Data Science Fundamentals – CSE 519

Final Project Report

Do Popular Songs Endure?

Content

1. Introduction

1.1 Problem Statement

Surprisingly, most of the popular enduring songs nowadays (as measured by current sales or airplay) were not actually ranked to the top of the charts when they first came out back in 1960s. The main objective of this project is to find the endurance of popular songs, and to formulate if songs which were popular at the time of their release actually remain their popularity over time. By investigation of a variety of datasets, the popularity of songs can be defined by various methods such as ranking functions which relate directly to number of views, Billboard rankings, etc, and other characteristics as well like number of sales, etc. Some other interesting factors that might contribute to the songs popularity can be explored more in depth with Million Songs Dataset for example, which include a rich features set about any particular songs such as its danceability, song_hotttness, analysis sample rate, types of genre, etc. Based on these features, we will build a model to predict the endurance of songs' popularity over time from the time they were released.

1.2 Available Resources

The first important step is to collect dataset. Based on our experience of listening to the songs and searching on the internet, we have found that there are various sources of relevant data available online. We will do data cleaning step and combine these relevant datasets together to make modeling task become easier later on.

- **Billboard dataset**

We began to collect data from Billboard dataset by using Python API for accessing music charts from Billboard.com. This dataset contains weekly Hot 100 singles chart from Billboard.com. Each row of data represents a song and the corresponding position on that week's chart. The dataset contains chart entry instance (typically a single track), which is type of ChartEntry and have attributes like title, artist, peek position, last position, week, rank. The advantage of using this dataset is that it is relatively easy to use and scrape using Beautiful Soup.

- **Youtube scraping**

So we have the rank for all the songs based on Billboard dataset, but in order to build a model to predict and investigate why popular songs nowadays were not in top chart when they were initially came out, we need to know the number of views, as well as the number of likes and dislikes for each song along with its artist information. To do this, we will scrape songs data from Youtube, by also using Beautiful Soup.

- **Lastfm scraping**

Similarly to Youtube, we can also scrape data from Lastfm. The advantage of Lastfm is that it contains accurate results and given a searching query of song name and artist name, it can produce a single correct output information about number of listeners and scrobbles. With Youtube, one the problem might rise when a video appears in an album list and this can cause duplicate counting of songs and artist when we try to collect our data. Thus Lastfm can alleviate this issue.

- **Million Songs Dataset**

This dataset contains a rich set of audio specific features such as loudness, beats, artist hotness, etc. These features might be related to measuring the trend or popularity of songs, thus it can be useful in future once we deep dive more into building a better model with more features from this dataset to understand more about the audio aspect of the songs and predict songs' durability.

- **Spotify Score Dataset**

This dataset contains a rich set of audio specific features such as loudness, beats, artist hotness, etc. Although this is a good resource for gaining information about songs' popularity. One downsize of this source is this is not primary data, but secondary data. Because it has been processed by Spotify's algorithm.

2. Selected Dataset

Studying all above-mentioned datasets. We pick out the most important and most informative datasets for our projects below.

2.1 Songs' Metadata – One Million Dataset

One Million Songs dataset contains many useful information about a song that if used appropriately, can predict the song's durability through time. Some of the most important features include: tempo, loudness, durations, artists hotness. These features will be analyzed to know whether they have influence on song's performance or not, and if it does, to which extent.

2.2 Past Songs' Popularity – Billboard Rankings

The Billboard ranking datasets are chosen as records' popularity at the time they were released. This is a very trustful source of data and reflect very well how popular the records were when they were released. For each record, the Billboard Top100 chart returns its ranking and also the time (in weeks) the record was released.

- For the time feature (the precise week when the record enters Billboard Top100 chart), we interpolate them to years so that we have time as a continuous features of year.
- For the ranking feature, we focus the peak position – the highest ranking the record can get while it was on Billboard chart. Because this peak position does reflect how popular the song was.

2.3 Current Songs' Popularity - Youtube Views Count

Although there are many sources to measure popularity of songs nowadays (Spotify score, Lastfm listeners…), we choose Youtube views count as our main popularity measurement. This choice is made based on two observations:

- Youtube is very ***popular channel*** for music listeners. Indeed, as a part of Google, Youtube is now the most famous video sharing websites that has reached every corner of the internet world. Therefore, the performance (measured by views count) of a song on Youtube is closely related to how popular it is.
- Data collected from Youtube can be considered ***primary data***. Primary data has many interesting characteristics that other secondary data, such as Spotify score, does not own. For example, the views count from Youtube does obey power-law distribution with years and ranks, which we will discuss detaily below. Data from Spotify, however, does not have this characteristic because it has been processed by Spotify own formulas, for its own purpose of measuring songs' popularity.

For every song in the list of Billboard top100, we search and count views of videos having titles containing the exact song's name and song's artist, so that we are evaluated popularity of a record (a specific song by a specific singer). Because we want to avoid counting as one song for songs that are covered by famous singers. And also, different songs may have same name as well.

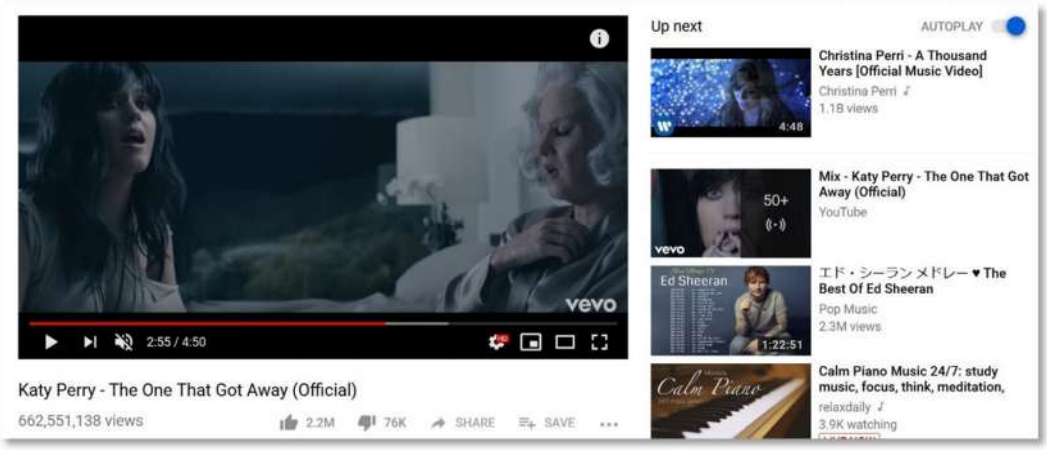


Figure 1. Youtube views count used as current popularity

2.3.1 Sniff-test

We first conduct some sniff-test to see if there is anything unusual, out of our expectation. The first thing we do is computing statistics for all the views count scraped. The results are shown in Table 1.

Statistics	
Count	**22,045**
Mean	**37,336,322**
Std	**197,279,353**
Min	**0**
25%	**67,702**
50%	**654,232**
75%	**8,293,671**
Max	**5,712,222,342**

Table 2. Statistics of Youtube views count scraped

The first unusual thing we find is that there are a number of songs having 0 view count. This is unreasonable because even if a song is not popular as expected, its views count should be higher than that. We take a look at these songs, some examples are shown in Table 2. We realize that the main reasons for this abnormality are:

- Song's name or artist name have some unique/uncommon characters. One example is Ke$ha, using the dollar symbol $ instead of S. In Billboard chart, her name is notated as "Ke$ha". However, on Youtube, we can only find her name notated as "KeSha".
- Collaborative works. Billboard uses too many symbols (and therefore inconsistent) to denote collaborative works. For example, using word "feat", "ft.", "featuring", "X", "and",… . This is very hard for us to detect the correct version of Youtube video for the song. And there could be a chance that the algorithm cannot find the collaborative record with the artist group named as "Barry Manilow *with* Kid Creole *&* The Coconuts" from Youtube video's title.

Title	Artist	Year
Alvin's Boo-Ga-Loo	Alvin Cash & The Registers	1966
One Day At A Time	Tupac With Eminem Featuring The Outlawz	2004
Praying	Ke$ha	2016
Wake Up In The Sky	Gucci Mane X Bruno Mars X Kodak Black	2017

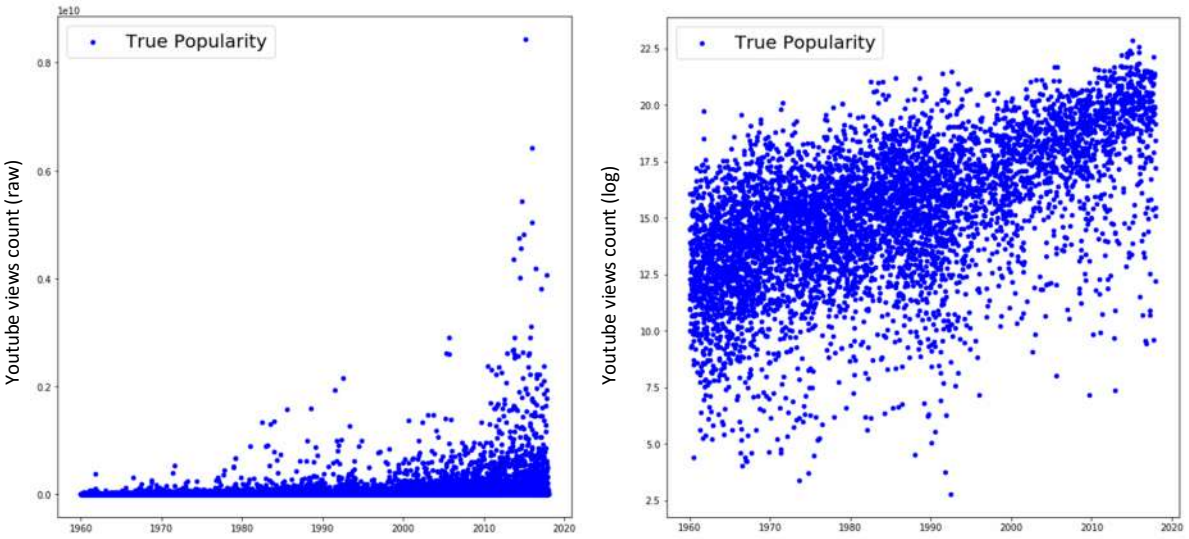
Table 2. Songs having zero Youtube views count

In order to not let these noises and errors affect our model, we eliminate these data points out of our dataset.

2.3.2 Dataset Analyzing

Youtube views count as a function of year released

We then start exploring, trying to have some intuitive understanding of the dataset. First, a very intuitive expectation comes to our minds, that is the song's popularity (in terms of views count) should obey ***power-law distribution*** with respect to time. Meaning that the popularity of a song will decrease through time in an exponent manner. We test this hypothesis by plotting Youtube views count of all songs from 1960 to 2000 versus the year the song was released. Here, we make the released year feature continuous by interpolating using the song's released weeks and months within a year. In Figure 2a, every blue dot represents a song from 1960 to 2000 in Billboard Top100 chart. Based on the plot, we does see that the distribution does has the shape of a power-law distribution. To illustrate the hypothesis better, we take log of the popularity and then plot it again in Figure 2b. The plot looks very much like a linear shape. This means the original views count does obey power-law distribution.



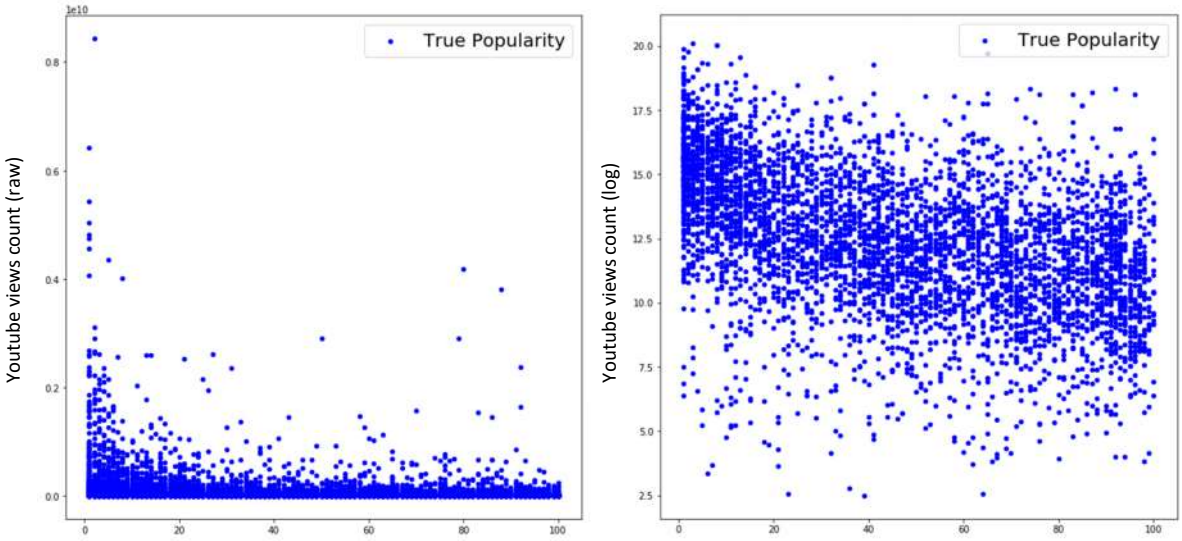
(a) Raw Youtube views count of all Billboard Top100 songs against years released

(b) Log Youtube views count of all Billboard Top100 songs against years released

Figure 2. Popularity of all Billboard Top100 songs against years released

Youtube views count as a function of Billboard rankings

We also wondering that whether the power-law distribution is also applicable with respect to songs' ranking on Billboard chart? Because as Professor Skienna as well as Teaching Assistant Sahil suggested: the 1st-rank song will be much better than be 2nd-rank song, which will be much better than the 3rd-rank song. The 98th-rank song, however, is not different much from 99th-rank song and also not much from 100th-rank song. We test this hypothesis by plotting Youtube views count of all songs from 1960 to 2000 versus the highest ranking (peak position) of the song in Billboard chart in Figure 3a. We does see a ***power-law distribution*** shape. We then take log of the popularity and plot again in Figure 3b. The plot does show a nice linear shape, proving that the original views count is indeed, obey power-law distribution.



(a) Raw Youtube views count of all Billboard Top100 songs against ranks

(b) Log Youtube views count of all Billboard Top100 songs against ranks

Figure 3. Popularity of all Billboard Top100 songs against ranks

So by observing graphs, we can see that songs' popularity does obey power-law with respect to years and rankings on Billboard. From here, we will just focus on the ***log-popularity of Youtube views count***, because the log views count plots (Figure 2b and Figure 3b) are much easier to observe and figure out many interesting information.

Graph's shape analysis and explanation

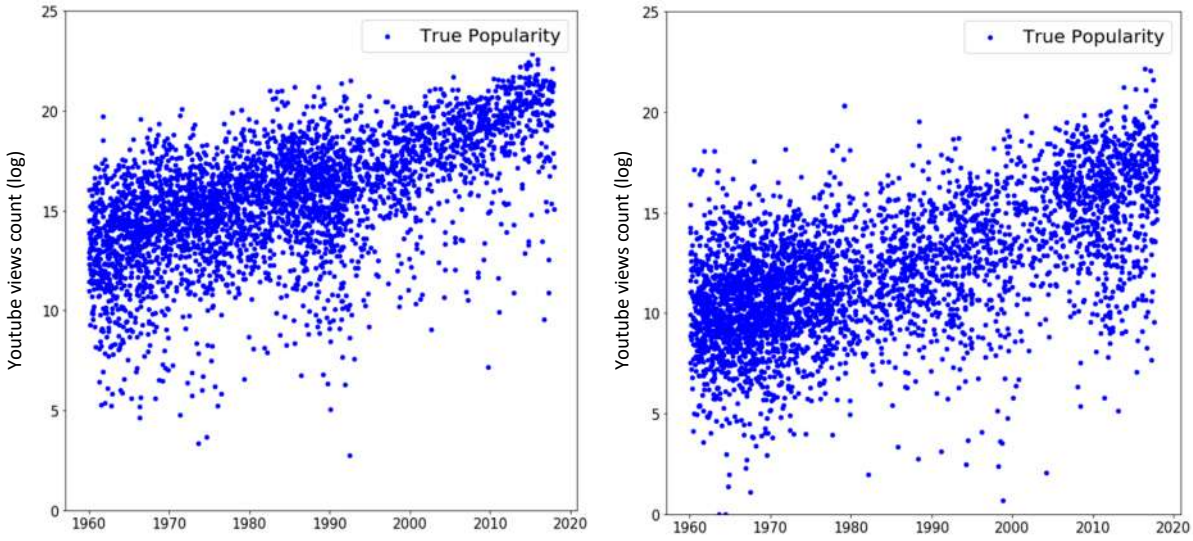
We then analyze the plot's shape and try finding reasonable explanations for it.

- First, we see that, as in Figure 4a, the popularity (log-Youtube views) is ***higher*** for recent songs (released in 2010s) than old songs (released in 1960s). This is reasonable because a song's popularity wears out over time. The farer from now, the less popular it is comparing to recently released songs.
- Second, we find that, songs with higher ranks (e.g. top 1-10) lies ***higher*** than songs with lower ranks (e.g. top 80-100). We can observe this from Figure 4a and Figure 4b. This is also understandable. Because given the same released years, songs that were ranked 1-10 are now still perform better than songs that were ranked 80-100 (measured by Youtube views count).

From these two observations, we can have a conclusion about the function of popularity by year released and Billboard rankings as follow:

- Popularity is increasing with year released (the higher, more recent the year, the higher the popularity).
- Popularity is decreasing with Billboard ranks (the larger the ranks, e.g. 95th, the lower the popularity)

The next section goes deeper into how we will form a function for predicting songs' popularity.



(a) Log-Popularity of songs ranked [1-10] against years

(b) Log-Popularity of songs ranked [80-100] against years

Figure 4. Popularity of songs at different Billboard ranks against years

3. Building Model for Popularity Prediction

3.1 Intuitive Formula

We now build a function for predicting songs' popularity based on year released and Billboard rankings. Using the analysis from section 2.3, we have guessd that our Youtube views count function should have these properties:

- Obeying ***exponential rules*** with respect to years released and rankings.
- ***Increasing with year released*** and ***decreasing with Billboard rankings***.

From these two observations, our function should have the form:

$$P = c.e^{-\alpha.Y}.e^{-\beta.R} \tag{1}$$

With P is predicted Youtube views count, Y is the distance between current year (2018) and released year (in other words, $Y = 2018 - released\_year$), R is the highest ranking of the song on Billboard chart. c, α and β are three parameters that we have to find.

We now taking log of formula (1), we then have:

$$\log P = \log c + (-\alpha).Y + (-\beta).R \tag{2}$$

We can denote again the parameters as $c_0 = \log c$, $c_1 = -\alpha$ and $c_2 = -\beta$, we then have a linear form function with respect to *year distance* ($Y = 2018 - released\_year$) and R ranking:

$$\log P = c_0 + c_1.Y + c_2.R \tag{3}$$

Therefore for the next step, we will conduct Linear Regression algorithm to find the parameters that fit our data the most.

3.2 Linear Regression

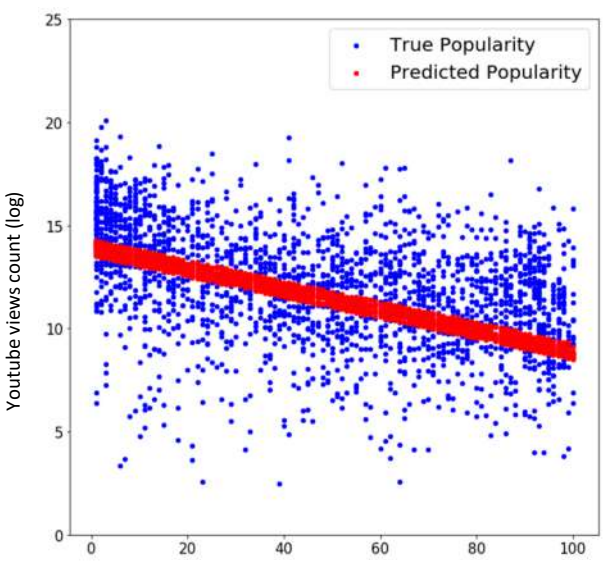
Based on formula (3), we will conduct Linear Regression algorithm with variables are $X = (Y, R)$ and predicting popularity value is $\tilde{Y} = \log P$.

Using Python's *sklearn* module for machine learning, we can find the fit values for parameters:

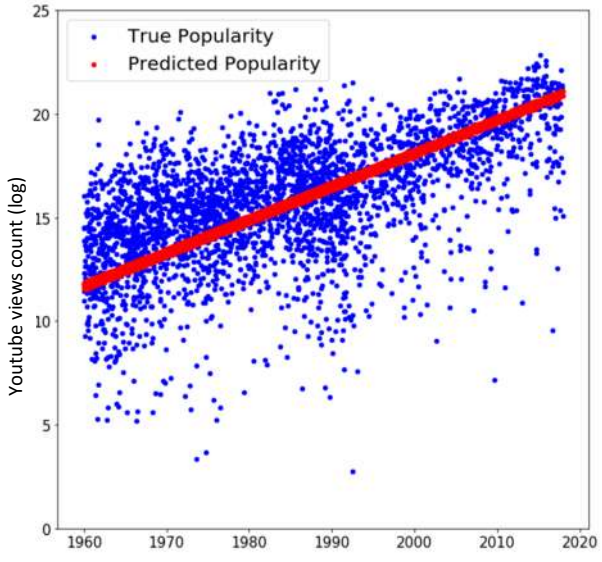
$$\begin{aligned} c_0 &= 21.1999 \\ c_1 &= \text{-}0.1601 \\ c_2 &= \text{-}0.0502 \end{aligned} \tag{4}$$

To test how correct our regression parameters are, we plot the ground truth popularity versus predicted popularity. Notice that because our formula shows that popularity is calculated based on two variables, year and ranking, in order to plot in 2D, we have to fix one the variable and plot the popularity versus the other variable. In the Figure 5 below:

- Figure 5a: ***fix year*** Y in a small range of 5 adjacent years, and ***plot popularity across all ranks*** from 1-100.
- Figure 5b: ***fix ranking*** R in a small range of 10 adjacent ranks, and ***plot popularity across all years*** from 1960-2018.



(a) Fix years range, plot ground truth and predicted popularity against ranks



(a) Fix rankings range, plot ground truth and predicted popularity against years

Figure 5. Plotting predicted popularity against true popularity

Observing the plots of ground truth popularity versus predicted popularity, we find that the linear regression fit the data pretty well. From this Figure, we can also find that there are many songs that lie high above prediction, and also many songs lie drastically below prediction. We can tell that these are ***overperform and underperform songs***. Indeed, in Figure 6 we zoom in from Figure 5b closely at a range of columns (songs ranked 1-10 released from

1970-1972). We can ***assign*** songs that are too far (higher or below) from the average as over/underperform songs respectively.
In the next section, we will build a formula explicitly to detect over/underperform songs and analyze deeper about their characteristics that make them perform much better or worse usual songs.

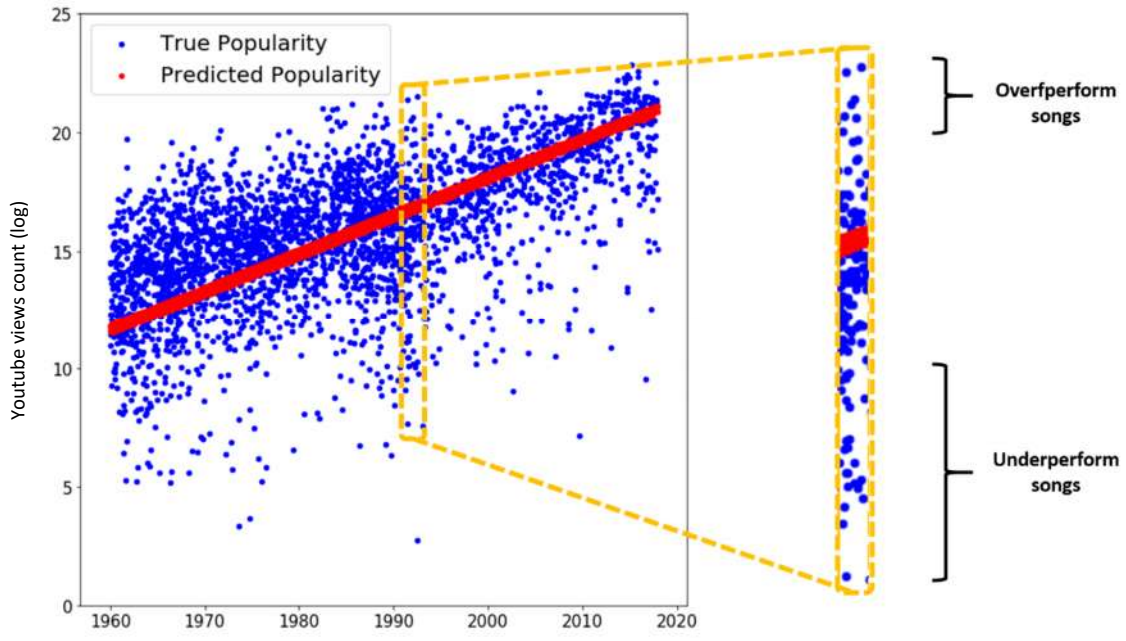


Figure 6. Determining overperform and underperform songs from truth popularity and predicted popularity

4. Detecting Over/Underperform Songs

4.1 Rules for Detecting Over/Underperform Songs

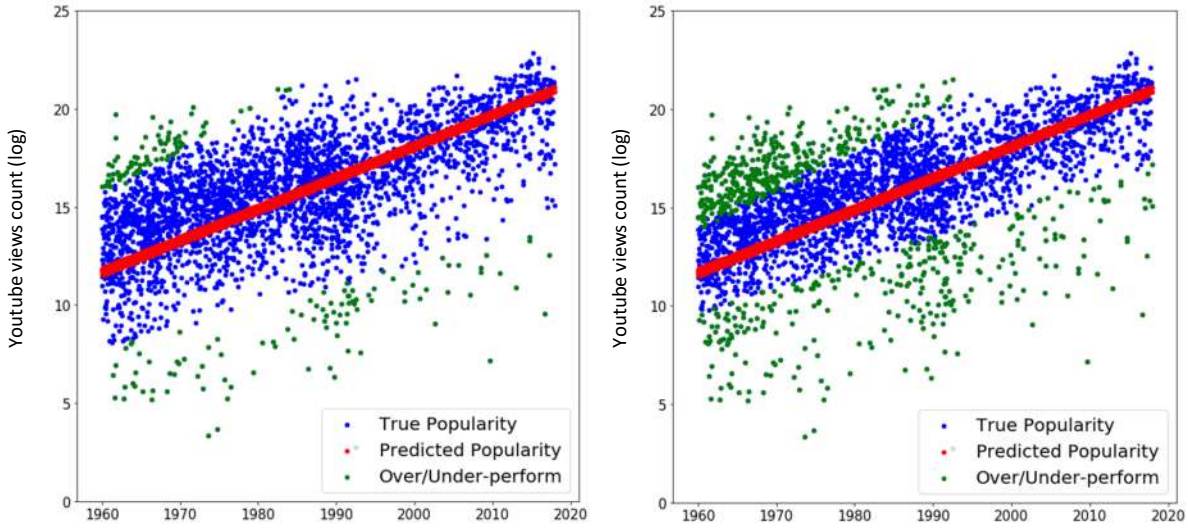
As suggesting by Professor Skiena, we define whether a song is over/underperform or not by measuring the ratio between the song's ground truth popularity and its predicted popularity. To be more particular, for every song, we will compute:

$$r = \frac{\log P_{ground\_truth}}{\log P_{prediction}} \tag{5}$$

If this ratio r is exceed or less a specific threshold θ, the song will be considered over/underperform song accordingly:

- If $r < \theta_{underperform}$, the song is considered underperforming.
- If $r > \theta_{overperform}$, the song is considered overperforming.

For example, Figure 7a illustrates our chosen threshold of $\theta_{overperform} = 1.3$ for overperform and $\theta_{underperform} = 0.7$ for underperform songs.



(a) Thresholds chosen as $\theta = \left(\theta_{overperform} = 1.3, \theta_{underperform} = 0.7\right)$

(b) Thresholds chosen as $\theta = \left(\theta_{overperform} = 1.2, \theta_{underperform} = 0.8\right)$

Figure 7. Effects of threshold values to number of over/underperform songs detected

4.2 Determining Threshold

Notice that, how we choose $\theta = \left(\theta_{underperform}, \theta_{overperform}\right)$ directly relates to the number of over/underperform detected. Indeed, the higher $\theta_{overperform}$ is, the less songs are assigned as overperform and reverse. On the other hand, the less $\theta_{underperform}$, the less songs are assigned as underperform and reverse. Figure 7a and 7b illustrate this idea. In Figure 7b, the $\theta_{overperform}$ is less and $\theta_{underperform}$ is greater than that of Figure 7a, therefore, more songs are assigned as over/underperform.

For analyzing purpose, we set the threshold $\theta = \left(\theta_{underperform} = 1.32, \theta_{overperform} = 0.25\right)$ to only choose out about 100 over/underperform songs (50 songs for each type). We think this set is large enough (but not too large) for us to analyze more thorough the characteristics of over/underperform songs. Figure 8 illustrates our choice of thresholds.

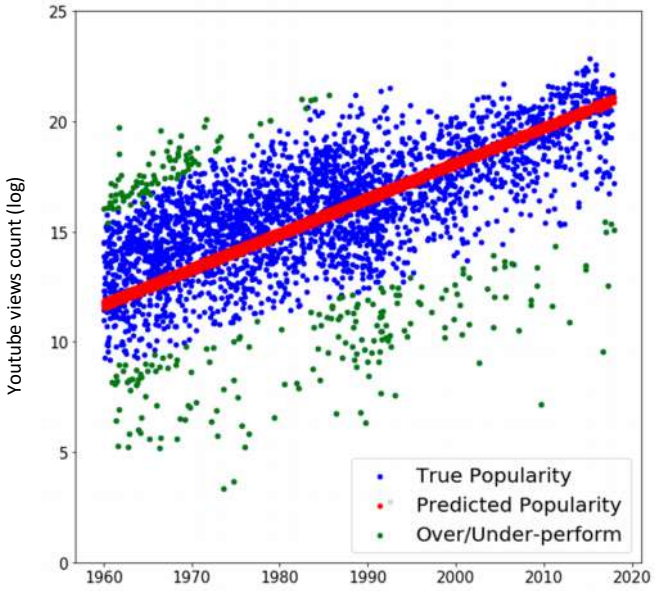


Figure 8. Chosen threshold to determining over/underperform songs

$$\theta = \left(\theta_{overperform} = 1.32, \theta_{underperform} = 0.25\right)$$

After have chosen out over/underperform songs, in the next section, we will analyze these songs to see if they have any characteristics that are different from usual songs. Below are two tables of 20 most overperform and most underperform songs detected.

List of most overperform songs

No.	Title	Artist	Year	Peak Position	Ground Truth Views (log)	Predicted views (log)	Ground Truth/Predicted Ratio
1	Let Me Entertain You	Ray Anthony	1963	96	18.04	7.2	2.51
2	Here Comes The Night	Ben E. King	1961	81	18.06	7.75	2.33
3	Milord	Edith Piaf	1961	88	16391	7.31	2.32
4	That's When I Cried	Jimmy Jones	1960	83	17.13	7.45	2.3
5	Ring Of Fire	Duane Eddy	1961	84	17.11	7.54	2.27
6	Sweet Thursday	Johnny Mathis	1962	99	15.66	6.94	2.26
7	Hot Cakes! 1st Serving	Dave "Baby" Cortez	1973	91	16.5	7.5	2.2
8	These Arms Of Mine	Otis Redding	1963	85	17.1	7.81	2.19
9	I Can't Explain	The Who	1965	93	16.68	7.71	2.17
10	A Letter To Dad	Every Father's Teenage Son	1967	93	17.54	8.13	2.16
11	Smokie-Part 2	Bill Doggett	1960	95	14.26	6.77	2.11
12	The Son Of Rebel Rouser	Duane Eddy	1964	97	15.31	7.31	2.09
13	Riverboat	Faron Young	1960	83	15.41	7.37	2.09
14	It's All Right	Sam Cooke	1980	93	14.87	7.15	2.08
15	Viva Las Vegas	Elvis Presley	1964	92	15.88	7.64	2.08
16	Night Time	Pete Antell	1963	100	14.47	6.99	2.07
17	Don't Stop The Wedding	Ann Cole	1962	99	14.48	7.03	2.06
18	For Your Precious Love	Jerry Butler	1970	99	15.48	7.57	2.05
19	For Your Precious Love	Dinah Washington	1962	98	14.47	7.08	2.04
20	Where Were You When I Needed You	Jerry Vale	1965	99	15.24	7.47	2.04

Table 3. List of overperform songs

List of most underperform songs

No.	Title	Artist	Year	Peak Position	Ground Truth Views (log)	Predicted views (log)	Ground Truth/Predicted Ratio
1	Mary Did You Know	Jordan Smith	1960	24	8.95	19.29	0.46
2	Careless Whisper	Wham! Featuring George Michael	2016	1	9.56	20.54	0.47
3	Caroline	Amine	2016	11	9.44	20.06	0.47
4	Rockin' Around The Christmas Tree	Brenda Lee	2017	14	9.58	20.05	0.48
5	Blood	Kendrick Lamar	2017	54	8.73	18.01	0.49
6	Clique	Kayne West, Jay-Z, Big Sean	2012	12	6.66	19.37	0.5
7	Gangsta Lovin'	Eve Featuring Alicia Keys	2002	2	9.06	18.23	0.5
8	Ooh!	Mary J. Blige	2003	29	8.69	17.07	0.51
9	Changed It	Nicki Minaj & Lil Wayne	2017	71	8.73	17.14	0.51
10	Since Way Back	Drake	2017	70	8.75	17.19	0.51
11	Stuntin' Like My Daddy	Birdman	2006	21	9.27	17.95	0.52
12	Girls Just Want To Have Fun	Glee Cast	2012	59	8.75	16.89	0.52
13	6 Foot 7 Foot	Lil Wayne Featuring Cory Gunz	2011	9	9.92	19.24	0.52
14	One Of Us	Glee Cast	2010	37	9.22	17.81	0.52
15	Wishing On A Star	The Cover Girls	1992	9	8.64	16.26	0.53
16	I Hate U	Prince	1995	12	8.83	16.64	0.53
17	Stranded (Haiti Mon Amour)	Jay-Z, Bono	2010	16	9.86	18.74	0.53
18	All I Know	The Weekend Featuring Future	2017	46	9.78	18.34	0.53
19	Great Is Thy Faithfullness	Jordan Smith	2016	30	10	18.98	0.54
20	Moves Like Jagger / Jumpin' Jack Flash	Glee Cast	2012	62	9.11	16.76	0.54

Table 4. List of underperform songs

5. Analyzing Over/Underperform Songs

By studying songs that are classified as over/underperform songs as defined by our method in section 4, we find many interesting characteristics of these songs. Analyzing about 100 songs, 50 of them are overperform songs, the other are underperform songs, we find that songs's durability can be explained by two main sets of factors:

- *Special occasions, temporary trends*. Example: famous cover of songs, death of artist, movies about artists released… These factors are unique for each song and therefore, are very hard to be used to predict durability of other songs.
- *Song's metadata*. Example: tempo, genre, loudness or artist hotness,… These are features that can be shared between different songs and therefore, can be used to predict another song's durability.

We will analyze each of these two sets of factors.

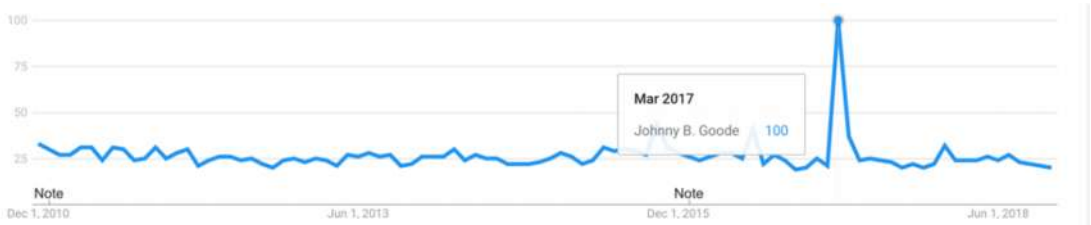
5.1 Songs' Durability Explained by Special Occasions, Temporary Trends

Studying over/underperform songs, we find there are some main special occasional/temporary trending reasons affecting songs' durability:

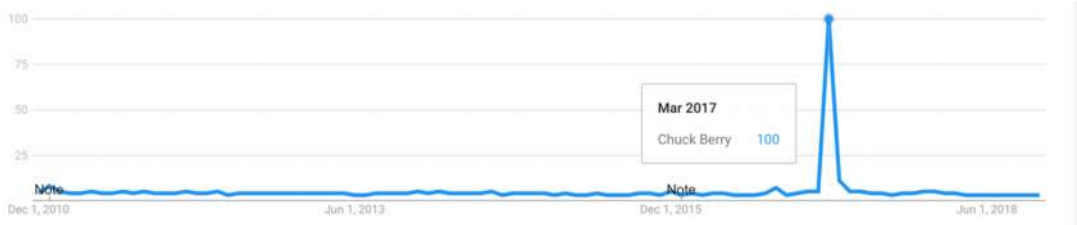
- **Artist's passing**

  When a well-known artist passes away, his death usually will be notified through many means of media to acknowledge many people. People listened to this news tend to want to know more as well as pay respect to the artists by looking up for them, reading about them and listening to their music. This temporary trend pushes Youtube views of the artist's records significantly. Below are two examples of a death of an artist making his songs become *overperform songs*.

- *Johnny B. Goode* by Chuck Berry
  Predicted views (log): 11.09
  Ground truth views (log): 7.443
  Predicted/Ground truth Ratio: 1.49
  By using Google Trends API analysis, we noticed that there is a surge searching Google for the artist name as well as his records around 18th March, 2017 which was the day Chuck Berry past away. This event helps pushing Youtube views count significantly and makes his records overperform.



a) Google Search trending for record "Johnny B. Goode"



b) Google Search trending for artist "Johnny B. Goode"

Figure 9. Google Search trending for record "Johnny B. Goode"by Chuck Berry

- *Starman* by David Bowie
  Predicted views (log): 10.27
  Ground truth views (log): 5.93
  Predicted/Ground truth Ratio: 1.73
  Similarly, because of David's death in January 2016, many of his records suddenly becomes highly interested by many music listeners. And this helps his record *"Starman"* overperform other songs released at the same time, having the same rankings on Billboard at the time.

✧ **Movies about artist's career**

Many artists' careers are made into movies by their admirers. Some of them are are outstanding, talented artists, the other have very inspirational stories. When these movies released, there is a great trend searching for the artists as well as their works. This helps pushing up Youtube views count of their records and make them overperform. Below are two examples.

- *Milord* by Edith Piaf
  Predicted views (log): 7.31
  Ground truth views (log): 3.15
  Predicted/Ground truth Ratio: 2.32

Édith Piaf was a French vocalist, songwriter, cabaret performer and film actress noted as France's national chanteuse and one of the country's most widely known international stars. Despite her success, Piaf's life was filled with tragedy. In 2007, a movie made about her life was released, helping people understand more about the life of this talented woman. Figure 10 illustrates the peak in year 2007 that helps her record -"Milord" become a overperform record.

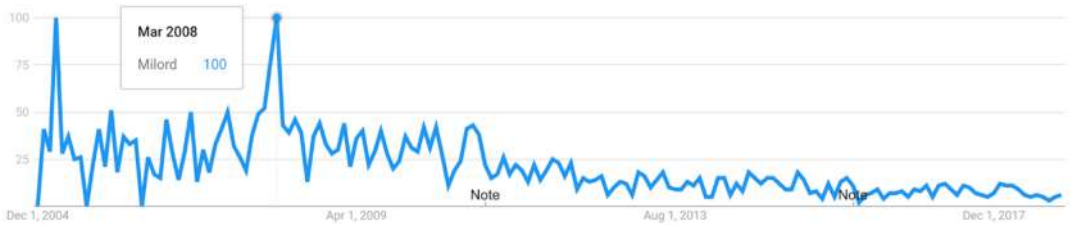


Figure 10. Google Search trending for record "Milord" by Edith Piaf

- *Don't Stop Me Now* by Queen
  Queen – the band considered the best rock band of all time has had a great influence on the rock records later generations. Recently (late 2018), the movie "Bohemian Rhapsody" about the band was released and created a wave of exploring the band, helping push their records popularity to be overperform.

✧ **Holidays occasions**

Songs relating to special holidays such as Christmas or New Year are usually played again and again every year, making their popularity goes up through time. Hence, overperform other songs that are released at the same time but not tying to a particular holiday occasion. One example is *The Christmas Song* by Nat King Cole. Figure 11 shows Google Search Trend of the song through years. We can see that every year at Christmas, his name as well as his records are searched by listened to.

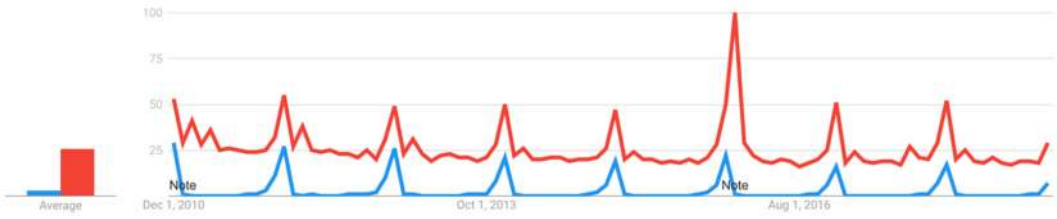


Figure 11. Google Search trending for record "The Christmas Song"(blue) by Nat King Cole (red)

✧ **Famous cover of a record**

Many times, a famous cover of an original record makes people listen to the original one and therefore, pushes the original's popularity. For example, "*The Twist*" by Chubby Checker was released in 1960 and came on top of Billboard rank then. In 1988, *"The Twist"* again became popular due to a new recording of the song by The Fat Boys featuring Chubby Checker. This version reached number 2 in the United Kingdom and number 1 in Germany. In 2014, Billboard magazine declared the song the "biggest hit" of the 1960s. The success of "*The Twist*" by The Fat Boys help pushing the popularity of the original record by Chubby Checker.

5.2 Songs's Durability Explained by Metadata of Songs

Analyzing the One-Million Song Dataset, we test the effect on durability of each important feature of a song, including:

- ✓ Tempo
- ✓ Duration
- ✓ Song hotness
- ✓ Loudness
- ✓ Artist hotness
- ✓ Genre

We find out that the most important factor is *artist hotness*. This is the main factor deciding whether a song will be over/underperform. We will analyze as well as show examples for this factor's influence on songs' durability below. On the other hand, other factors (tempo, duration, genre,…) are found to not have significant impact on a song's performance through time. Indeed, as Figure 13 shows, there is not much differences in these features of an over/underperform song with a usual song.

✧ Songs' durability explained by Artist hotness - effect of artist "Brand"

When analyzing songs' durability, we find that there is a great effect of artist's "Brand" to the performance of their songs. Indeed, for some artist, although one of their work might have a low Billboard ranking at its released time, but over the time, thanks to the artist's reputation, the song become overperform (Youtube views count higher than expected). On the other hand, there are some artists that although having highest numbers of songs in Billboard chart, turns out to be underperforming over time. Maybe due to their "Brand" is not good enough.

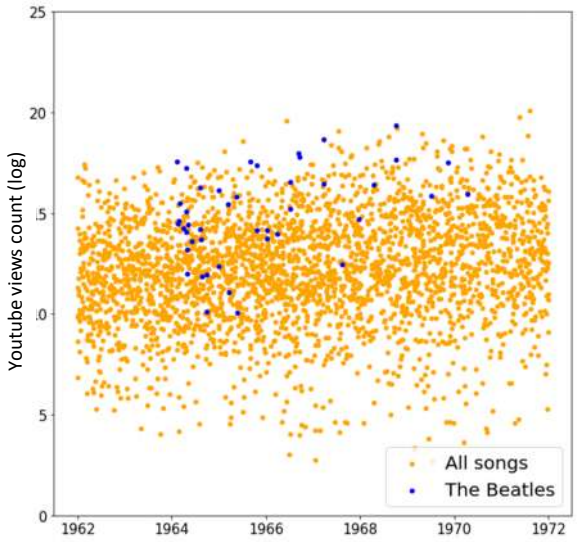
To choose artists to analyze, we first find a set of artists having highest number of songs in Billboard Top100. Then we analyze each of these artists and look for interesting relationship between artists' "Brand" and their songs' performance over time. Below are some illustrative examples explaining the true effect of artists' "Brand".

The Beatles are an English rock band formed in Liverpool in 1960. With members John Lennon, Paul McCartney, George Harrison and Ringo Starr, they became widely regarded as the foremost and most influential music band in history. And until now they are still considered one the best bands. Thanks to this fact, most of their songs, even songs that have low Billboard rankings when they were released perform very well through time. Indeed, many of their songs lie higher to much higher than average comparing to songs released at the same time. Figure 12a shows this fact. The explanation for this observation is that the "Brand" of "The Beatles" have been so great that boost all of their songs to be well-overperform comparing to their peers, even for songs that were not top Billboard chart in the past.

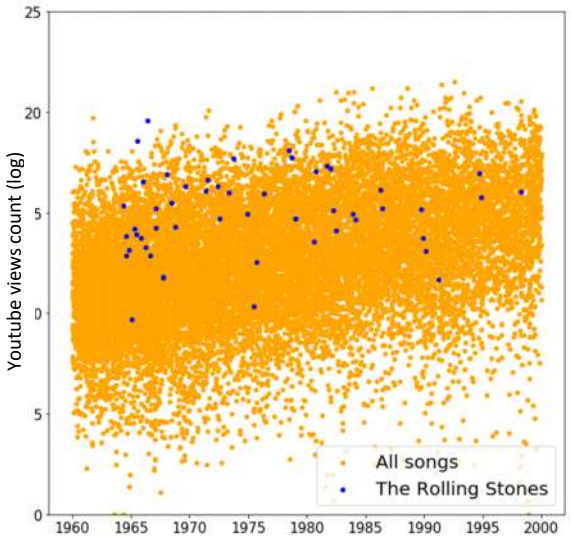
The Rolling Stones is another famous English band formed in 1962 with members Brian Jones (guitar, harmonica), Mick Jagger (lead vocals), Keith Richards (guitar, backing vocals), Bill Wyman (bass), Charlie Watts (drums), and Ian Stewart (piano). They are the forefront of a wave called "British Invasion" in the US in 1964. They are identified to be famously youthful and rebellious. This "Brand" helps their songs not only survive but overperform comparing to their peers until now. Figure 12b shows the band's records mostly above average to a good extent. *However, it seems that their earlier songs (from 1962-1975) are better overperforming than their latter (from 1985-2000) songs*. Indeed, their most overperform song is "Paint It, Black" which is released as early as 1966. While other latter songs (from 1985-2000) mostly stays average.

The Glee Cast is a band originally came from the musical series "Glee", which is very popular with US and world teenagers from 2009-2014. The series bring to teenagers a lively colorful musical life that inspired students to pursue stage art. The show has won many awards such as Golden Globe, Teen Choice

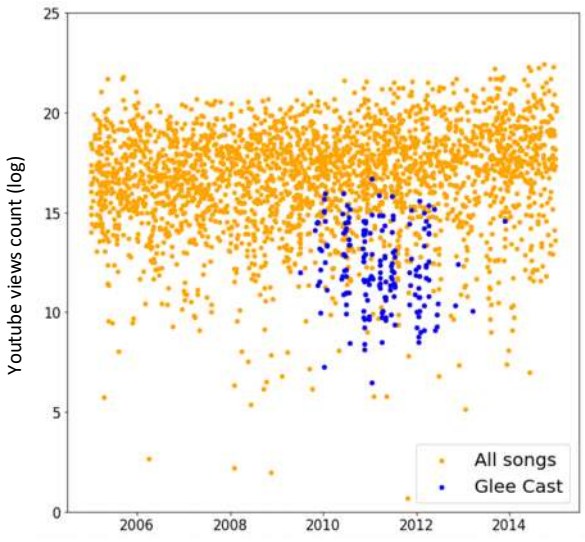
Award for many consecutive years. In Billboard chart, this band has the highest number of records which come up to 174 records. However, usually these records do not stand overtime. Indeed, most of their songs are well-underperform as seen in Figure 12c. The reason is that this band generally does not have their "original songs". Because as a musical, they mainly cover famous songs from famous singers (Lady Gaga, Katy Perry, Madona,…). Although this helps them to get into Billboard chart at the time they released the cover, it does not help them stand over time. Therefore, their records turn out to be underperforming over time.



(a) Popularity of The Beatles versus All songs



(b) Popularity of The Rolling Stones versus All songs

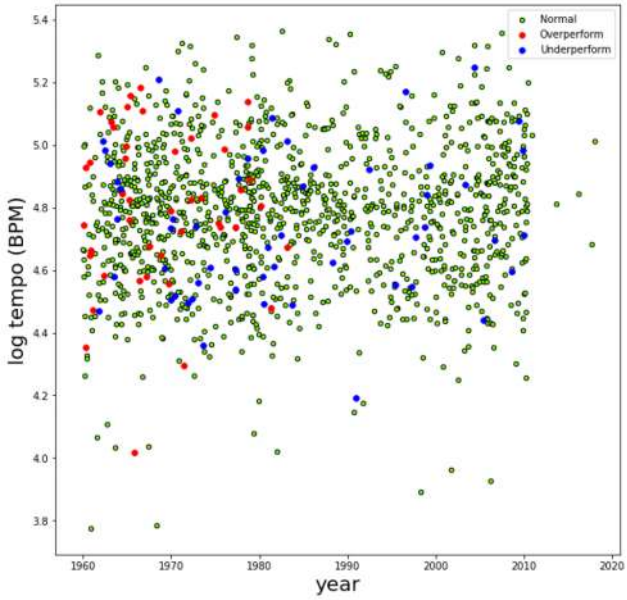


(c) Popularity of Glee Cast versus All songs

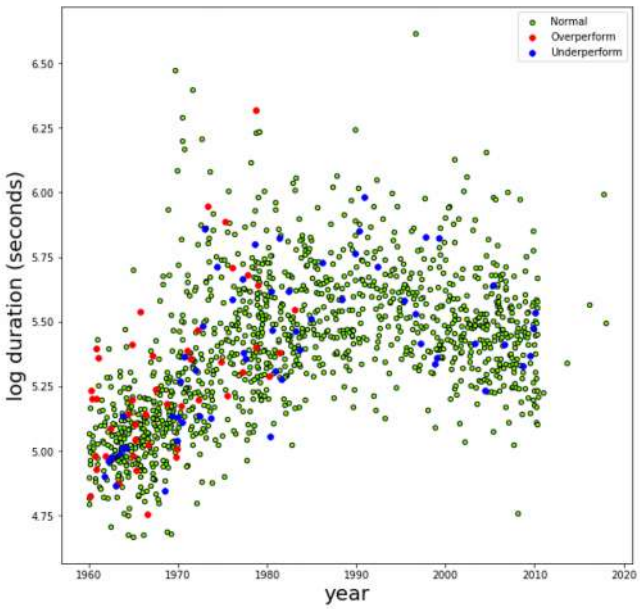
Figure 12. Effects of artist "Brand" on songs durability

✧ **Songs' durability explained by other metadata**

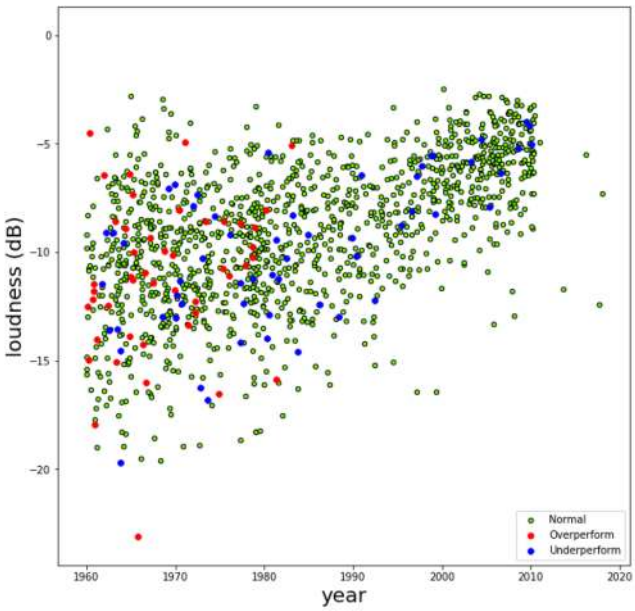
As mentioned above, other metadata (tempo, loudness, duration, genre…) seem not to have a significant effect on songs' durability. In general, these features of over/underperform songs are not different much comparing to usual songs. Figure 13 illustrates this statement. We can see that there is not any separation between over/underperform songs with usual songs.



(a) Tempo of over/under/normal songs through years



(b) Duration of over/under/normal songs through years



(c) Songs hotness of over/under/normal songs through years

Figure 13. Effects of other metadata (tempo, duration, loudness, songs hotness) on songs durability

6. Conclusion

By analyzing Youtube views count, Billboard Top100 chart and One-Million Songs dataset, we gain a better understanding about factors that affect a song's durability through time. Our findings point out that the most important factors for a song's performance is the "brand" of the artist. This observation can be illustrated by many examples. Other factors also very important are special occasion and temporary trending such as death of an artist, releasing of a movie relating to the artist. All of the factors combining together helps a song over/underperform comparing to other songs released at the same time, at the same specific Billboard ranking.

7. References

[1] Million Song Dataset: https://labrosa.ee.columbia.edu/millionsong/

[2] Billboard ranking: https://www.billboard.com/charts

[3] Grammy Awards: https://www.kaggle.com/theriley106/grammyawardsinnumbers

[4] S. Homan, "Popular music and cultural memory: Localised popular music histories and their significance for national music industries: data," 2012.

[5] Y. Kim, B. Suh, and K. Lee, "# nowplaying the future billboard: mining music listening behaviors of twitter users for hit song prediction," in *Proceedings of the first international workshop on Social media retrieval and analysis*. ACM, 2014. pp. 51-56.

[6] J. Berger, and G. Packard, "Are Atypical Things More Popular?"