Hits on Spotify: Anatomy of Tracks, Artists and Streams

Chen Xu
RWTH Aachen University
Aachen, Germany
chen.xu1@rwth-aachen.de

ABSTRACT

Music is the preferred leisure activity for most individuals. While with over millions of songs, tiny fraction of them are spread widely and can make the Spotify global top ranking lists. This project aims to investigating the properties of hit songs and artists and streaming pattern of hit songs on Spotify. A two-process data collection is done to get the data, following that are the hit songs, hit artists and dynamics of stream analysis. The results show that majority of hit songs appear only once on the list, while a few do spend an extremely long time on the ranking list. Looking at patterns characterizing artists, the results show that some artists are significantly more productive than others while vast of the artists only have one hit song. Finally, there are some patterns on the dynamics of streams identified. The hits analysis on Soptify are not only interesting but also crucial for our understanding for hits properties, and also knowing the power of idea "data has a better idea".

KEYWORDS

Spotify, Hits, Visual Analysis

1 INTRODUCTION

Spotify is one of the most popular audio streaming platforms in the world. They have an API for developers to utilise their huge database of music to build interesting applications and uncover insights into our listening habits. Spotify currently holds 70 million songs in its library, with 60,000 added every day [1]. That's nearly one per second. but only a tiny fraction of these songs can make the Spotify Global Top 200 ranking list. I am interested to know which songs made it into the Global Top 200 charts on Spotify over the past five years from 2017 to 2021, which artists were most popular and what trends I could identify from the hit songs stream. I believe we can get some new discoveries from the old crowd favourites. Let's get started the anatomy journey by research framework.

1.1 Research Framework

The research is divided into 4 stages, starting with stage 1 research proposal including define the research objectives and research questions, stage 2 data collection from Spotify Charts website and data extraction from Spotify API. Audio features and artists information are collected in this part and three main data sets are consequently created and preprocessed. The stage 3 consist of analysis of hit songs. hit artists and streams and the results and conclusions are finalized in stage 4. These stages are summarized in Figure 1 Research Framework.

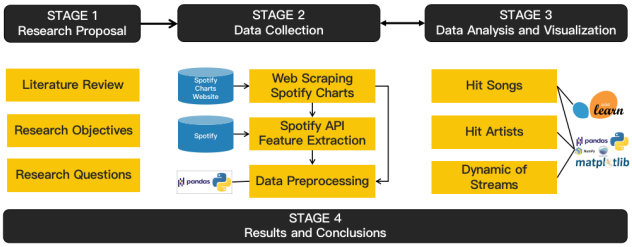

Figure 1: Research Framework

1.2 Research Questions

The analysis of this project is guided by research questions. The purpose of this research is to answer the following research questions:

(1) What is the pattern of longevity of hit songs?
(2) What is the temporal trend of the audio features?
(3) What are the characteristics of hit songs?
(4) What patterns of repeated success among the artists?
(5) Musical genre analysis over Genre Network based on artists.
(6) What are the relations between position, length of stay of a song on the list and song streams?
(7) What is the temporal changes in streaming, can we capture the observed stream patterns?

2 DATA

In order to spur the research, a data collection process is done. My data sets are composed of data from two sources: the global top 200 charts on Spotify Charts website [4](Figure 2, this website is no longer available starting June 3, 2022) and Spotify API [5]. So the collection of data in a two-step process. First scraping Global Top 200 Charts from the Spotify Charts website and then use the results of that scrape to pull data from spotify API.

2.1 Spotify Global Top 200 Charts Collection

Spotify does not provide data on the number of streams through its API, so the stream data for the songs on the Global Top 200 charts come from the Spotify Charts website [4](Figure 2). The website allows you to filter the chart by region, time period and date. For the purpose of my analysis, I am using the global, weekly ranking data from January 1st, 2017 to December 31st, 2021 (in total 261 weeks). Although Spotify Charts provides a handy "download to CSV" link at the top of the page, manually downloading each week for the full year is too time-consuming, and will also result in 261 files to be imported for analysis. Instead, I used Python to write a script which would scrape each week's chart, process the file, and join the data together to create a single CSV file for analysis. The data scraped from the website does not contain features for the date(e.g.

month and year), so I needed to add them during processing. I also extracted the song ID from each song's URL and stored it as a new feature. These features will be needed later to get the audio features for each song from the Spotify API. After processing, the streams data are stored in a data frame, with column features including song name, position, artist, streams, URL, ID, year, month, etc.. In total, the scrape yields in total 52200 position entries that between January 1st, 2017 to December 31st, 2021 (in total 261 weeks).

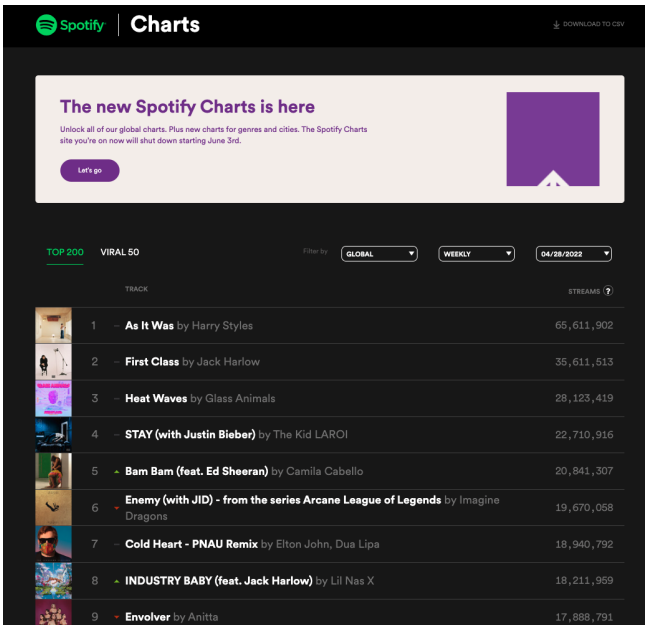

Figure 2: Spotify Charts Website

2.2 Spotify Features Extraction

With Spotify, they provide developers access to some of their data regarding playlists, users, and artists through their Web API. It can be used for data extraction and analysis purposes. For example, When grabbing each song from an album, we can obtain song information such as song name, album, release date, length, and popularity. A collection of music songs consists of various types of data. for example, data could consist of music audio files or metadata such as track title and artist name. More importantly, Spotify's API allows us to extract a number of "audio features" such as danceability, energy, instrumentalness, liveness, loudness, speechiness, acousticness, and tempo. To add some more context to the analysis, we can use the Spotify API to get the audio features for each song that appears in the Global Top 200, as well as some information about the artists. What I used is a wrapper utility around Spotify API called SpotiPy to make handsome one-line-long requests instead of explicitly reaching the endpoints. The Spotipy library provides a Python interface to the API, which makes it easy to get started. The first step is to log into the Spotify Developer site [5] and register an app, the result is shown in Figure 3. This will generate a client ID and client secret used to authenticate API requests. After registration, I then set up the connection to the API.

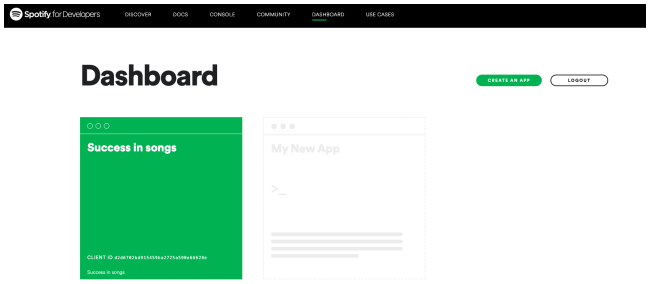

Figure 3: App on Spotify Developer Site

Getting the audio features of a song is simply a call to the audio_features() function, passing in the song ID as a parameter. There is a description of each of the features in the API documentation on Spotify developer website[3], see the Appendix A Spotify Audio Features in the end. Using the song ID from the streams data set collect in the last part, I looped through all songs ID and pulled the audio features, adding them to a new data frame, which was then written out to a CSV file. The final song features data frame including features: song id, energy, liveness, tempo, speechiness, acousticness, danceability, key, duration_ms, loudness, valence. Finally, to get the artists information, I used the search feature of the API. The streams data does not contain the artists ID, so to find the artist information, I needed to search on the artist name. Some tracks did not have an artist, so I also had to filter the null values from the list. I again used the approach of looping through the list and calling the API to pull the information for each artist, then writing the results out to CSV. Because we are using a search function to find the artist information, Spotify often returns more than one result (as it does not perform an exact match search). After searching a few artists and checking the results, the API seemed to be returning the correct artist first, so I used the assumption that the first match would be the artist I wanted. I also chose to only keep the artist ID, name, genres, popularity and number of followers for my analysis. In summary, using a combination of the Spotify Charts website and the Spotify API, I was able to generate three data frames containing data on positions, streams, song audio features and artists information, process and write them out to CSV files.

2.3 Data Description

The "Spotify Global Top 200 Data Set" is a data set that contains ranking of the 52200 position entries. For each song, it includes information such as the Position, Track Name, Artist, Streams, URL, Week Start, Week End, ID, URI. Track Name, Artist, URL, ID and URI are categorical variables, the others are quantitative variables.

The "Song Features Data Set" is a data set that contains various audio statistics of the 4918 hit songs on Spotify. For each song, it includes information such as song id, energy, liveness, tempo, speechiness, acousticness, danceability, key, duration_ms, loudness, valence. Song id is categorical variables, the others are quantitative variables.

The "Artist Information Data Set" is a data set that contains various basic statistics about total 955 artists. For each artist, it

includes information such as artist ID, name, genres, popularity and number of followers. Popularity and number of followers are quantitative variables, the others are categorical variables.

3 HIT SONGS

3.1 What is the pattern of longevity of hit songs?

A hit song is a song which appears on the Spotify global top 200 ranking list for at least one week. There are great differences between these hit songs. Some hit songs appeared on the list for a singe week while some retain their status for hundreds weeks. To illustrate this, I measured the length of stay on the list for all hit songs (Figure 4 and Figure 5).

The result shows that majority of hit songs appear only once on the list, while a few do spend an extremely long time there. Distribution of Longevity of hit songs based on the number of weeks they stayed on the list. The difference of Figure 4 and 5 is the scale of y axis. The number of songs in Figure 5 is shown in logarithmic scale to account for both the large number of short stays and the few exceptionally longer presence.

The longest stay during the observation period is Shape of You by Ed Sheeran which remained on the list for a record of 257 weeks. Other examples of long-lasting hit songs are Say You Won't Let Go by James Arthur (255 weeks), Perfect by Ed Sheeren (250 weeks), Goosebumps by Travis Scott (224 weeks) and Jocelyn Flores by XXXTENTACION (215 weeks).

Figure 6 shows the best rank a hit song achieved on the Spotify global top 200 ranking list vs. the number of weeks it stayed on the list. The size of the dots indicates the number of hit songs with the same attributes. Overall, the better a song's best ranking, the higher is its probability of staying longer on the list.

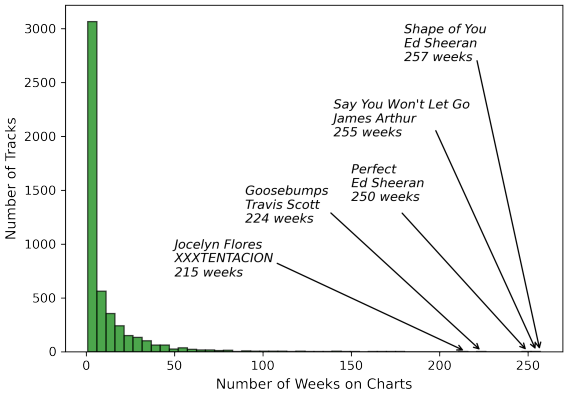

Figure 4: Longevity of hit songs (1)

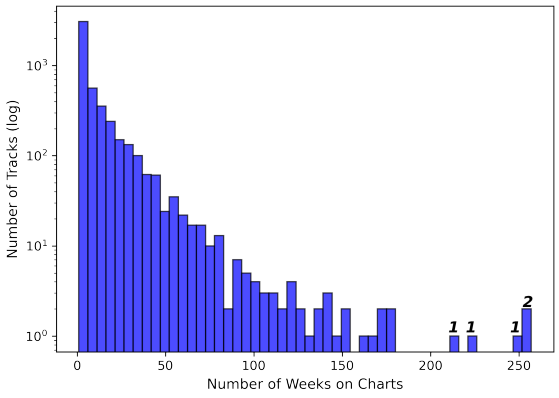

Figure 5: Longevity of hit songs (2)

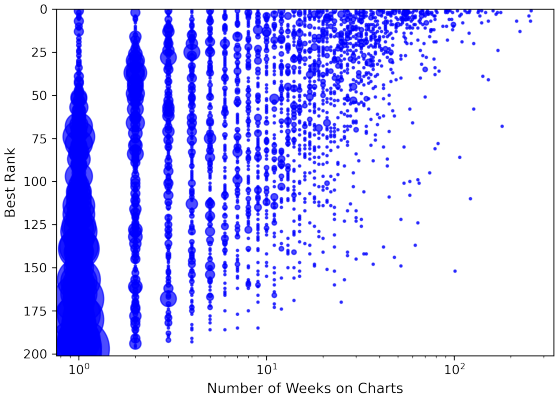

Figure 6: Longevity of hit songs (3)

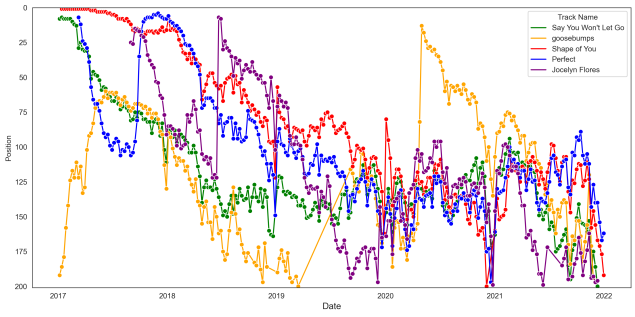

Figure 7: Top 5 songs on the ranking list

Each song is a story, and those widely circulated songs, but also resonate with many people. Figure 7 show the trend of the top 5 songs according the length they stay on the ranking list. I deep into two hit songs to see what stories we can find there. Figure 8 and Figure 9 shows the calendar map of the best position of Shape of You and Jocelyn Flores in every month. The number on the figure indicate the best rank of the song in the corresponding month. Shape of You became the most-streamed song when it was released as a single in January 2017 and keep a high position the whole year.

The music and it's rhythm just makes people want to get up and dance along which maybe the main reason why it gets so famous. Even though it has a downtrend for the past five years but still alive and actively appeared on the ranking list. Jocelyn Flores is a song written by XXXTENTACION for his friend who suffered from depression and committed suicide in May 2017. In June 2018, he was murdered in an apparent robbery. We can see that the song was expected to return to the top the month of his death following his murder.

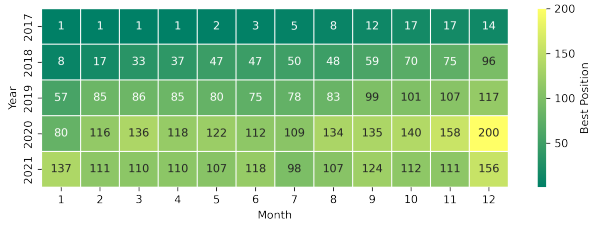

Figure 8: Best Position of Shape of You

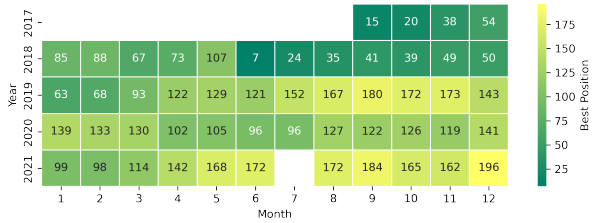

Figure 9: Best Position of Jocelyn Flores

3.2 What is the temporal trend of the audio features?

In this part, I am going to analysis the temporal trend of the audio features over the past 5 years. Many questions interest me. Do tracks become more danceable or more relaxed over the past 5 years based on the audio attribute levels? Is there a seasonal mood change? How COVID-19 restrictions influenced song features on Spotify Global Top 200 charts? The time-series box-plot for 10 different numeric audio features, including energy, liveness, tempo, speechiness, acousticness, danceability, key, duration_ms, loudness, valence, are shown in season level (Figure 10,11) and year level (Figure 12, 13).

We can see that there are not so much differences between seasons. Summer have a higher acousticness compare to other seasons. There are also not too much differences between different years. The Year 2019 is quite special, it has lower energy, higher acousticness and lower valence. Valence have slightly increased in year 2020.

3.3 What are the characteristics of hit songs?

The following is going to analysis the hit songs characteristics based on the audio features. A K-means++ clustering is done on

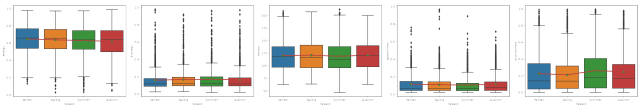

Figure 10: Seasonal audio features (1)

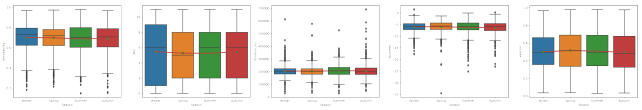

Figure 11: Seasonal audio features (2)

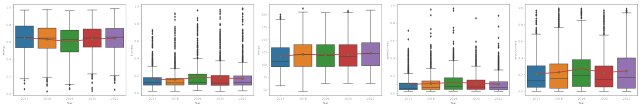

Figure 12: Yearly audio features (1)

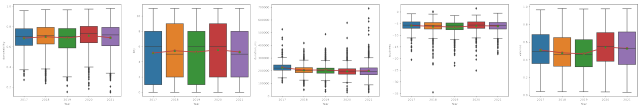

Figure 13: Yearly audio features (2)

the audio features. All the 12 audio features that can get through Spotify API are listed in Appendix A Spotify Audio Features. In order to effectively reduce the complexity of the problem by using the Principal Component Analysis (PCA) to make the dimension reduction. From the Figure 14 Cumulative Variance Plot after PCA, we can observe that each of the principal components explain a pretty considerable amount of variance. And it is a good rule of thumb to preserve around 80% of the variance. Therefore, I select the 6 most important principal components to incorporate in the k-means++ algorithm.

In order to implement k-means++ clustering, I also need to select a number of clusters, k, which distinctly splits the data. I use the elbow method, which is arguably the most popular technique, to help me select the number of clusters. Looking for a elbow in the WCSS graph(Figure 15) in which before the elbow would be steeply declining, while after it would be much smoother. The number of cluster 6 is the ideal number of clusters.

The Figure 16 and Table 1 shows the cluster analysis results. We can see that cluster 1 has high speechiness, cluster 2 has high liveness, cluster 3 has low valence, cluster 4 has low instrumentalness, cluster 5 has high acousticness, high instrumentalness, high loudness and cluster 6 has high key.

4 HIT ARTISTS

4.1 What patterns of repeated success among the artists?

To understand the patterns of repeated success among hit artists, I used all the unique hit songs released by them in the past five

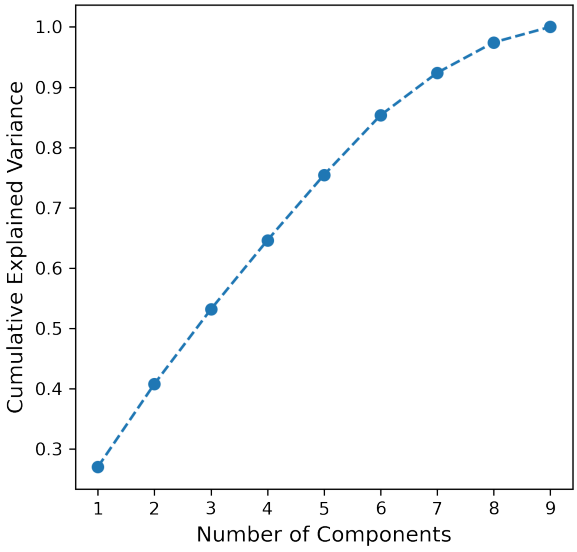

Figure 14: Cumulative Variance Plot after PCA

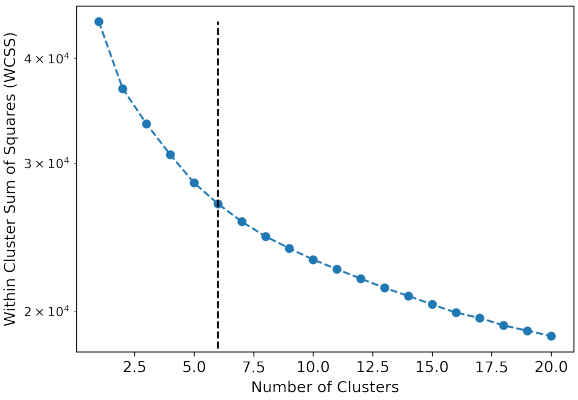

Figure 15: WCSS curve used for selecting the optimal number of clusters

years. After some data processing, I ended up with 4918 hit songs by 955 artists, indicating that the ranking list is dominated by a small number of artists with multiple hit songs. The difference of Figure 17 and 18 is the scale of y axis. As indicated by Figure 17 and Figure 18, some artists are significantly more productive than others while the vast majority of the artists only have one hit song.

Taylor Swift takes the lead in repeated success with 127 songs on the ranking list over the past 5 years. Similarly, Drake with 118 hit songs, Juice with 76 hit songs, BTS with 74 hit songs and Post Malone with 73 hit songs are also especially successful.

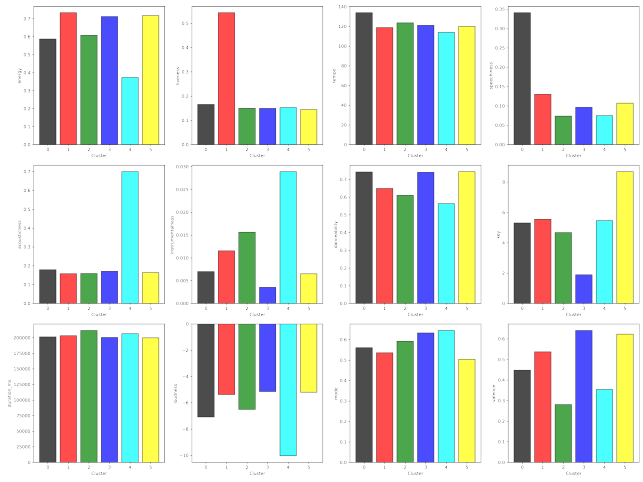

Figure 16: Clusters analysis results

Table 1: Clusters analysis results

Cluster #	Significant Attributes
1	High speechiness
1	High liveness
3	Low valence
4	Low instrumentalness
5	High acousticness, High instrumentalness, High loudness
6	High key

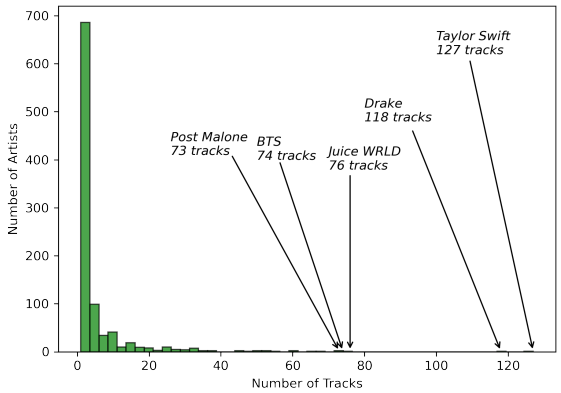

Figure 17: Repeated success of artist (1)

4.2 Musical genre analysis over Genre Network based on artists.

The genre perspective is important when analyzing the success of music, as each genre has its own unique taste audience. This part aims to better understand the collaboration between different

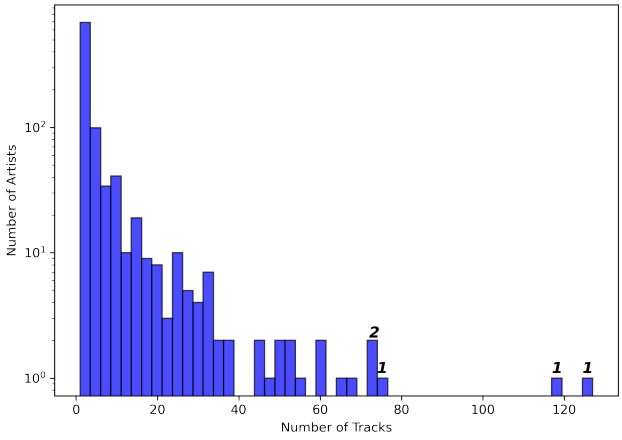

Figure 18: Repeated success of artist (2)

artists in a genre perspective and musical success. It may reveal the importance of how artists from different genres collaborate to produce a new hit song.

A genre network is modeled as a graph of nodes that are connected by edges. The nodes represent are genres. Oliveira et al. [2] claimed a Genre network building from tripartite model. As shown in the Figure 19, reduction from the tripartite (a) to the one-mode Genre Collaboration Network (c). The intermediate step is an Artist Network with genre information (b). Artists and genres are linked when hit songs involve both nodes. The example data used for building the genre network is shown as Table 2 Genre network example metadata. When two genres appear in a genre type of one artist at the same time, they are connected and these edges are undirected. For example. the artist Bryce Vine belongs to the genres ['electropop', 'pop', 'pop rap'], then 'electropop', 'pop', 'pop rap' are nodes that are fully connected with 3 edges between 'electropop' and 'pop', 'electropop' and 'pop rap', 'pop' and 'pop rap'. Mambo Kingz belongs to the genres ['reggaeton flow'], there will be only one node and no edges formed.

Figure 20 Genre network based on artists shows the resulting Genre Network, which has 521 nodes and 2202 edges. What I am interested is the giant connected component of the graph as shown in Figure 21 Giant connected component and the corresponding degree distribution. We can see that the rank of the degree follows the power law distribution. The top 5 genres (Table 3) according the degree is pop (154), dance pop (114), pop rap (80), rap (73) and edm (61).

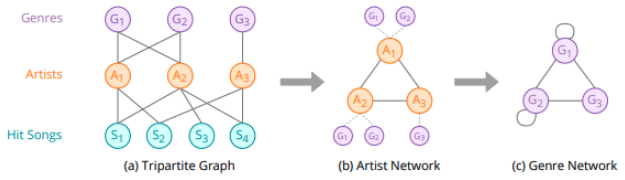

Figure 19: Genre network from tripartite model

Table 2: Genre network example metadata

Artist	Genres
Melanie Martinez	['alt z', 'electropop', 'pop']
Bryce Vine	['electropop', 'pop', 'pop rap']
Mambo Kingz	['reggaeton flow']
LX	['german hip hop', 'hamburg hip hop']

Figure 20: Genre network based on artists

Table 3: Top 5 genres according to degree

Rank	Genre	Degree
1	pop	154
2	dance pop	114
3	pop rap	80
4	rap	73
5	edm	61

5 THE DYNAMICS OF STREAMS

5.1 What are the relations between position, length of stay of a song on the list and song streams?

We can get the distribution of the stream for all hit songs from Figure 22 Dynamics of streams (1). To understand the dynamics of streams, I looked into the relationship between the average stream of each hit song and best ranking the hit song achieved on the ranking list (Figure 23) and the length of stay of a hit song on the list (Figure 24). Obviously, the more higher the stream, the better is its ranking on the ranking list. For most songs, we can also observe

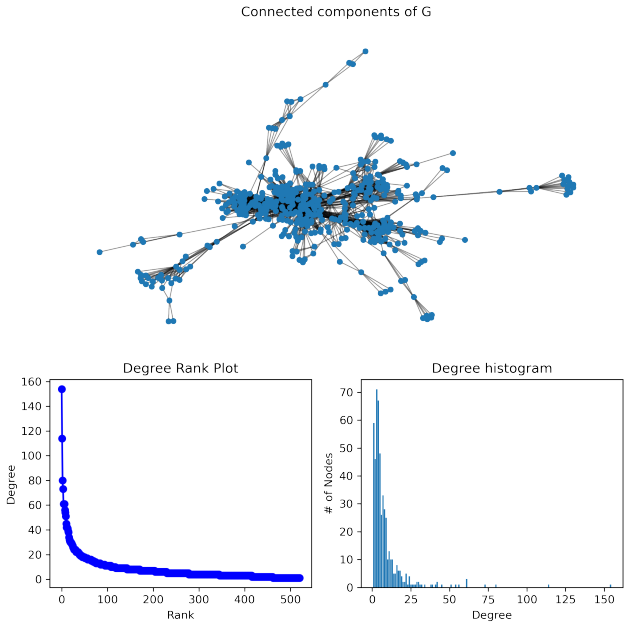

Figure 21: Giant connected component

a correlation between the best ranking and the stream. Surprisingly, I do not observe a strong correlation between the length of stay on the ranking list and the song stream (Figure 24). In summary, the best rank a song achieves on the ranking list is a good predictor of its stream: the better the rank, the higher the stream. The length of stay of a song on the ranking list can not be a good indicator of the stream yet.

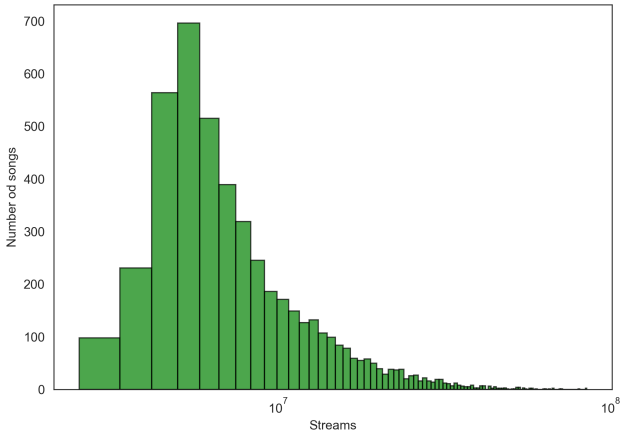

Figure 22: Dynamics of streams (1)

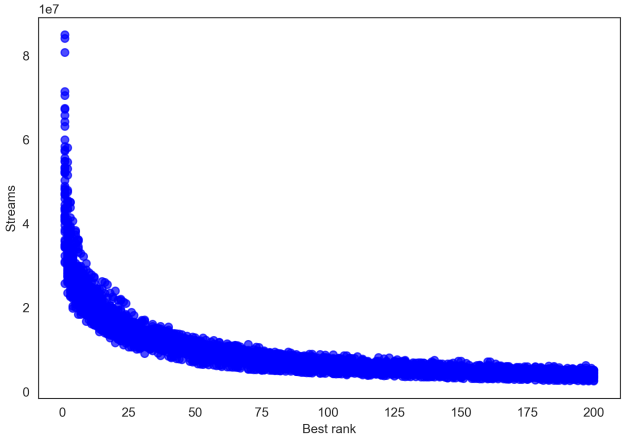

Figure 23: Dynamics of streams (2)

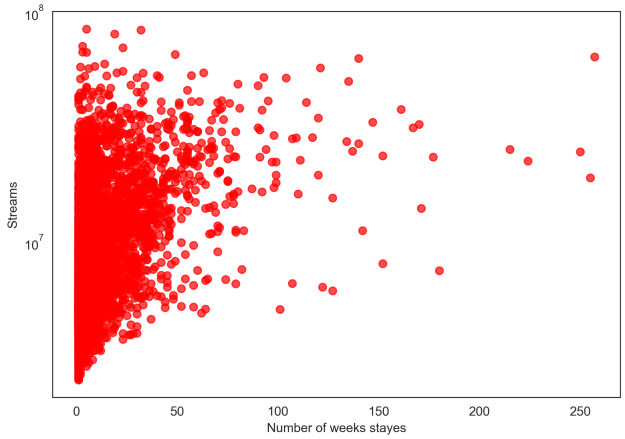

Figure 24: Dynamics of streams (3)

5.2 What is the temporal changes in streaming, can we capture the observed stream patterns?

To look into the temporal changes in streaming, I looked into the weekly(Figure 25), seasonal(Figure 26) and yearly(Figure 27) fluctuations in the streaming patterns during the past five years. To explore how these fluctuations affect the ranking list, I measured the median number of stream that got the songs on the list at different times. The dots correspond to the median stream of all songs on the ranking list in the corresponding temporal level. The reason why focus on the median instead of the average given the high variability in streaming.

Overall, the results show that there is a uptrend over the past five years(from Figure 25 and Figure 27) and there is also a significant increase in stream late-December during holiday(from Figure 25 and Figure 26). In summary, it can be seen that songs temporal fluctuations happen within a year, songs get higher streams during the holidays.

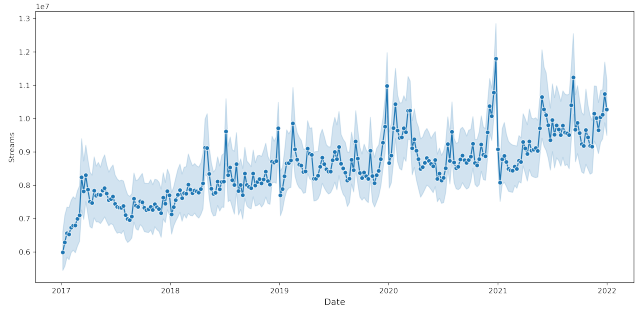

Figure 25: Dynamics of streams (4)

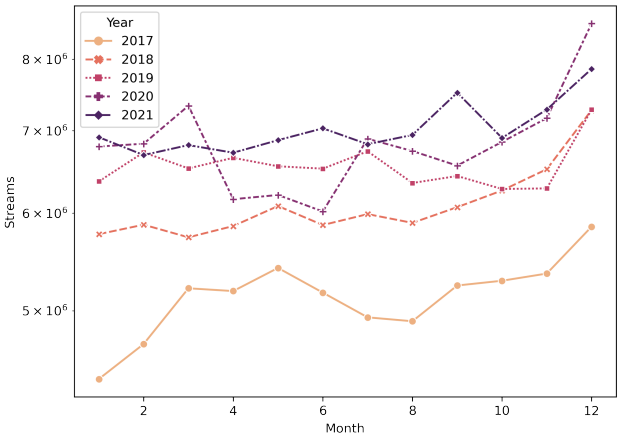

Figure 26: Dynamics of streams (5)

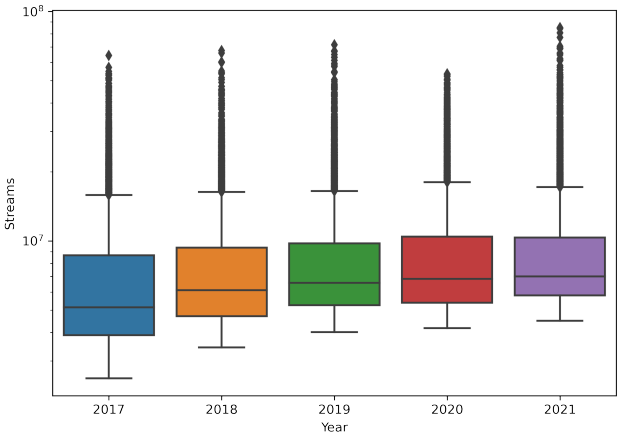

Figure 27: Dynamics of streams (6)

6 CONCLUSIONS

The goal of this project is to consider different perspectives to analysis the properties or discovering existing pattern of hit songs, artists and dynamics of stream. The results show that majority of hit songs appear only once on the list, while a few do spend an extremely long time on the ranking list. The better a song's best ranking, the higher probability of this song stays longer on the list. There are no significant audio feature differences among different seasons and different year. Looking at patterns characterizing artists, the results show that some artists are significantly more productive than others while vast of the artists only have one hit song. When building the Genre Network based on network, we can see that artists from pop and rap are more likely working together to make hit songs. Finally, we can identify there are some temporal patterns on the streams, people tend to listen to more music on holidays.

There are some limitation of the study. The first is the data limitation, the data sets are small only cover 5 years data and do not cover any songs that are not on the ranking list. So the observed properties may not be the unique properties that the hit songs, artists hold. For this reason future expansion of this work could be to include more data, consider negative samples, i.e. the songs that are not on the ranking list. What's more, the theoretical limitations of this project. Research questions cannot be answered by empirical data alone without theoretical frameworks for guidance. All the analysis and their conclusions can be optimised more insightful. I expect my findings on hits on Spotify can offer a starting point and inspiration to investigate the hit songs and artists further. Future research could also look at Hit Songs Prediction and Song Sentiment Analysis.

ACKNOWLEDGMENTS

I wish to thank Professor Jürgen Lerner for giving support and guidance.
The data sets used to the project can be found here: https://git.rwth-aachen.de/chenxu/projectspotify.git

REFERENCES

[1] Business of Apps. 2022. *Spotify Revenue and Usage Statistics (2022).* Retrieved June 1, 2022 from https://www.businessofapps.com/data/spotify-statistics/
[2] Gabriel P Oliveira, Anisio Lacerda, and Mirella M Moro. 2020. Musical Genre Analysis Over Dynamic Success-based Networks. In *Companion Proceedings*. 9.
[3] Spotify. 2022. *Get Track's Audio Features.* Retrieved August 1, 2022 from https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/
[4] Spotify. 2022. *Spotify Charts - Spotify Charts are made by fans.* Retrieved June 1, 2022 from https://spotifycharts.com/regional
[5] Spotify. 2022. *Spotify for Developers: Home.* Retrieved August 1, 2022 from https://developer.spotify.com/

A SPOTIFY AUDIO FEATURES

Table 4: Spotify Audio Features

Feature	Range	Description
`Acousticness`	Float number between 0 and 1	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
`Danceability`	Float number between 0 and 1	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
`Energy`	Float number between 0 and 1	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
`Instrumentalness`	Float number between 0 and 1	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". 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.
`Key`	Integer from -1 to 11	The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 2 = D, and so on. If no key was detected, the value is -1.
`Liveness`	Float number between 0 and 1	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.
`Loudness`	Float number between -60 and 0	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks.
`Mode`	Integer, either 1 or 0	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
`Speechiness`	Float number between 0 and 1	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording, the closer to 1.0 the attribute value.
`Tempo`	Float number	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
`Time Signature*`	Integer from 3 to 7	An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".
`Valence`	Float number between 0 and 1	A measure from 0.0 to 1.0 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).