* by Mark Ronson and Bruno Mars, and *Cake By The Ocean* by DNCE, and the tracks with highest values of liveness *Rap God* by Eminem, *Dancing Queen* by ABBA, and *The Box* by Roddy Ricch where their values of liveness are less than 0,799 which is less than 0,8 what is a Spotify defined limit of track actually being live. The valence of most of the tracks falls between 0,327 and 0,665. With *September* by Earth, Wind & Fire, *There's Nothing Holdin' Me Back* by Shawn Mendes, *Pumped Up Kicks* by Foster The People having the highest value of valence, and *Falling* by Harry Styles, *HIGHEST IN THE ROOM* by Travis Scott, and *Lose Yourself* by Eminem having the lowest value. The tempo of a majority of the tracks is situated between 98,007 and 136,041. But the ones that pop out are *changes* by XXXTENTACION, *Make You Feel My Love* by Adele, and *I Wanna Be Yours* by Arctic Monkeys with the lowest values, and *FourFiveSeconds* by Rihanna, Kanye West, and Paul McCartney, *Animals* by Maroon 5, and *Back In Black* by AC/DC with highest values.

Table 1: Summary statistics of out data set

Variable	Min.	1st Quartile	Median	Mean	3rd Quantile	Max
Duration(in ms)	95466,000	195373,000	19573,000	219344,306	241693,000	417920,000
Popularity	45,000	78,000	83,000	81,782	86,000	99000,000
Danceability	0,188	0,578	0,674	0,660	0,755	0,949
Energy	0,111	0,533	0,662	0,642	0,773	0,987
Key	0,000	2,000	5,000	5,267	8,000	11,000
Loudness	-18,064	-7,463	-5,866	-6,283	-4,662	-2,261
Speechiness	0,023	0,038	0,054	0,090	0,099	0,481
Acoustiness	0,000	0,035	0,117	0,219	0,310	0,978
Instrumentalness	0,000	0,000	0,000	0,008	0,000	0,657
Liveness	0,023	0,093	0,116	0,169	0,193	0,799
Valence	0,059	0,327	0,479	0,498	0,665	0,979
Tempo	64,934	98,007	117,996	118,961	136,041	205,846

Time signature	1,000	4,000	4,000	3,961	4,000	5,000

It should be noted that in this analysis the Spotify API does not provide a release date for a specific song, so the release date of an album is used as also the release date of a song. We also got the data for all the artists and albums for songs. From there we got the data for labels, popularity of artists, and genres. Since the data was taken from the playlist which was made on the 21st of July 2021, for the dates before that we can't know when the songs achieved a billion streams, therefore those are excluded from any analysis including the time of when a billion streams was achieved.

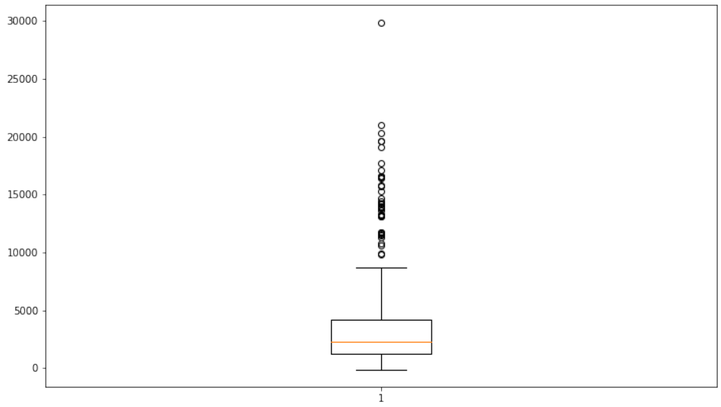


3.2. Songs though time that got onto the list and the time it took them to get onto the list

In this chapter, we will look at the distribution of songs that got a billion streams through months and years(*RQ1)*. In Image 1 we can see that from October of 2021, there's a constant adding of around 10 songs each month, but in the recent few months that number has increased.

Image 1: Graph of number of songs through months and years

We have also analysed how long it takes for songs to get the achievement. Out of 349



songs(the ones that got a billion streams after the 21st of July 2021). On average it took

3,913.994 days(10 years and 264 days) for songs to reach the achievement but a standard deviation of 4,496.741(12 years and 116 days). The median stands at 2,268 days(6 years and 78 days), 1st quartile at 1,214(3 years and 119 days) and 3rd quartile at 4,427(12 years and 47 days). We can also see that data visualised in Image 2, in the box plot it is even more clearer how many outliers there are.

Image 2: Box-plot of days it took for songs to achieve billion steams

Talking about the outliers in Table 2 we can see the songs that have the lowest difference between the release date of the album and reaching a billion streams on Spotify. Most notable is Flowers with a negative value, but as mentioned before we are looking at album release dates, not song release dates, and nowadays artists release one or a few songs from the album before the release of the actual album. So Miley Cyrus' Flowers reached a billion streams even before the release of the album it's on. It should also be noted that Olivia Rodrigo has two songs with some of the shortest times between release date and reaching billion streams. Both of the songs, *drivers licence* and *good 4 you* are on the same album *SOUR*. It is also interesting that only two songs that reached a billion streams in 2023 are also on the list, those being *Flowers* by Miley Cyrus and *Seven(feat. Latto)(Explicit Ver.)* by Jung Kook and Latto.

Table 2: Songs how have a shortest time from release date and reaching billion streams

Name	Artist	Added at	Album release date	Difference in days
Flowers	Miley Cyrus	2023-05-04	2023-08-18	-106
MONTERO (Call Me By Your Name)	Lil Nas X	2021-10-04	2021-09-17	17
As It Was	Harry Styles	2022-07-28	2022-05-20	69
drivers license	Olivia Rodrigo	2021-08-11	2021-05-21	82
Seven (feat. Latto) (Explicit Ver.)	Jung Kook, Latto	2023-10-30	2023-07-14	108
STAY (with Justin Bieber)	The Kid LAROI, Justin Bieber	2021-11-04	2021-07-09	118
Bad Habits	Ed Sheeran	2022-03-31	2021-10-29	153
good 4 u	Olivia Rodrigo	2021-10-26	2021-05-21	158
Kill Bill	SZA	2023-06-05	2022-12-09	178
Enemy (with JID) - from the series Arcane League of Legends	Imagine Dragons, JID, Arcane, League of Legends	2022-12-27	2022-07-01	179

Looking at Table 3 we can examine songs that have the longest time from the release of the album to achieving billion streams. It is fascinating how the oldest song with a billion streams is a Christmas song from the 40's, and that the song achieved a billion streams just in the begging of autumn. We can also see how the older songs that have achieved the milestone are mostly from ‘classic’ bands or artists like Queen, Fleetwood Mac, ABBA, The Beatles, and The Rolling Stones. The performers we maybe don’t listen to usually, but still know either their names or even songs. Those are the artists and songs that have become timeless and established themselves as everlasting.

In summary, our analysis of songs getting over a billion streams on Spotify reveals evolving trends in music releases and highlights the enduring appeal of timeless classics. The monthly distribution showcases a recent surge in additions. Notable outliers, including pre-album releases and entries from 2023, add nuance to our findings.

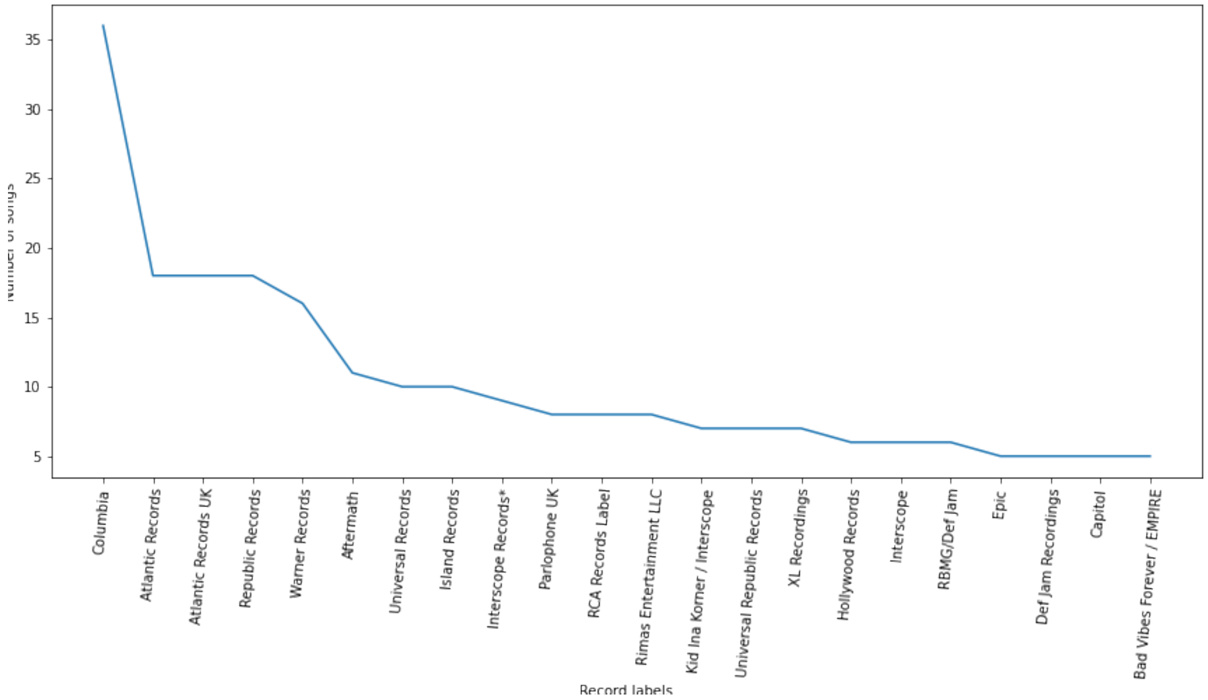
Table 3: Songs how have a longest time from release date and reaching billion streams

Name	Artist	Added at	Album release date	Difference in days
We Will Rock You - Remastered 2011	Queen	2023-02-24	1977-10-28	16555
Dreams - 2004 Remaster	Fleetwood Mac	2022-06-27	1977-02-04	16579
Dancing Queen	ABBA	2023-07-11	1976-10-11	17074
Sweet Home Alabama	Lynyrd Skynyrd	2022-11-07	1974-04-15	17738
Have You Ever Seen The Rain	Creedence Clearwater Revival	2023-03-02	1970-12-07	19078
Fortunate Son	Creedence Clearwater Revival	2023-06-13	1969-11-02	19581
Here Comes The Sun - Remastered 2009	The Beatles	2023-05-09	1969-09-26	19583
Ain't No Mountain High Enough	Marvin Gaye, Tammi Terrell	2023-05-17	1967-08-29	20350
Paint It, Black	The Rolling Stones	2023-10-19	1966-04-15	21006
White Christmas - 1947 Version	Bing Crosby, Ken Darby Singers, John Scott Trotter & His Orchestra	2023-09-27	1942-01-01	29854

3.3. Analysis of record labels with most songs with billion steams

In this section, we will identify the most popular record labels based on the number of songs they have that have reached more than a billion streams on Spotify(*RQ2)*. In Image 3 we can see that Columbia is by far the most popular one, with Atlantic Records, Atlantic Records UK, and Republic Records following with almost half the number of songs.

Image 3: Record labels with number of their songs

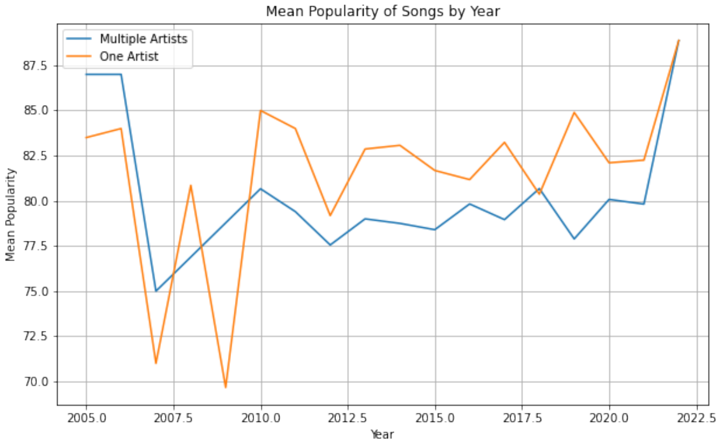


3.4. The comparison between songs with only one artist vs. the songs with multiple artists

This section examines the prevalence of songs with multiple artists and compares their popularity to songs with only one artist(*RQ3*). It utilises histograms, Mann-Whitney U tests, and graphs to analyse the mean popularity over the years. This analysis can provide insights into the impact of collaborations on the popularity of songs and the trends in the music industry.

Firstly, we looked at the mean popularity of the songs of the two groups over the years(Image 4). The most interesting period is from 2010 and onwards. There's a trend of songs with one artist being more popular. With the biggest difference in 2019 with a difference of 7.004 where songs with one artist are more popular, and the smallest in 2018 with a difference of 0.315 where songs with multiple artists are more popular. This is particularly interesting because logically, if a song has multiple artists, the song would be listened to by both of their audiences. Therefore song would have a wider reach and bigger audience and would be more popular. But we can also look at it from a different perspective. If two artists making a song together make music in different genres, the song they make together does not appeal to either of their audiences. We have excluded 2023 from this statistic because, as already mentioned, only two songs released in 2023 reached billion streams.

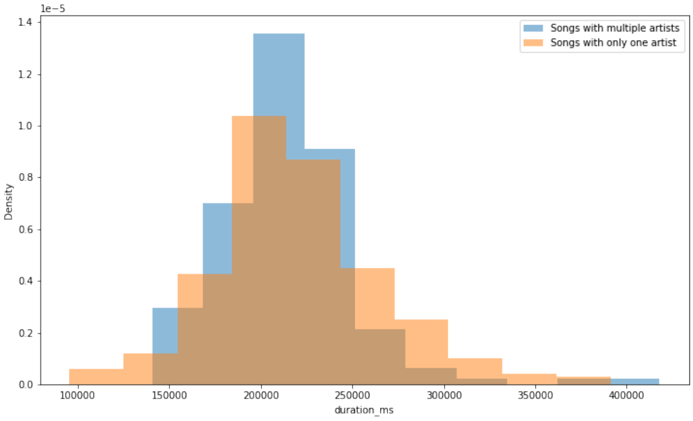
Image 4: Plot of mean popularity of songs with multiple artists and one artist



Next, we observed numerical values of songs with multiple and only one artist, plotted them into histograms to visualise the data better, and then we tested the hypotheses(*Ha - Hg*) with the Mann-Whitney U test.

- ***Ha0****:* There is no significant difference in the distribution of the duration of songs with multiple artists and songs with only one artist.
- ***Ha1****:* There is a significant difference in the distribution of the duration of songs with multiple artists and songs with only one artist.

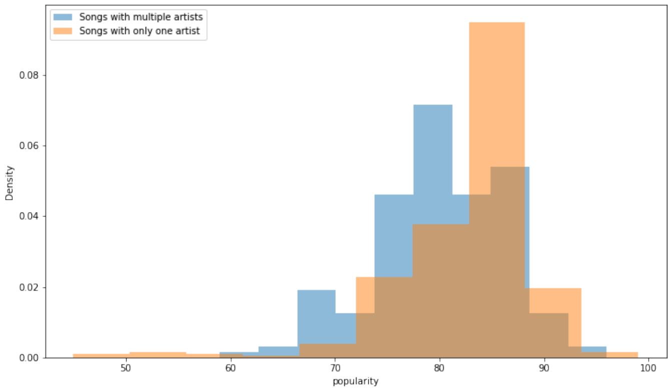
Image 5: Density histogram of duration of songs with multiple artists vs. songs with one artist



In Image 5 we can see that the graph is a bit more narrow and that the duration of songs with multiple artists has a bit less dispersion in its values. We perform the Mann-Whitney U test, and we get a statistic with the value 26,847.5 and a p-value of 0.2088. On the level of significance α=0.05 we fail to reject the null hypothesis. So we conclude that there's no difference in the distribution of duration of songs with multiple artists and songs with only one artist.

- ***Hb0***: There is no significant difference in the distribution of popularity of songs with multiple artists and songs with only one artist.
- ***Hb1**:* There is a significant difference in the distribution of popularity of songs with multiple artists and songs with only one artist.

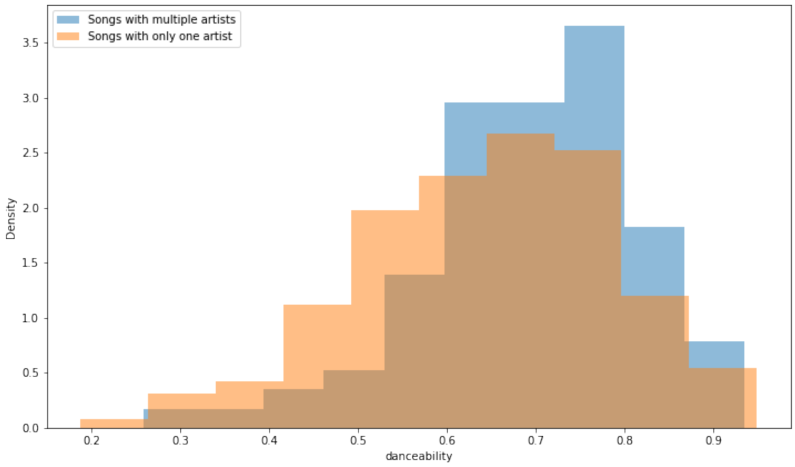
Image 6: Density histogram of popularity of songs with multiple artists vs. songs with one artist



In Image 6 we can see a histogram of popularity of songs with multiple artists vs. songs with only one artist. We test the hypotheses *Hb0* vs *Hb1* with Mann-Whitney U test and we get the statistic 20,490.0 and p-value of 9.895e-08. On level α=0.05 we reject the null hypothesis. So we can conclude that there is a significant difference in the distribution of popularity between songs with multiple artists and songs with only one artist.

- ***Hc0****:* There is no significant difference in distribution of danceability of songs with multiple artists and songs with only one artist.
- ***Hc1****:* There is a significant difference in distribution of danceability of songs with multiple artists and songs with only one artist.

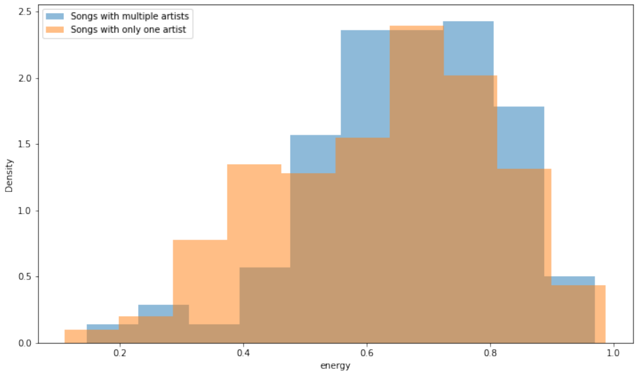
Image 7: Density histogram of danceability of songs with multiple artists vs. songs with one artist



Observing Image 7 we can observe the density of danceability of songs with multiple artists vs. the songs with only one artist. We test the hypotheses *Hc0* and *Hc1* with the Mann-Whitney U test, from which we get the values of statistic 35,433.000 and p-value 2.35e-05. From these values, we can reject the *Hc0* on significance level α=0.05 and conclude that there's a significant difference in the distribution of danceability between songs with multiple artists and songs with only one artist.

- ***Hd0***: There is no significant difference in the distribution of energy of songs with multiple artists and songs with only one artist.
- ***Hd1***: There is a significant difference in the distribution of energy of songs with multiple artists and songs with only one artist.

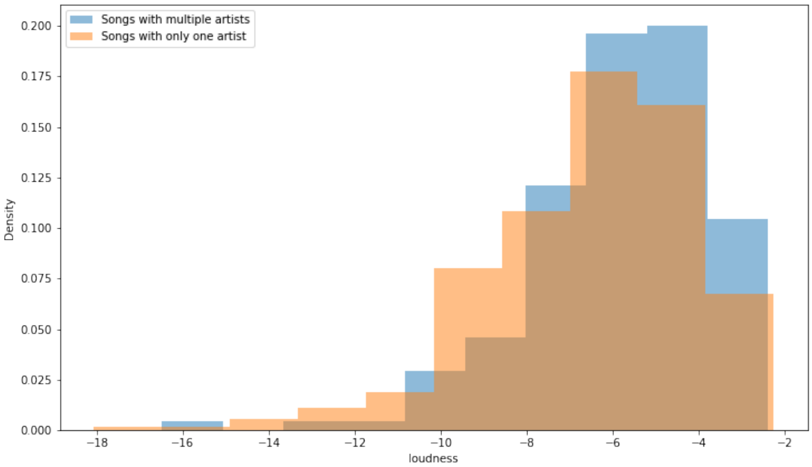
Image 8: Density histogram of energy of songs with multiple artists vs. songs with one artist



In the visual depiction of Image 8, we can see the distribution of energy of songs with multiple artists vs. the songs with only one artist. We test the hypotheses with the Mann-Whitney U test. From the test, we get the statistic of 32,684.5 and a p-value of 0.013. On the significance level of α=0.05, we can reject the null hypothesis *Hd0* and conclude that there is a significant difference in the distribution of energy between songs with multiple artists and songs with only one artist.

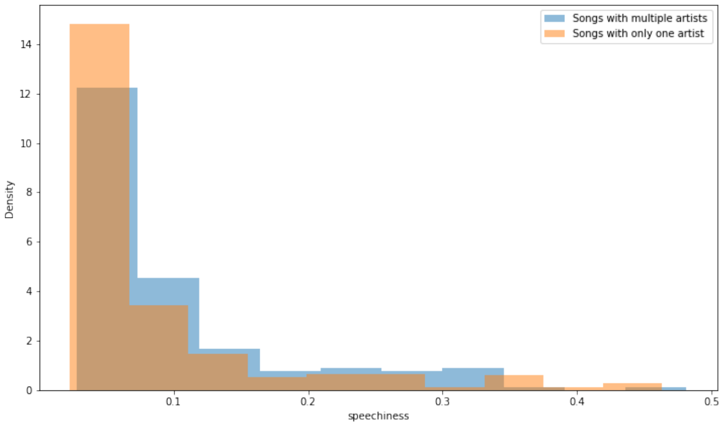
- ***He0****:* There is no significant difference in the distribution of loudness of songs with multiple artists and songs with only one artist.
- ***He1***: There is a significant difference in the distribution of loudness of songs with multiple artists and songs with only one artist.

Image 9: Density histogram of loudness of songs with multiple artists vs. songs with one artist



Notably, the graph in Image 9 depicts the distribution of loudness of songs with one artist compared to songs with multiple artists. We, again, use the Mann-Whitney U test to test the hypotheses *He,* and we get the statistic of 33,519.5 and a p-value of 0.003. From these values, we can reject the *He0,* at the level of significance α=0.05, and conclude that there's a significant difference in the distribution of loudness between songs with multiple artists and songs with only one artist.

Image 10: Density histogram of speechiness of songs with multiple artists vs. songs with one artist

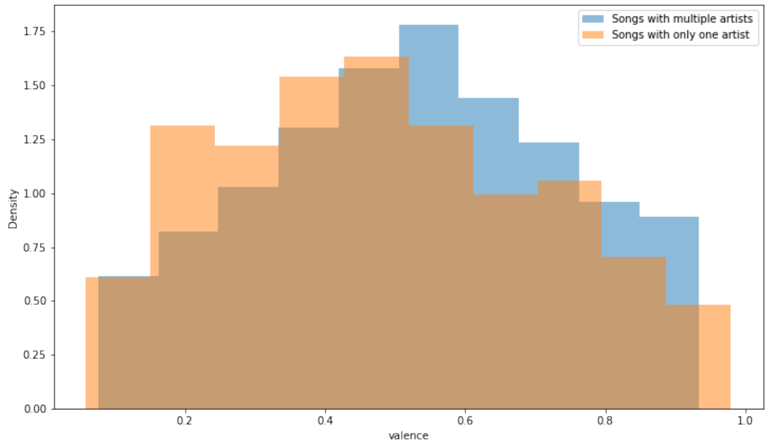


- ***Hf0***: There is no significant difference in the distribution of speechiness of songs with multiple artists and songs with only one artist.
- ***Hf1***: There is a significant difference in the distribution of speechiness of songs with multiple artists and songs with only one artist.

Evident from the chart in Image 10 we can see the distribution of speechiness between songs with multiple artists and songs with only one artist. We use the Mann-Whithey U test to test the *Hf0* and *Hf1*, and we get the statistic 35,040.0 and a p-value of 6.97e-05. With those values, we can reject *Hf0* at the level of significance α=0.05. We conclude that there is a significant difference in the distribution of speechiness between songs with multiple artists and songs with only one artist.

- ***Hg0***: There is no significant difference in the distribution of valence of songs with multiple artists and songs with only one artist.
- ***Hg1***: There is a significant difference in the distribution of valence of songs with multiple artists and songs with only one artist.

Image 11: Density histogram of valence of songs with multiple artists vs. songs with one artist



Depicted in the graph in Image 11 we see a graph of the distribution of valence between songs with only one artist and songs with multiple artists. We use the Mann-Whitney U test to get the statistic 31,794.0 and p-value 0.057. At the level of significance α=0.05 do not reject the *Hf0,* and conclude that there's no difference in the distribution of valence between songs with multiple artists and songs with only one artist.

3.5. Where on the album are the songs mostly?

Firstly, we have to take only the songs that are on the albums, which leaves us with 414 songs. We constructed a data frame with percentages of where on the album songs are, and as we can see on the histogram in Image 12 more than half of the songs are in the first third of their respective albums. On the box plot in Image 13, we can also see how the median is 0,300, 1 quartile is 0,154 and 3rd quartile is 0,571.

- ***Hh0:*** The average track ratio is equal to or greater than 1/3.
- ***Hh1:*** The average track ratio is less than 1/3.

We perform a one-sample t-test on the data to test the hypotheses *Hh0* and *Hh1*. By performing the test we get the values of statistic 3.597 and a p-value of 0.00036, with 413 degrees of freedom. With that, we can reject the null hypothesis *Hh0* at the level of significance α=0.05. So, we conclude that the average track ratio is significantly less than 1/3.

Image 12: Histogram of count of songs by where they are on the album

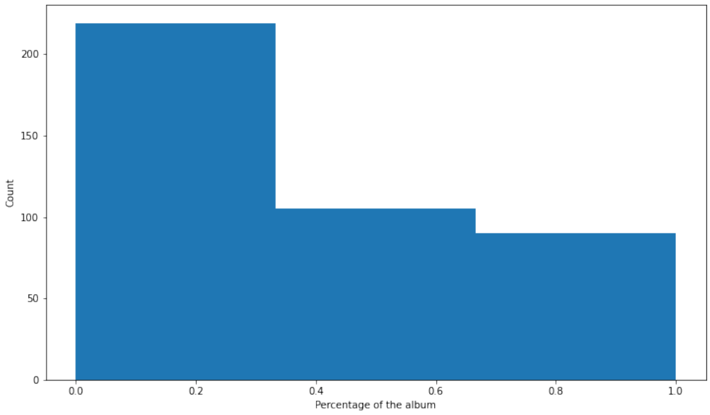
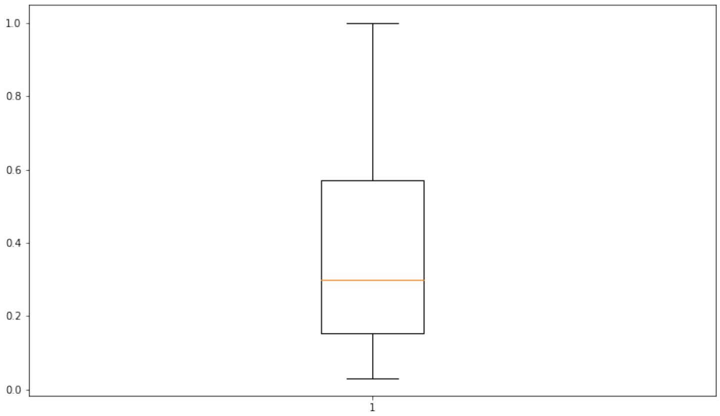


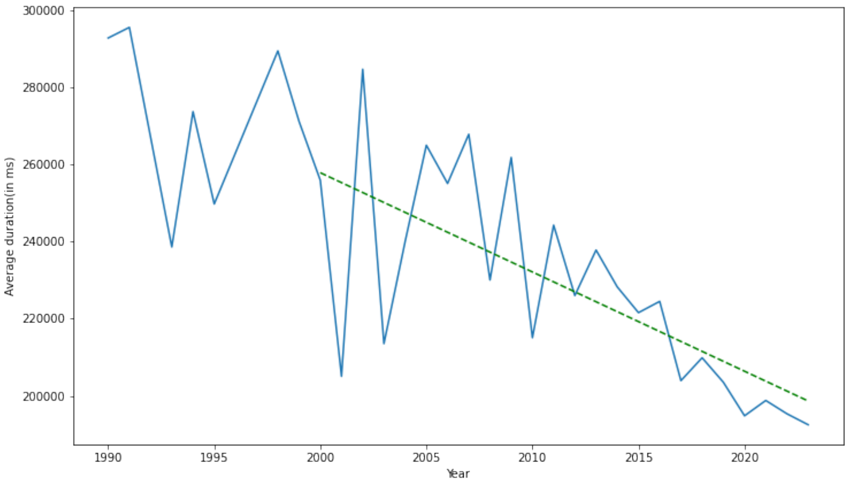
Image 13: Box plot for percentages of songs on the album



3.6. Trends in song duration and energy

In this section, we will investigate the relationship between the duration and energy of songs over time(*RQ4*). We utilise linear regression and examine the average energy and average song duration over the years.

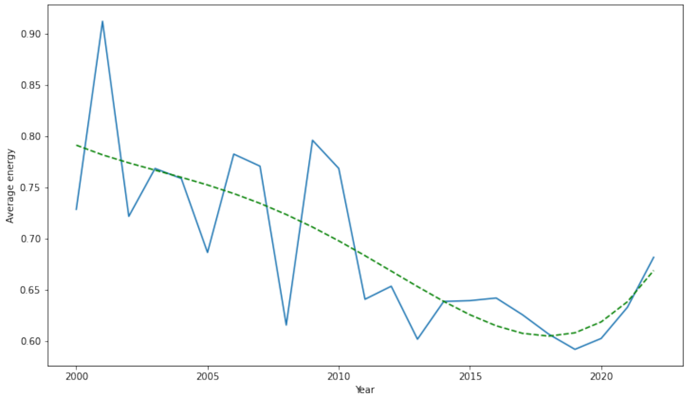
Image 14: Trend of average duration of songs through years, with a linear regression line



In this study, we used a simple method called linear regression to see how the duration of songs changes over the years. The results showed that the average duration is going down, with a noticeable decrease of -2,576.37 milliseconds per year. To check how well our method explains the changes, we used a determination coefficient, and it came out to be 0.487. This means our method explains about 48.7% of why song durations are changing. While that's quite good, it also suggests that there are likely other reasons influencing song durations that our method doesn't consider. For a visual representation, please refer to Image 14.

Secondly, a polynomial regression model of degree 4 was used to get the association between the release year of songs and their average energy levels. The resulting coefficients, [5.64040642e-06, -4.52949648e-02, 1.36401309e+02, -1.82558675e+05, 9.16253299e+07], result in a complex polynomial equation of the trend in average energy over the years. The graphical representation is shown in Image 15. The R-squared score was computed as 0.595. This score suggests that the polynomial regression model does a good job explaining about 59.5% of the changes on average over the 24 years. In the graph, we can see that there's a clear point of change in trend of average energy in songs. With 2018 as the lowest point with an average energy of 0.605, and the average energy turning upwards after that point.

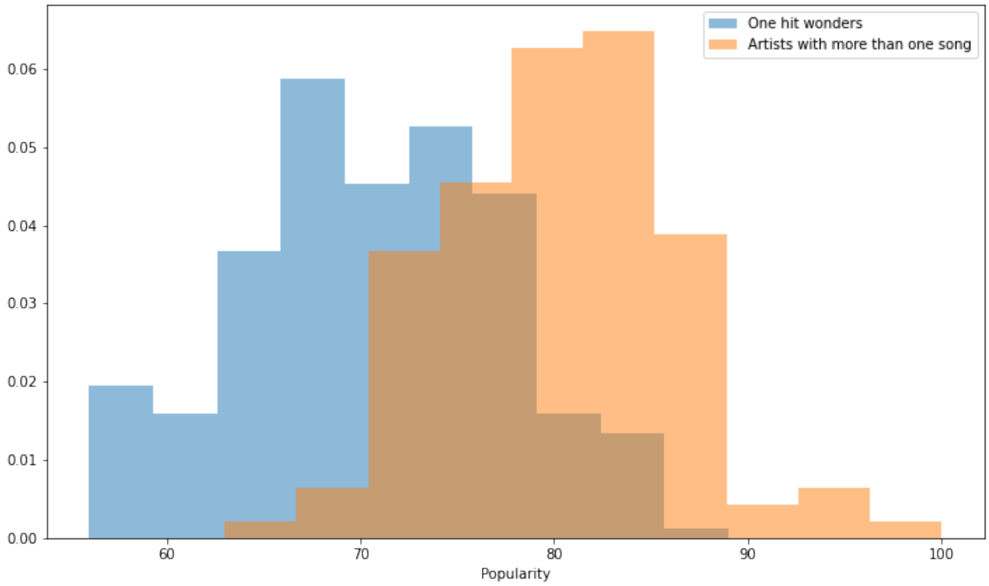
Image 15: Trend of average energy of songs through years, with a linear regression line



3.7. Comparing popularity between “one-hit wonders” and multi-hit artists

Here we analysed of artists’ popularity, distinguishing between "one-hit wonders" and those with multiple hit songs(*RQ6*). For “one-hit wonders”, having 248 artists, the mean popularity score stands at approximately 70.83, with a standard deviation of 6.72. The range extends from a minimum of 56 to a maximum of 89, and the interquartile range (IQR) spans from the 25th percentile (66) to the 75th percentile (75). In contrast, artists with more than one hit song, totaling 125 artists, exhibit a notably higher mean popularity of around 80.15, accompanied by a standard deviation of 6.02. The popularity scores for this category range from a minimum of 63 to a maximum of 100, with an IQR between the 25th percentile (76) and the 75th percentile (84). This analysis, coupled with visualisation in the plot in Image 16, provides insights into the distinct popularity patterns between “one-hit wonders” and artists with lasting success, helping us understand the dynamics of longevity in the music industry.

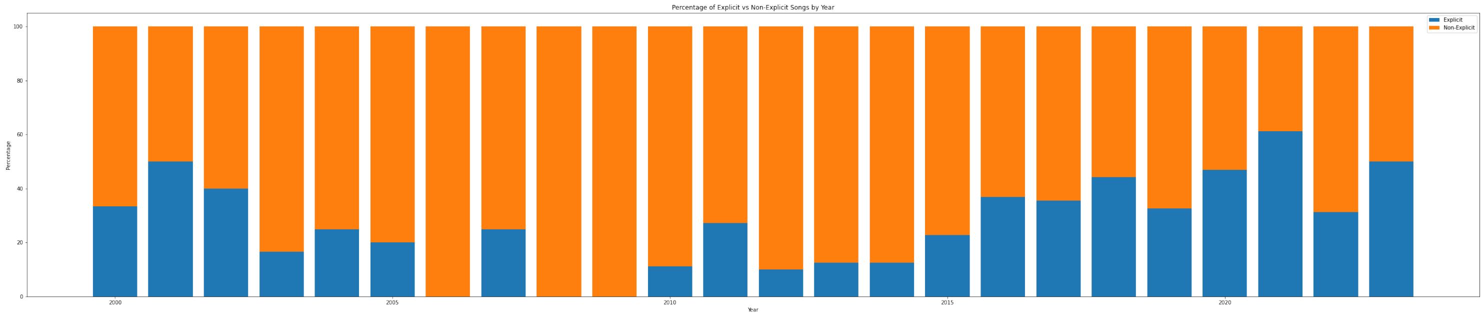
Image 16: Popularity distribution comparison between “one-hit wonders” and multi-hit artists



3.8. Examining the popularity of explicit songs over time

Following, we took a deeper look into involving prevalence of explicit songs over the years(*RQ7*), examining the percentage representation of explicit and non-explicit content within the most streamed songs. The data reveals fluctuating trends, providing a view into social attitudes and musical preferences. In the early 2000s, explicit songs held a relatively lower percentage, with notable increases in the following years. The explicit percentage peaked in 2021 at 61.29%, signifying a potential shift in the acceptance or production of explicit content. On the other hand, the percentage of non-explicit songs displayed a dynamic pattern, reaching its peak in 2006, 2008 and 2009 at 100%, suggesting a contrasting trend. The accompanying Image 17 visually presents a graph depicting the changing percentages over time, offering a comprehensive illustration of the evolving landscape of explicit content in music.

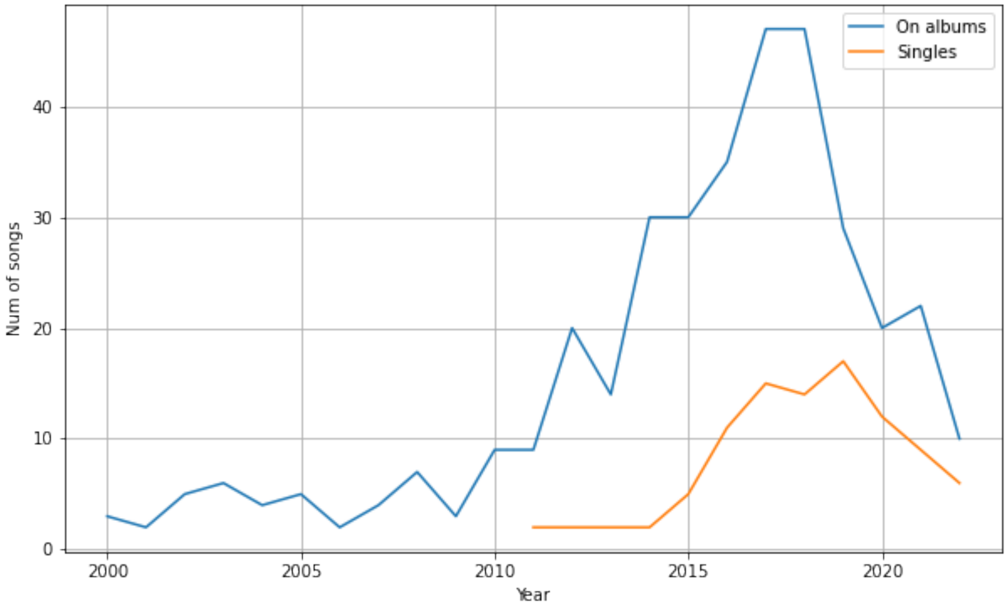
Image 17: Percentages of explicit songs through years



3.9. Singles vs. albums trends over the years

In this section, we will take a closer look into the yearly production of songs on albums versus the single(*RQ8*). The data reveals a dynamic pattern in the number of songs from albums, hitting its peak in 2018 and 2017 with 47 songs released each year. Yet, the count of singles demonstrates more variability, reaching its peak at 15 in 2017. To better understand these trends from 2000 to 2022, refer to Image 18, where a graph does a good visual representation of the data.

Image 18: Plot of count of singles and songs from albums released through years



3.10. What music genres dominate in hits

Genres, as defined by Spotify for songs with over a billion streams, give us a peek into the different kinds of popular music(*RQ9*). Looking at the graph in Image 19, it’s clear that pop is the most prevalent genre, with a lead of 298 songs falling into this category. Right behind it, we have genres like rap, dance pop, and hip hop, all contributing to the mix. There are some more specific genres like Canadian contemporary R&B, Miami hip-hop, and Barbadian pop, but also some main ones like rock, pop, rap, and hip-hop.

Then, we decided to look at the presence of ‘main’ genres so we counted the presence of pop for example as well as pop dance, and dance pop, for hip-hop we counted in Miami hip-hop, etc. With that change, we got the graph shown in Image 20. In this analysis, we can see that pop is even more prevalent with 385 songs. In this graph, we can see metal, soul, k-pop, country, grunge, and indie, which were not present before. Trap has surpassed EDM, with 65 songs, while EDM has only 39.

Here we can explore the diversity of the music on the streaming platform and different tastes, that all combine into a unique worldwide music scene.

Image 19: Top 30 Spotify genres and their count

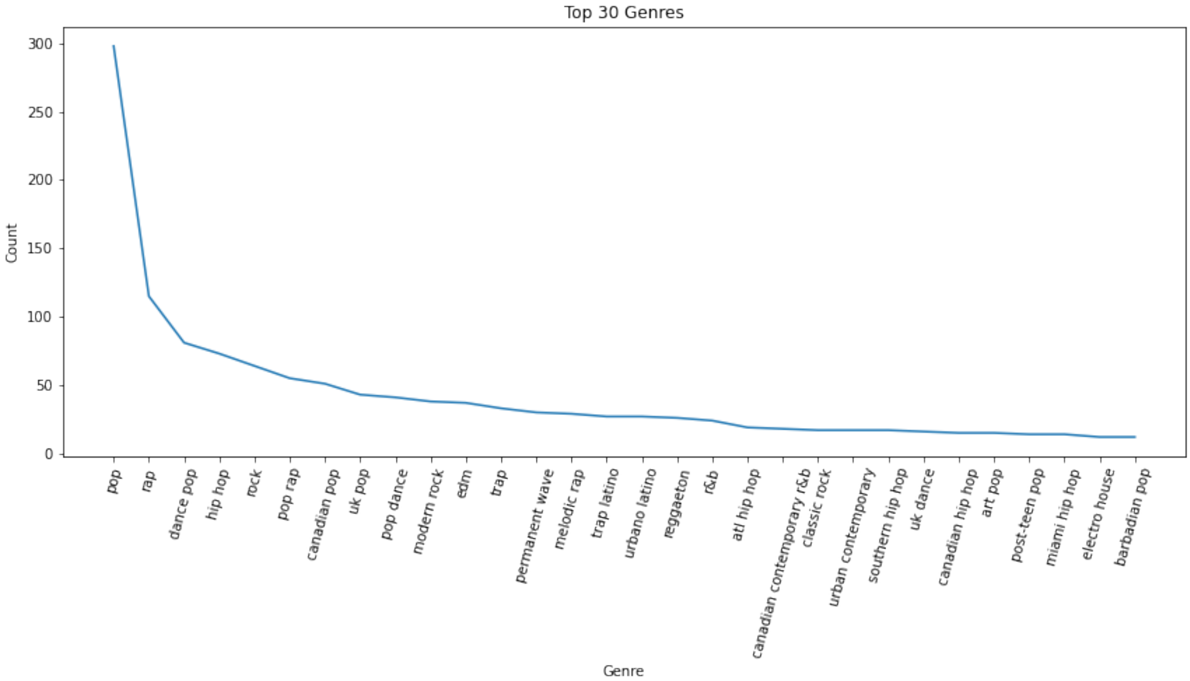
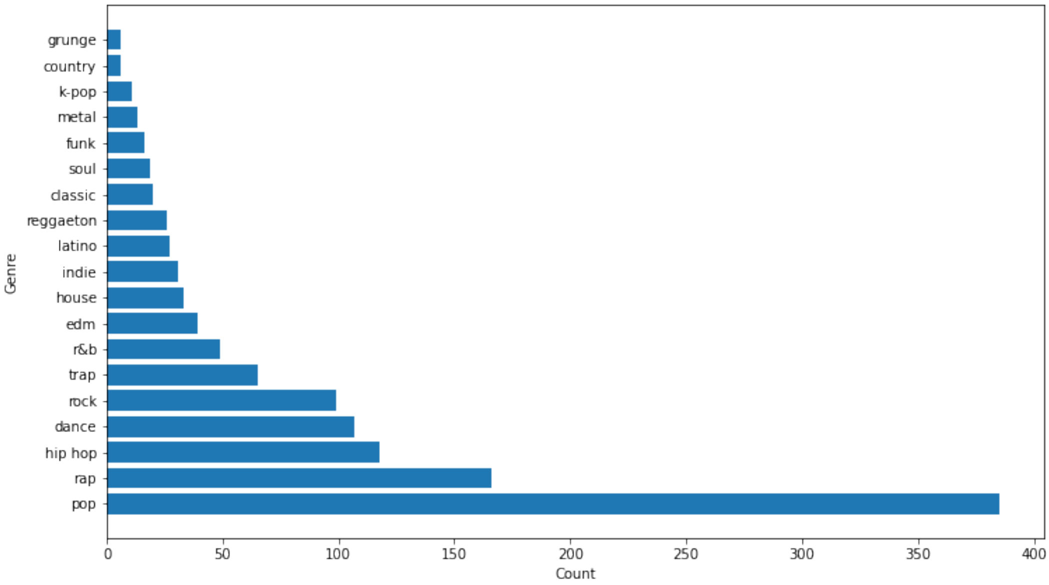


Image 20: Most prevalent 'main' genres and their count



3.11. Artists with the most billion-stream songs and how popular they are

This graph(on Image 21) shows the artists who have achieved remarkable success in the realm of billion-stream hits on Spotify*(RQ10)*, showcasing the most prolific contributors to this exclusive club. Justin Bieber leads with 14 songs that have surpassed a billion streams, followed closely by Drake with 13, Rihanna and Bad Bunny with 12 each, and Ariana Grande and The Weeknd with 11 each. The list further includes renowned artists such as Ed Sheeran, Post Malone, Bruno Mars, and XXXTENTACION, each boasting 10 songs that have garnered over a billion streams. Diverse in genres and styles, these artists have left a significant mark on the global music scene, attaining widespread popularity and recognition for their contributions to the billion-stream club on Spotify.

This section presents a statistical description of the popularity of artists based on a dataset of popularity scores(*RQ5)*. The data reveals a diverse distribution, with a mean popularity score of approximately 73.95. The histogram in Image 22 visually represents the spread of artist popularity, showcasing how scores are distributed across the dataset. The minimum popularity is 56, the maximum is 100, and the majority of artists fall within the interquartile range (IQR) between the 25th and 75th percentiles. The standard deviation of 7.85 indicates a moderate level of variability around the mean.

Some of the most popular artists with songs streamed more than a billion times include Taylor Swift(100), Drake(95), Bad Bunny(95), The Weekend(93), Travis Scott(90), contrary to some of the least popular artists are Jack Ü(56), Bipolar Sunshine(57), Nyla(57), Jawsh 685(57) and WATT(57).

Image 21: Most prevalent artists and the number of their songs with over billion streams(with more than 5 songs)

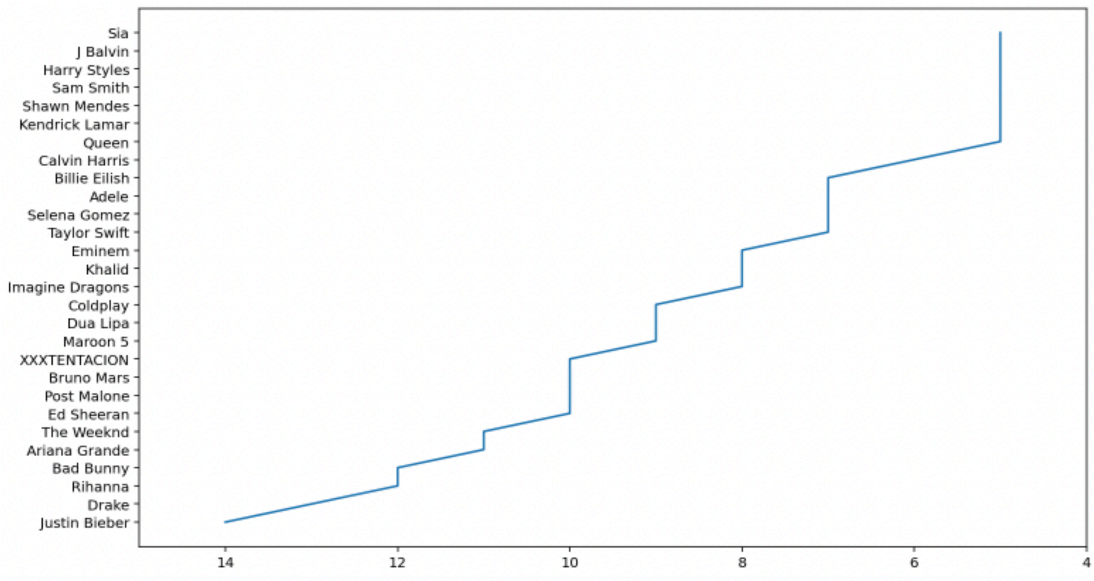
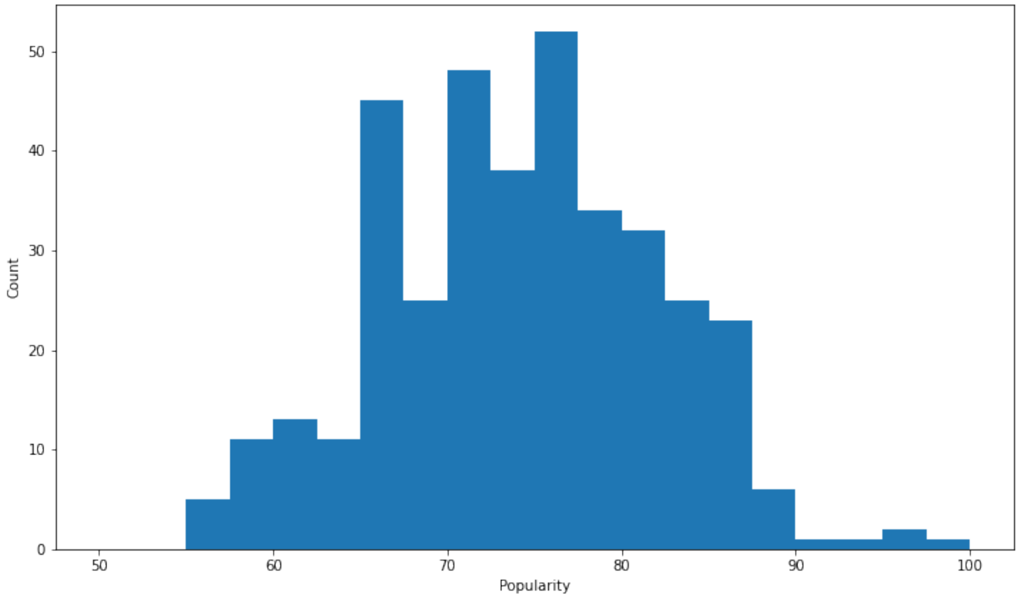


Image 22: Popularity distribution of artists

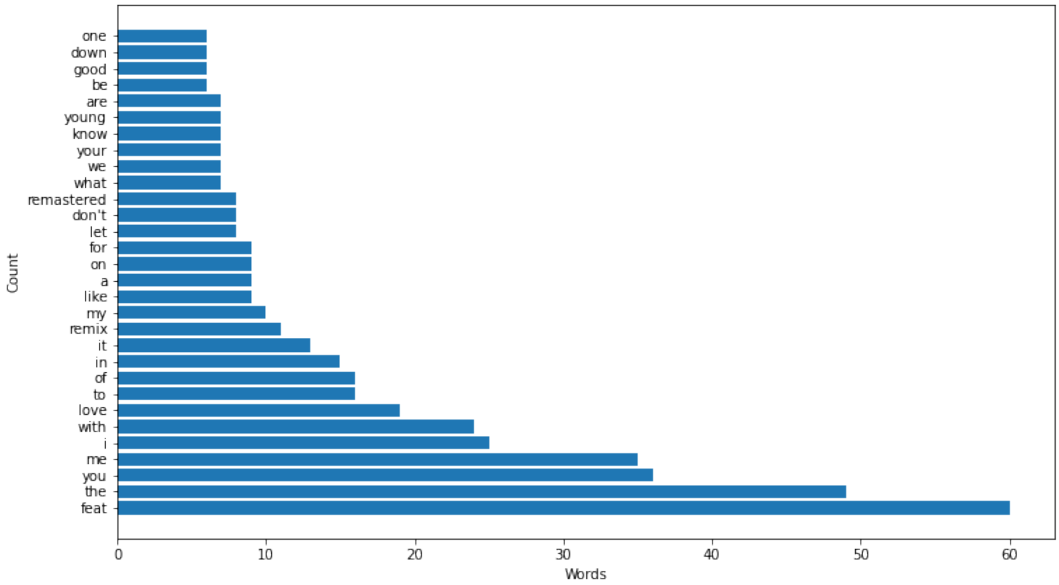


3.12. Most common words in popular song titles

Lastly, we analysed the most used words in the song titles*(RQ11)*. The analysis, shown in the graph in Image 23, indicates that words like feat. (featuring another artist), ‘the‘ ,’you,’ ‘me’, and ‘I’ are some of the most used in song titles. Additionally, high occurrences of words like ‘love‘ ,’to‘ ,’in‘ ,’your’, and ‘my’ suggest themes related to personal experiences and emotions. The presence of ‘remix’ and ‘remastered’ indicate modified or updated versions of songs.

Moreover, the inclusion of ‘let‘ ,’don’t’, and ‘know’ suggest communication or expression in song titles. The appearance of ‘young’ could indicate a focus on youth or a specific demographic. Overall, the words reflect that some of the most streamed songs are a mix of personal, social, and creative themes.

Image 23: Count of 30 most prevalent words in song titles



4. Discussion

4.1 Conclusions

Through our data analysis, we have researched several questions to gain a deeper understanding of songs that have achieved over a billion streams on Spotify. We examined the time it took for songs to reach a billion streams and identified the distribution of songs over months and years. The average time it takes to get a billion streams is 10 years and 264 days, but the median stands at 6 years and 78 days. There are many ‘outliers’ in the form of older songs which make the distribution skew to the right.

We also analysed the record labels with the most songs that have achieved over a billion streams. Columbia Records emerging as the front-runner, highlighting its significance in the music industry.

Contrary to popular belief, we concluded that songs with only one artist tend to be more popular than songs with multiple artists. We also concluded that songs with multiple artists and songs with only one have no significant difference in the distribution of their durations and valence, but there is a significant difference in the distribution of popularity, danceability, energy, loudness, and speechiness.

Furthermore, we have seen a trend of decreasing average song duration over time. We also observed variations in the average energy levels of songs, with a decrease until 2018, and an uprise in recent years.

Additionally, we examined that artists with more than one hit song tend to have a higher mean popularity score than “one-hit wonders”, suggesting a lasting impact and longevity in the music industry. The percentage of explicit songs has increased over time, reflecting changing social attitudes.

By exploring the genre presence in the “BILLIONS CLUB” playlist we can see pop as the most prevalent genre, followed by rap, dance pop, and hip hop. The analysis of “main” genres further highlights the prevalence of pop.

Moreover, we’ve seen that Justin Bieber leads the list of artists with the most songs surpassing a billion streams, closely followed by Drake, Rihanna, and Bad Bunny. The current popularity of artists is diverse with a mean of 73.95. Some of the most popular are Taylor Swift, Drake, and Bad Bunny, while some of the least popular ones are Jack Ü, Bipolar Sunshine, and Nyla.

Last, we analysed the most popular words in song titles, with them indicating the presence of a lot of features on songs and the presence of modified and updated versions of songs. Some words suggest themes of conveying personal experiences and emotions through songs.

Our analysis provides valuable insights into the enduring appeal of timeless classics, the impact of collaborations on popularity, and the diverse representation of genres among highly streamed songs.

4.2 Theoretical implications and practical implications

Our study challenges common belief in the music industry by providing new insights into the journey of songs to billion streams on Spotify. The findings contribute to the understanding of the temporal dynamics of music popularity, showing that the process takes a longer time than commonly assumed. The dominance of Columbia Records highlights the enduring role of established labels in the digital era, challenging notions of decentralised music distribution. The unexpected popularity of solo artist performances challenges prevailing beliefs about the collaborative nature of hit songs. Additionally, the study introduces nuanced insights into the distribution of musical attributes, providing a deeper understanding of audience preferences and the factors influencing music appreciation. This information can be leveraged by stakeholders to make informed decisions about collaboration choices, marketing strategies, playlist curation, and artist development, enhancing their chances of success in the competitive music landscape.

4.3 Limitations and future research

While our analysis provides valuable insights into music trends, there are limitations that should be considered. Firstly, our study focused solely on songs that have reached over a billion streams on Spotify, which may not capture the full spectrum of music trends and preferences. Future research could explore the streaming patterns of songs with lower stream counts to provide a more comprehensive understanding of music consumption and popularity. Additionally, our analysis primarily relied on quantitative data, and incorporating qualitative research methods such as interviews or surveys could provide deeper insights into the subjective experiences and motivations of listeners. Understanding the emotional and psychological aspects of music consumption would enhance our understanding of the drivers behind song popularity.

Overall, our analysis contributes to the existing body of knowledge on music streaming and provides valuable insights into the factors influencing the success of songs on digital platforms like Spotify. By understanding the evolving trends in music releases, the enduring appeal of timeless classics, and the impact of collaborations, artists and industry professionals can make informed decisions and strategies to navigate the ever-changing music landscape and effectively engage with their audiences.

References

Nijkamp, R. (2018). Prediction of product success: explaining song popularity by audio features from Spotify data. University of Twente. Available: http://essay.utwente.nl/75422/1/NIJKAMP_BA_IBA.pdf

Luo, K. (2018). Machine Learning Approach for Genre Prediction on Spotify Top Ranking Songs. School of Information and Library Science of the University of North Carolina. Available: https://cdr.lib.unc.edu/concern/masters_papers/ns064961b

Nugroho, A., Manongga, D., & Purnomo, H. D. (2023). Analysis of Spotify Top Songs During the Covid-19 Pandemic. Universitas Semarang, Indonesia: Universitas Kristen Satya Wacana, Indonesia. Available: https://journals.researchsynergypress.com/index.php/ijmadic/article/view/1565