An Analysis of Songs' Survival on Billboard Top 100 Weekly Chart

Introduction

It might sound somewhat absurd using math and statistics to study music, but have you ever wondered what makes a hit song? What is the trend of the music genre? How do people's taste of music change over time? This paper will show you the types of music that are most likely to be popular.

We assume a piece of music is "popular" if it is on the Billboard Top 100 Weekly Chart. Billboard, as a famous entertainment media brand, tracks the most popular songs and albums in different genres, including tempo, danceability, acousticness, loudness and many more in the past several decades. We build survival models to compare the survival of the music pieces with different identities to see what kind of music can stay longer on the chart. Besides, we develop KM models to investigate people's music preferences or attitudes toward music over the years and how that change in preferences might have an impact on a song's survival on the list. It's worth mentioning that by looking at the data, we discover some hidden criteria for rankings and how digitalization of music is affecting the length of survival for each piece of music.

Our paper will focus on answering the following 2 questions:(1) What could be the factor within a song that influences its survival? (2) What could be the factor that is among songs that are similar? We will look at numeric covariates that have the potential to influence a song's survival on the chart. Plus, we will show how the categorical covariates of music influence the song's survival.

Methods and Data

To better answer these two questions, we introduced several correlated datasets that contain the songs and the artists that appear on the Billboard Hot 100 charts throughout the past decades. We filtered the datasets and added covariates by calculation. The datasets we focus on are `Billboard60yrs`, `Billboard20yrs`, and `Spotify`. To guarantee the consistency of our research, each song has a unique SongID in all these datasets.

`Billboard60yrs` originates from a data.world dataset which includes the weekly top 100 songs from 1958 to 2017, as well as their position, previous peak position, and weeks on chart of that given week[1]. This dataset is created by Sean Miller. We filter it into `Billboard60yrs` so that it now only includes a song's name, artist, ID, and maximum weeks on chart(maxWeeks), i.e. our time variable, and a variable called decade. The variable presents which period a piece of music is in.

`Billboard20yrs` originates from a dataset created by Daniel DeFoe which includes the recent 20 years (2000-2019) of weekly top 100 songs on Billboard[2], somewhat overlapping with the first raw dataset but with more features like lyrics, genre, and writing credits, which allow us to do more diverse research topics. Based on the original data, we then added columns including maxWeeks (the survival indicator), binary classifiers of whether a specific song is Pop, Hip Hop, Country, etc. or not, as well as categorical covariates that assign each of the songs into a specific category of time period. Because the dataset is limited to recent 20 years, we defined the temporal categories as 4 five-year periods, from 2000 to 2019. We then put the songs into one of these four categories, according to their year of release. We also added another binary classifier that evaluates if a song is created by one of the Artists of the Decade, whose names we acquired from Billboard.

`Spotify` comes from a dataset created by George McIntire that features top 20 songs of Billboard weekly top 100 charts, irrespective of their temporal order[3]. We joined the original dataset with `Billboard20yrs` and made `Spotify` in order to get more information on the songs on this chart.

Song attributes' effects on survival:

The non-parametric estimation by Kaplan-Meier does not make a good estimation, so we test different models and find out that the lognormal model makes a good estimation. The variables we used are acousticness: a confidence measure from 0.0 to 1.0 of whether the track is acoustic; liveness: the presence of an audience in the recording, where higher liveness values represent an increased probability that the track was performed live; tempo: the overall estimated tempo of a track in beats per minute (BPM); and instrumentalness: predicts whether a track contains no vocals. The four variables are all numeric. Log TR meaning...

3 independent univariate models...

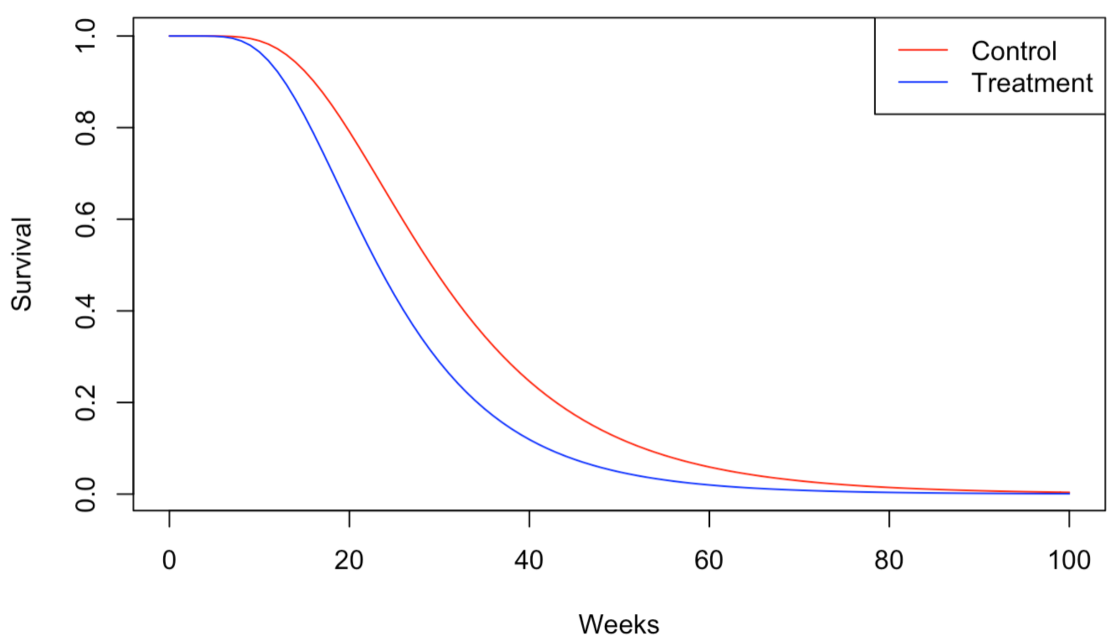
	log(Time Ratio)	Std. Error	P value
Model 1			
tempo	-0.001868	0.000819	0.023
Model 2			
acousticness	0.3026	0.1220	0.013
Model 3			
liveness	-0.3905	0.1532	0.011

We then put all three variables into a single model and the results are shown below:

```
                        Value Std. Error      z      p
(Intercept)          3.371244   0.109799  30.70 <2e-16
Spotify$liveness    -0.344448   0.152884  -2.25  0.024
Spotify$tempo       -0.001585   0.000818  -1.94  0.053
Spotify$acousticness  0.233280   0.122929   1.90  0.058
```

	log(Time Ratio)	Std. Error	P value
tempo	-0.001585	0.000818	0.053
acousticness	0.233280	0.122929	0.058
liveness	-0.344448	0.152884	0.024

Three variables here are or very close to being significant. Based on this lognormal AFT model, we can assume two songs' attributes and compare their survival. The first song will be the control group with 0 liveness, tempo and acousticness. The second song will have 0.19 liveness, 121.82 tempo, and 0.130 acousticness. The values are the mean liveness, tempo and acousticness respectively. Below is the survival graph for the two songs:



And by calculating the area under the survival curve, we get the mean survival time for the two songs, 32.4 for control and 25.8 for treatment, which is a song with mean tempo, liveness, and acousticness. One thing to mention here is that this model is based on Billboard top 20, not top 100.

We can also use a non-parametric alternative (Cox PH model) to the lognormal model mentioned above:

	log(hazard ratio)	Hazard ratio	P value
tempo	0.002821	1.002825	0.123
acousticness	-0.342780	0.709794	0.214
liveness	0.132080	1.141200	0.695
Likelihood ratio test=5.06 on 3 df, p=0.1678			

```
                          coef exp(coef)  se(coef)      z     p
Spotify$liveness      0.132080  1.141200  0.337364  0.392 0.695
Spotify$tempo         0.002821  1.002825  0.001828  1.544 0.123
Spotify$acousticness -0.342780  0.709794  0.276045 -1.242 0.214
Likelihood ratio test=5.06  on 3 df, p=0.1678
```

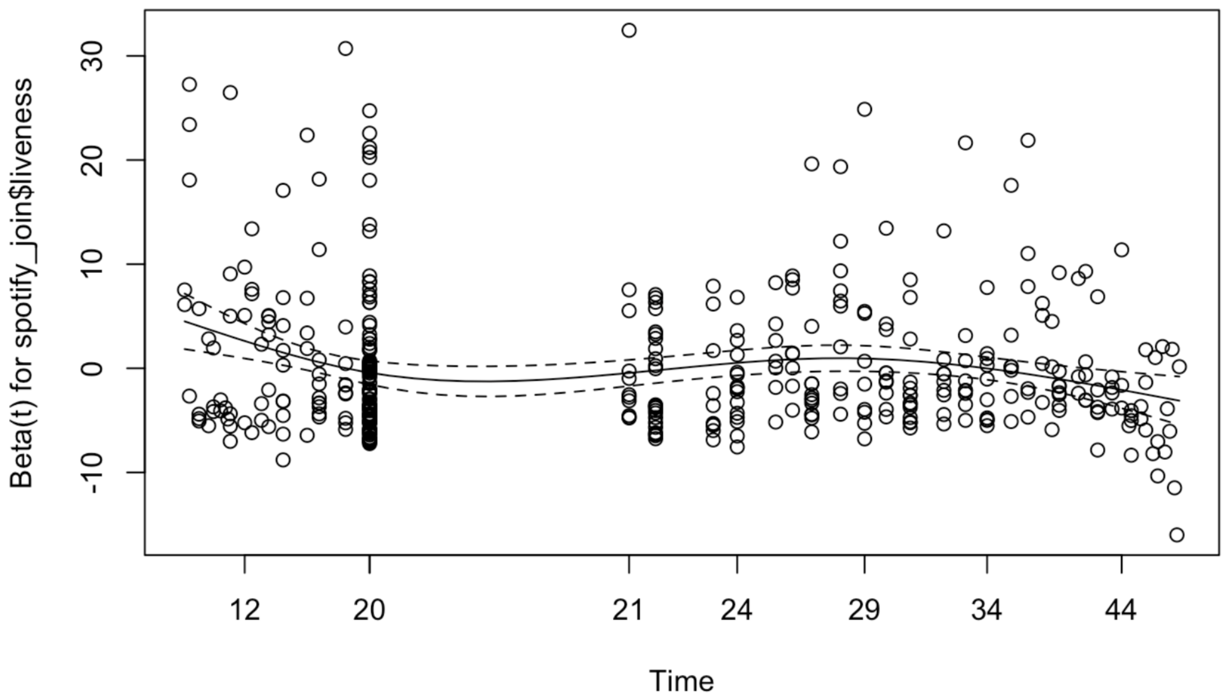
If we fit another variable instrumentalness into the model, the model improves significantly.

	log(hazard ratio)	Hazard ratio	P value
tempo	0.002702	1.002706	0.1393
acousticness	-0.316719	0.728536	0.2512
liveness	0.152575	1.164830	0.6514
instrumentalness	2.796143	16.381334	0.0467
Likelihood ratio test=7.84 on 4 df, p=0.09746			

```
                              coef exp(coef)  se(coef)      z      p
Spotify$liveness          0.152575  1.164830  0.337700  0.452 0.6514
Spotify$tempo             0.002702  1.002706  0.001828  1.478 0.1393
Spotify$acousticness     -0.316719  0.728536  0.276004 -1.148 0.2512
Spotify$instrumentalness  2.796143 16.381334  1.406094  1.989 0.0467
Likelihood ratio test=7.84  on 4 df, p=0.09746
```

The sign of the coefficients correspond to what we have for the AFT model. But none of the covariates attain statistical significance, except for the newly added instrumentalness. By calculating the hazard ratio, we find out that a song with average tempo,liveness,acousticness and instrumentalness is 1.38935 times more likely to fall out of the list than the control group song.

The PH assumption of this model is tested by plotting the Schoenfeld Residuals and a formal test:



	P value
tempo	0.231
acousticness	0.119
liveness	0.032
instrumentalness	0.360

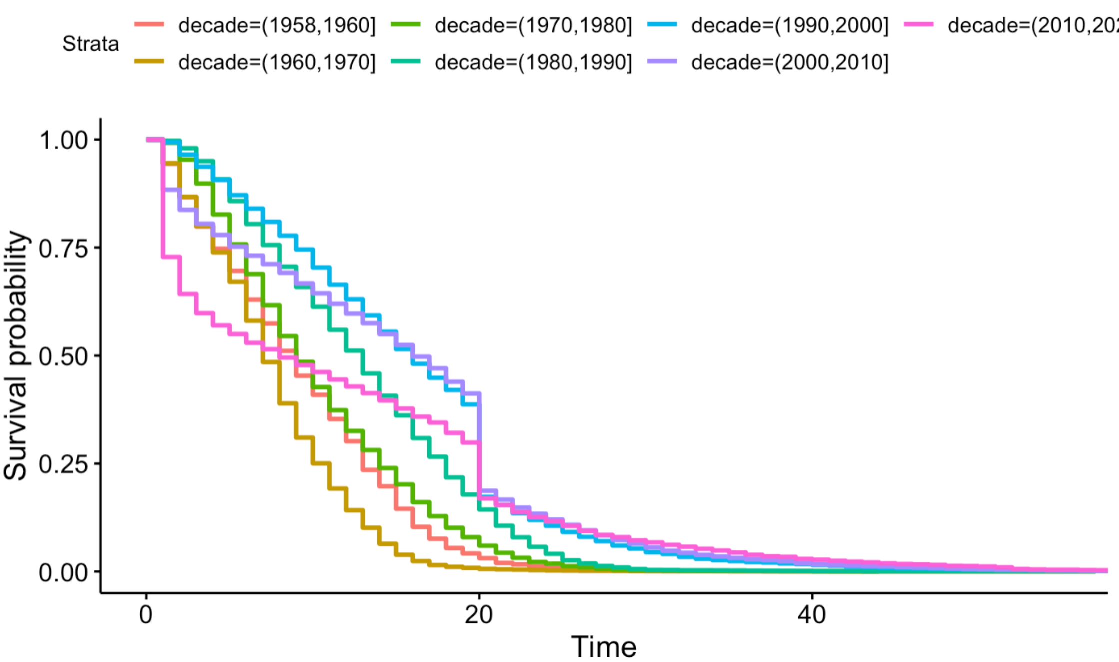
```
                         chisq df     p
Spotify$liveness         4.614  1 0.032
Spotify$tempo            1.435  1 0.231
Spotify$acousticness     2.433  1 0.119
```

```
Spotify$instrumentalness 0.837  1 0.360
GLOBAL                   7.887  4 0.096
```

The Schoenfeld residual plots generally show no violation of the PH assumption, except for the liveness variable graph above. This also agrees with the formal test with a small p-value for liveness, as well as a large p-value of 0.6514 in the model.

In conclusion, a piece of music that lasts longer on the top 100 is generally associated with slow tempo, less acousticness, less instrumentalness and more acousticness.

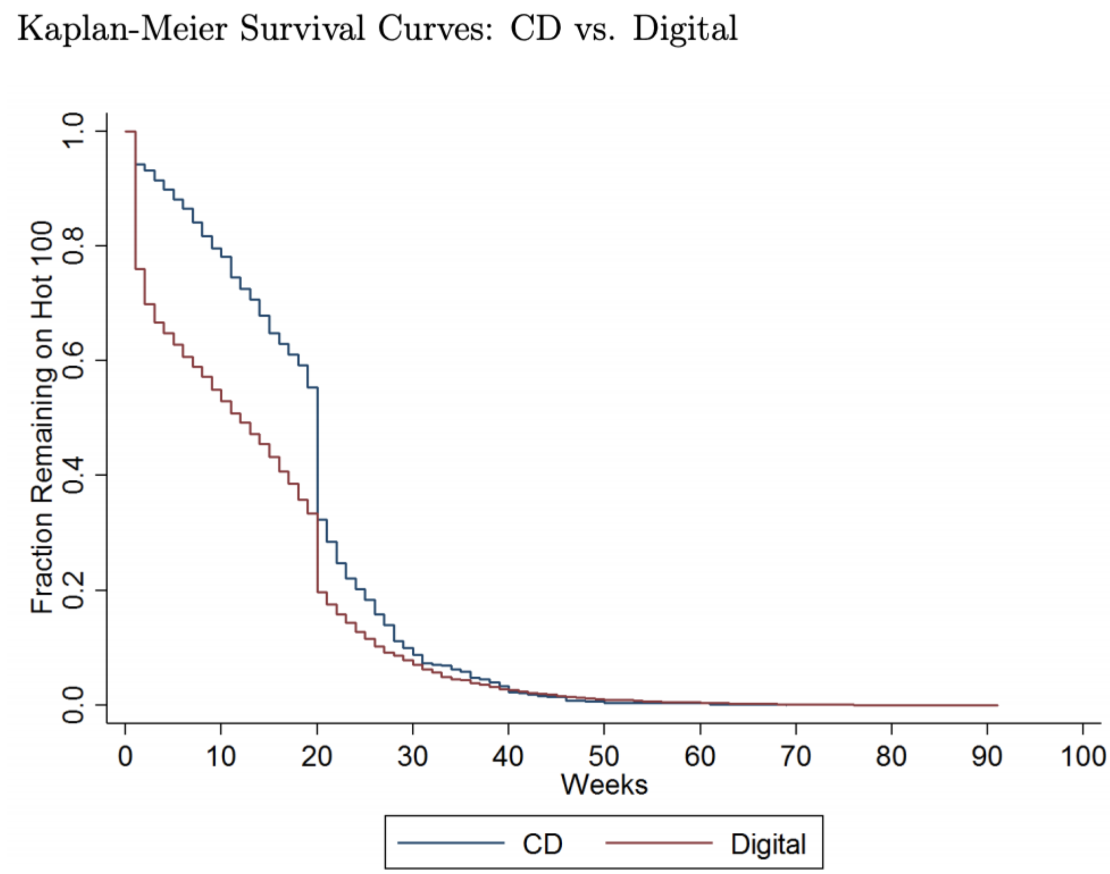
Having researched the factors within a song that may contribute to the increase in survival, we want to ask another question: could a song's length of survival also be related to the categories it fits in? We want to know how similar the survival experience a song is compared to other songs that are similar in some ways, so we started by defining some categories that may help. This includes genres or the time period of the song.



The first thing we notice on the graph (and many other graphs) is the sudden drop of survival at around week 20. We investigate and find out that there is a hidden criterion for the weekly chart. If a song has been on the chart for 20 weeks, once it falls out top 50 the next week, it is counted as out. We see this sudden drop only on the 90s, 00s, and 10s curves, so we infer that this is an adjustment made in the late 80s. There is also a mechanism for songs lasting more than 52 weeks -- if it falls out of top 25 the week after the 52nd, it is counted as out. Very few songs can last for 52 weeks, so its effect on the graph is less discernible. These rules also state that a song that is kicked out can only return to the Top 100 chart if it eclipses its previous peak position.[4] We have not been able to find the exact motive for the 20-week rule, but an unconfirmed source states that the 52-week rule is established in response to the viral hit *Uptown Funk* in 2015. [5]

The next thing we notice is the sudden drop of the purple curve (2010s) at around week1: about 25% of the songs in the 2010s cannot survive more than 1 week, whereas compared to ~10% for the rest of the decade groups. The runner-up decade with this phenomenon is the 2000s, which has a survival of ~85% after week 1. This means that more songs in the 2010s and 2000s stay shorter on the chart. This could be due to the online streaming of music, the popularization of social media and smartphones and its resulting marketing campaigns and/or strategies, as well as people's change of preferences.

Besides the purple line, we can tell from the graph that old songs tend to survive longer. Below is a graph we find from a research on Billboard charts done by Lao and Nguyen[6], which shows significant longer survival for CD.



One possible explanation is the number of music created each year, or decade. More music means more competition. Also, the listeners are becoming more diverse and larger in population, which also leads to more competition.

Move on to binary covariates

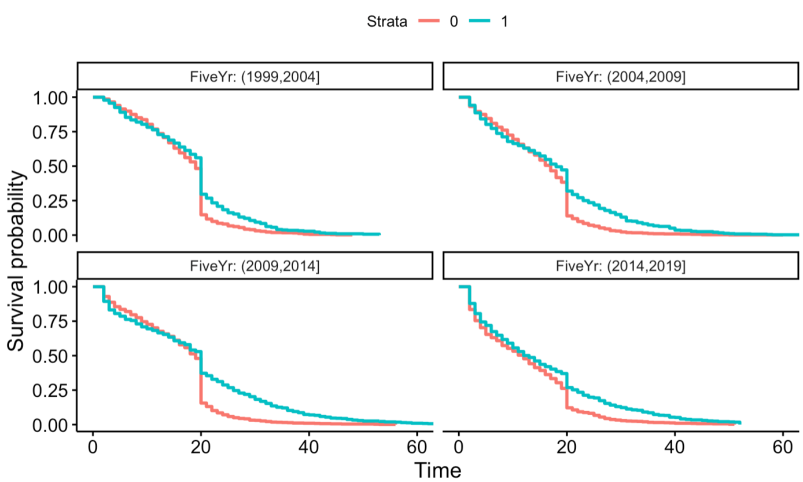
- How do they work?
    - Yes or No as the value
    - Explain 0 and 1
    - Eg. genre
- What did we do
    - Grepl - search specific texts
    - Eg. electro pop - isPop, isElectronic

- Did the same thing for artists of the decade

What did we do

- Left: % of that genre vs fiveYr time period
- Right: Survival curve of that genre vs not that genre, stratified by fiveYr time period

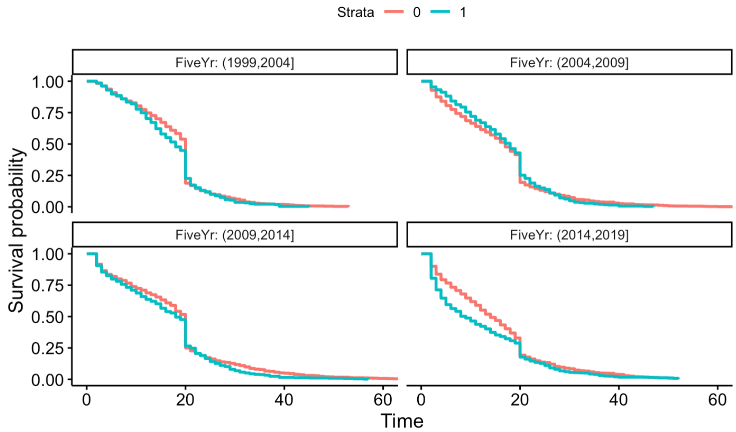
isPop



Pop music in every one of the four five-year time periods adapt to the 20-week rule better than the non-Pop music genre. Although they tend to have similar survival experiences pre-week 20, a smaller proportion of songs drop out of the charts after week 20, meaning that a larger proportion of Pop songs stays in the top 50 range after week 20.

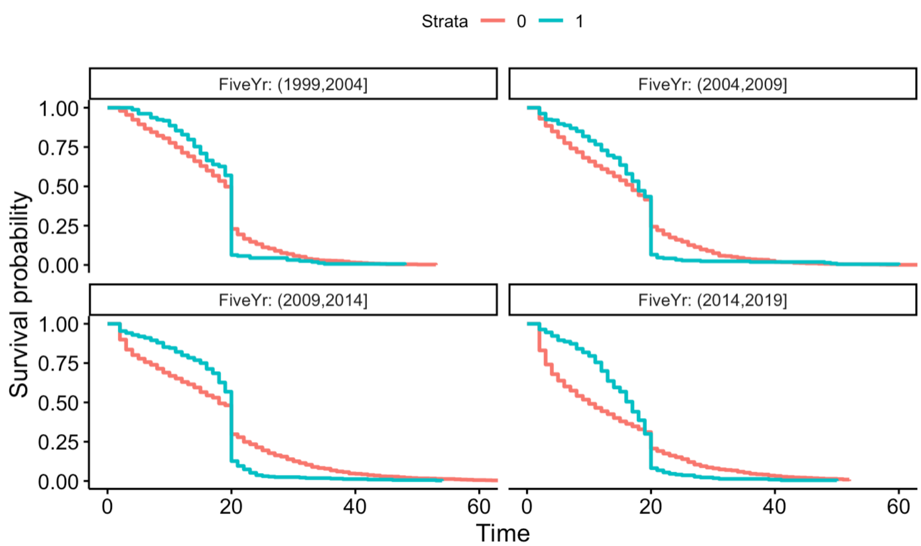
- Pop music is a preferred genre by more people.

isHipHop



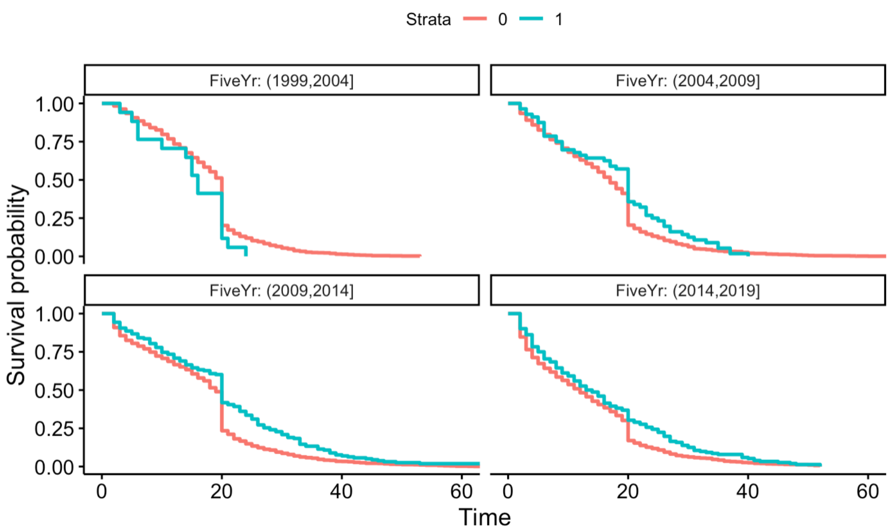
- Generally similar survival experiences
- From 2015 to 2020, hip hop songs tend to have worse survival than other songs, indicated by the gap between the blue and red curves in the last graph.
- Steeper - higher hazard (chance failure at this time point, given it has not yet occurred)
- Larger percentage in industry brings more competition
- More competition = higher turnover rate = more easily swapped by other popular songs.

isCountry



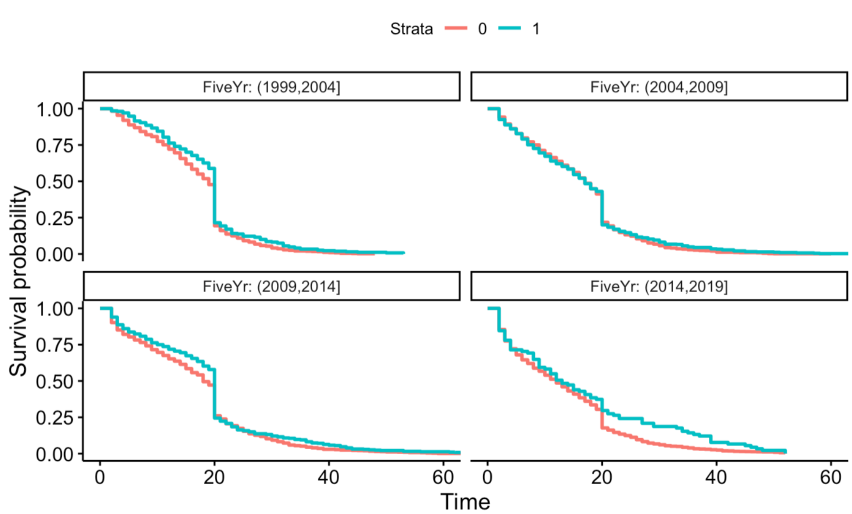
- Gap larger as fiveYr period changes
    - Stays rather stable in the first few weeks. In the chart, not so easy to be out
    - better overall survival experience compared to non-country before week 20
- More significant drop at week 20 in all FiveYr periods:
    - Higher proportion of Country music ranked on lower 50.
    - Popular, but less well-received compared to Pop

isEDM



- for each five-year period since 1999, the survival of EDM music has been increasing.
    - 1) In a longitudinal comparison, for the first five-year period (2000-2004), almost no EDM music survives more than 20 weeks. Since then, its survival has been in steady increase up to (2010-2014), where the largest maxWeeks is off the charts.
        - Due to the limited sample size, perhaps?
        - Corresponds to the popularization of EDM in American music industry since 2010 - rave culture
    - 2) In a cross-sectional comparison, within each five-year period, the performance of EDM music begins with slightly beneath the Kaplan-Meier curve of non-EDM music, and eclipses the curve of non-EDM music in the following five-year periods.

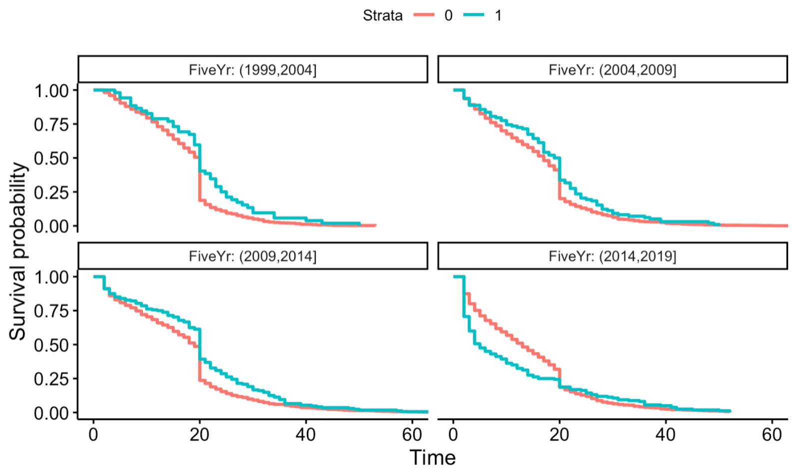
Rock:



*

- % in decline over the years
- Survival experience gets better (also, due to limited sample size?)
- Due to the emergence of EDM, substituting Rock as the iconic subculture.

Artists of the decade:



Songs involved in Artists of the decade generally perform better than those that are not in the first three five-year time periods. In the last five-year period, the survival of the songs that are from an Artist of the Decade is worse than those that are not.

- more hits by famous singer enter the charts, but less stays long
- Emergence of more grass-root music creators and new ways of publicizing the music (instagram, tiktok, musical.ly, etc)
    - Eg. Old town road.
- More artists to share the charts, so less chance for the top to stay in - fierce competition

Concl:

- knowing the survival of some specific categorical covariates is helpful for us to understand the larger trend
- Though no models, we used KM to simply reflect what is happening.

Conclusion:

By fitting different models, we find that a piece of music that lasts longer on the top 100 is generally associated with slow tempo, less acousticness, less instrumentalness and more acousticness. For recent songs, they tend to have a harder time surviving on the list, initially, than old songs: only about 25% of songs from 2010-2020 can survive more than 1 week, compared to roughly 10% for the other decades. This is especially evident after online streaming starts, which brings more exposure and thus more competition.

Among multiple genres, we see that as time progresses, some are being accepted more by the public while others are not. Also, we noticed that as the technology develops over the years, it inevitably affects the music industry and establishes new and more dynamic trends. Similarly, with the advent of online social networks and video platforms, grass-root artists find themselves closer to being recognized by a larger audience and competitive against other famed artists.

Sources:

1. Miller, Sean. 2017, https://data.world/kcmillersean/billboard-hot-100-1958-2017/workspace/file?filename=Hot+Stuff.csv. Accessed 10 March 2021.
2. DeFoe, Daniel. *Data on Songs from Billboard 1999-2019*, 3 March 2020, https://www.kaggle.com/danield2255/data-on-songs-from-billboard-19992019. Accessed 10 March 2021.
3. McIntire, George. *Spotify Song Attributes*, 5 August 2017, https://www.kaggle.com/geomack/spotifyclassification. Accessed 10 March 2021.
4. Billboard. *Billboard Charts Legend*, 31 October 2019, https://www.billboard.com/p/billboard-charts-legend. Accessed 10 March 2021.
5. Zhihu User. Comment for Answer for *How are Billboard Charts Made?* 10 February 2020, https://www.zhihu.com/question/20040905/answer/1003179333, Accessed 10 March 2021.
6. Lao, Jerry, and Kevin Hoan Nguyen. "One-Hit Wonder Or Superstardom? The Role Of Technology Format On Billboard'S Hot 100 Performance". *Web.Stanford.Edu*, 2016, https://web.stanford.edu/~xhnguyen/BillboardandTechnology.pdf. Accessed 10 Mar 2021.