

DATA SCIENCE PROJECT: MUSIC STREAMING POPULARITY ANALYSIS

Group 11

Nguyễn Tuấn Dũng - 20194427

Phùng Quốc Việt - 20194463

Vũ Quốc Việt - 20194464

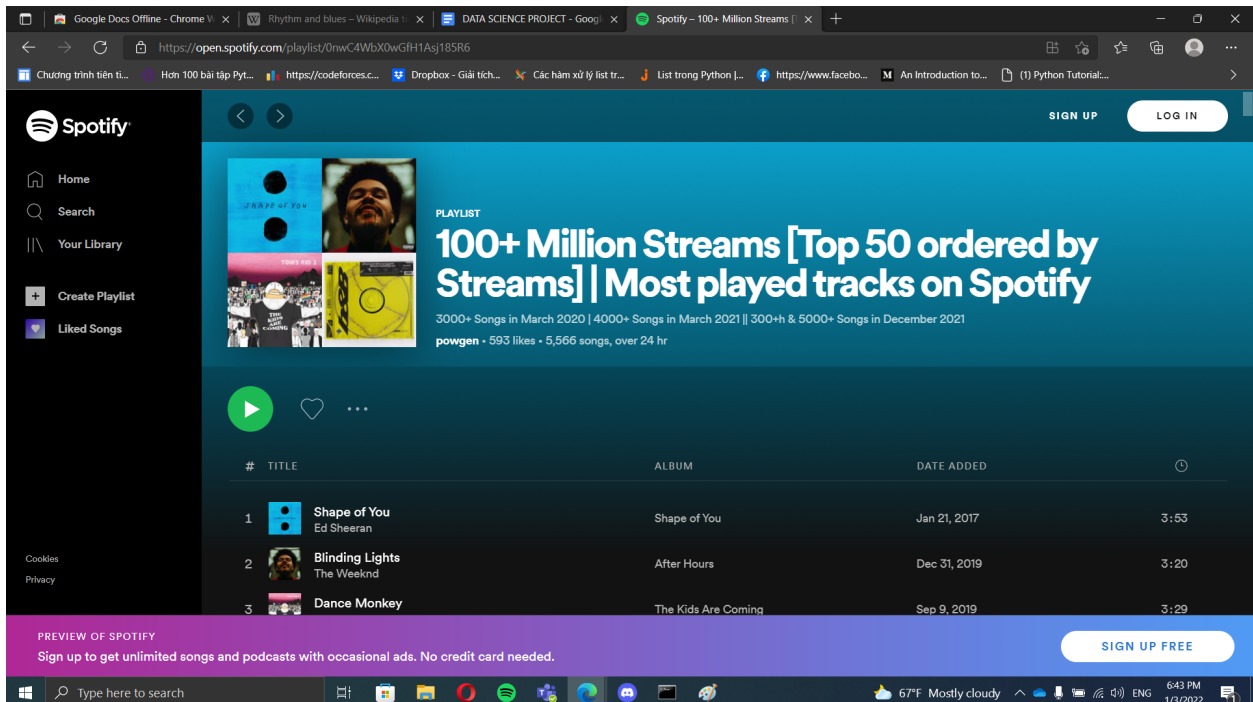
I. Introduction

The music industry currently plays a huge role in today's society, generating billions of dollars in profit and distributing millions of songs and albums every year. One of the most popular ways to distribute music currently is through streaming platforms, such as Spotify or YouTube. These platforms have millions of songs uploaded to their catalog, and expect millions of users annually. The performance of a song on these streaming platforms is an important aspect to its overall success and profit. In this project, we will analyze the most popular songs on these platforms based on their various features. We hope we'll be able to identify important trends, patterns and insights that can impact a songs' performance and popularity.

II. Data scraping

A. Spotify scraping

Firstly, we retrieve a Spotify compilation playlist that contains 4460 songs with over 100 million play counts on the platform: [Link \[1\]](https://open.spotify.com/playlist/0mwC4WbX0wGfH1Asj185R6)



The screenshot shows a Spotify web interface in a Chrome browser. The playlist is titled "100+ Million Streams [Top 50 ordered by Streams] | Most played tracks on Spotify" by user "powgen". It has 593 likes and over 5,566 songs added over 24 hours. The playlist description mentions "3000+ Songs in March 2020 | 4000+ Songs in March 2021 | 300+h & 5000+ Songs in December 2021". The track list is as follows:

#	TITLE	ALBUM	DATE ADDED	
1	Shape of You Ed Sheeran	Shape of You	Jan 21, 2017	3:53
2	Blinding Lights The Weeknd	After Hours	Dec 31, 2019	3:20
3	Dance Monkey Tones and I	The Kids Are Coming	Sep 9, 2019	3:29

At the bottom, there is a "PREVIEW OF SPOTIFY" banner with the text "Sign up to get unlimited songs and podcasts with occasional ads. No credit card needed." and a "SIGN UP FREE" button. The Windows taskbar is visible at the bottom of the browser window.

Then, we use the Spotify API for developers [2] to retrieve the songs' basic information and audio features that we will specify later in this report.

The Spotify API doesn't return to us the play count of the songs, or the record label that owns it. Also, the Spotify front page nor the HTML doesn't display the play counts of the song, and as far as we are aware, we can't scrape the views using libraries like *BeautifulSoup*, *Scrapy* ... Another problem was that Spotify doesn't have a separate page for a single song, only for albums (for songs that don't belong on any albums, it will be in a special type of album: Single with it being the only song).

Fortunately, Spotify still sends the actual play count to the browser's web traffic. We can retrieve the songs' play count by capturing the network's activity of our browser while loading the Spotify page of the song's album, downloading the HAR file (short for HTTP Archive, is a format used for tracking information between a web browser and a website) generated by our browser. So, to scrape the view and the record label, we did the following steps

- Retrieve the album's id of each song. Concatenating the song's id behind the "https://open.spotify.com/album/" prefix will return the URL to that album's page.
- Next, we utilized the *Selenium* library and its *Chrome Webdriver* to automatically load the albums' pages.
- To automatically retrieve the HAR file, we used a library called [browsermob proxy](#). Sometimes, the page might take too long to return the HAR file, so we check if the HAR file has been returned every 1 second. After a long amount of time, we will close the driver and skip that song to save time. We will keep a record of failed songs to scrape them again later.
- When the browser has returned the HAR file, we can either download all of the HAR file and extract the information we need later, or extract all the information directly without downloading the HAR file. We chose to download first and extract later, so that if any problems were to occur we will still have the files on our machine. Also, since the HAR file is of the songs' album's page rather than the song itself, the HAR file will contain the records of songs on the album that are not on our original playlist. These records will be filtered once we perform merging in the later steps.

After this scraping step, we ended up with a csv file containing the following information: *track id*, *track's names*, *play count* for each song.

B. YouTube Scraping

Our objective is to find out how many Youtube views these songs have. From the list of songs obtained in the previous steps, we tried to find the corresponding music videos on Youtube and retrieve its view count. This is done by utilizing the youtube keyword search api [3]. The keyword used for searching is a combination of the name of the track and the name of the artist. Special characters such as “#”, “-”, etc and noisy keywords such as “remastered”, “radio edit”, etc are removed from the keyword by a filter. Using this API, we obtained the list of Youtube video id corresponding to the original track on Spotify.

Using the video id, we can reconstruct the link to that video on Youtube and request the html file of the website, and then extract the view count. Initially we tried using the requests library to get the html content, but it was quite inconsistent: at times it took up to 13 seconds to return the response (we later found out that this was the ad wait time), other times it froze indefinitely without any response. As a solution to this we used Selenium instead and it ran smoothly, only taking roughly 1-2 seconds per song.

One problem that we encountered while finding the corresponding video on Youtube was that this keyword searching method was not 100% reliable. Sometimes the song obtained from the Youtube api would be different from the original song. We tried our best to minimize this by improving our keyword filter to remove as many irrelevant words as possible. Another problem was that some songs on Spotify do not have their counterparts on Youtube. Therefore a song different from the referenced Spotify song will be obtained from the Youtube api. We have no solution to this apart from accepting a small number of noise in our dataset.

C. Wikipedia Scraping

We retrieved the genres of a song by scraping its Wikipedia page (or the page of the album it belongs to). To do this, first we used the Wikipedia API [4] to search for the page using the song’s title and artist. However, Spotify’s titles of some songs might contain unnecessary elements that can prevent us from getting good search results. For example, the title “*Bohemian Rhapsody - Remastered 2011*” returns the Wikipedia page to Queen’s discography rather than the song because of the “- *Remastered 2011*” part. So, we turned all of our titles into lowercase letters, filtered out elements like these before searching. It’s worth noting that some songs either don’t have a Wikipedia page, or that our search keywords might not be enough to return the correct results, so the search result returns the Wikipedia article of something else instead.

After this step, we ended up with the URL for the search results. Next, we want to retrieve the genres of the songs from these URLs. Since this is a fairly simple scraping task, we used the BeautifulSoup to achieve this.

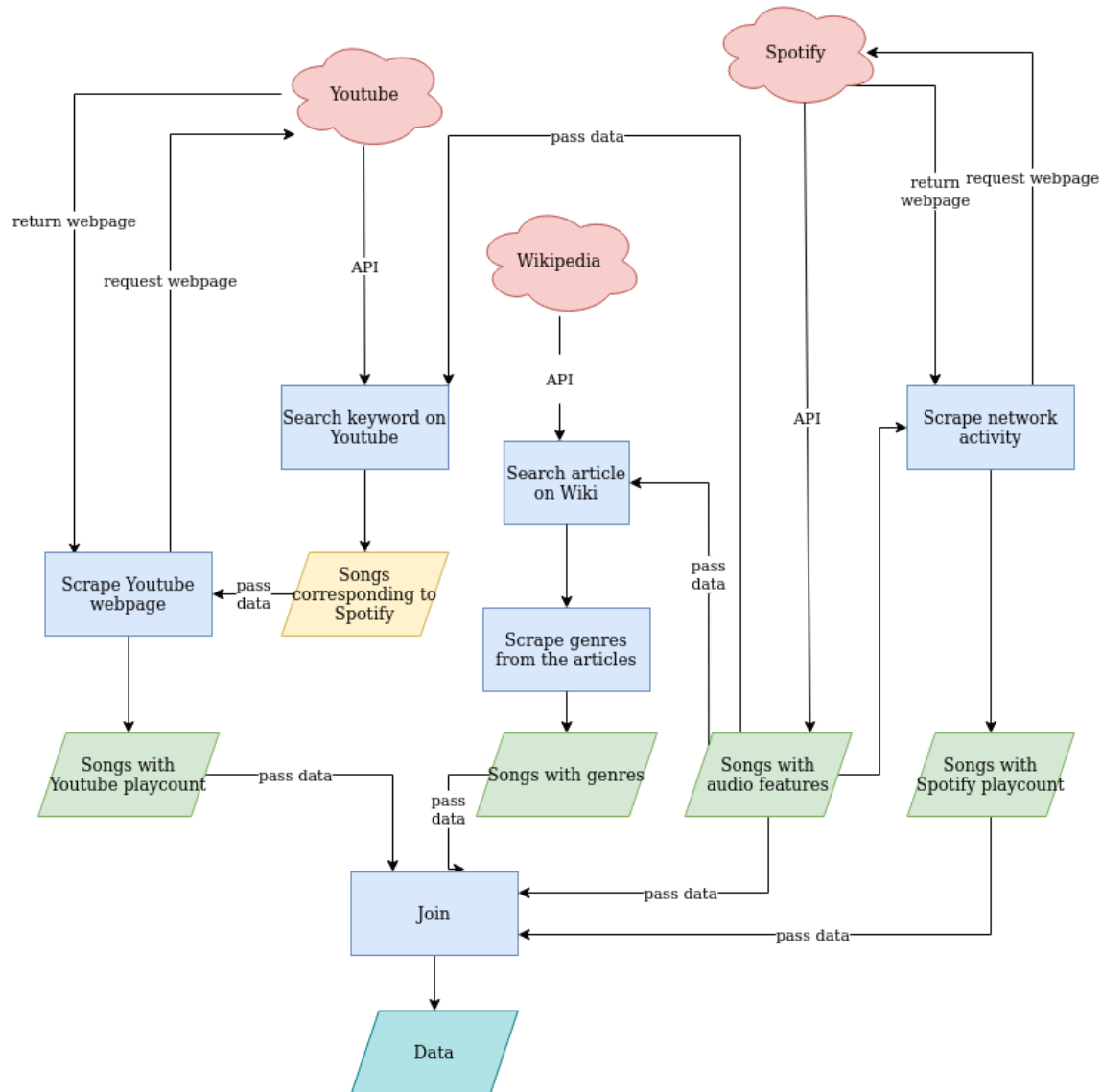
Most wikipedia articles for songs or albums have an info-box element containing basic information, including genres on the right side of the front page like this. If it's a song or an album, then this element class will be *"infobox vevent"* or *"infobox vevent haudio"*. This will help us filter out if an article is not of a song or an album.

Single by Queen	
from the album <i>A Night at the Opera</i>	
B-side	"I'm in Love with My Car"
Released	31 October 1975
Recorded	August–September 1975
Studio	Rockfield (Monmouthshire) · Roundhouse · Sarm · Scorpio Sound · Wessex Sound (London)
Genre	Progressive rock · hard rock · progressive pop

Finally, we ended up with a csv file with the columns: track's spotify id, track's name, artist, and the genres.

III. Data Preprocessing

A. Data Integration



After scraping data from 3 different sources (Youtube, Spotify and Wikipedia), we obtained 4 csv files shown in green in the flow diagram. Each one of them contains one or more features that are unique to them. As the last step of our integration process, we left-joined “Songs with audio features” with the other 3, one by one, on the “track_id” column to obtain the final dataset. We chose “Songs with audio features” as the base data to join with the others because it contains all the songs we intended to use for building our popular songs dataset.

B. Data cleaning

As we have mentioned in the Youtube Scrapping section, the song we obtained from Youtube does not always correspond to the referenced song from Spotify. Such cases usually result in low Youtube play counts. To minimize the amount of noise in our dataset, we remove songs that have less than 10 million Youtube views (which is 429 songs - 1/10 of our dataset). Beside wanting to minimize noise in our data we dropped these songs also because our main objective is to analyze the popular songs on both platforms, so we don't want to keep songs with too few play counts on one platform.

The total number of unique genres in our dataset is 456, most of which only have a handful of instances in our data. Because of these problems, we feel the need to group them up to perform data analysis effectively. Most subgenres can be traced back to 10 major genres: pop, edm, hip hop, jazz, r&b, rock, folk, country, reggae and metal. From online research and our own knowledge, we hand-built a dictionary of subgenres in order to map them to the major genres.

The audio features of our dataset was directly obtained from the Spotify API. Therefore these data are rather clean so there was not a lot of cleaning we needed to do here.

The remaining steps of the cleaning process goes as follows:

- Change "duration" from minute:second format to second. E.g. 03:20 -> 200
- Change "mode" to its actual name:0 to minor, 1 to major
- Change "key" from integers to key name (C#, B, D, etc)
- Change "time signature".E.g 4 -> 4/4
- Format "date" to yyyy/mm/dd

C. Data description

Feature	Type	Description
explicit	categorical	Whether or not the track has explicit lyrics (true = yes it does; false = no it does not).
time_signature	categorical	An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure)
key	categorical	The key the track is in.
mode	categorical	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived.
danceability	numeric	Danceability describes how suitable a track is for

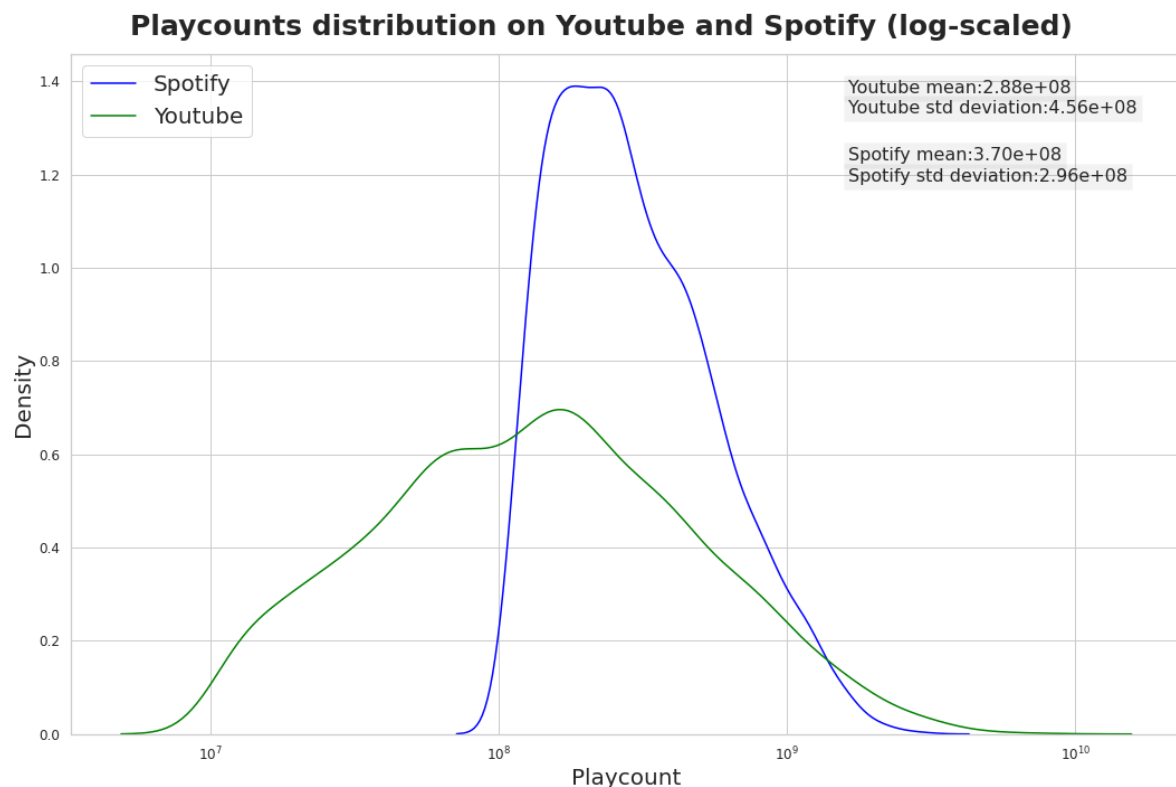
		dancing. A value of 0.0 is least danceable and 1.0 is most danceable.
energy	numeric	Energy represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
loudness	numeric	The overall loudness of a track in decibels (dB). Values typically range between -60 and 0 db.
duration	numeric	The duration of a track in seconds
speechiness	numeric	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
instrumentalness	numeric	Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
liveness	numeric	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
valence	numeric	A measure describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
tempo	numeric	The overall estimated tempo of a track in beats per minute (BPM)
genres	categorical	The genre of the song

playcount	numeric	Spotify's play count of a track
view	numeric	The number of views of a song on Youtube

IV. EDA and Data visualization

A. Target variables analysis

To start off with the analysis, we first look into our target variables, which are YouTube and Spotify play counts of the songs. Firstly, we examine the playcount distribution on each platform.



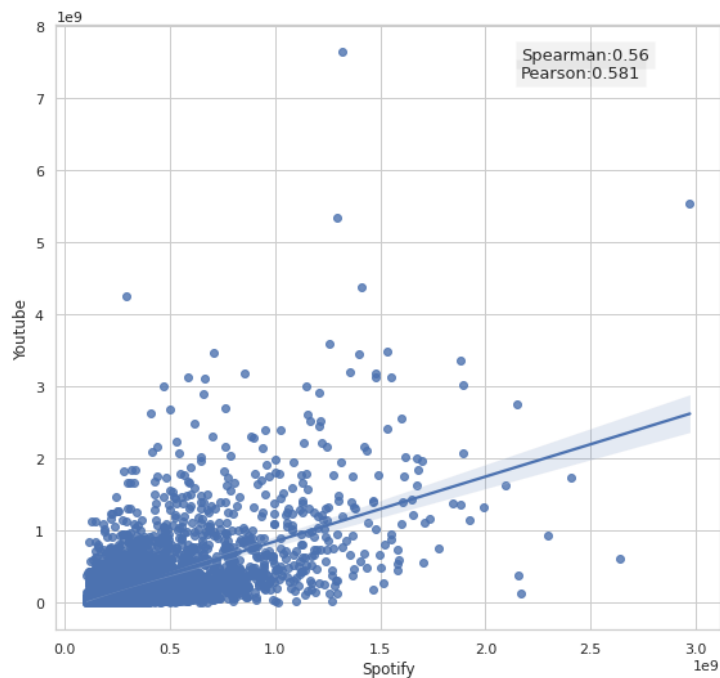
The Spotify play counts distribution starts at around 100 million play counts which is the minimum number of playcount a song must have in order to be extracted from Spotify. The corresponding Youtube play counts have a wider range, especially on the lower end of the playcount spectrum. Unlike on Spotify, many songs on Youtube lie below the 100

million play counts mark. Incorrect song alignment between Spotify and Youtube is perhaps one of the reasons behind this distribution.

The mean number of play counts on Spotify is a lot higher than that on Youtube, at 370 million and 288 million respectively. This supports the fact that compared to Youtube, Spotify is the preferred medium for online music streaming. Spotify play counts also has a more compact distribution with standard deviation 1.5 times less than Youtube.

While the YouTube play counts have lower mean compared to Spotify, YouTube has much more extreme outliers where some songs tend to have a play count of almost 8 billion.

Scatter plot of Spotify playcounts and Youtube views

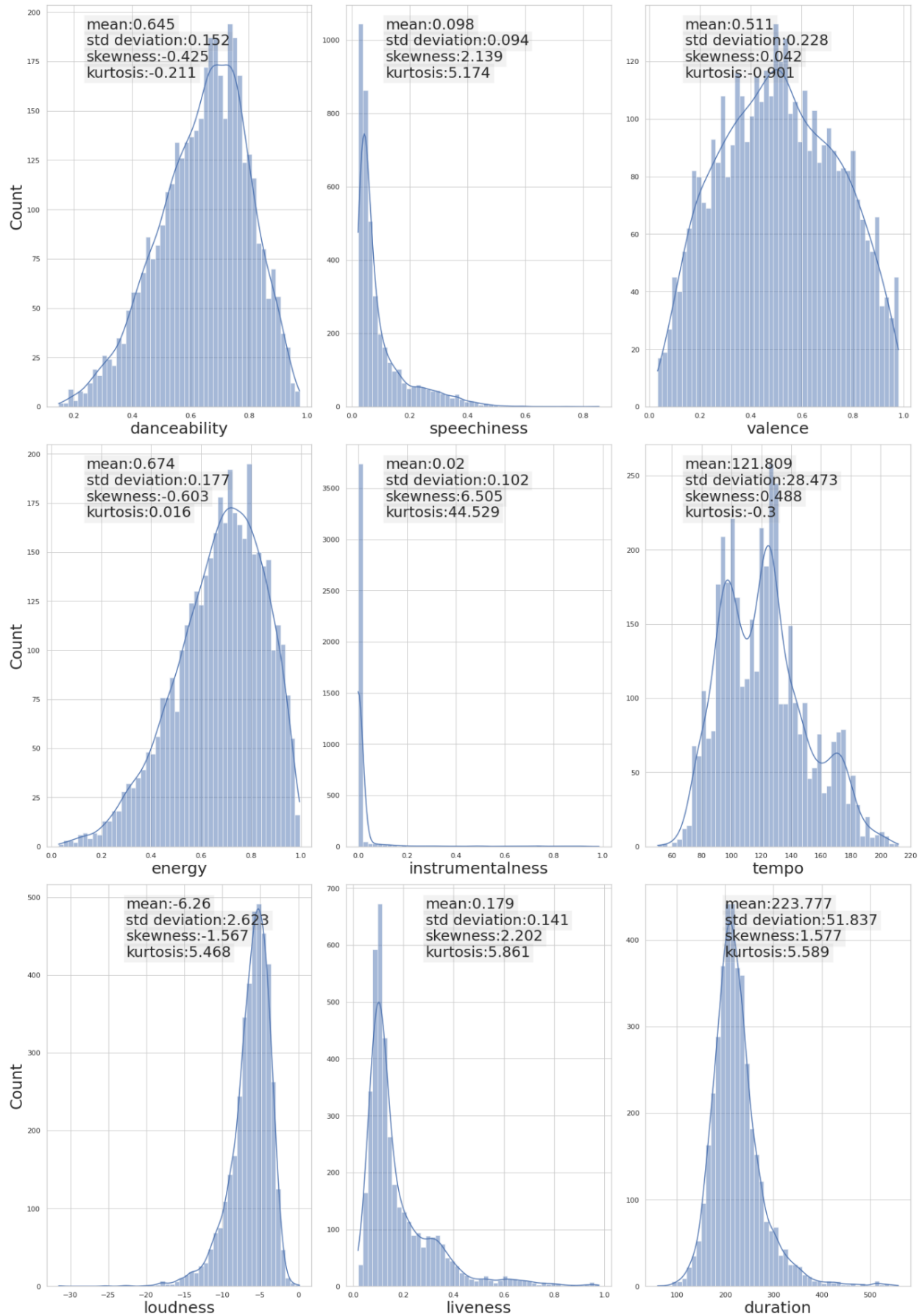


Both pearson and spearman correlations between Youtube and Spotify play counts are about 0.57, which suggests a positive monotonic relationship between these two variables. This is expected as a popular song on one platform tends to perform well on the other.

B. Numeric audio features analysis

Next, we examine a few numeric audio features in our dataset. Firstly, we plotted the distribution of these features, as well as computing some of their basic statistics: mean, standard deviation, skewness and kurtosis.

Distribution of some numeric variables

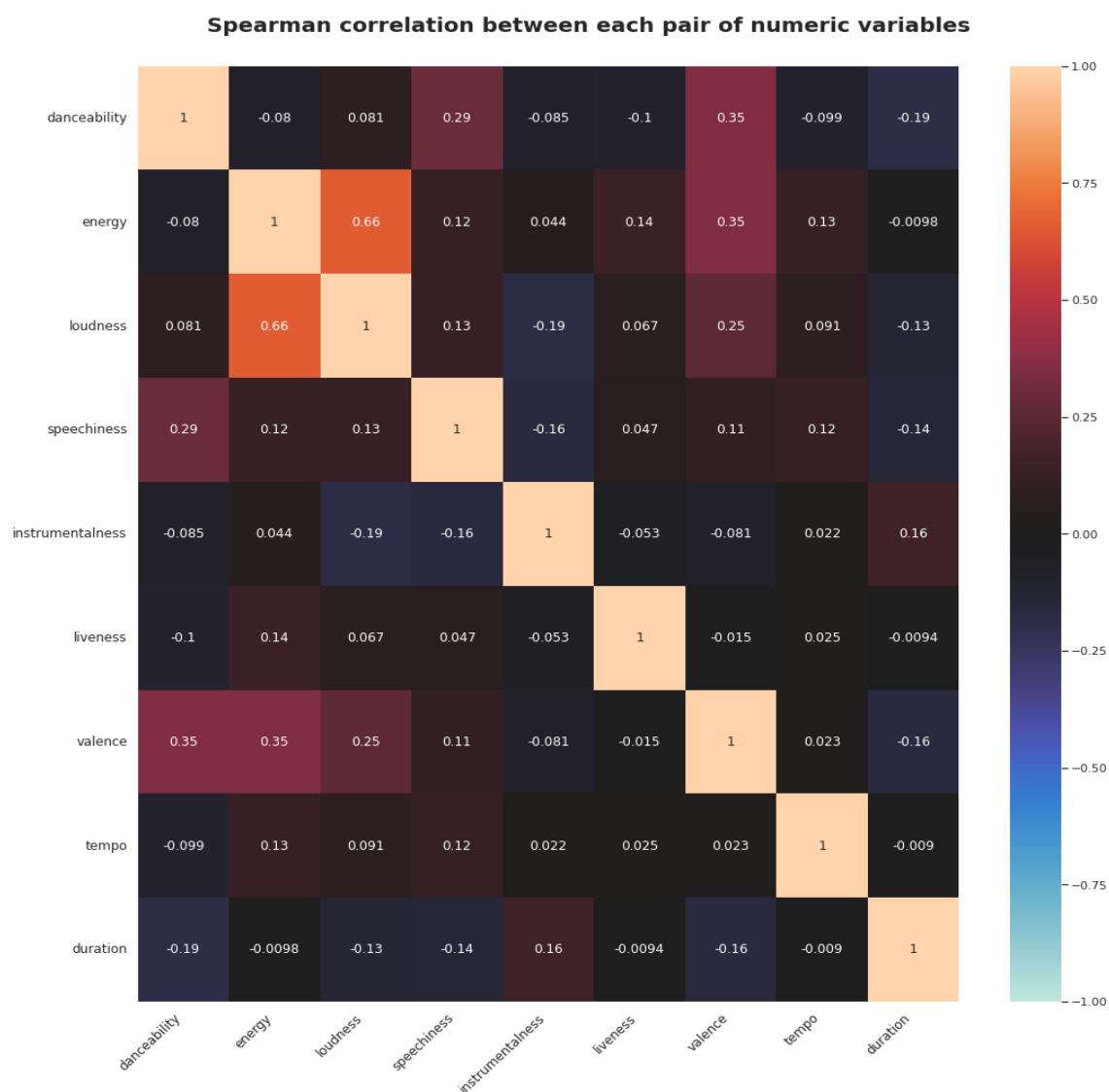


- Danceability (describes how suitable a track is for dancing from a measure of 0 to 1): Of all the distribution of our numeric features, that of danceability seems to be the closest to a normal distribution. The mean of this feature is around 0.65, which is moderately higher than the average 0.5. This indicates that the popular songs we've collected can vary in danceability, but overall the songs tend to be a bit more "danceable".
- Energy (a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.): The energy in our dataset is slightly skewed towards the left, with the mean being 0.674. From the distribution, we see that popular songs tend to be a little bit more energetic, however plenty of songs have lower energy as well.
- Loudness (The overall loudness of a track in decibels, typically between -60db and 0db): As we can easily see from the distribution and the statistics, the loudness is heavily skewed. This indicates that most popular songs are likely to be loud in volume.
- Speechiness (detects the presence of spoken words in a track, value ranges between 0.0 to 1.0): The general distribution speechiness for popular songs are quite low, with the mean of 0.098 and are heavily skewed, extremely "pointy" distribution with a kurtosis of 44.5. This shows that the presence of spoken words in popular songs is quite rare, and almost every popular song contains musical content.
- Instrumentalness (Predicts whether a track contains no vocals, The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks): Most popular songs' instrumentalness feature is quite low, with the mean being only 0.02 and heavily skewed. From this, we conclude that generally almost all of the popular songs contain singing vocals.
- Liveness (Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.): The liveness features' distribution mostly focuses around less than 0.4, which means that the majority of the popular songs are studio recordings rather than live performances.
- Valence (A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track): The valence is not heavily skewed as some of the features above. The distribution is distributed densely around the middle area, with the mean equals 0.511 and the standard deviation equals 0.228. This shows that the musical positiveness of popular songs vary greatly with no noticeable trends, and

a lot of the songs fall somewhere between positivity (happy) and negativity (sad), but a handful of them can fall on the extreme sides as well.

- Tempo: The tempo in popular songs can vary greatly, and is densely distributed around the 80 to 160 area, which is a standard tempo range, with a few outliers that can go up to 200.
- Duration (ms): based on the distribution, most popular songs last around 150-300 seconds, which is around 2.5 to 5 minutes. Once again, this is a pretty standard time range for the music industry in general.

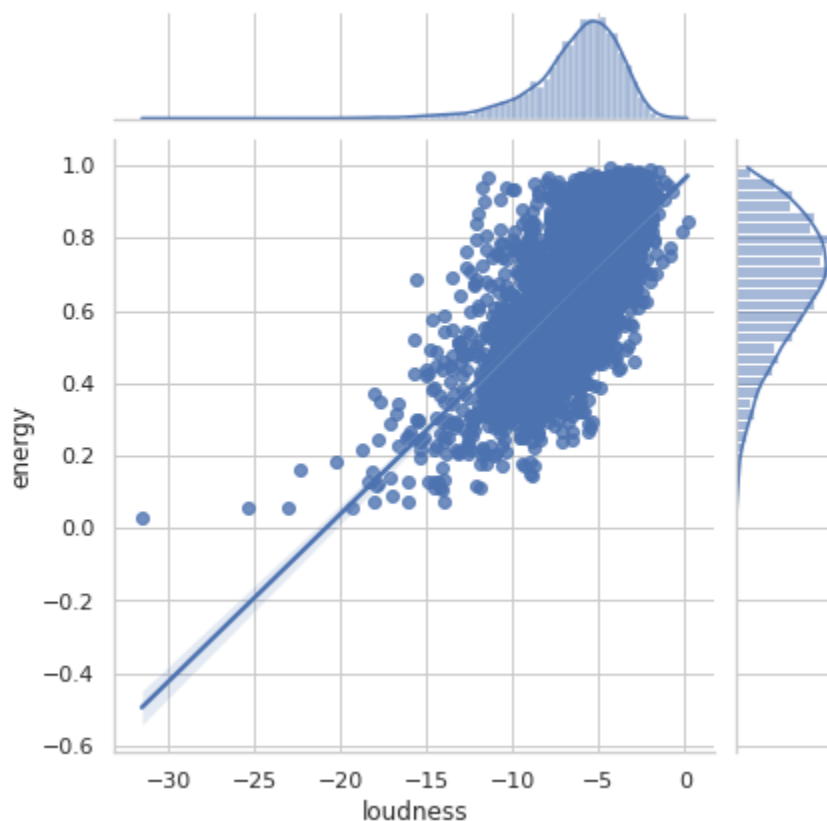
Now, we will examine the relationship between each pair of numeric variables. We compute Pearson and Spearman correlation between them and store it in a matrix like so:



Across the correlation matrix, most variables have relatively low correlations with another. There are some more noticeably higher correlations, such as valence and energy or valence and danceability. The valence is the measure of musical positiveness, the energy being the intensity and activity of the song, and danceability measuring how “danceable” a song is. So it makes sense that a song that is danceable is more likely to be happy and positive, or a song that is positive tends to be louder and more fast paced.

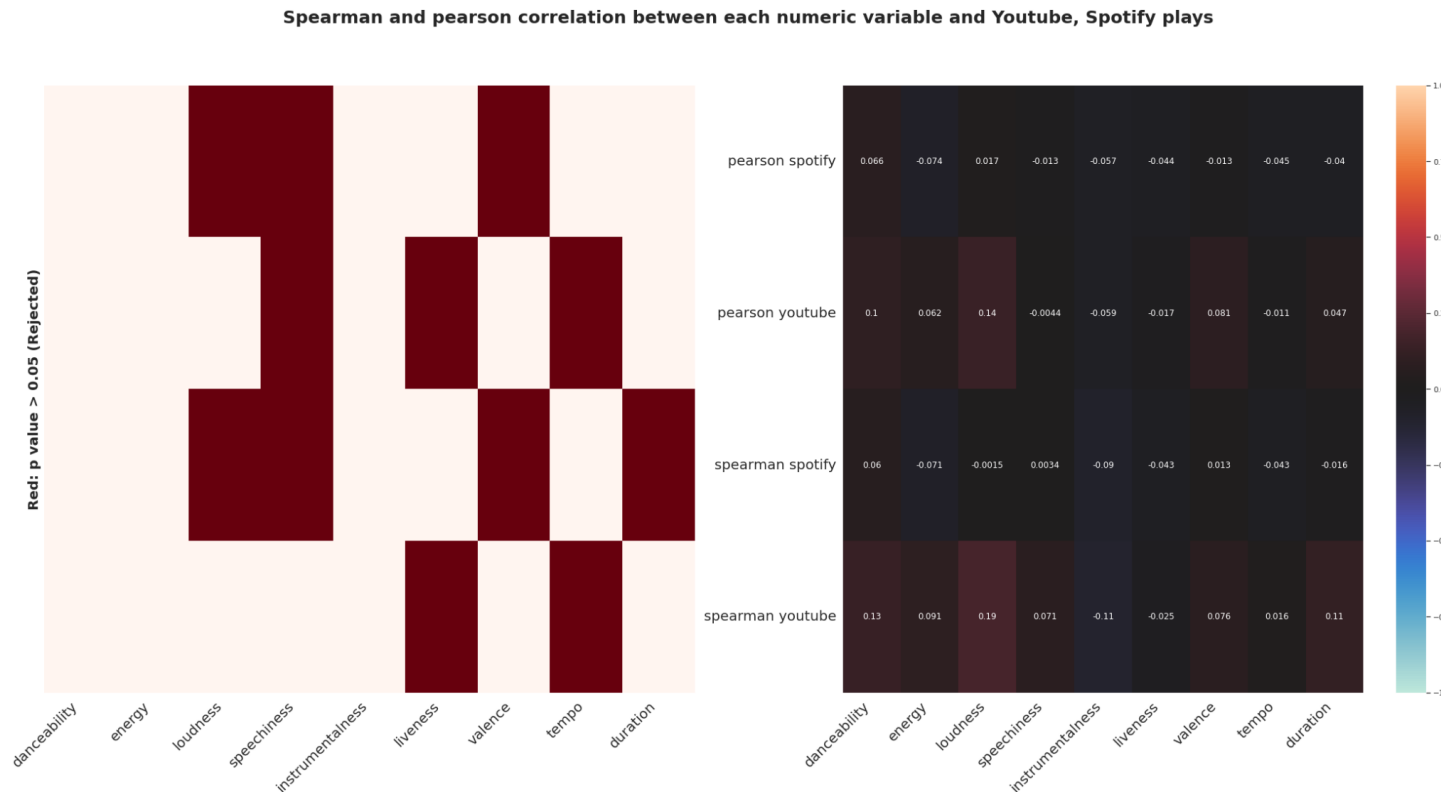
However, there are two variables here that have a fairly strong correlation compared to the rest, which are “loudness” and “energy” with a correlation of 0.66. This relationship can be justified by the nature of these variables: an energetic track usually feels fast, loud and noisy. In order to observe this relationship, we visualize these two variables using a scatter plot:

Scatter plot with marginal between energy and loudness



The positive linear correlation between “loudness” and “energy” can be seen clearly from this plot. A song becomes more energetic as its volume increases.

Now we investigate the correlation between each exploratory numeric variable and the response variables (Youtube and Spotify play counts):



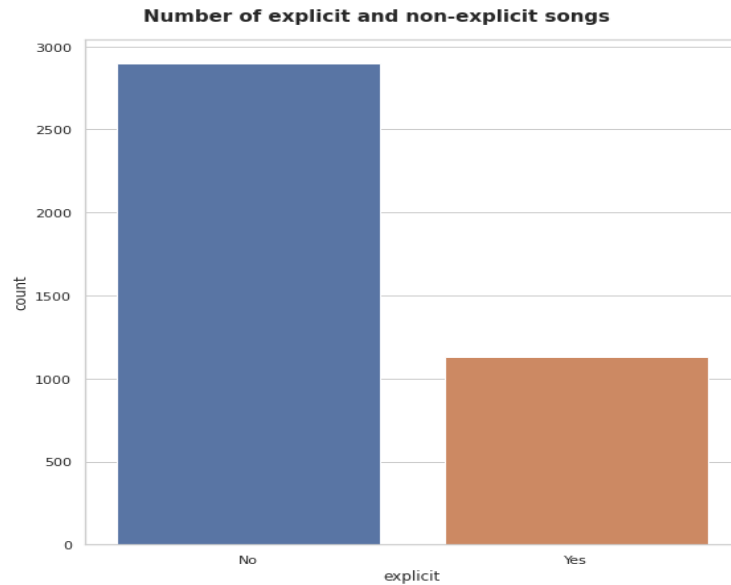
The plot on the left shows how reliable the corresponding correlations on the right are. Red cells mean that they have a p value of than 0.05, which suggests that the null hypothesis: “the correlation between these variables is 0” is not rejected. All numeric features have low correlation with playcount, being spearman or pearson. This result is to be expected. Take “danceability” as an example, in real life, the fact that a song is more “dancy” does not guarantee its popularity. The same can be said for the other features.

Despite this, there are still some weak correlations between loudness , danceability and the number of Youtube play counts.

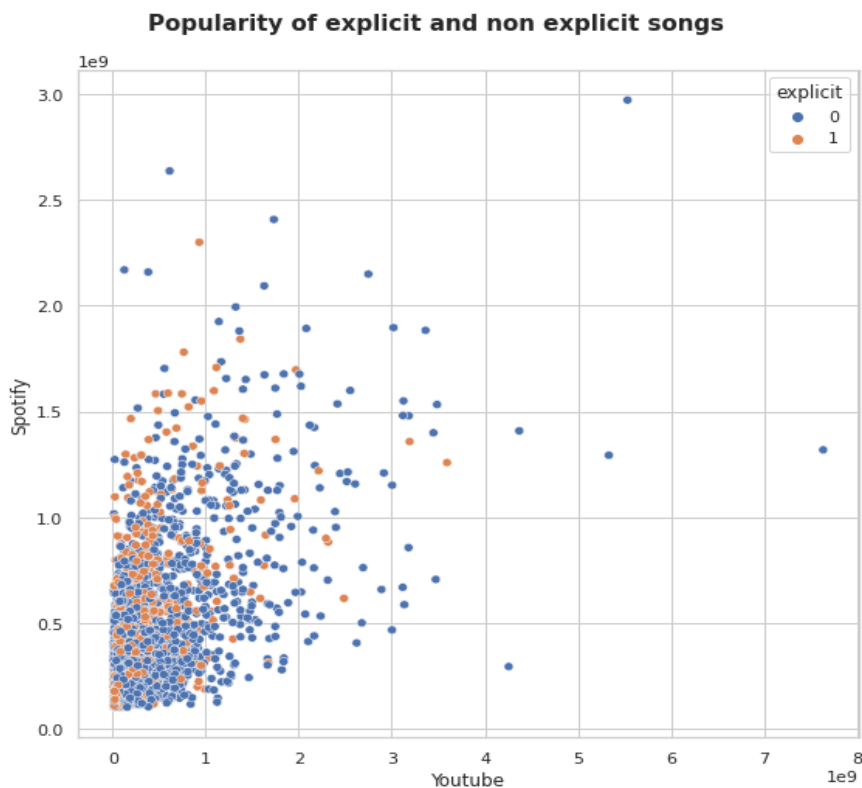
C. Categorical features analysis

Explicit lyrics

This is a boolean variable that denotes whether or not a track has explicit lyrics.



As we can see from the bar charts above, the majority of the songs have non-explicit lyrics (about 2800-2900 out of 4031 songs) compared to about over 1000 songs with explicit lyrics. This makes sense, since the explicitness in lyrics might limit the potential audience range for the song, particularly children



To visualize the relationship between the explicitness and the YouTube and Spotify play counts, we plotted a scatter plot between these variables, since scatter plot is a great way to visualize the relationship between numeric features and categorical features with small modalities. As observed, most songs regardless of explicitness tends to cluster to the lower left part of the plot, which are the songs with about 10 million to 2 billion views on YouTube and 100 million to 1.5 billion play counts on Spotify. However, it seems like most great outliers in our data are non-explicit in lyrics. All of the songs with over 4 billion views on YouTube are non-explicit, and only one explicit song, compared to about 7 non-explicit one, achieves 2 billion play counts on Spotify. It seems that for regular popular hits, the explicitness of the lyrics doesn't have much impact on the song's popularity, but most extremely successful songs tend to be more "clean" when it comes to the lyrics.

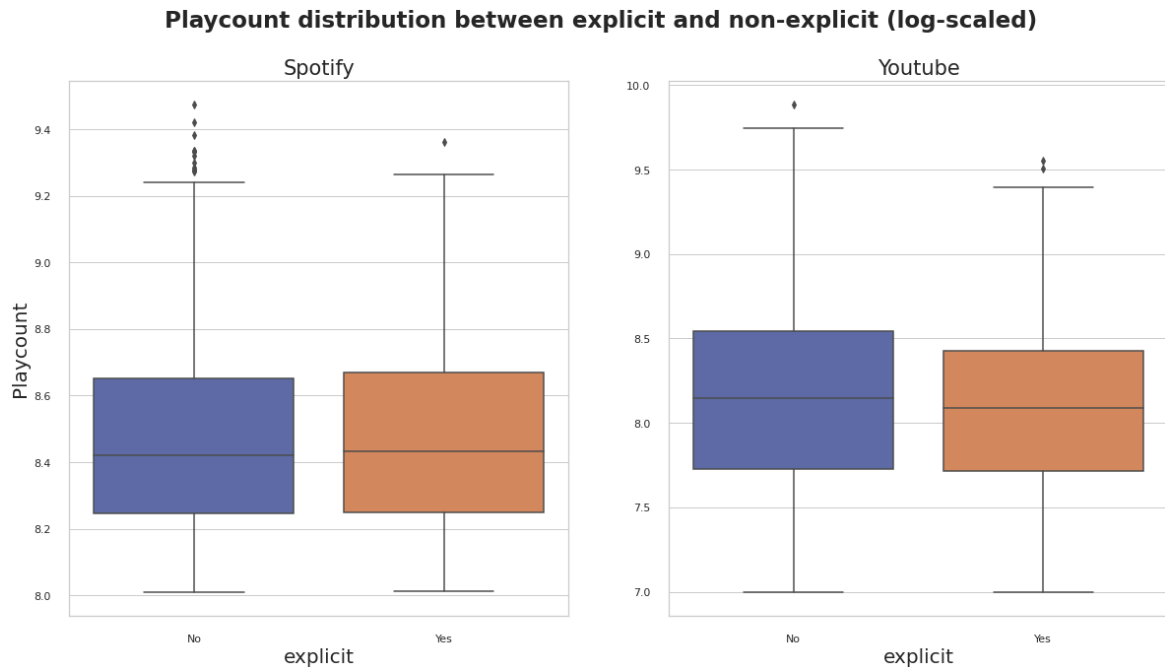
To further investigate whether the explicitness of a song impacts its performance on these two streaming platforms, we performed a simple z-test.

- Z-test:

We want to test whether a song having explicit lyrics has any impact on its popularity. We apply the two tailed z-test to two populations: songs in major and songs in minor with the null hypothesis being the mean between these two populations are equal:

Platform	p-value
Spotify	0.316
YouTube	5.05×10^{-7}

From the result we obtained, on Spotify, the difference in between two means is not statistically significant. However, explicitness does have an impact on a song's popularity on Youtube. We also plot a boxplot as a visual verification of this result.

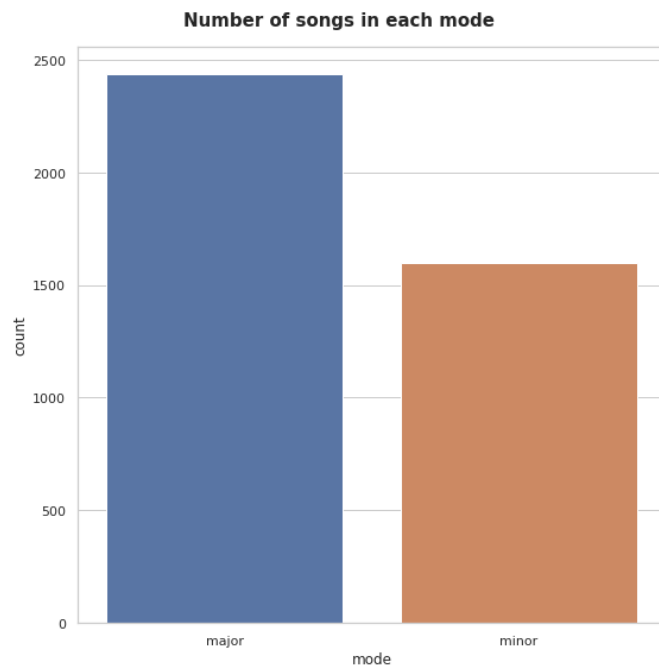


Indeed, while the popularity between explicit and non-explicit songs on Spotify are nearly identical, non-explicit songs tend to perform better than explicit ones on Youtube.

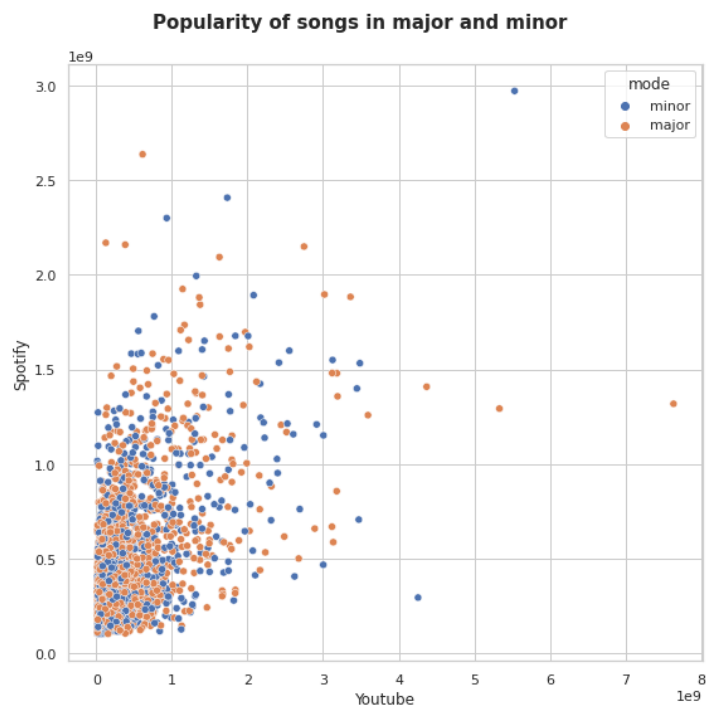
One of the reasons for this result could be that YouTube is a more “family-friendly” platform than Spotify. YouTube has a stricter policy when it comes to content monitoring, and favors more family friendly videos than explicit ones. So this might explain why on YouTube, non-explicit songs tend to have better performance

Mode

The mode is a categorical feature, which indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Further information can be referenced here: [Link \[5\]](#). We will not go into detail on the music theory behind this feature, but to summarize: songs written in major modes sound upbeat, strong, and somewhat happy, while those written in minor modes might sound more serious and sad.



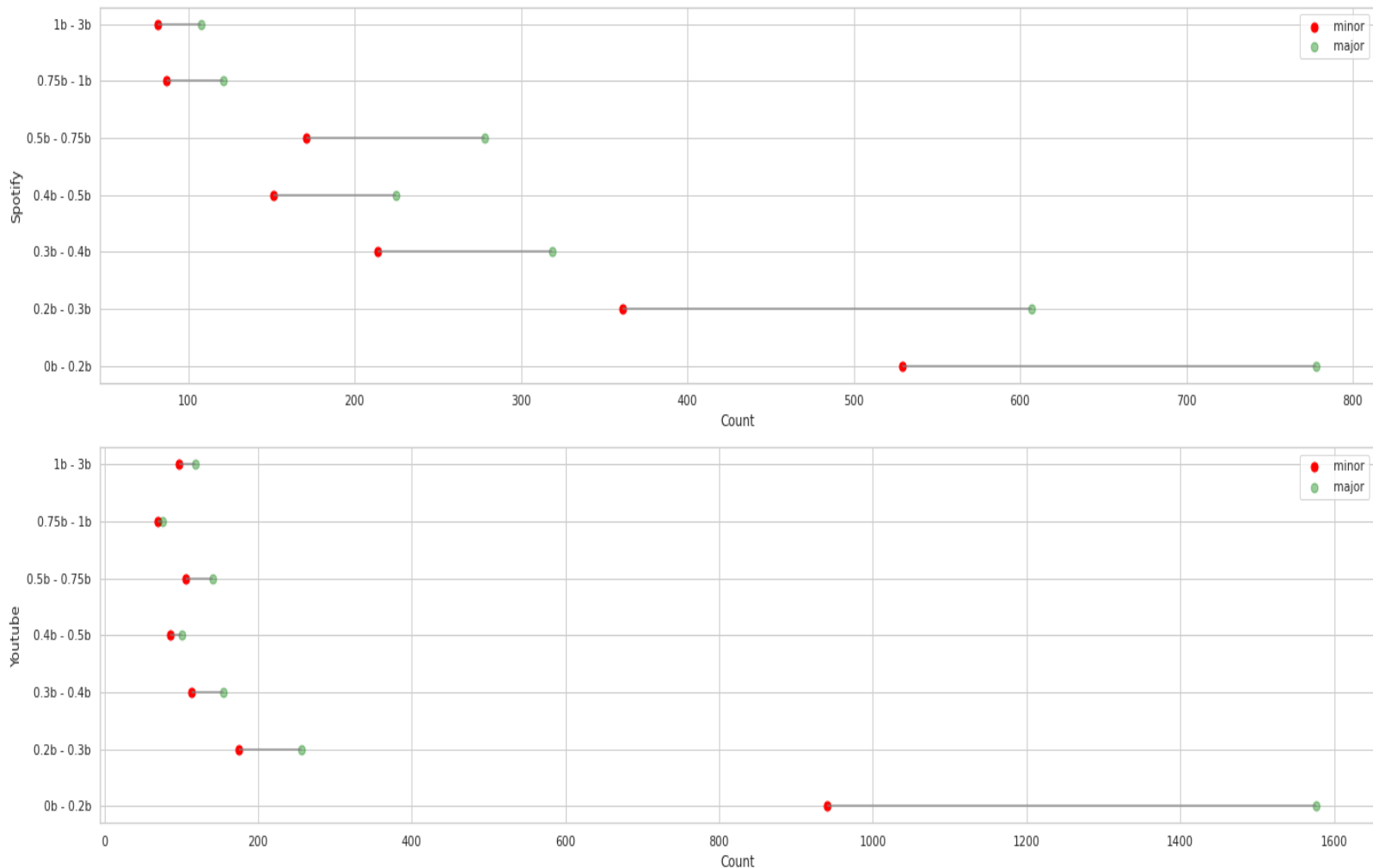
As observed, the major modes are more popular for popular songs (almost 2500 records).



Similar to the explicitness above, we used a scatter plot to visualize the mode's relationship between the target variables. However, the density and distribution of both of the major and minor mode songs don't seem to be too informative and indicate no clear patterns on this feature.

We decided to utilize a lollipop plot to visualize the number of songs in each mode in different play count ranges.

Number of songs in major or minor in different playcount ranges

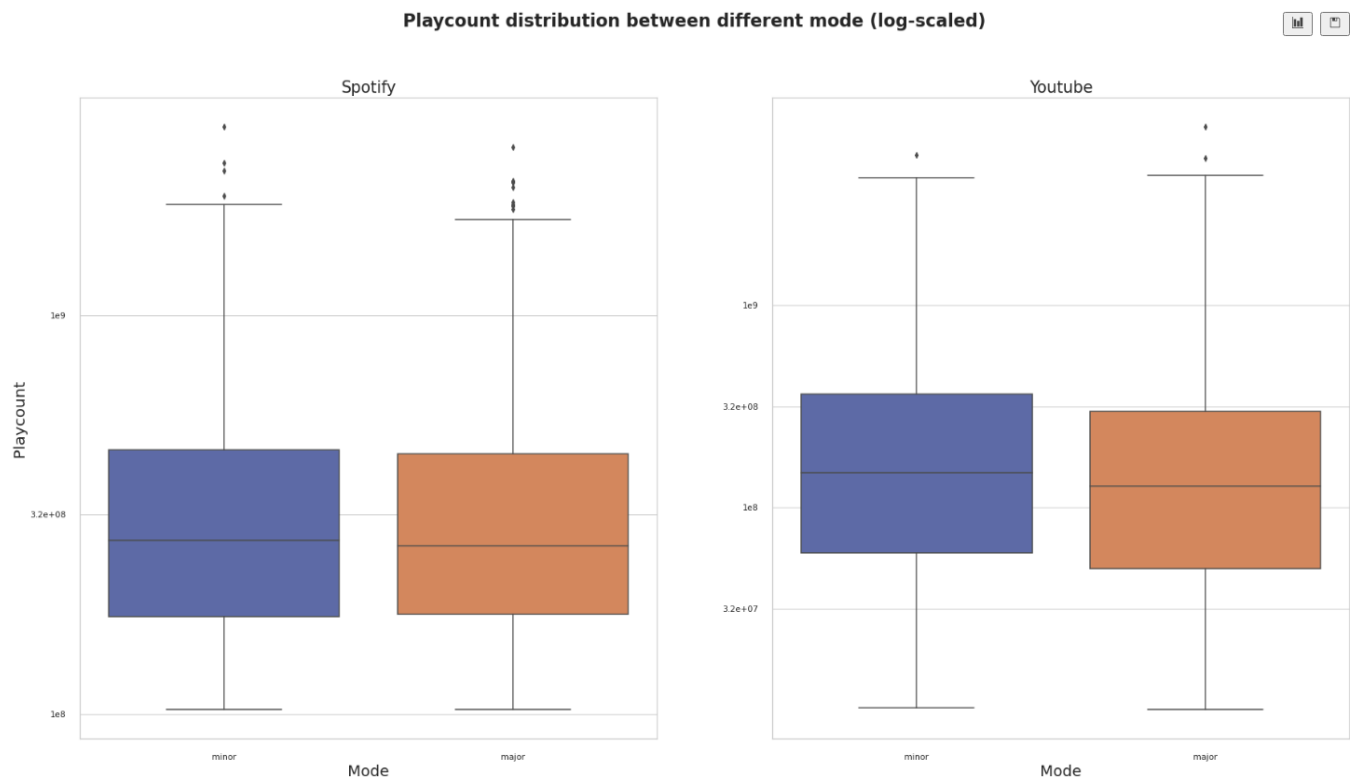


Compared to Spotify, in higher view ranges, namely more than 200 million playcounts, the difference between the number of hits in minor and the number of hits in major is greatly reduced. On Spotify, songs in major accounts for a much higher percentage of songs across all play count ranges. On Youtube on the other hand, the significance of a song's mode in reaching high playcounts is vastly reduced.

- **z-test**

We also conduct the z-test on this variable the same way we did on the “explicit” variable:

Platform	p-value
Spotify	0.409
YouTube	0.003

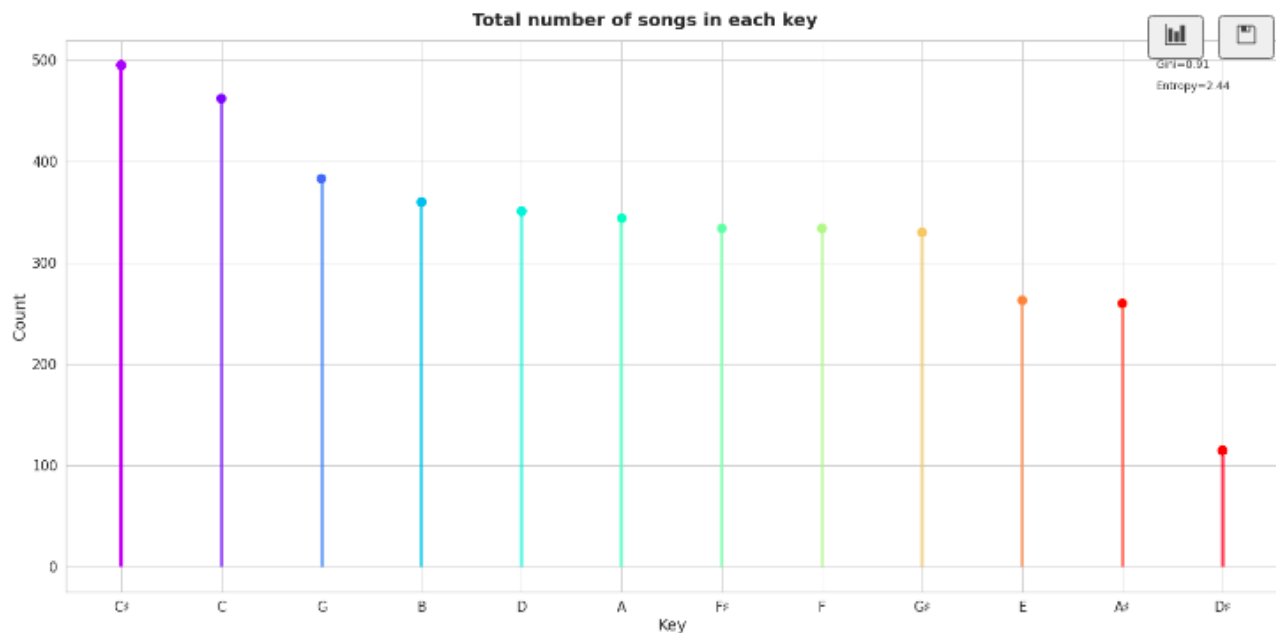


Using the z-test, we obtain a similar result as the “explicit” variable. Again, the test shows that the mean number of play counts between two modes are similar on Spotify. , With the p value less than 0.05, the mode of a song seems to affect only its popularity on Youtube. Songs in minor in general are more popular on Youtube than songs in major, even though “mode” does not have any statistically significant impact on the success of a song on Spotify.

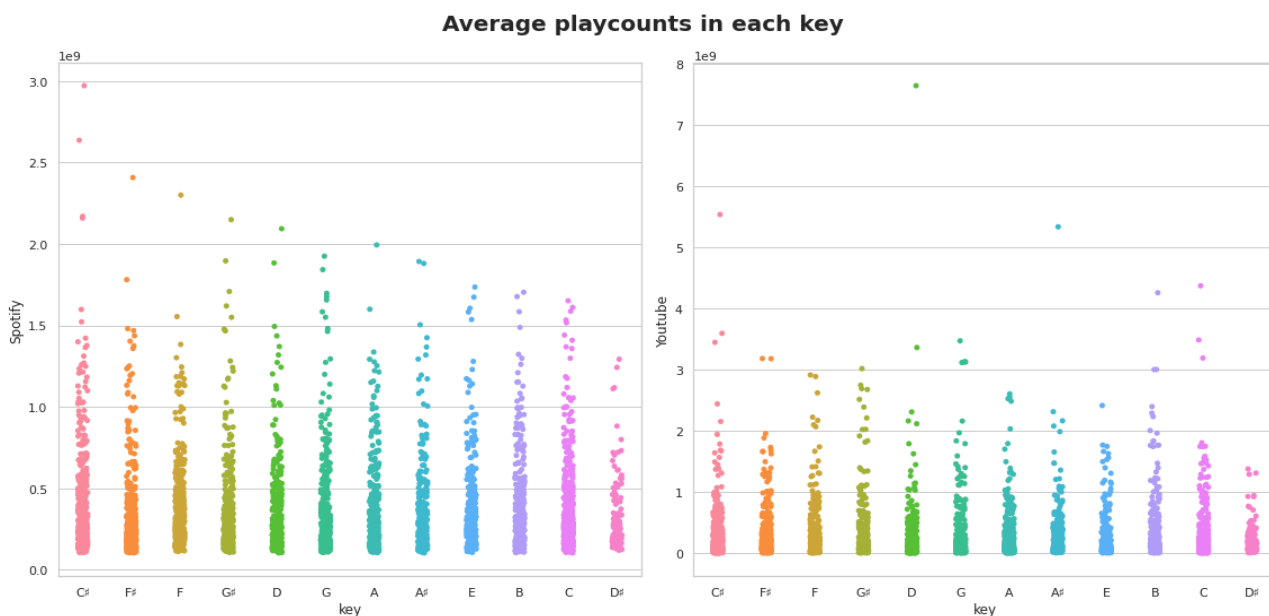
Key

This variable denotes the key the track is in (C, C#, D, D#....).

Reference link: [Key \(music\) - Wikipedia](#) [6]



From the Gini (0.91) and Entropy (2.44) calculated, and from the chart above, we can see that the key variable is quite evenly distributed. The keys based on the C note (C and C#) are the most popular keys, while D# fall a bit short compared to the others (only about over 100 songs). While there are keys that are slightly more popular than the others, there doesn't seem to be a clear trend in terms of popularity.

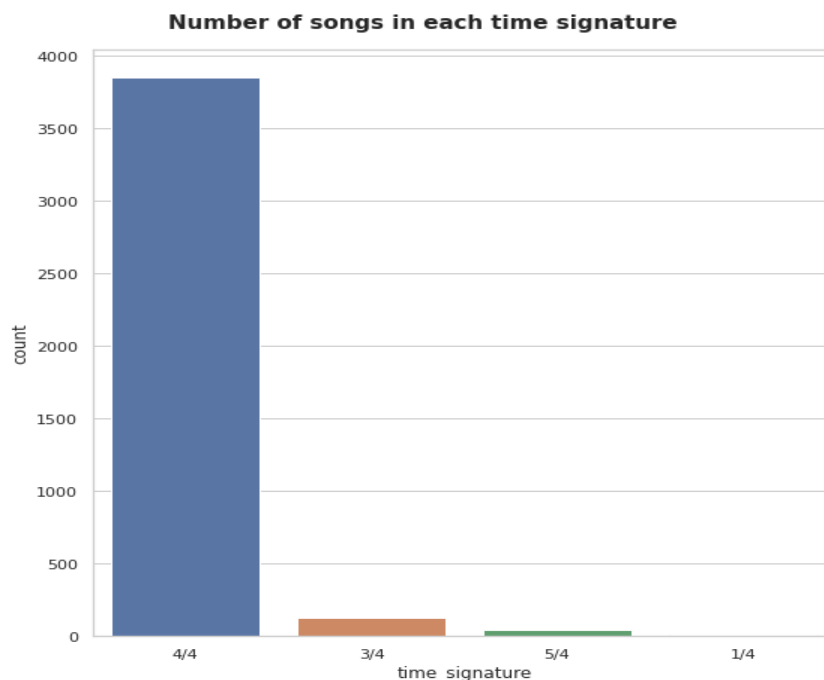


We decided not to use a single regular scatter plot to visualize the Spotify and YouTube play counts like we did for the other two categorical variables, since this feature has quite a lot of modalities, so the scatter plot we tried looked quite messy and uninformative.

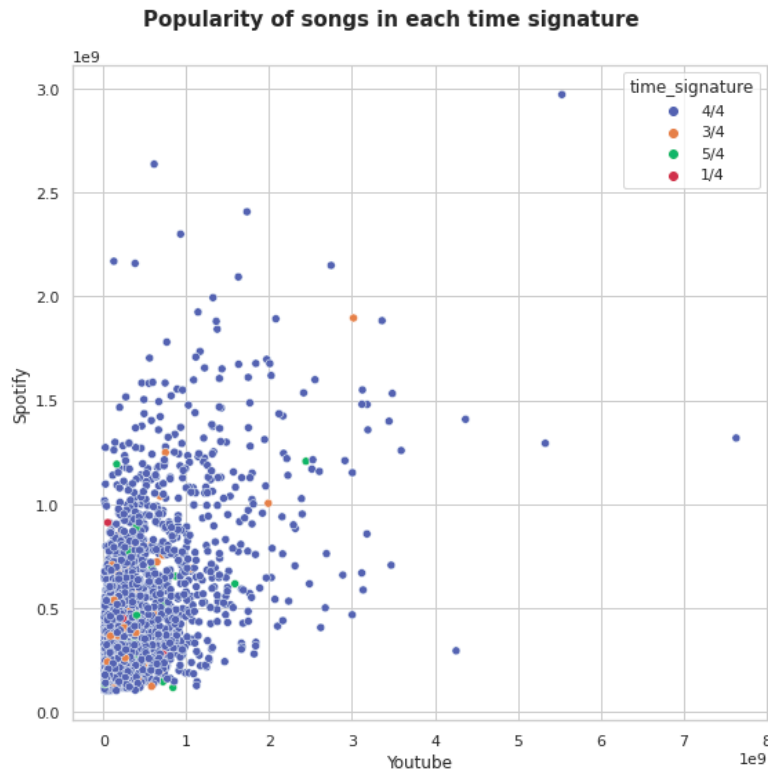
From the strip plot, we can see that most of the keys' clusters are fairly similar. The majority of the population in every key concentrates around the under 1.5 billion play counts for Spotify and under 2 billion play counts for YouTube, which is largely similar to the general dataset. However, we can observe that some keys might be more likely to contain massive hit songs, for example C# containing three over 2 billion songs on Spotify and one song with over 5 billion views on YouTube, or D containing the most popular song on YouTube and a song with over 2 billion plays on Spotify... But overall, it seems that the song's key doesn't seem to have much correlation or impact on their number of play counts on both platforms.

Time signature

The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). [Reference link](#) [7]



Predictably, the 4/4 time signature, which is often considered the most popular, or the “default” time signature for modern music, is overwhelmingly popular among the popular songs as well.



It seems that there is not a lot of information to be gained when we take a look at the scatter matrix. The 4/4 time signature dominates in every single YouTube and Spotify play counts ranges, and almost all of the popular outliers are written in 4/4 as well. Overall, the 4/4 seems to be the safest, and most popular option for time signature for artists when writing their songs.

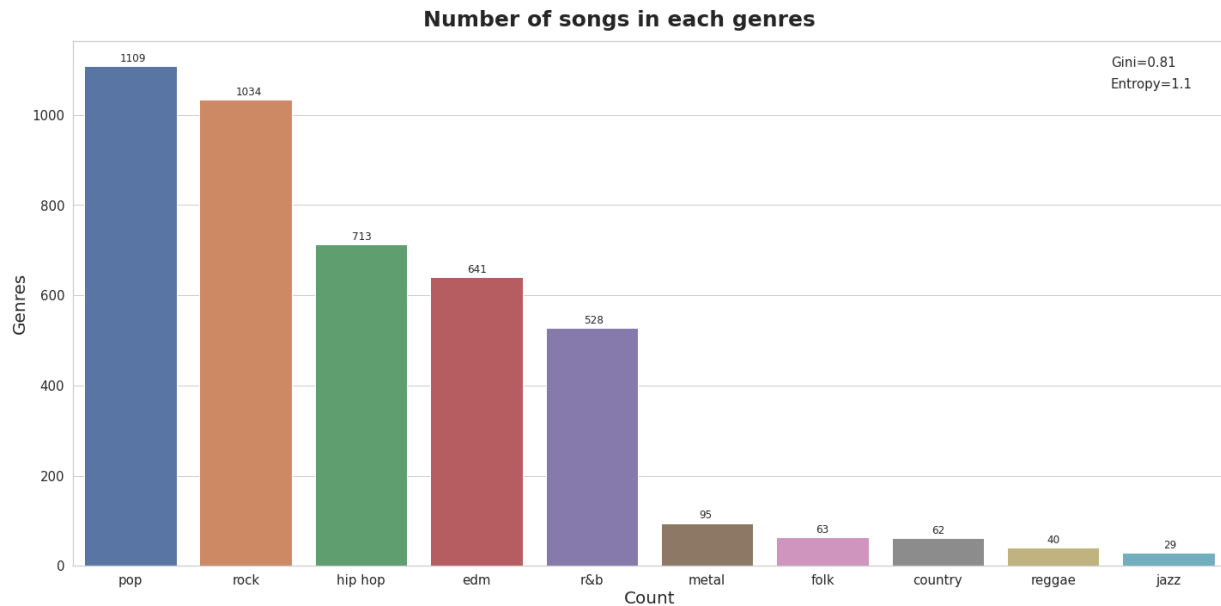
D. Genre analysis

Genre is an important aspect for a song’s popularity. In this section, we want to analyze the most popular genres right now, the genres’ relationship to their performance on the streaming platforms, as well as the musical features of each genre to get a sense of the current music trend among the popular songs on YouTube and Spotify.

Like we specified above in the Preprocessing section, for the genres we will group the subgenres that we scraped into 10 larger genres: Pop, Hip hop, Electronic - Dance, R &

B, Rock, Metal, Folk, Country, Reggae, Jazz. It's worth noting that the genre is a multi-valued attribute.

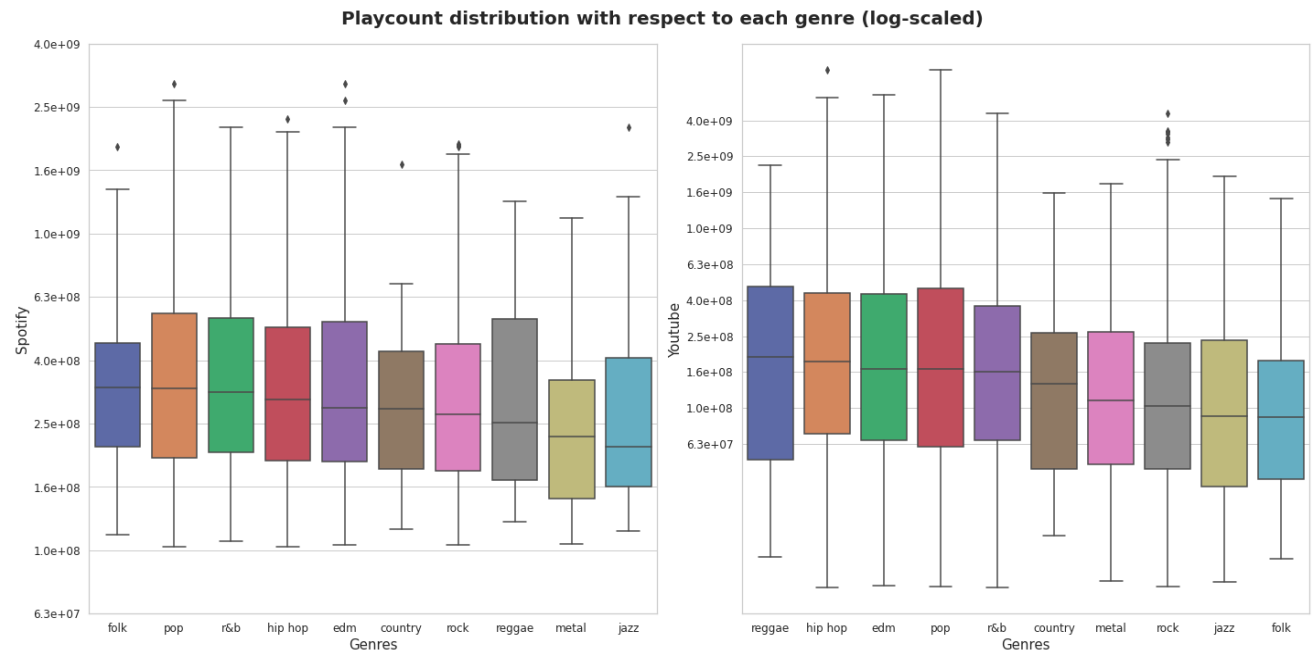
Genre counts



Like we specified above in the Preprocessing section, for the genres we will group the subgenres that we scraped into 10 larger genres: Pop, Hip hop, Electronic - Dance, R & B, Rock, Metal, Folk, Country, Reggae, Jazz. It's worth noting that the genre is a multi-valued attribute.

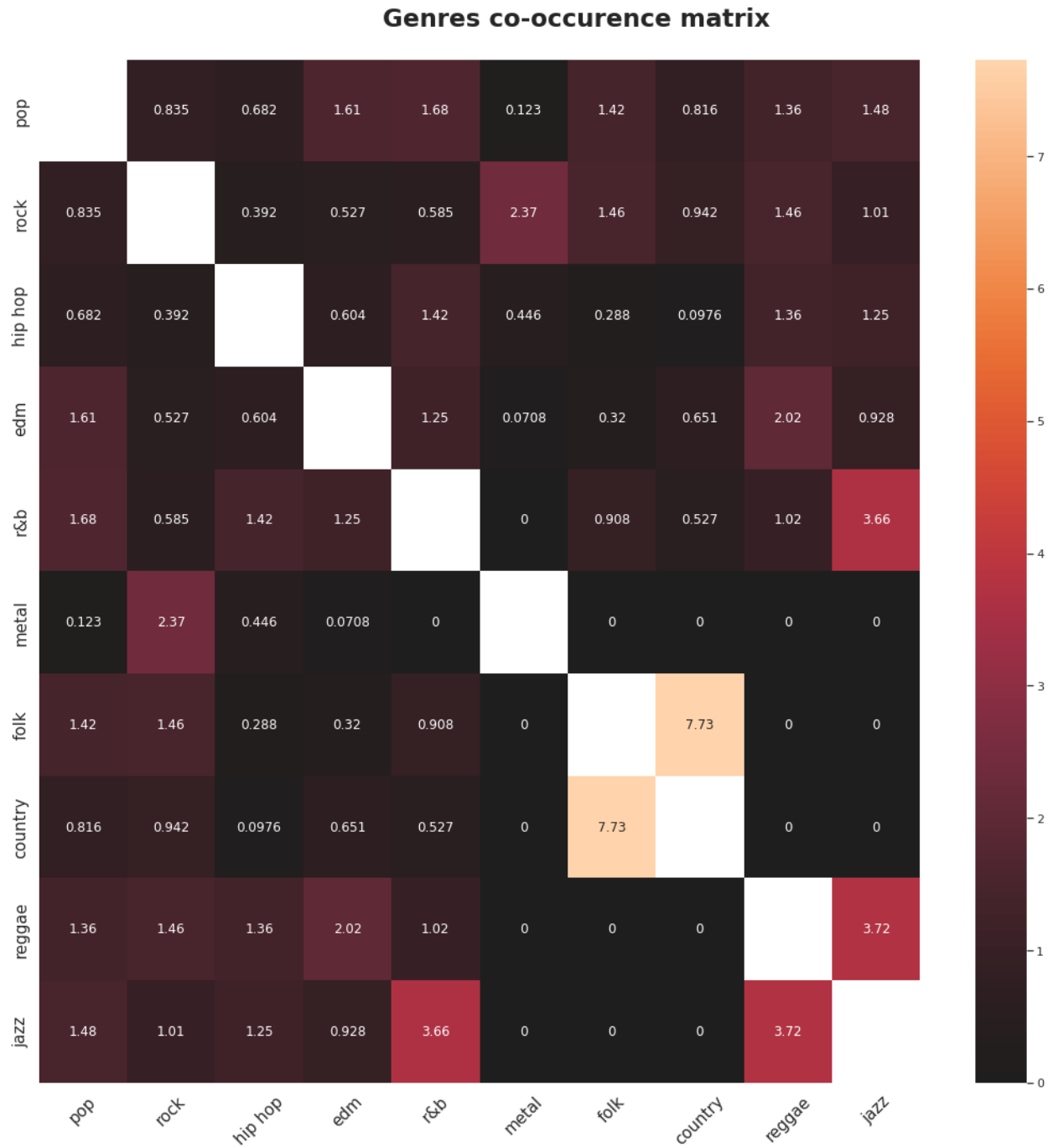
We can see that pop is the most popular genre (1109 out of 4031 songs), with rock being a close second. Hip hop, Electronic - Dance, and R&B are fairly popular as well. Predictably, less “mainstream” genres such as Metal, Folk, Country, Reggae and Jazz are significantly less popular compared to others.

Play counts distributions with respect to genres



We observed that the median value of the play counts for every genre are not significantly different for both Spotify and YouTube, with reggae and jazz may be falling a little behind. The minimum lower quartile between the songs are not quite similar too, since we are doing analysis on the most popular songs and collecting songs with over 100 million views on Spotify. However, the genres pop, hip hop, r & b, edm, which are the most mainstream genres of the music industry currently, tend to have a greater upper quartile range and a higher maximum compared to the other genres. This indicates that these genres tend to have a lot more super popular hit songs that stand out from the others.

Genre co-occurrence



Next, we want to analyze the frequency of co-occurrences between different genres. Above is the co-occurrence matrix between the genres. The coefficient is calculated as below:

The value in each cell is the normalized number of times the genre in the corresponding row and the genre in the corresponding column appears together in a song. The higher the cell value, the more often two genres are present in the same song. Since the raw co-occurrence number is heavily affected by how often one genre appears in the dataset, we normalize it by dividing the raw value with the expected number of times two genres appear together. The expected value is computed by the following formula:

$$Expected = \frac{Count_{first\ genre} \times Count_{second\ genre}}{Total\ count\ of\ all\ genres}$$

We can see that most “mainstream” genres like Pop, Rock, Electronic are generally more diverse, having co-occurrence with every other genre, but not too correlated with one particular genre. Genres like Pop and Rock are big genres that borrow a lot of elements from other genres, so them having co-occurrence with every other genre is understandable.

The first thing we can notice is the exceptionally high co-occurrence (7.73) between country music and folk music. This makes sense realistically, since these two genres are very similar in terms of style, as well as their grassroot and working class origin.

We can see the co-occurrences between Jazz and R & B, Jazz and Reggae are quite high as well. Once again, this reflects the similarities in sound and in origin of these genres. Both Jazz and R & B are of African - American origin, with Jazz being a predecessor for R & B to take a lot of inspirations from. Both of these genres are usually associated with “bluesy” and soulful sound, taking inspiration from Blues music as well.

For in depth information, please reference these sources: [Jazz \[8\]](#), [R&B \[9\]](#)

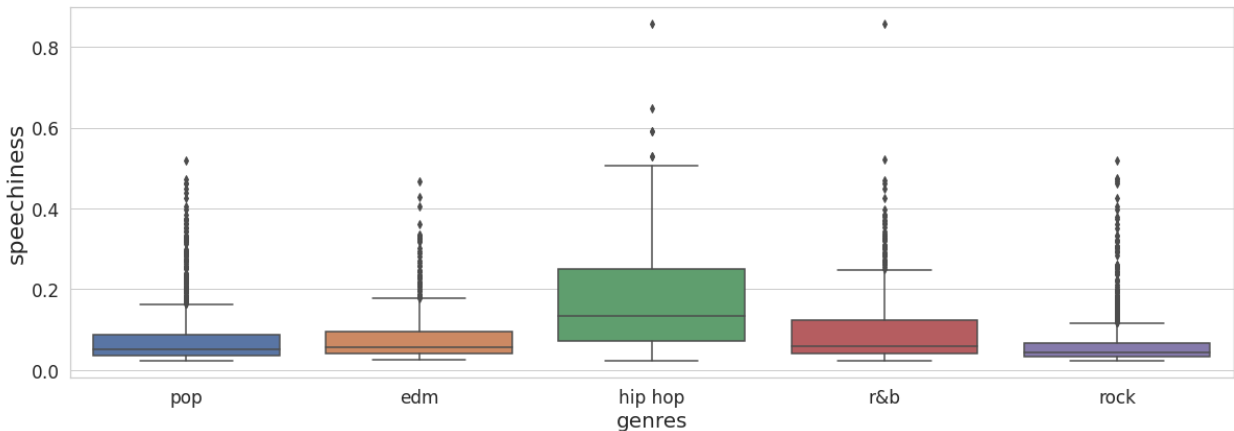
For the relationship between Jazz and Reggae, we have to admit our knowledge on these subjects are severely limited, and we don’t have enough time to research more thoroughly into this. Regardless, the high co-occurrence observed between these two genres seem quite interesting and informative to us, and can potentially be investigated further.

The co-occurrence between Rock and Metal is quite high as well. Once again, this makes sense since Metal can essentially be considered a subgenre of Rock music, with both genres are known for their signature electric guitar sounds.

Some genres don’t correlate at all, mostly genres with different backgrounds such as: reggae and metal,

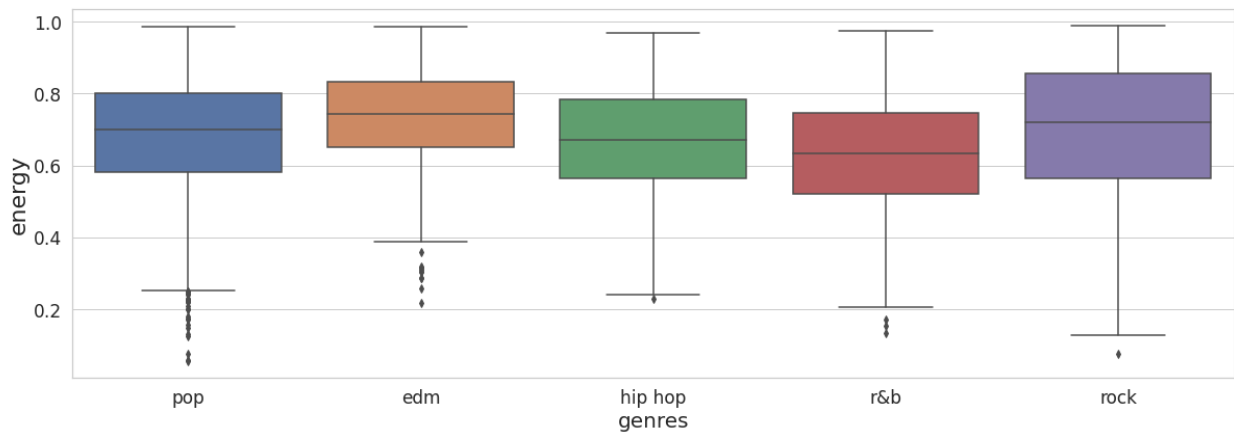
Genres and numeric audio features

We investigate some numeric audio features of songs in the five most popular genres: pop, rock, hip hop, r & b and electronic - dance, to see the musical patterns across these genres.

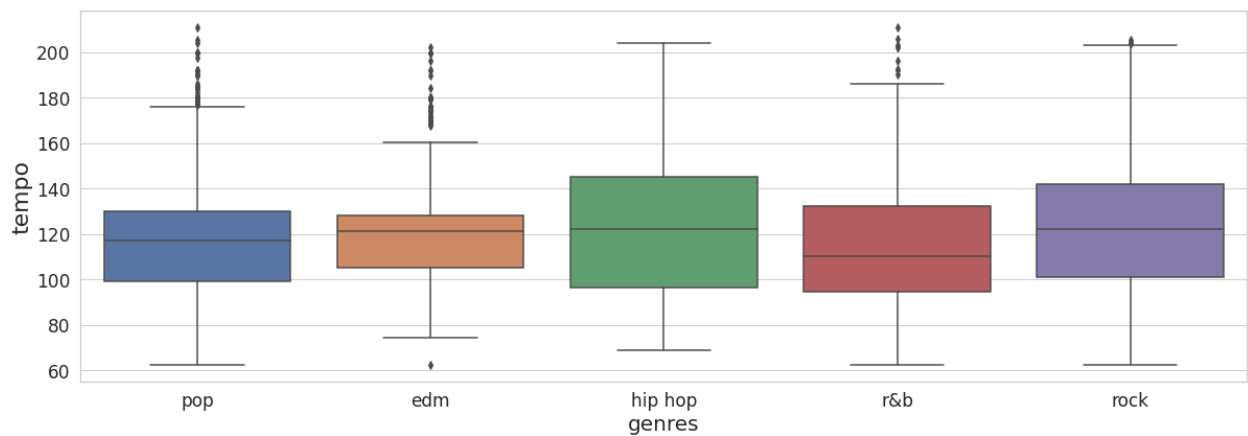


In terms of danceability, it seems that hip hop is the most “danceable” genre compared to the rest of the other 4. Hip hop and rap are characterized by catchy beats, so this comes off as no surprise. However, we expected the danceability of EDM to be higher, since it is a genre that is quite popular in parties, festivals, night raves... Rock seems to be the least danceable genre of these 5.

Most genres seem to be quite low on speechiness, which means that they rarely feature spoken vocals. Hip hop, of course, is generally much higher in speechiness compared to others, since hip hop is essentially speaking or chanting over rhythmic beats.



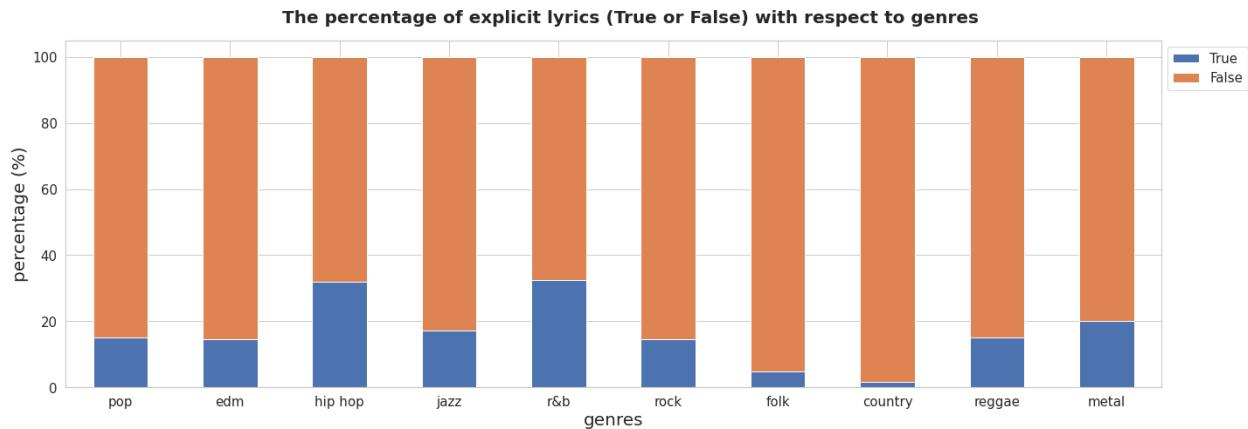
Popular genres tend to be quite high in energy, with rock being the most energetic genre. R & B seems to be a bit lower compared to other genres, indicating that the genre is more “chill” and laid back.



Hip Hop and Rock seem to have generally faster tempos compared to the other 3.

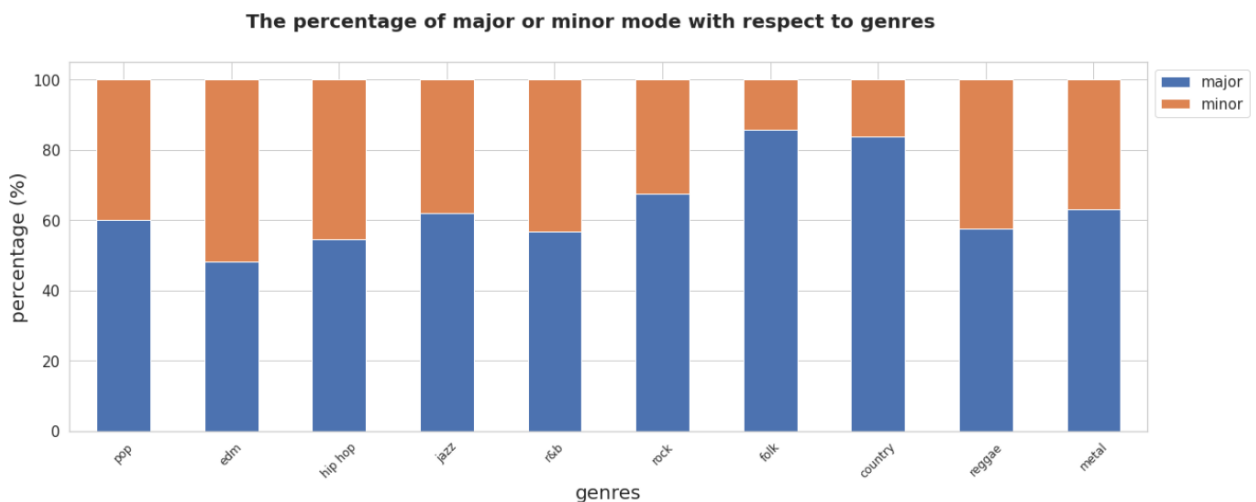
Genres and some categorical features

We will next perform analysis of the song genres and other categorical features in our data.



Unsurprisingly, hip hop (particularly rap music) is the genre with the most popular songs that contain explicit lyrics (about 40%). Hip hop has always been considered a “raw” and “ghetto” genre of music that originates from lower class African American communities and frequently deals with subjects such as: gangs, crimes, racism, injustices... R & B is also a genre originating from African-American community, and many rap songs share elements from R & B and vice versa, which explains the high frequency of explicitness in their lyrics. We are quite surprised that R & B's frequency of explicitness is almost as high as Hip hop. Although they share a lot of similarities concerning the origin, the lyrical themes, rap borrowing a lot of elements from R & B... , normally R & B doesn't strike us as an overly explicit genre. Once again, our knowledge and experience is limited, so it's best we don't hypothesize anything that we're unsure of.

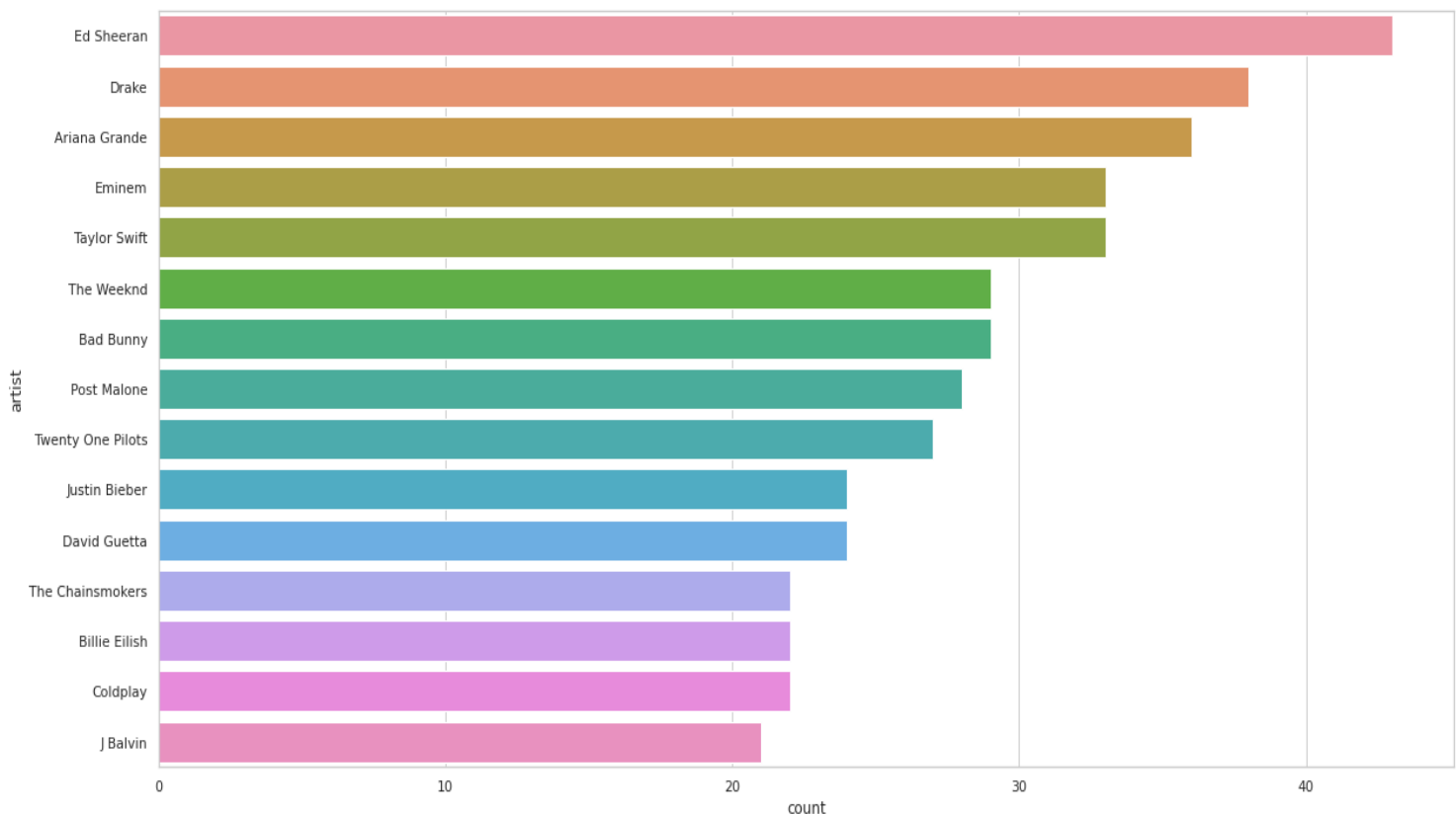
For more information on hip hop, please reference following the wikipedia article: [Hip hop music - Wikipedia](#) [10]



There is not quite a lot to analyze when it comes to the mode of the song with respect to genres. The only thing that caught our attention was the high percentage of major mode (over 80%) for the genres folk and country compared to the rest. This might indicate these genres tend to be more upbeat and joyful compared to other genres.

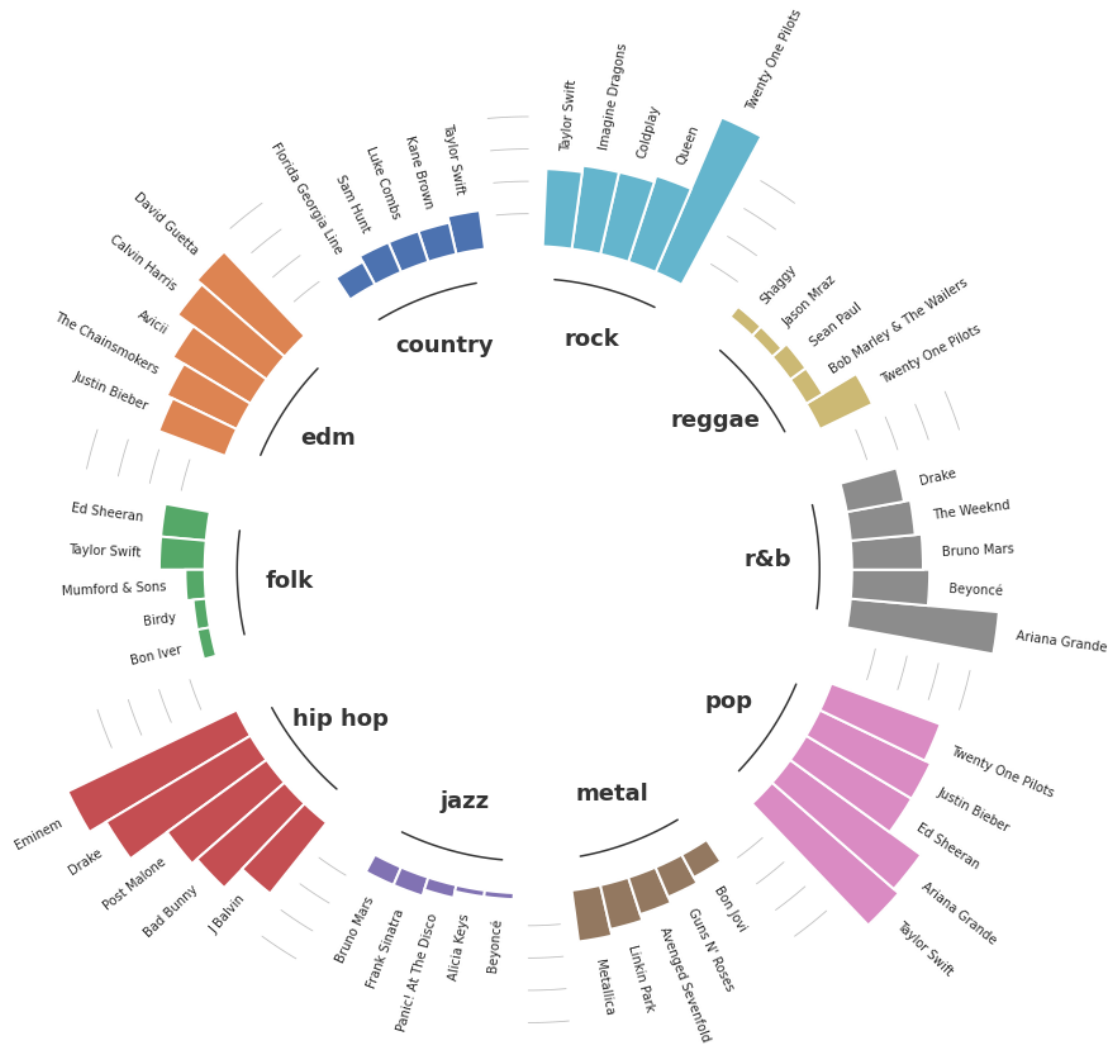
E. Artist analysis

The main artist of the song might be an important feature that can impact a song's popularity.



Above are the most frequent artists to have a popular song that appears on our dataset. Ed Sheeran leads as the most popular artist, with more than 40 songs that make the cut. Others like Drake, Ariana Grande, Eminem are not too far behind.

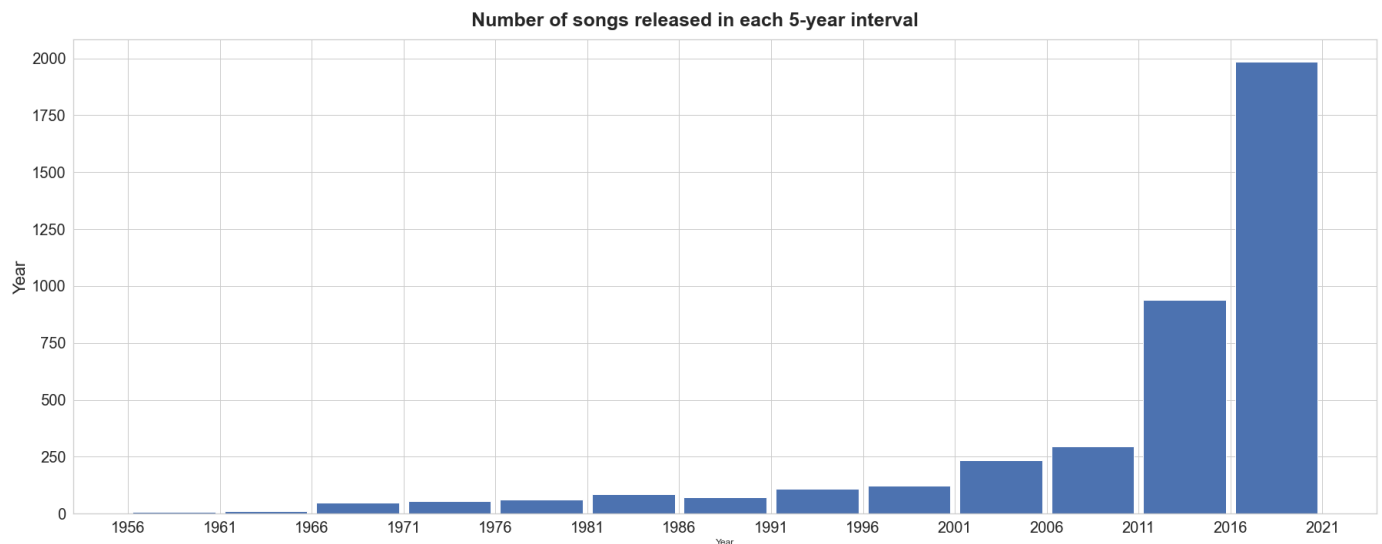
Most popular artists with respect to each genre



We visualized the 5 most popular artists for each respective genre. Recounting on the most popular artists earlier, almost all of these popular artists appear in the top 5 for either Pop, Hip Hop, Rock or Electronic-Dance. This proves once again the popularity of these genres in the music industry. A lot of the most popular artists appeared in the top 5 for multiple genres, such as Twenty One Pilots (rock, pop), Ariana Grande (pop, R & B), Taylor Swift (pop, country, folk)...

F. Release date analysis

The release date of the song might have an important impact on their popularity on the streaming platforms, since music trends evolve and change constantly throughout time. Also since music streaming platforms are relatively new compared to the lengthy history of the music industry, it's likely that newer songs tend to be more popular, given that a lot of old songs, even if super popular at the time, might have fallen out of popularity by now, while newer songs are more likely to gain popularity.

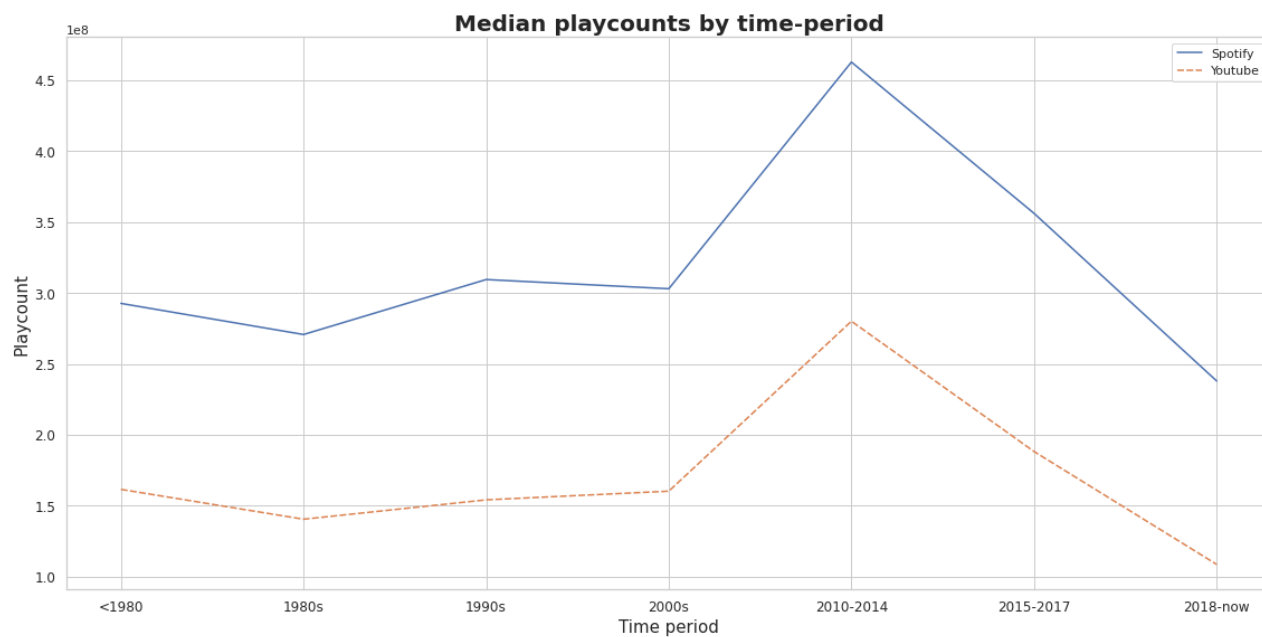
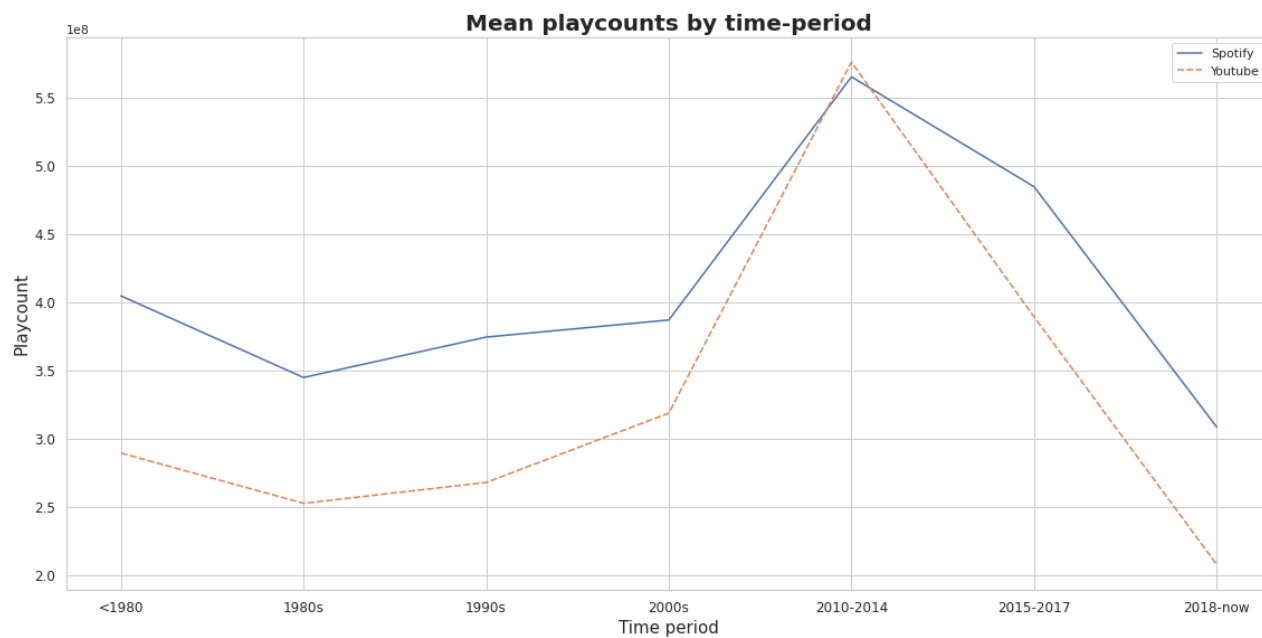


The plot above displays the number of hit songs based on their released year, broken up into 5-year intervals. Just as expected, the number of popular songs increased as time progressed. This is because we are analyzing popular songs on streaming platforms, which are songs with a high number of play counts. So, it's normal that songs from older eras, particularly before 2011 tend to get significantly fewer views on modern streaming platforms. Newer songs go more in line with the current music trends, are more advertised, and generate more attention from the public and the media, so it makes sense that modern songs are more likely to generate a lot of plays on streaming platforms.

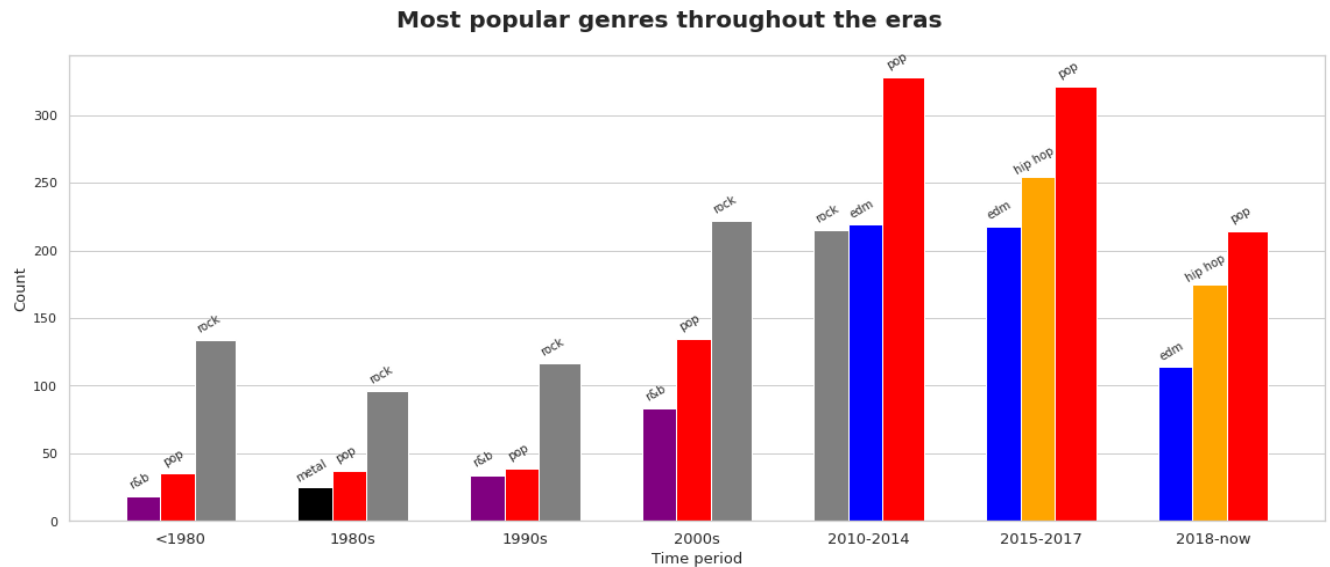
We observe a big jump from the pre-2011 eras to the 2011-2016 era. The 2011-2016 is also the era when sites like YouTube and Spotify developed rapidly, thus, songs released from this era are more likely to be popular on these platforms.

There is a big jump in numbers of popular songs from the 2011-2016 era to 2016-2021 era as well. This period of time saw massive increase in user base and popularity for these already famous platforms. With the increase in number of users, the newer famous songs tend to gather much more plays than songs just from a few years before. For example, the song "*Gangnam Style*" was a huge hit for its time, and was famous for being the first song to achieve 1 billion views on YouTube, which was deemed a

massive achievement at the time. But by now, a lot of famous songs had already crossed the 1 billion mark.



By plotting the mean and median play counts for each time period, we noticed a trend that applies for both YouTube and Spotify. Both the mean and the median number of play counts dropped from pre-1980 eras going into the 1980s, and slowly increased going into the 1990s and 2000s. Then, we noticed a huge spike going into 2010-2014 era for both YouTube and Spotify, and from then on decreasing for later eras. While the 2010-2014 era has less popular songs compared to its later time periods, on average the songs from this era seem to be more popular.



The above bar charts display the most popular genres across time periods. Rock steadily remains the dominant genre with pop as its second and R & B (or metal for the 1980s) as the third. This reflects the trends of the time, with rock bands from many time periods dominating the music industry. This trend remained until the 2010-2014 era where rock was overtaken by pop and the rise of electronic-dance. From then on, Pop consistently remained as the most popular genre, with hip hop and electronic-dance being popular as well.

V. Conclusion

In this project, we scraped and analyzed popular songs on YouTube and Spotify with the target variables being their respective play counts on these streaming platforms. We investigated multiple audio features to gain insights: their general distribution, how impactful are they to the songs' number of play counts... We also analyzed the genres

of the songs to gain information on popular genres, the typical trends across genres and how the songs' popularity might be impacted by its genres. Finally, we analyzed the trends of music across different eras, how a songs' release date might affect its popularity...

VI. References

- [1] [Spotify playlist used](#)
- [2] [Spotify API for Web Developers](#), Spotify
- [3] [YouTube API](#), YouTube
- [4] [Wikipedia API](#), Jon Goldsmith
- [5] <https://hubguitar.com/music-theory/major-versus-minor>, Hub Guitar
- [6] [Key \(music\) - Wikipedia](#), Wikipedia
- [7] [Time Signature | Music Appreciation 1 \(lumenlearning.com\)](#), Natalia Kuznetsova, Tidewater Community College.
- [8] [Jazz - Wikipedia](#), Wikipedia
- [9] [Rhythm and blues - Wikipedia](#), Wikipedia
- [10] [Hip hop music - Wikipedia](#), Wikipedia

Contributions

- Subject proposal: Divided between all members
- Scraping:
 - Spotify Scraping
 - Nguyễn Tuấn Dũng (50%)
 - Phùng Quốc Việt (50%)
 - YouTube Scraping
 - Phùng Quốc Việt
 - Wikipedia Scraping
 - Vũ Quốc Việt (50%)
 - Nguyễn Tuấn Dũng (50%)
- Data cleaning and integration
 - Phùng Quốc Việt
- Exploratory Data Analysis and Data Visualization
 - Analyze target variables
 - Programming: Nguyễn Tuấn Dũng
 - Analysis: Vũ Quốc Việt
 - Analyze numeric features
 - Programming
 - Plot distributions: Vũ Quốc Việt
 - Statistics calculation: Vũ Quốc Việt
 - Correlation matrix between numeric features, scatter plot: Nguyễn Tuấn Dũng
 - Correlation P value evaluation: Phùng Quốc Việt
 - Analysis: Phùng Quốc Việt (50%), Vũ Quốc Việt (50%)
 - Analyze categorical features
 - Programming:
 - Count plot of each variable: Vũ Quốc Việt
 - Popularity scatterplot of each variable: Vũ Quốc Việt
 - Hypothesis testing, boxplot: Phùng Quốc Việt
 - Mode, key lollipop: Nguyễn Tuấn Dũng

- Analysis: Phùng Quốc Việt (50%), Nguyễn Tuấn Dũng (50%)
- Genre analysis
 - Programming:
 - Genre count: Nguyễn Tuấn Dũng
 - Playcount distribution of each genre: Nguyễn Tuấn Dũng (50%), Vũ Quốc Việt (50%)
 - Genre co-occurrence matrix: Nguyễn Tuấn Dũng (50%), Phùng Quốc Việt (50%)
 - Analysis: Nguyễn Tuấn Dũng (70%), Vũ Quốc Việt (15%), Phùng Quốc Việt (15%)
- Artist analysis
 - Programming:
 - Most popular artists bar plot: Vũ Quốc Việt
 - Most popular artists across genres bar plot: Phùng Quốc Việt
 - Analysis:
 - Phùng Quốc Việt (40%)
 - Vũ Quốc Việt (60%)
- Release date analysis
 - Programming:
 - Number of songs in time intervals, Mode and Median play counts: Vũ Quốc Việt
 - Most popular genres throughout the eras: Phùng Quốc Việt
 - Analysis:
 - Nguyễn Tuấn Dũng (50%), Phùng Quốc Việt (50%)
- Slides: Vũ Quốc Việt