Do Popular Songs Endure?

Introduction

The aim of this project is to develop a prediction model that applies the concepts of data science and predicts the popularity of a recording 'w' weeks from now, given that the current popularity is 'p' by analyzing the dataset of songs and artists starting from the last six to seven decades by taking into consideration a number of parameters to predict the current popularity of a particular song recording.

Data:

We obtained data from two primary sources, details of which are as follows:

1. Billboard Data Top 100 Per Week

Billboard publishes a list of Top 100 songs each week throughout the year. So we have dataset available for more than 60 years using the Billboard. We then used the Billboard API from the source(Link: https://github.com/guoguo12/billboard-charts) and we were able to get the list of songs for more than 60 years along with their artist name and current position in the billboard. Along with that, we got a number of other attributes as well. The dataset has the following attributes:

- title – The title of the track.
- artist – The name of the artist, as formatted on Billboard.com.
- peakPos – The track's peak position on the chart at any point in time, including future dates, as an int (or None if the chart does not include this information).
- lastPos – The track's position on the previous week's chart, as an int (or None if the chart does not include this information). This value is 0 if the track was not on the previous week's chart.
- weeks – The number of weeks the track has been or was on the chart, including future dates (up until the present time).
- rank – The track's current position on the chart.
- isNew – Whether the track is new to the chart.

2. Spotify Songs Popularity Data

Spotipy is a lightweight Python library for the Spotify Web API. Using Spotipy Web API we are getting access to all of the music data provided by the Spotify platform. We then used the Spotipy API from the source(Link: https://github.com/plamere/spotipy) and got the popularity of the songs we got from the billboard. Then, we got the popularity of the songs from billboard by using the artist and track name and hitting them on the Spotipy API. Then we are using the popularity data to compare the actual result and predicted the result and we were able to depict using the Power Law function.

3. Discogs API:

Discogs API provides the genre and style of the song. We used these features to identify the popular music of that decade and why a particular song was underperforming and what led to a decrease in popularity.

4. MusicBrainz API:

Music Brainz is an open music encyclopedia which contains music metadata. Using the MusicBrainz API we got the release year of every song. Along with music year, we also extracted features like Ratings, Tags, Instruments used, Place recorded using the MusicBrainz API since we wanted to identify the reason why some songs are performing as outliers and some songs are under/overperforming. We also identified the Genre and Style of the song using Discogs, since we needed features to identify why songs are underperforming and overperforming.

5. Youtube API:

Youtube API provides detailed statistics for any song. Using the Youtube API we get the following features: Number of Views, Likes, Dislikes, comments, and favorites. We are using these features to identify the trends in songs and how Youtube has been influential in increasing and decreasing the popularity of the songs. Using the number of views, likes, dislikes for a song, we can identify why the popularity of the song was decreased.

6. Wikipedia:

We used Wikipedia pages of the outlier songs and their artists to gain insights into any other external factors other than any features mentioned above. For example, using Wiki pages of the songs, we could identify external factors like the splitting of the band, the death of an artist etc, which resulted in the popularity of the song to reduce over a period of time.

Data Preprocessing:

The dataset had various uncleaned features such as the weeks and current position fields in the billboard data had zero value. The reason the dataset was having zero values was that there were inconsistencies while calling the Billboard API for a few weeks in the billboard data. All such data items have been removed from the dataset and got the dates in the standard format. From the popularity data, we discarded any values which were zero. Also, we removed all the duplicate names for a particular rank while finalizing the data set. In short, we considered only those songs that were at a particular rank of the billboard at any point of time and we considered them only once so that our data analysis doesn’t become too cluttered with too many data points.

Song Popularity Analysis

We are first aggregating data for a period of 60 years from the Billboard API. This dataset contains the song name and the week for that year and a number of other attributes. One of these includes the rank of that song in that week for a particular year. In each year, we intend to get all the song's name with its particular rank for all the weeks. Now that we have all the songs name, we found all the song's release year using the MusicBrainz API. Then we find the corresponding Spotify popularity for each song. We then find the log of the popularity for that song since that popularity needs to be clubbed together for better understanding of the data. Also, when we tried to plot without using the log values, we weren't able to identify any particular insights because there was a lot of data congregation. We plot the graph along the axis in such a way that y-axis denotes the Spotify's popularity in log value for that song and the x-axis denotes the year on which song was released. Then, we found out a linear regression line defining the points for the particular rank.

Similarly, we found out the regression line for the 2nd, 50th, 75th and 99th ranked songs for all the years. Once we had all the regression line, we could easily figure out the overperforming songs from the graph as these are seen as the songs excessively over the line. Similarly, we could figure out songs which are underperforming. The plot for the model can be shown below:

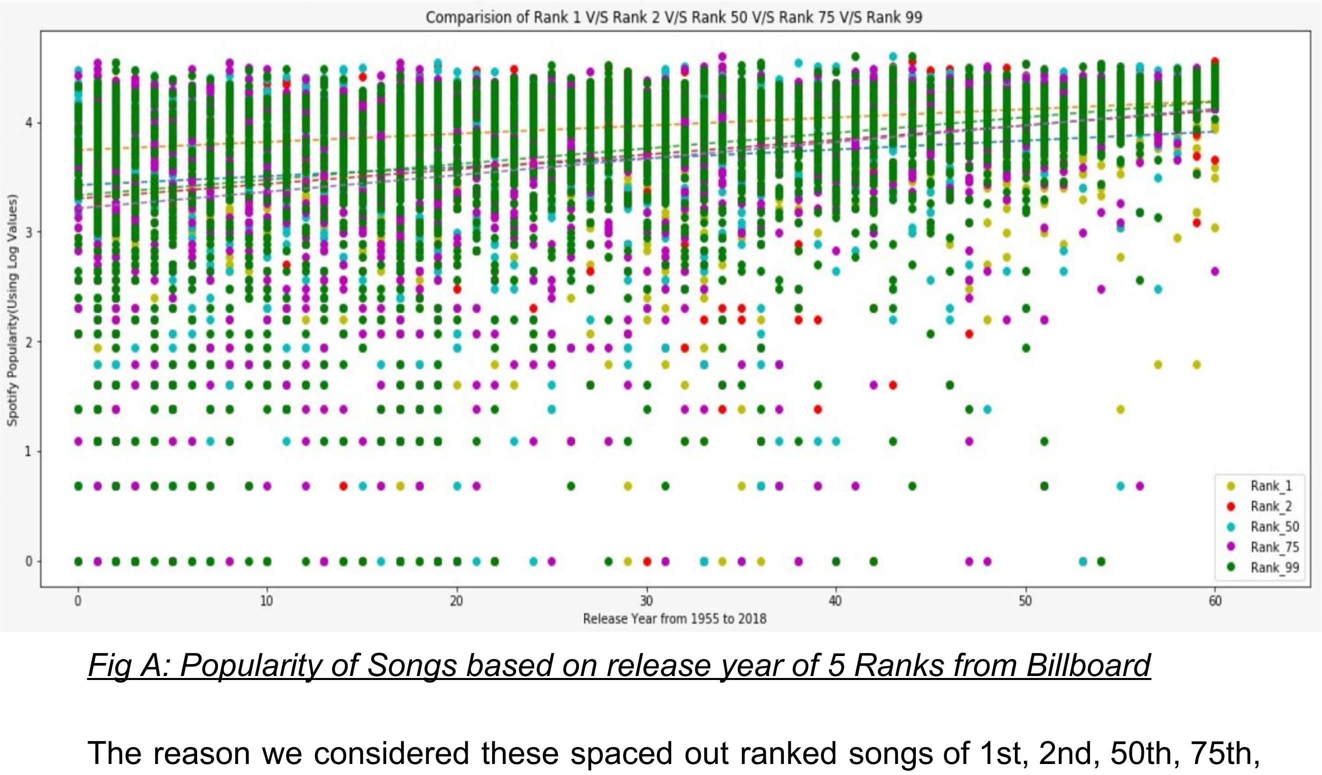


Fig A: Popularity of Songs based on release year of 5 Ranks from Billboard

The reason we considered these spaced out ranked songs of 1st, 2nd, 50th, 75th, and 99th rank is so that we can get more detailed insights for the top songs and the bottom songs of the billboard and how their performance and outliers are present. The regression line describing the rank is used to determine the outliers. From the above figure, we see a lot of songs which were of rank 99 (based on their Spotify Popularity) having a less popular as compared to the other ranked songs, which infers that many 99 ranked song faded away over time, rather than improving

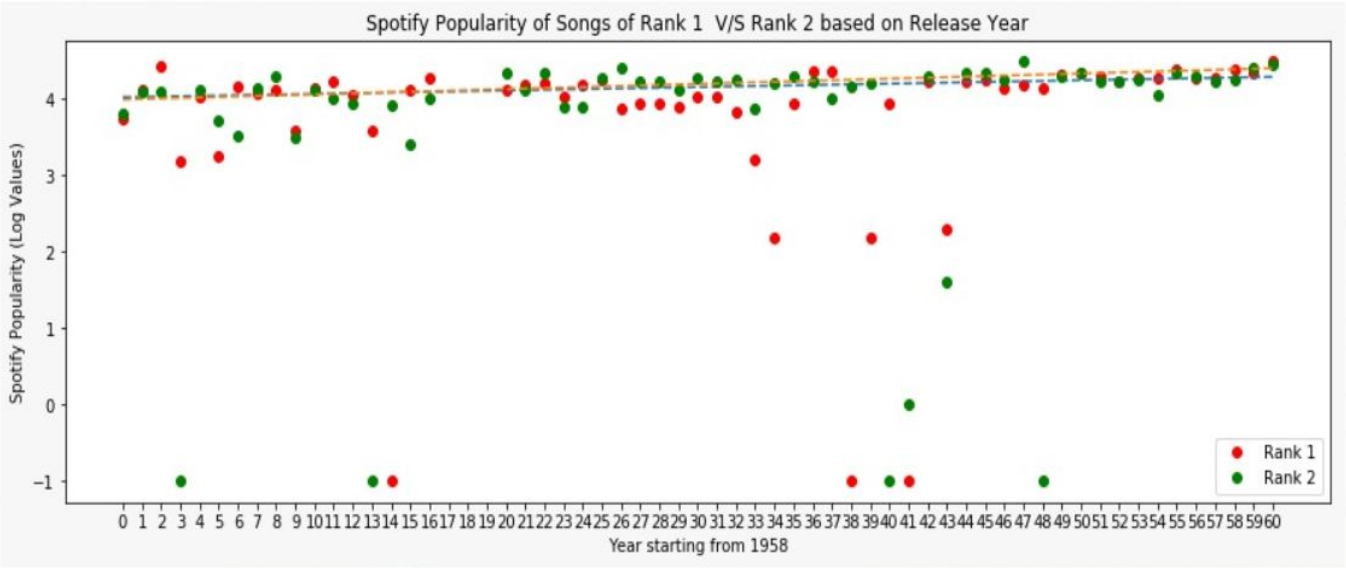


Fig B: Popularity of Rank 1 vs Rank 2 Songs based on release year

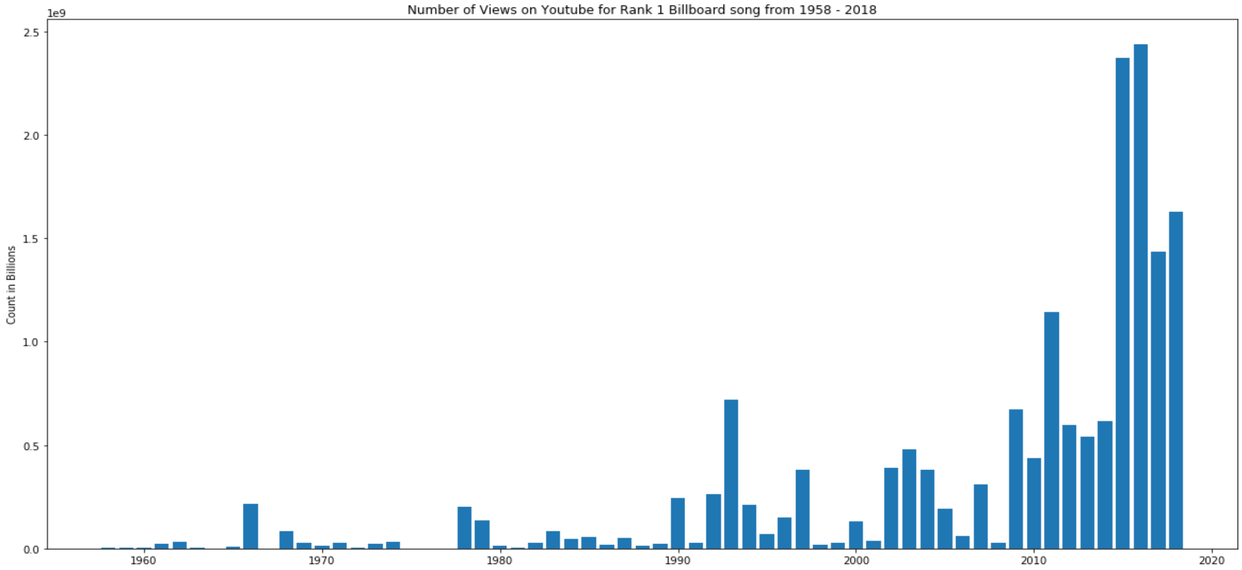
In order to get more detailed insights into the songs and the outliers, we made another plot. In the above plot, we have considered songs of Rank 1 and rank 2, from one particular week of each year from 1958 to 2018, and took their release year and their Spotify popularity in order to make the above plot. So for example, here we have considered rank 1 and 2 songs released in week 5 of each year from 1958 to 2018. Then we found those song's release year and current Spotify popularity. Here, we can see exactly which all songs are the outliers that are deviating far out from the regression line. We have made detailed insights into the above outliers below.

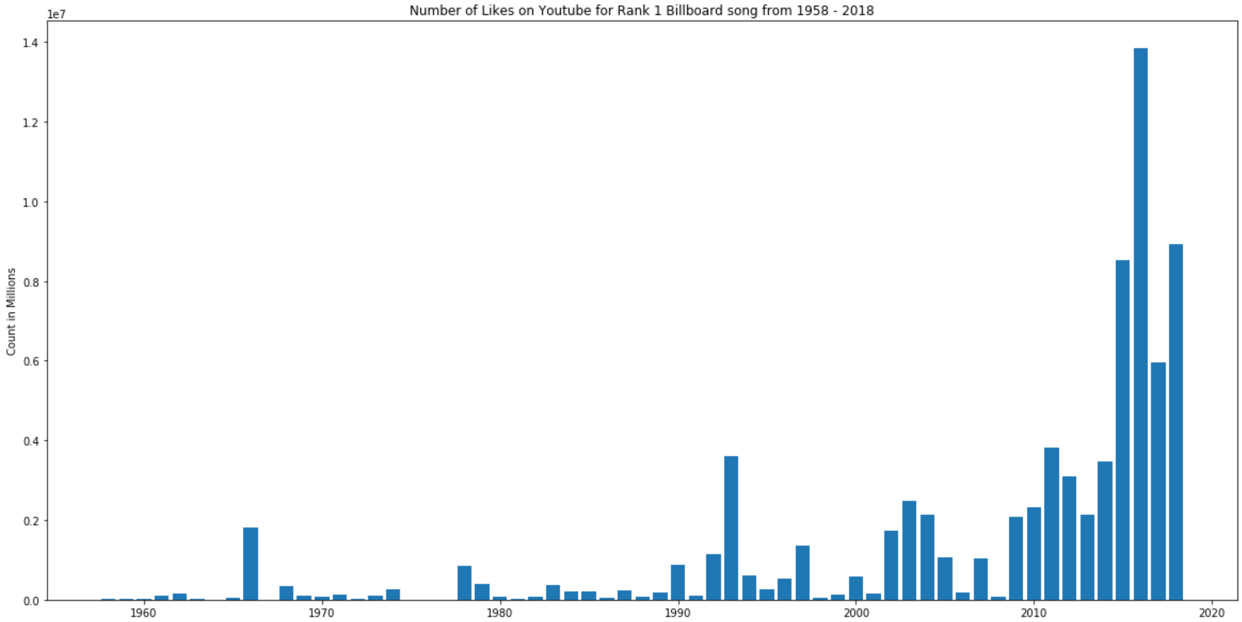
Insights into the figure:

One interesting insight that we can see in the above figure (A), is that the outliers in the right side are of less in proportion than that of the outliers in the left side of the figure. This means that after 2000, the outliers proportion was less as compared to that of before. One of the reason could be because of the introduction of Streaming website's popularity such as Youtube, Hulu, Apple Music, Google Play Music, Deezer, and others. Because of these streaming websites, the popularity has been in increasing for the songs which were in billboard rank 1. The reason could be since the views of these songs were ever increasing.

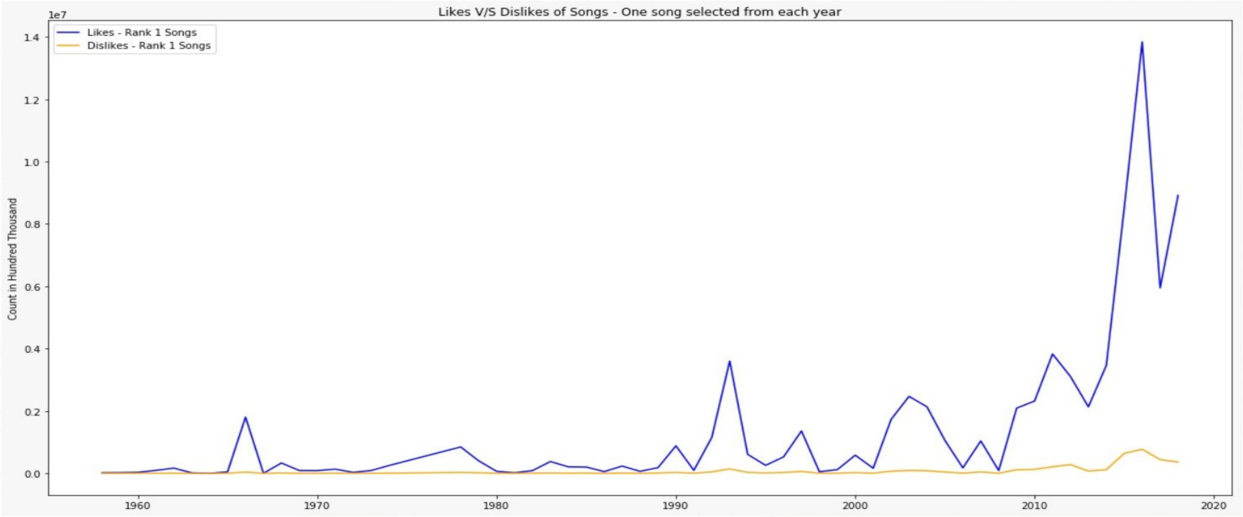
In order to get the detailed effect of streaming services, we found a detailed analysis of songs using Youtube API. We used Youtube API to get the following statistics: Number of Views, Likes, Dislikes, and Comments.

First, we are plotting the number of views and the number of views for one 1st Rank song from each year of the Billboard. We can clearly see the spikes on the right side after the year 2005 as compared to the decades before that. So this is one of the reasons why songs from the previous decade were underperforming as compared to the new-generations.

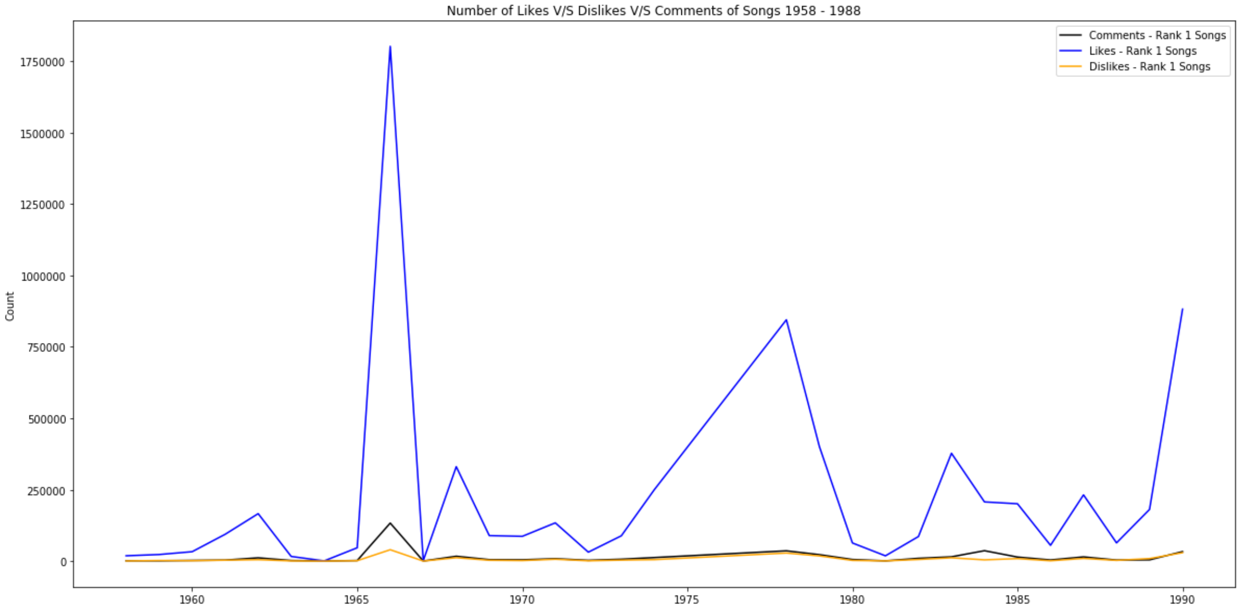


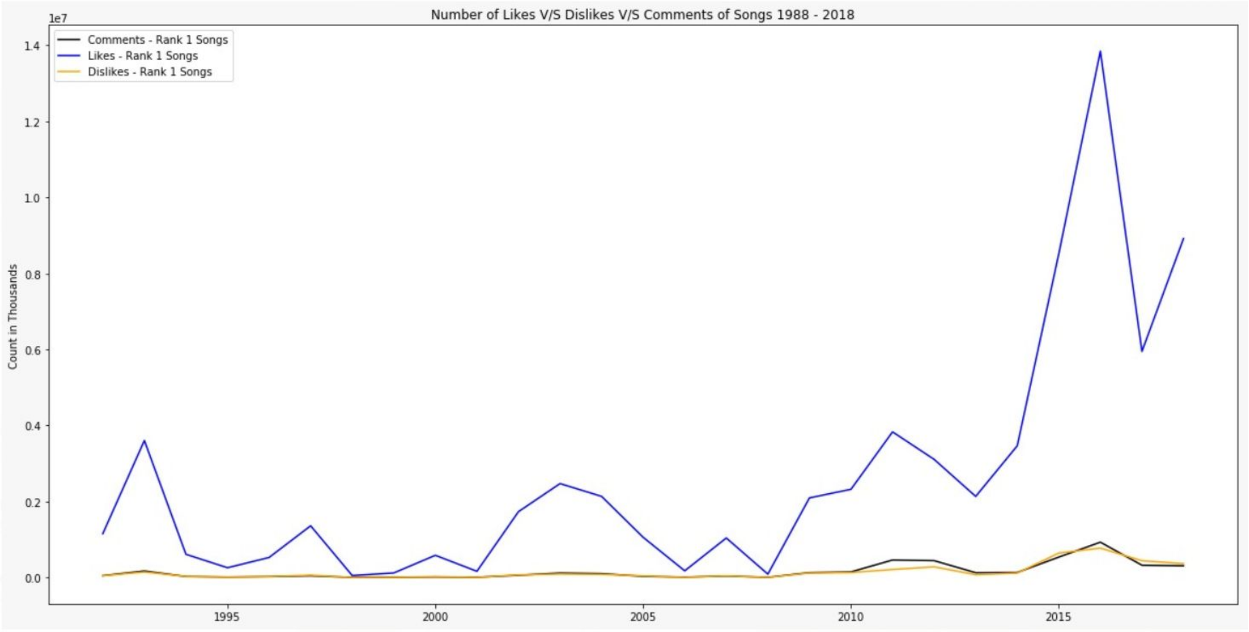


From the above plot, we can clearly infer that the songs number of views is directly proportional to the number of likes.



In order to get a detailed insight into the effects of the internet and the streaming services have on the popularity of a song, we have made two plots which compare the number of Likes, Dislikes and the Comments. The first plot, shows the songs and the counts of the three parameters, before the age of the Internet(that is 1988 - assumption), while the second plot shows the count of the three parameters post Internet boom. We can clearly see, that the later plot has almost all the songs having a lot of likes. And we know(in general) that the number of likes is proportional to the number of views of a song.

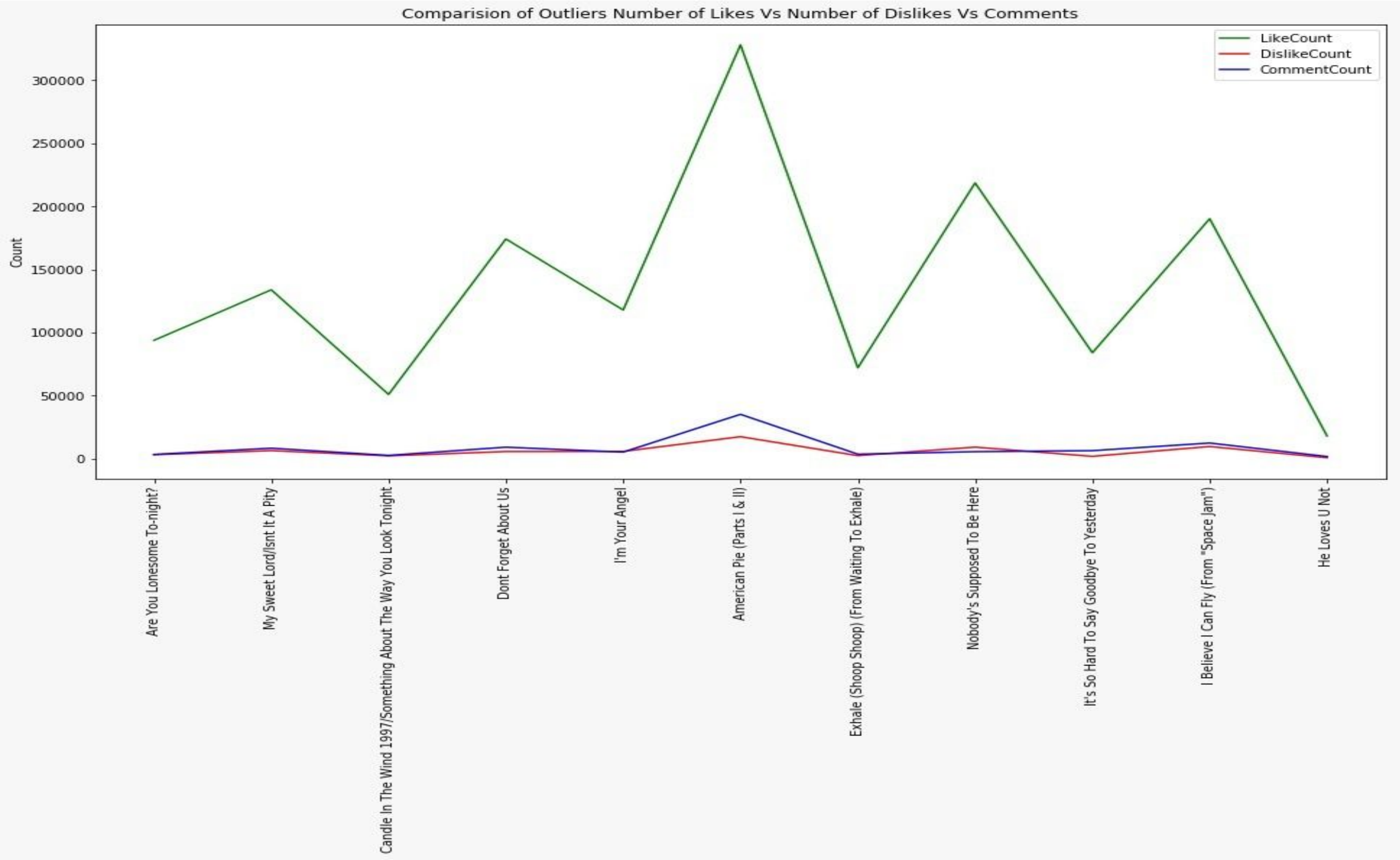




From the above, we can see the effect of Youtube(and similarly other streaming apps) on the popularity of the songs before and after the Internet Boom. So we can clearly infer that Youtube has a role in affecting the popularity of the songs, and that's why we can see that the newer songs(from figure A) have a high popularity.

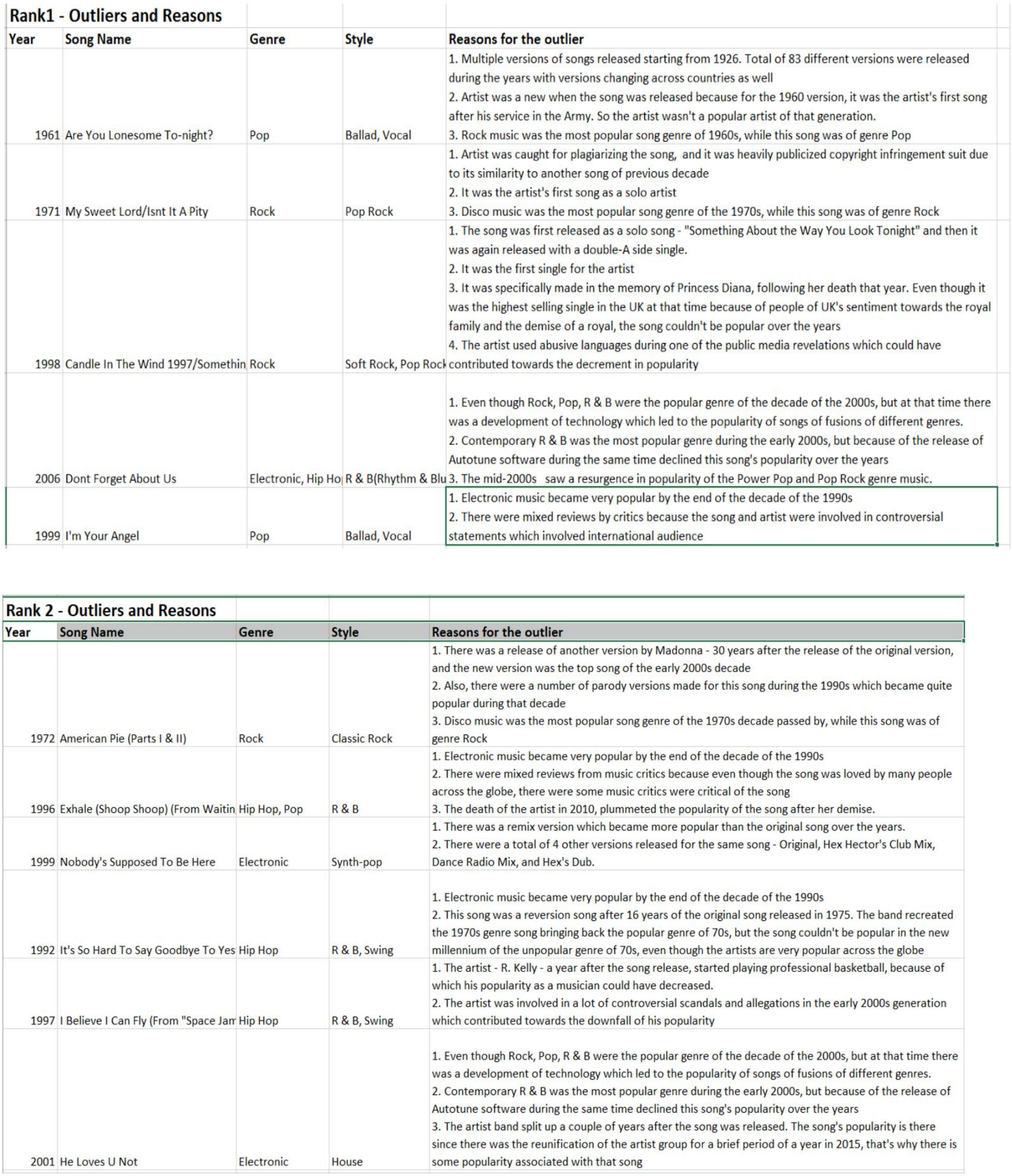
Other insights:

In order to identify any other effects which could reduce the popularity of songs, we considered the songs from Figure B(They consist of underperforming songs). Now first we tried to get the number of likes, dislikes, and comments from the Youtube API for those songs. So we had twelve songs to compare their number of likes, dislikes, and comments on Youtube. We already know that those twelve were either Rank 1 or at Rank 2 on Billboard at some point in time from 1958 to 2018. So that means they were popular songs during that time. But over a period of time, we see that their Spotify popularity has been low because of some reason. In order to analyze why their popularity decreased over time, we first took into consideration the data from the Youtube API. We considered the number of likes, dislikes, and comments. We could get the following visualization plot:



From the above plot, we can see that we have a lot of likes for the songs, which is justified since all the songs in the above plot were either rank 1 or rank 2 at some point of time in the billboard. But at the same time, we can see an alarming number of dislikes as well for the same songs. This could be one for the reasons that the songs had a lot of dislikes(on Youtube) which could have contributed to its downfall over the years. This happens with a lot of songs, that they receive a lot of flak or criticism over time because of some reason related to the song or with the artist or any other reason.

But along with the song name, artist name and the release year, we used the Discogs API in order to get the Genre and Style of the songs. Along with that, we got the number of versions released for the song and when were the versions released and the country of origin for each version release. So we did a detailed analysis of the underperforming Rank 1 and Rank 2 songs and identified reasons for the drop in popularity over a period of time as follows(Since the data is too large we have shown the data in an image rather than using a Table):



Rank1 - Outliers and Reasons

Year	Song Name	Genre	Style	Reasons for the outlier
1961	Are You Lonesome To-night?	Pop	Ballad, Vocal	1. Multiple versions of songs released starting from 1926. Total of 83 different versions were released during the years with versions changing across countries as well 2. Artist was a new when the song was released because for the 1960 version, it was the artist's first song after his service in the Army. So the artist wasn't a popular artist of that generation. 3. Rock music was the most popular song genre of 1960s, while this song was of genre Pop
1971	My Sweet Lord/Isnt It A Pity	Rock	Pop Rock	1. Artist was caught for plagiarizing the song, and it was heavily publicized copyright infringement suit due to its similarity to another song of previous decade 2. It was the artist's first song as a solo artist 3. Disco music was the most popular song genre of the 1970s, while this song was of genre Rock
1998	Candle In The Wind 1997/Somethin	Rock	Soft Rock, Pop Rocl	1. The song was first released as a solo song - "Something About the Way You Look Tonight" and then it was again released with a double-A side single. 2. It was the first single for the artist 3. It was specifically made in the memory of Princess Diana, following her death that year. Even though it was the highest selling single in the UK at that time because of people of UK's sentiment towards the royal family and the demise of a royal, the song couldn't be popular over the years 4. The artist used abusive languages during one of the public media revelations which could have contributed towards the decrement in popularity
2006	Dont Forget About Us	Electronic, Hip Ho	R & B(Rhythm & Blu	1. Even though Rock, Pop, R & B were the popular genre of the decade of the 2000s, but at that time there was a development of technology which led to the popularity of songs of fusions of different genres. 2. Contemporary R & B was the most popular genre during the early 2000s, but because of the release of Autotune software during the same time declined this song's popularity over the years 3. The mid-2000s saw a resurgence in popularity of the Power Pop and Pop Rock genre music.
1999	I'm Your Angel	Pop	Ballad, Vocal	1. Electronic music became very popular by the end of the decade of the 1990s 2. There were mixed reviews by critics because the song and artist were involved in controversial statements which involved international audience

Rank 2 - Outliers and Reasons

Year	Song Name	Genre	Style	Reasons for the outlier
1972	American Pie (Parts I & II)	Rock	Classic Rock	1. There was a release of another version by Madonna - 30 years after the release of the original version, and the new version was the top song of the early 2000s decade 2. Also, there were a number of parody versions made for this song during the 1990s which became quite popular during that decade 3. Disco music was the most popular song genre of the 1970s decade passed by, while this song was of genre Rock
1996	Exhale (Shoop Shoop) (From Waitin	Hip Hop, Pop	R & B	1. Electronic music became very popular by the end of the decade of the 1990s 2. There were mixed reviews from music critics because even though the song was loved by many people across the globe, there were some music critics were critical of the song 3. The death of the artist in 2010, plummeted the popularity of the song after her demise.
1999	Nobody's Supposed To Be Here	Electronic	Synth-pop	1. There was a remix version which became more popular than the original song over the years. 2. There were a total of 4 other versions released for the same song - Original, Hex Hector's Club Mix, Dance Radio Mix, and Hex's Dub.
1992	It's So Hard To Say Goodbye To Yes	Hip Hop	R & B, Swing	1. Electronic music became very popular by the end of the decade of the 1990s 2. This song was a reversion song after 16 years of the original song released in 1975. The band recreated the 1970s genre song bringing back the popular genre of 70s, but the song couldn't be popular in the new millennium of the unpopular genre of 70s, even though the artists are very popular across the globe
1997	I Believe I Can Fly (From "Space Jam	Hip Hop	R & B, Swing	1. The artist - R. Kelly - a year after the song release, started playing professional basketball, because of which his popularity as a musician could have decreased. 2. The artist was involved in a lot of controversial scandals and allegations in the early 2000s generation which contributed towards the downfall of his popularity
2001	He Loves U Not	Electronic	House	1. Even though Rock, Pop, R & B were the popular genre of the decade of the 2000s, but at that time there was a development of technology which led to the popularity of songs of fusions of different genres. 2. Contemporary R & B was the most popular genre during the early 2000s, but because of the release of Autotune software during the same time declined this song's popularity over the years 3. The artist band split up a couple of years after the song was released. The song's popularity is there since there was the reunification of the artist group for a brief period of a year in 2015, that's why there is some popularity associated with that song

Fig C: Rank 1 and 2 Underperforming songs issues

From the above data we can clearly see that, in many of the songs we have the issue that songs of another genre became more famous in the decade after the release of that song. One other common reason that we found amongst the songs is that, there was some controversy(external factor) related to either the artist or the song, which contributed to the decline in popularity of the song.

Similarly, we did an analysis for the overperforming songs as well. First, we did an analysis for the song which was in Rank 1:

Release Year	**Song Title**	**Genre**	**Style**	**Reasons**
2005	Let me Love	Electronic, Hip Hop	R & B	1. It was a fusion of 2 genres - Electronic and Hip Hop and during the mid-2005s, the fusion of two or more genres was the trend 2. The artist - Mario, won a Grammy award for this song and was the eighth most successful single of the decade 3. In 2008, the song was listed on the Billboard's All-Time Top 100 Hot 100 singles during the first 50 years of the chart.

Now for Song of Rank 2, we saw the following analysis for its overperformance:

Release Year	**Song Title**	**Genre**	**Style**	**Reasons**
1960	Wonderland by Night	Jazz	Easy Listen	1. The song was on billboard's number 1 hit for five continuous weeks 2. Bert Kaempfert, the artist of this song, was an orchestra leader and songwriter, who was a very popular figure in the 1960s for his songs and many songs were well acclaimed. 3. One year after the release of this song he hired "The Beatles" for a song and he was quite popular for his songs

So we can see the artist’s popularity was one of the reasons why the songs were overperforming and were performing better than the songs of the same time. Similarly, there were other features as well which contributed to the songs over or underperforming.

Resources used:

1. Youtube API (Link: https://developers.google.com/youtube/v3/)
2. Music Brainz API (Link: https://musicbrainz.org/doc/Developer_Resources)
3. Discogs API (Link: https://www.discogs.com/developers/)
4. Spotify API (Link: https://developer.spotify.com/documentation/web-api/)
5. Wiki Pages (Link: https://www.wikipedia.org/)