DTU Compute
Department of Applied Mathematics and Computer Science

The dynamics of pop music charts: Is it getting harder to make a hit?

Master Thesis

Marta Ewa Lech

The dynamics of pop music charts: Is it getting harder to make a hit?

Master Thesis
June, 2023

By
Marta Ewa Lech

Cover photo: Dmitry Demidov, 2020
Published by: DTU, Department of Applied Mathematics and Computer Science, Richard Petersens Plads, Building 324, 2800 Kgs. Lyngby Denmark
https://www.compute.dtu.dk/

Approval

This thesis has been prepared for five months at the Department of Applied Mathematics and Computer Science, at the Technical University of Denmark, DTU, in fulfillment of the requirements for the degree Master of Science in Computer Science and Engineering, MSc Eng. The project period was from 2nd January 2023 to 2nd June 2023 and the thesis corresponds to 30 ECTS. The supervisors were Sune Lehmann Jørgensen and Jonas Lybker Juul.

It is assumed that the reader has a basic knowledge in the areas of statistics.

Marta Ewa Lech - s212481

Marta Lech

Signature

02.06.2023

Date

Abstract

Music plays a big part in our lives. It is a powerful tool for expressing emotions and influencing our mood. During the last few decades, decreasing album sales and the increasing popularity of streaming services have shaken the music industry. Songwriters and artists have argued over this development's negative impact on the creative community. Some suggest the industry has gotten more competitive due to the increasing amount of music uploaded to streaming platforms; in particular, some argue that making a hit has gotten more challenging. Until now, however, these claims have largely remained claims, and large-scale data analyses have not been used to investigate whether it has indeed become harder to make a hit. In my thesis, I confront this knowledge gap. To investigate whether it has become more challenging to make hits, I scraped the influential Billboard Hot 100 music chart and used modern data science tools to investigate how the song chart dynamics have changed since 1958, when the Billboard Hot 100 first started. I find hints that the industry has become more competitive, making it more difficult for some songs to achieve better rankings: Nowadays, most songs last shorter on the charts and experience more abrupt changes than in the 20th century. On the other hand, other songs have significantly improved their lifetimes: The highest-ranked songs maintain their top positions in the charts for longer time spans nowadays than in the past. The study also reveals other changes in chart dynamics, such as songs starting at significantly better ranks nowadays as compared to the previous century. Moreover, most charts seem occupied by constant hitmakers, giving less chance for new artists to enter.

The changes in the dynamics in the last few decades, which I document in this thesis, support the artists' claims: The dynamics of hit songs have changed over time; it seems that it has grown harder to make a hit.

Acknowledgements

I want to thank my supervisors, Jonas Lybker Juul and Sune Lehmann Jørgensen, for their guidance, curiosity, and support in the last five months. Their interest in the results of my study has greatly encouraged me throughout the research process. Special thanks to Jonas for our weekly meetings, valuable input, and for helping me steer this study in new and promising directions.

Lastly, I would like to thank Mikołaj, my friends, and my parents for supporting me in the last five months.

Contents

Preface . . . ii
Abstract . . . iii
Acknowledgements . . . iv

1 Introduction **1**

2 Literature Overview **3**
2.1 Science of science . . . 3
2.2 Success in artistic careers . . . 3
2.3 Social influence and collective memory . . . 4
2.4 Impact of digital technologies on music . . . 4
2.5 Predicting hit songs . . . 5
2.6 Career boosting . . . 5
2.7 Dynamics of rankings . . . 5
2.8 Summary . . . 6

3 Dataset **7**
3.1 Other charts . . . 10

4 Methods **13**
4.1 Exploratory data analysis . . . 13
4.2 Environment . . . 13
4.3 Scraping . . . 14
4.4 Data preprocessing . . . 15
4.5 Data segmentation . . . 15
4.6 Probability theory and statistics . . . 15
4.7 Distance measures . . . 17
4.8 Ranking measures . . . 17
4.9 Curve-based trajectories analysis . . . 19
4.10 Filtering . . . 19
4.11 Archetypal analysis . . . 20
4.12 Machine learning . . . 20

5 Results **23**
5.1 Lifetime of songs is increasing for the best-performing ones and decreasing for others . . . 23
5.2 Position change distribution gets more balanced . . . 29
5.3 Song trajectories get broader and start higher for best-performing pieces . . . 35
5.4 Lower-ranking songs are moving more abruptly on the charts . . . 45
5.5 Fewer new artists get to be in the charts and the number of collaborations increases. . . . 49
5.6 Hitmakers dominate the rankings, but their songs tend to perform worse. . . . 53

6 Discussion **57**
6.1 Increasing rivalry in the music industry . . . 57
6.2 Impact of streaming . . . 58
6.3 Collaborations as a way of boosting the performance . . . 59

6.4 Limitations of this study . . . 59
6.5 Future work . . . 60

7 Conclusion **61**

Bibliography **63**

A Datasets **69**
A.1 Billboard Hot 100 . . . 69
A.2 iTunes Podcasts . . . 69
A.3 New York Times best sellers . . . 71

B Results for other datasets **72**
B.1 Podcasts . . . 72
B.2 Books . . . 74

C Code samples **77**
C.1 Billboard Hot 100 spider . . . 77

D Position change distribution **78**
D.1 Pearson's first and second skewness . . . 78

E Trajectory analysis **79**
E.1 Average and median trajectories . . . 79
E.2 Start and end positions . . . 81
E.3 Hamming distance . . . 82
E.4 Clustering . . . 83

F Ranking dynamics **86**
F.1 Displacement probability . . . 86

G Hitmakers **87**
G.1 Top positions of selected famous performers . . . 87

1 Introduction

The last few decades of technological development have vastly influenced the music market. With the increasing popularity of streaming services, such as YouTube and Spotify, the industry drastically changed its marketing and release strategies [29]. The album sales importance was quickly dominated by songs' digital sales and airplay around the 90s and early 2000s [4, 40]. One of the most prestigious music charts, Billboard Hot 100, was forced to change its ranking strategy numerously throughout its history to follow the trends in music listening [44]. In 2012 Nielsen Media Research provided 625 million streams in only one week for Billboard to incorporate into their ranking rules. It is estimated that in 2020, 2.17 trillion streams were played worldwide [1], which gives a tenfold increase. Moreover, in 2022 on average, 49000 songs a day were uploaded to Spotify [93]. All these numbers show how streaming sites have dominated the music industry. Consequently, record companies have adapted to using these services to their advantage [29].

The changes in the music market have shaken up the artistic community. David Hesmondhalgh has summed up the criticism of streaming in his paper [42]. The central claims were that the new system financially damaged the musicians, unjustly boosting the major labels and the distribution of music awards. One of the most successful songwriters, Ryan Tedder, has said in an interview for BBC that the popularity and capabilities of streaming services have made it harder for artists to create new hits [79]. First, newly released songs are up against every other song released in the past 70 years. Apart from that, with tens of thousands of songs uploaded to Spotify daily, the industry has become more competitive than ever. Furthermore, social media platforms, such as TikTok, can generate a large amount of interest quickly, making a song "viral" unexpectedly, sometimes years after its release. According to Tedder, it is nearly impossible now to predict which songs will perform well on the charts and when.

Research on the impact of streaming on music has shown that currently, it is the users' favorite way of exploring and listening to music [29]. While some artists believe that they are losing a lot of money, it is undeniable that streaming will continue to grow and control the music industry. There is still a lot unknown about its impact on the performance of songs. Have the last few decades of listening to music online changed how the tunes move on the charts? Are there significant differences in how songs performed in the 70s compared to the 2010s? Has it been fair for both the top artists and the new artists? I believe that the answers to these questions lie in the changes in the dynamics of charts throughout the years.

All things considered, the primary hypothesis I want to answer is: "Is it getting harder to make a hit?". To achieve that, I will investigate the changes in the dynamics of the Billboard Hot 100 weekly music charts. I chose Billboard, one of the most recognized music magazines worldwide and a standard of music ranking in the United States. The company has been putting a lot of effort into capturing the current trends and accurately reflecting the popularity of songs. It combines album sales, airplay, and streaming, to select the best of the best [15]. There are different ways of measuring the difficulty of making a hit using the Billboard Hot 100. A few significant features to compare are the length of the lifetime on the charts, the proportion of new artists, and many more. In my thesis, I will look at the different aspects of songs' performance and compare the results over time to show that it has gotten more challenging to make a hit.

Although considerable research has been conducted, we still do not understand the complex phenomena behind achieving artistic success. Combining the changes in dynamics with possible causes could help people working in the music industry understand the market and improve their release strategies. They could adjust their promotion schemes to match those of the best-performing songs in the charts, possibly extending the lifetime and peak position. It could also help them pick the artists to collaborate with to make as many songs enter the prestigious charts as possible. Understanding the dynamics could also help music magazines like Billboard improve their ranking systems. It could help reduce bias created by streaming services or social media and even the chances for all artists to enter the charts. After all, rankings play a big part in shaping the success of a song and should reflect the industry and listeners' experience as well as possible.

In my work, I will show clear hints indicating that making a hit song got more challenging. To achieve that, I will present the differences in the features of the Billboard Hot 100 pop music chart. I will focus on the following characteristics:

- lifetime of songs,
- weekly position changes,
- trajectories of songs,
- proportion and performance of new artists and hitmakers on the charts.

In the Literature Overview section, I will outline the current research on artistic success and other relevant literature. I describe the process of obtaining the data in the section Dataset. Next, in the section Methods, I explain the techniques used to get the relevant features from the data. The sections Results and Discussion contain the findings and the explanations of the results, as well as arguments supporting the thesis. In the last section, Conclusion, I will summarize the work and highlight the most important results.

2 Literature Overview

Creative industries have been studied in various contexts. Some researchers have focused on the science behind artistic success [51, 98] and finding ways to predict it [61, 92]. Others focus on economic [42, 78] and social [21, 22, 80] aspects. In the following sections, I present some of the most important research areas related to my work.

2.1 Science of science

"Science of science" (SciSci) is a general term used to describe research on the processes ruling scientific discoveries [34]. The main goals of these types of studies are to understand and help accelerate science. However, many of these findings can be mirrored in other industries, such as film, art, or music. Some overlapping topics include reputation and success.

One of the topics in SciSci is career dynamics. Fortunato et al. [34] have shown no patterns in the occurrence of the highest-impact work among scientists. Moreover, the highest-impact works often combine the prior work with new, unusual ideas. The studies in SciSci have also revealed that the distribution of citations for scientific papers is highly skewed. This means some papers are never cited, while others accumulate thousands of citations. The same effects can be spotted in the music industry [4].

Other researchers have also tried to detect productivity patterns and impact in scientific careers [27, 92]. They have shown that the impact is distributed randomly in scientists' careers. Sinatra et al. [92] have created a model that suggests that the papers of the highest impact require a combination of both luck and the author's ability to publish high-impact papers.

Lastly, Chu et al. [25] have suggested that the increasing number of papers might lead to scientific stagnation. They have shown that the growing number of publications increases the disproportions between articles with the most citations and those that will likely never be cited.

In summary, the science of science is a rapidly growing research field where scholars explore and quantify various aspects of academia. Many findings about careers can be mirrored in the music industry.

2.2 Success in artistic careers

Researchers have been trying to predict the events of highest impact work and hot streaks in the artistic industries [10, 51, 61]. One important insight from this line of research is that most of the processes occurring in careers are random [27, 98]. However, Janosov et al. [51] showed that there are careers more and less prone to luck. They have indicated that the success of electronic music is greatly influenced by uncertainty, while classical music is more resilient to this factor. Fraiberger et al. [35] showed that the success of artists is correlated with access to prestigious institutions.

Williams et al. [98] investigated the productivity of actors and showed that while they could not predict their efficiency (proportion of active years to the career length), they identified some patterns in their careers. The most famous actors are more likely to attract new jobs, expressing the "rich-get-richer" phenomenon. They also showed that artists' careers exhibit bursts of activity and inactivity (hot and cold streaks).

Lastly, many researchers have been trying to create models predicting whether a product will be successful [10, 18, 49, 101]. In the music industry, some methods use sonic features of songs to measure their quality and impact on success [8, 10, 49]. The results suggest that performers' portfolios and the features of products play a significant role in shaping creative careers. Bradlow et al. [18] took a different direction and tried to create a model for the Billboard Hot 100 ranking. They focused on a small sample of the whole dataset, which was one year. While they managed to capture some of the trajectories' statistics, they showed that it is difficult to estimate the exact positions, with only 10% accuracy.

2.3 Social influence and collective memory

Whether a piece of art or a song will become a success highly depends on the human behavior aspects: e.g., social influence [6, 21, 30, 71, 80] and collective memory [22, 62].

Popularity rankings are known to affect how people perceive product quality. However, social influence has been shown to increase inequality and instability in charts [21, 80]. It increases the gap between popular and unpopular products [31]. It has also been shown to increase the unpredictability of the market [80]. Consequently, the best products are unlikely to perform poorly, and the contrary is observed for the worst products, but all the other outcomes are possible. This might suggest that social influence contributes to the strengthening of the "winner-take-all" effect [19]. There is a growing demand for charts to be fair and impartial to meet people's expectations.

In the Internet era, social media also play a massive role in shaping trends in different fields, including music. Many people use websites like Twitter to spread information, which has also been a powerful marketing tool [52]. Cosmiato et al. [30] created a highly accurate model (97%) for predicting album positions in ranking based on the traffic and sentiment from social media posts. Moreover, they showed that the most relevant features of their training came from Twitter.

Another important factor of product performance is collective memory. Researchers have shown that it consists of an initial phase of great attention, followed by a gradual decrease (forgetting) [22, 36]. However, others have also shown that collective attention spans are becoming shorter, presumably due to the increasing amount of information available and its quick spreading [62]. Digital media have also increased our consumption, leading to faster exhaustion of interest in a given topic.

2.4 Impact of digital technologies on music

The digital era of music has reduced the cost of creating songs, as well as the costs of listening to them. Platforms like YouTube have been proven to speed up the process of making a song a hit [37, 58]. The first signs of the impact of digital access to music have been investigated for MP3 music sharing [11, 12]. Researchers investigated how the websites allowing to download MP3 copies of music affect the survival of albums on music charts. They came to the conclusion that peer-to-peer technologies enhanced the performance of high-debuting songs and superstars. On the contrary, it decreased the lifetime of low-ranked albums and less famous artists. That article gave some first hints on how the digital era impacts music performance.

Easier access to music made it easier for lesser-know artists to be heard [4, 70] but also increased the overall competition for attention [33]. The position of superstars might be threatened due to the increasing number of artists and songs they have to compete with [5, 38, 102].

Though the appearance of streaming platforms has been controversial, some research indicates that the industry has experienced more positives than negatives [29]. Consumers can easily and freely discover new songs. Artists can reach bigger audiences and promote their products [78]. Moreover, streaming platforms have been shown to reduce piracy in the music industry [3].

2.5 Predicting hit songs

Predicting the positions of songs based on their features could help record companies and songwriters achieve better performance in times of overflowing music markets. Researchers have been developing hit prediction models, and recent findings have shown promising results.

Cosmiato et al. [30] created a highly accurate model based on social media data. They used a Random Forest classifier on quantitive features, such as the number of fans and sentiment features from posts. They could predict albums' positions in the Billboard 200 album chart with 97% accuracy (94% on real-usage test).

Even though classification and regression have been widely used in hit prediction, other researchers decided to try different methods. Some used deep learning techniques, such as convolutional neural networks (CNN) on audio features [99, 100]. Moreover, they showed that CNN achieved better results than simple linear regression. Pham et al. [74] utilized Gradient Decision Boosting Trees (GDBT) on the metadata of songs, lyrics, tonal, and rhythm features. Their solution outperformed many other predictive models, but they suggested that further improvements of that method are required.

Though song prediction was not in the scope of my study, the mentioned research shows that considerable improvements have been achieved with more complex machine learning algorithms. However, these models are relatively new, requiring further investigation to understand the learning process.

2.6 Career boosting

Growing prestigious networks (access to many prestigious institutions) is often a massive boost to careers in creative industries. Fraiberger et al. [35] took on the difficult task of quantifying reputation among artists. They showed that those with access to more prestigious networks were more likely to maintain long and successful careers. Sekara et al. [90] showed that experience also plays a significant role in achieving scientific success. Papers from new authors have, on average, less scientific impact.

Filippo [77] suggested that the number of comments (publications aiming to correct or criticize a paper) might influence the future impact of a paper. They observed that high publicity, on average, indicates the chances that an article will be highly cited, despite criticism and controversy.

Another way known to boost a career and a piece of work is collaborating. It has been shown that authors collaborate more often and in larger teams than before [34]. Moreover, the results of larger teams usually have a higher impact than those of small research groups [34, 45]. This shows the power of collaborating and how it can impact the performance of a product.

2.7 Dynamics of rankings

Rankings are a way of presenting hierarchies in elements. They are a popular approach to describing prestige and popularity, e.g., in sports, academia, film, or music. Scientists

have been studying the statistical properties of rankings for many years [67]. Others tried to find the relationship between talent and position in hierarchy [75]. Another popular area of research in this field is the dynamics of rankings [7, 48].

One of the essential pieces of research on rankings in recent years was conducted by Iñiguez et al. [48]. They investigated the dynamics of charts for different systems. They provided evidence for temporal patterns of rank dynamics and divided charts into two types - open and closed. Open rankings have stable top positions (meaning that the songs ranked at the top only changed their positions rarely), while closed ones also have stable bottoms. Moreover, they discovered two types of movements that shape the charts - abrupt and smooth/diffusive (small position changes around the initial rank). Their mathematical model produces many essential features of real-world rankings.

Pósfai et al. [75] focused on the correlation between talent and position in the ranking. They showed that while more talent does not necessarily mean better rank, a sufficient talent difference is needed to overtake someone in the chart. In the paper, an artificial Bonabeau hierarchy model was used, which determines the rank of an element by its ability to defeat others in pairwise competitions [75]. They found that the only way for elements with smaller talent to overtake those with far greater talent is for the better element to be removed (due to ranking rules). However, for smaller differences, talent plays a more substantial role.

2.8 Summary

Judging from most of the articles on creative careers, predicting the song's performance and whether it will make a hit still remains in active development. Recent models have been able to successfully predict only specific examples and given a large amount of historical data. There are still no methods to forecast success real-time and using real-world setups. The music industry highly depends on social aspects, which often lead to skewness in the distribution of attention. Moreover, the digital era has shaken up the music world by allowing more and more new artists to be heard.

In my thesis, I want to fill the gap between predicting success and the research on rankings. I focus on understanding the change in dynamics of pop music charts over time and how it impacts the difficulty of creating a hit song.

3 Dataset

The data on the pop music charts was taken from *The Billboard Hot 100* music charts [17]. The website contains information on the top 100 most popular songs each week, starting 4th August 1958. To acquire archival charts, a date in the format "yyyy-mm-dd" can be appended to the link[1]. The date indicates the first day of the week of the desired ranking. The charts are updated each Tuesday morning, except for holidays, when they are released on Wednesdays [16]. I show an example chart from Billboard Hot 100 in Figure 3.1.

I chose to fetch data from Billboard Hot 100, as it was one of the few charts allowing users to browse through historical data. It also contained all the information necessary to investigate the chart dynamics: position, last week's position, and total weeks on the chart. Moreover, Billboard's website structure has already been investigated by other researchers. A lot of examples of scraping this site can be found on the Internet [86]. There are a few user-made solutions available online for extracting weekly charts from Billboard, but to have better control over the data and its format, I wrote my own. I explain the technique in section 4.3. I also considered and scraped other websites that I did not end up using in this project, such as `Hitlisten.NU` [43], `Official Charts` [69], and `Acharts` [96].

After choosing the most appropriate website, I scraped Billboard's charts under Research License using Python's `Scrapy` package. The process and the spider I created are explained in section 4.3. In the end, I retrieved 3362 weeks of charts, starting 4th August 1958.

One of the main difficulties in extracting the data was finding the desired values in the chart's rows. The names of the HTML classes were not valuable in navigating over the songs' statistics. Therefore, using the trial and error method, I took all elements in a row and calculated the ordinal numbers for values like position, etc. Luckily, those numbers remained accurate for different pages, but the scraper would no longer produce valid results if the website's structure ever changed. I show an example of the website's HTML code containing data on one ranking entry in Figure 3.2.

Another thing I spotted when running the scraping was that some of the weeks' charts were incomplete. For example, 29th November 1979 misses entry at position 35[2]. I did not find any explanation on why some of the standings were missing. However, it was only the case for 13 weeks in 1979, and only one spot from each chart was missing. As the impact on quality was insignificant, I included these years in the data anyway. I show an example of missing data in a chart in Figure 3.3.

The successfully scraped rows were saved to a CSV file in a comma-separated format. The first row contained the column names. The final dataset consisted of 333887 rows[3] and ten columns:

- first_day_of_the_week – the first day of the week's ranking,
- artist – name(s) of the artist(s) performing the song,

[1] For example, 1958-08-04, making the link: https://www.billboard.com/charts/hot-100/1958-08-04

[2] https://www.billboard.com/charts/hot-100/1976-11-29/

[3] 3326 weeks with 100 entries + 13 weeks with 99 entries.

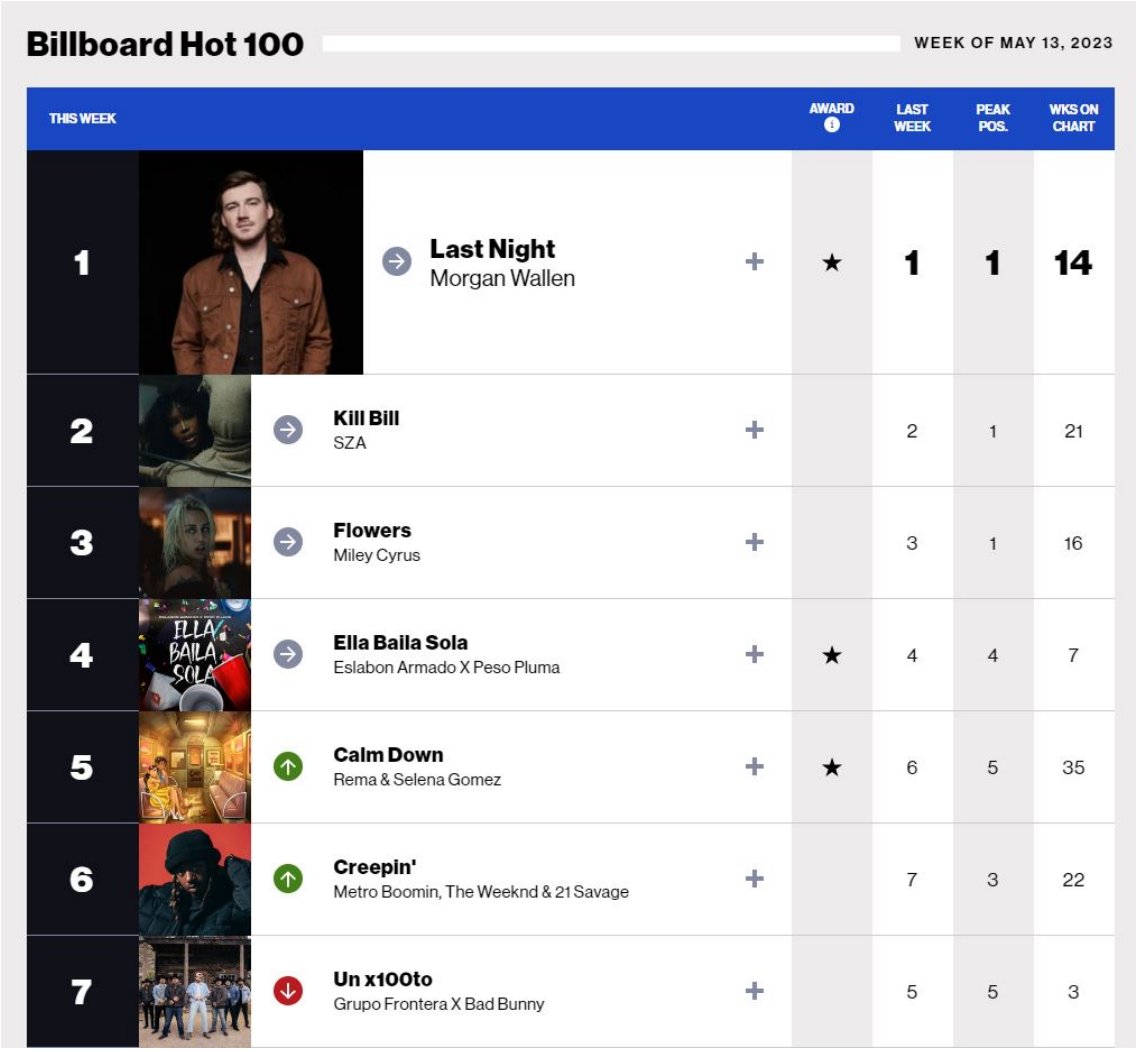


Figure 3.1: **Screenshot of the Billboard Hot 100 chart from the week of 13th May 2023.**
The picture shows a fragment of a Billboard Hot 100 chart. The chart contains information on this week's position, last week's position, peak position, and weeks on the chart. The arrows indicate the position change compared to the previous week's. The award column is Billboard's way of pointing out gains in the performance of particular songs. (Screenshot taken from the Billboard Hot 100 website)

- song_name – the name of the song,
- position – position of the song in the week,
- last_week_position – last week's position of the song ("–" if the song just entered the chart),
- peak_position – the highest position reached by the song until the week,
- weeks_on_chart – the total number of weeks a song has spent on the chart until the week.

Before exploring the dataset's features, I had to modify two columns with incorrect types. First, I converted the column "first_day_of_the_week" to `Pandas` date format. It allowed me

```
<ul class="lrv-a-unstyle-list lrv-u-flex lrv-u-height-100p lrv-u-flex-direction-col
umn@mobile-max"> flex == $0
  <li class="o-chart-results-list__item // lrv-u-flex-grow-1 lrv-u-flex lrv-u-flex-
  direction-column lrv-u-justify-content-center lrv-u-border-b-1 u-border-b-0@mobil
  e-max lrv-u-border-color-grey-light lrv-u-padding-l-050 lrv-u-padding-l-1@mobile-
  max">…</li> flex
  <li class="o-chart-results-list__item // u-width-66 u-width-30@mobile-max u-width
  -55@tablet-only lrv-u-flex lrv-u-flex-shrink-0 lrv-u-align-items-center lrv-u-jus
  tify-content-center lrv-u-border-b-1 u-border-b-0@mobile-max lrv-u-border-color-g
  rey-light lrv-u-order-100@mobile-max u-hidden@mobile-max">…</li> flex
  <li class="o-chart-results-list__item // a-chart-bg-color a-chart-color u-width-7
  2 u-width-55@mobile-max u-width-55@tablet-only lrv-u-flex lrv-u-flex-shrink-0 lrv
  -u-align-items-center lrv-u-justify-content-center lrv-u-background-color-grey-li
  ghtest lrv-u-border-b-1 u-border-b-0@mobile-max lrv-u-border-color-grey-light u-h
  idden@mobile-max"> </li> flex
  <li class="o-chart-results-list__item // a-chart-color u-width-72 u-width-55@mobi
  le-max u-width-55@tablet-only lrv-u-flex lrv-u-flex-shrink-0 lrv-u-align-items-ce
  nter lrv-u-justify-content-center lrv-u-border-b-1 u-border-b-0@mobile-max lrv-u-
  border-color-grey-light u-background-color-white-064@mobile-max u-hidden@mobile-m
  ax">…</li> flex
  <li class="o-chart-results-list__item // a-chart-bg-color a-chart-color u-width-7
  2 u-width-55@mobile-max u-width-55@tablet-only lrv-u-flex lrv-u-flex-shrink-0 lrv
  -u-align-items-center lrv-u-justify-content-center lrv-u-background-color-grey-li
  ghtest lrv-u-border-b-1 u-border-b-0@mobile-max lrv-u-border-color-grey-light u-h
  idden@mobile-max">…</li> flex
  <li class="o-chart-results-list__item // a-chart-color u-width-72 u-width-55@mobi
  le-max u-width-55@tablet-only lrv-u-flex lrv-u-flex-shrink-0 lrv-u-align-items-ce
  nter lrv-u-justify-content-center lrv-u-border-b-1 u-border-b-0@mobile-max lrv-u-
  border-color-grey-light u-background-color-white-064@mobile-max u-hidden@mobile-m
  ax">…</li> flex
  <li class="lrv-u-width-100p u-hidden@tablet">…</li>
</ul>
```

Figure 3.2: **Screenshot of the Billboard's Hot 100 website HTML code.**
The picture shows a Billboard Hot 100 website fragment, investigated using Chrome's developer tools. Each `li` class element contains, respectively, artist and song name, awards, last week's position, peak position, and weeks on the chart. There is no easy way of distinguishing which element corresponds to what information. (Screenshot taken from the Billboard Hot 100 website)

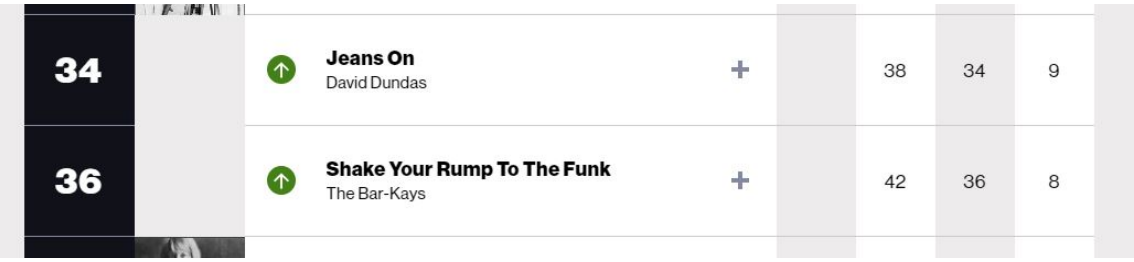


Figure 3.3: **Example of missing data for the chart from 29th November 1979.**
The picture shows a piece of the Billboard Hot 100 chart with a missing entry at position 35. (Screenshot taken from the Billboard Hot 100 website)

to create time-series plots using package-specific data manipulation methods. Another column, "last_week_position," contained both numerical and string values ("−" for the songs that entered the chart for the first time). I converted the string values to `NumPy's` NaNs (not a number) so that the whole column would be seen as numeric. Finally, I will only consider the total years from 1959 to 2022 in the calculations. The first few rows of the preprocessed CSV file can be found in Table A.1 of Appendix A. I calculated some basic statistics for the dataset and presented them in Table 3.1. The data contains information on weekly positions of 25026 songs from 10416 artists over 3339 weeks.

data	count
rows	333887
weeks	3339
artists	10416
songs	25026
years	64

Table 3.1: **Basic statistics for the Billboard Hot 100 dataset.**
The table shows some basic statistics for the Billboard Hot 100 scraped data. The measures are the number of distinct rows, weeks, artists, songs, and years. Most importantly, there are 333887 chart entries from 3339 weeks. The number of songs is over twice as high as the number of artists.

Various aggregation methods have been applied to the columns to fully utilize the data, including weeks on the chart before and after the peak and the position changes between two weeks. The new columns were created according to current needs to simplify plotting the data and further calculations.

Some entries in the "artist" column are collaborations/features between multiple artists. Without loss of generality, these rows are regarded as distinct artists unless specified otherwise.

The dataset prepared in the above ways was ready to be explored. The primary methods used throughout the thesis have been described in the chapter Methods.

3.1 Other charts

As part of the thesis, I also investigated whether the changes in dynamics in pop charts are reflected in other media. For that purpose, I chose books and podcasts. Another possible data choice included film but lacked accessible charts with historical entries. Moreover, the industry differs from the others in rankings, as movies currently at the top of IMDB's movie chart, e.g., "The Godfather," were released decades ago.

The process of obtaining the other datasets was the same as for pop music charts. The data was scraped using the `Scrapy` package but required writing new spiders to match different structures of the websites.

For podcasts, I used `iTunes Charts` [50]. It contains the daily top 100 iTunes podcasts in the United States. I show an example chart in Figure A.1 of Appendix A.

Since the podcats have not been around for as long as music, the data held rankings from only 24th November 2009. Another problem was that some of the first charts included only the top 10 podcasts, while later, it expanded to the top 100. Lastly, some of the daily charts

were missing, e.g., 4th December 2009[4]. However, I decided to use that source since, to my knowledge, no better alternatives existed. I present a part of the scraped data in Table A.2 of Appendix A.

The book charts were taken from `Hawes` [39]. The website archived years of *The New York Times Best Seller* rankings, starting from 1931. An example of a book chart was presented in Figure A.2 in Appendix A. The main challenge with these pages was that the charts were saved in a PDF format. It required an extra process of reading and parsing the files of inconsistent structures. Extracting the author and title was very time-consuming due to descriptions of the books included in the texts and irregular line lengths. Without loss of generality, I decided to save the entire text to one column, "author_title." The values were distinct for each author-title pair. The statistics, such as position, weeks on the chart, and last week's placement, were successfully saved to separate columns but also required careful extraction and parsing. The final rows have been presented in Table A.3 in Appendix A.

Obtaining the charts on other media was generally more challenging than for the music. There were hardly any websites with historical data available on the Internet. Most were inconsistent (e.g., every few days or weeks) or had significantly smaller list sizes than Billboard Hot 100 (e.g., top 10, top 20). All things considered, this reduced the chances of obtaining relevant results from comparing other sources of entertainment with pop music. However, I reproduced some of the graphs for books and podcasts and included the outcomes in Appendix B.

[4] https://www.itunescharts.net/us/charts/podcasts/2009/12/04

4 Methods

The crucial point of my thesis was to discover features, patterns, and anomalies in the pop music charts data that would help answer the hypothesis, “Is it getting harder to make a hit?”. The process of extracting the essential characteristics of the dataset is called exploratory data analysis (EDA). In this approach, I utilized several statistics and visualizations detailed in the following sections.

4.1 Exploratory data analysis

Exploratory data analysis (EDA) involves investigating the data to discover its main features. Often used are statistics and graphical visualizations but also machine learning. Some of the most fundamental steps in EDA include [57]:

1. Data collection and data preprocessing (data cleaning).
2. Feature extraction.
3. Investigating the relationships between different features.
4. Detecting abnormal observations.
5. Developing mathematical models explaining the data.

The above steps can be achieved by combining both graphical (e.g., histograms, heat maps) and non-graphical methods (e.g., median, variance). The results of the exploratory data analysis should be concluded to represent a meaningful explanation of the dataset.

EDA was an essential step in answering the hypotheses stated in my thesis. It was crucial to thoroughly explore the music charts dataset before suggesting possible reasons and explanations. All the techniques mentioned in the other sections were part of the EDA process.

4.2 Environment

The code was written entirely in Python 3.11 in the Jupyter Notebook integrated development environment (IDE). I chose Python, one of the most popular languages for data analysis, with easy syntax and many packages for data manipulation, plotting, and machine learning. Jupyter Notebook IDE was chosen as it provides an interactive and easy-to-use interface. Moreover, it allowed me to instantly show the output in the cell and store it in the same file as the code. It also allows adding cells with texts for simple note-keeping. For package management, I created a virtual environment and downloaded the following libraries:

- `NumPy` [68] - for working with arrays and various mathematical functions. I used this library mainly for converting the dataset to matrix and `Pandas` columns to `NumPy` arrays. Converting the complex data frames to matrices/arrays often optimized the performance or was necessary for machine learning algorithms.
- `Pandas` [72] - for data manipulation and analysis. It was the main library for reading and performing operations on the dataset due to its easy syntax and a vast range of predefined methods.

- `Matplotlib` [63] - for creating static plots. It was used to create the basic plots in the thesis. Plotting using `Matplotlib` was often faster than `Pandas` or `Seaborn` and was preferred for complex graphs.
- `Seaborn` [89] - for additional statistical functions and advanced visualizations. I mainly used it for plotting heat maps and contour maps.
- `Scikit-learn` [84] - tools for performing machine learning in Python. I used it for applying clustering and dimension reduction.
- `Scipy` [85] - a library for scientific computing in Python. In my thesis used for calculating the skewness of distributions.
- `Scrapy` [88] - a library for extracting data from the Web. I used it to gather the data for my research.
- `PyPDF2` [76] - for operations on PDF files. I used it for parsing PDFs with book rankings.
- `Tqdm` [95] - for adding progress bars to operations. I used this package to track the progress of longer computations. It also provides useful tools for keeping track of processed rows in `Pandas`' data frames.

Lastly, I used Git for version control. The code of my study is available on Github[1].

4.3 Scraping

Web scraping is a process of extracting data from websites. The process mainly involves downloading the pages (HTTP responses) and saving parsed HTML content. Scraping can be done manually or with the help of an automated process [103]. I used the Python `Scrapy` package in my thesis to conduct automated web scraping. The library provides an easy-to-use interface for web crawling. All the user has to do is define a "spider." A spider class describes which pages to scrape and how to extract data [87]. The code of my spider can be found in Listing C.1 in Appendix C.

The spider in my thesis would crawl over the pages from 4th August 1958 to 2nd January 2023, advancing weekly. For each date, it would download the contents of the Billboard page and find the container with the actual chart. Then, from each row in the charts container, it would extract the following values:

- artist name,
- song's title,
- this week's position,
- last week's position,
- weeks on chart.

In the end, the data was saved into a CSV file. Gathering the data from the Billboard charts took around 20 minutes. I created analogous spiders for iTunes podcasts and The New York Times Best Sellers rankings.

Creating a web crawler gave me a significant advantage in controlling the format and type of saved data.

[1] https://github.com/martalech/pop_charts_dynamics

4.4 Data preprocessing

Data preprocessing is a set of actions to increase the quality and readability of data. It is a crucial step before running machine learning algorithms and data analysis. The real-world data is often noisy, inconsistent, or incomplete [13]. Poor data quality can affect the results of machine learning models and data analysis. There are different ways to clean data: removing missing values, smoothing, filling in missing/incorrect values, and many more.

For the analysis, I removed the incomplete, non-full years data. One of the features of the Billboard dataset that required preprocessing was the column "last_week_position." Since the songs that entered the chart for the first time had the value "$-$," I had to convert these values to `NumPy` NaNs to keep the column in numerical format. Other changes included formatting the string column "first_day_of_the_week" to a date format.

4.5 Data segmentation

Data segmentation is dividing a large dataset into smaller groups of similar features. By dividing the data into smaller, more specific subsets, it becomes easier to identify trends that may not be apparent when analyzing the data as a whole. Segmentation helps answer questions about the dataset and find dependencies between different groups.

In my thesis, I divided the dataset of weekly charts into decades (from the 1960s to the 2010s) to aggregate the features and show their changes over time. I used decades in my analysis, as they are often used when discussing and comparing trends from different periods. They are also widely used to describe the evolution of music [66].

Another use of segmentation was clustering the songs' trajectories based on their features. It helped analyze the different types of flows on the charts and their changes over time. I describe the process of clustering in later sections.

4.6 Probability theory and statistics

I used various techniques from probability theory and statistics during the research process. I described the most crucial ones in the following subsections.

4.6.1 Probability distribution similarities

Distributional similarities are used to measure the divergence of two distributions. One of the most popular methods is the Kullback-Leibler (KL) divergence, defined as:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}, \tag{4.1}$$

where P and Q are compared distributions [53]. The measure can be explained as the information loss of P when approximating with distribution Q. The properties of the KL divergence are non-negativity, convexity, and additivity. However, the function is not symmetrical, so it cannot be considered a metric. This and other divergence methods are often used in machine learning or time-series analysis to calculate randomness and minimize error.

I used KL divergence in the thesis to measure similarities between position change distributions. The purpose of that was to compare different years and find patterns. Other methods considered included the Bchachattarya distance, but it was rejected, as it focuses on the amount of overlap between the distributions rather than closeness [14].

4.6.2 Probability distribution skewness

Skewness in probability distribution is a tool used to measure its asymmetry. One way to assess the centrality is by visually inspecting the mode, median, and mean. The possible configurations are presented in Figure 4.1.

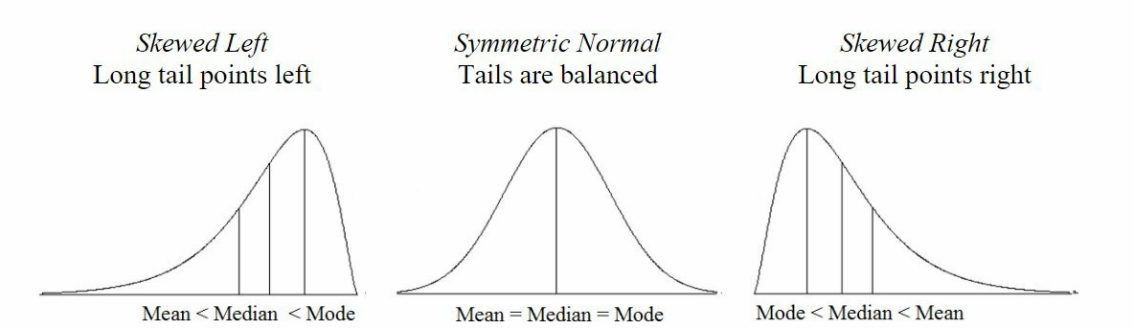


Figure 4.1: **Different types of skewness.**
Sketches show the general position of mean, median, and mode in a population. The relationship between these statistics can be used to classify a distribution's skewness. The possible types of skewness are left (negative), normal (symmetric), and right (positive). (Image taken from Doane et al. [32])

In statistics, there are a few ways to measure skewness. One of them is the Fisher-Pearson coefficient of skewness, which is defined as the fraction of the second and third moments around the mean:

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3/n}{(\sum_{i=1}^{n}(x_i - \bar{x})^2/n)^{3/2}}, \tag{4.2}$$

where $\bar{x}$ is the sample mean, and n is the number of values [32].

Simpler measures of skewness include Pearson's first skewness coefficient (mode skewness) and Pearson's second skewness coefficient (median skewness), defined in respectively Equation 4.3 and Equation 4.4.

$$Sk_1 = \frac{\bar{x} - m}{\sigma}, \tag{4.3}$$

$$Sk_2 = \frac{3(\bar{x} - m)}{\sigma}, \tag{4.4}$$

where m is sample mode and σ is the standard deviation. However, these coefficients are not as widely referenced in statistical textbooks as Fisher-Pearson's [32].

Skewness measures were used to investigate music chart dynamics to compare the position change distribution over the years. I focused on the results from the Fisher-Pearson skewness but included Pearson's first and second coefficients change in Figure D.1 in Appendix D.

4.6.3 Cumulative distribution function

Let X be a random variable. The cumulative distribution function (CDF) is defined as [81]:

$$F_X(x) = P(X \leq x), \tag{4.5}$$

so it is the probability that X will take a value less than or equal to x. The function is non-decreasing and takes values from the range $(0, 1)$.

To calculate the opposite, that is, the probability that X will take a value greater than x, a complementary cumulative distribution function (CCDF) can be used instead:

$$\bar{F}_X(x) = P(X > x) = 1 - F_X(x). \tag{4.6}$$

In statistical analysis, CDFs and CCDFs are used for comparing distributions, calculating probabilities, and finding frequencies of given values.

I used complementary cumulative distribution functions for comparing distributions of, e.g., position changes and start/end positions.

4.7 Distance measures

Distance measures are widely used in pattern analysis, clustering, classification, etc. In the real world, they quantify how far the objects are from each other in physical space. However, distance can also be interpreted as an abstract notion, representing a similarity of two elements [24]. One of the most fundamental metrics is Euclidean, which calculates the straight-line distance between two points. In N dimensions, it is defined as [73]:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{4.7}$$

where x_i and y_i are the coordinates of x and y in the ith dimension.

Another popular measure is the Manhattan distance. For two points x and y, the distance is calculated along axes at right angles. In other words, it is a sum of absolute distances between the coordinates, that is [73]:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|. \tag{4.8}$$

To compare two strings, Hamming distance can be used. For two arrays of equal length, it is the number of positions at which the corresponding symbols differ [73]:

$$d(x, y) = \sum_{i=1}^{n} 1_{x_i \neq y_i}. \tag{4.9}$$

For my thesis, I used Manhattan and Hamming distances to compare the similarities between the trajectories of songs. These results were later used to investigate the nearness of songs in different decades. The Hamming distance was not a good measure in that case, as it was doubtful for two tunes to have the same positions week by week. However, I included those results in Appendix E.

4.8 Ranking measures

The article written by Iñiguez et al. [48] provides many valuable metrics for measuring the stability of ranking systems. By comparing these measures, I could show the differences in how songs moved on the Billboard Hot 100 over decades.

A ranking list is a set of N_0 elements ordered by decreasing scores at times $t = 0, ..., T-1$. The Billboard Hot 100 dataset has $N_0 = 100$ and $T = 3339$ (weeks). The songs are ordered from positions 1 to 100. The best-performing element in the chart has rank $R_t = 1$ and the worst-performing $R_t = 100$.

The paper defines two types of ordering systems in charts: open and closed. Assuming that N_t is the number of distinct elements that have ever been in the ranking up to (and including) time t, the ranking is closed if for all t [48]:

$$N_t = N_0 \tag{4.10}$$

and open if:

$$N_t > N_0. \tag{4.11}$$

In the Results chapter, I show that Billboard Hot 100 is a type of open ranking with diffusion-like nature. However, in the early decades of the chart's history, it presented some features of the closed systems. I list measures used in the analysis of the dynamics in the subsections below.

4.8.1 Rank turnover

Rank turnover is defined as:

$$o_t = \frac{N_t}{N_0}, \tag{4.12}$$

where N_t is the number of elements ever seen in the ranking until time t, relative to the size of the ranking list N_0. The function is monotonic and increasing. It indicates how fast new elements make it to the list [48]. The mean turnover rate is expressed as:

$$\dot{o} = \frac{o_{T-1} - o_0}{T-1}. \tag{4.13}$$

For closed rankings $\dot{o} = 0$, and for the most open rankings, $\dot{o} = 1$.

4.8.2 Rank flux

Rank flux, defined as F_t, is the probability that an element enters or leaves the ranking at time t. The mean flux is calculated as follows:

$$F = \langle F_t \rangle = \frac{1}{T} \sum_t F_t. \tag{4.14}$$

Flux is roughly constant over time for most rankings, with occasional deviations and fluctuation. For closed charts, the enter/leave probability is 0.

4.8.3 Lévy jump

Lévy walk is a random walk in which the step lengths ℓ follow a power-law distribution [2]:

$$P(\ell) \sim \ell^{-\mu}, \tag{4.15}$$

where $\mu \in (1, 3]$ is a power-law exponent. They are mostly observed in natural movements of, e.g., insects and mammals.

In the context of movements on the rankings, the Lévy jump happens when a random element at rank R is placed to rank X following a Lévy flight regime. These moves can be sudden and abrupt due to the heavy-tailed nature of the distribution [48]. The most available rankings are dominated by long Lévy jumps.

4.8.4 Diffusion jump

Diffusion is a phenomenon of elements evolving randomly and continuously in time under certain conditions [47]. It occurs in many physical, biological, economic, and social processes.

In the rankings, a diffusion jump results from the rank change of another element (by Lévy jump). They are more local and smooth movements around the initial positions. The article showed that most charts follow a diffusion regime [48].

4.8.5 Replacement

Replacement happens in the ranking system when an element is substituted with a new one from outside the system. The original object is removed from the list. This process leaves the other ranks untouched [48]. The authors did not observe replacement-oriented charts in real-world data, but they can be simulated.

4.8.6 Displacement probability

The authors of the article introduced displacement probability to capture the three regimes of the rankings dynamics: Lévy walks, diffusion, and replacement. The probability that element with rank r gets displaced to x after a time t is defined as:

$$P_{x,t} = e^{-\upsilon t}(L_t + D_{x,t}), \tag{4.16}$$

where:

$$L_t = (1–e^{-\tau t})/N \tag{4.17}$$

is the probability that an element gets selected until time t and jumps to any other rank and:

$$D_{x,t} = D(x,t)\Delta x \tag{4.18}$$

is the probability that the element in rank r gets displaced to rank x after time t due to Lévy walks of other elements [48]. The remaining free parameters, τ and υ, are probabilities for respectively Levy/diffusion-like moves and replacement.

The displacement probability was calculated for Billboard Hot 100 for the whole dataset and different decades. The metrics calculation code was taken from the author's Github[2].

4.9 Curve-based trajectories analysis

Using fixed-time descriptive statistics such as mean and median can often be misleading when analyzing curves. I show the example in Figure 4.2.

Juul et al. [54] introduce a solution to summarize curve ensembles. They have proposed using curve boxplots to visualize trajectory confidence intervals. The procedure for creating these plots is as follows:

1. Rank curves from more central to less central.
2. Plot the envelope containing the most central curves.

I took the code of the calculations from the author's Github[3] and modified it slightly. I inverted the y-axis (as in the context of rankings 1 is the highest, and 100 - the lowest) and allowed to specify the axis to plot onto.

In the context of my thesis, they were used for visualizing song trajectories. I created those additional plots to support the findings from the average and median curves.

4.10 Filtering

Filtering is a process of transforming one time series into another [46]. It is mainly used to handle noise in the data. One of the simplest ways of smoothing time series is by using a moving average. The method calculates the mean for groups of data points (windows). It is defined as:

$$z_t = \frac{1}{k+1}\sum_{j=0}^{k} x_{t-j}, \tag{4.19}$$

[2]https://github.com/iniguezg/Farranks
[3]https://github.com/jonassjuul/curvestat

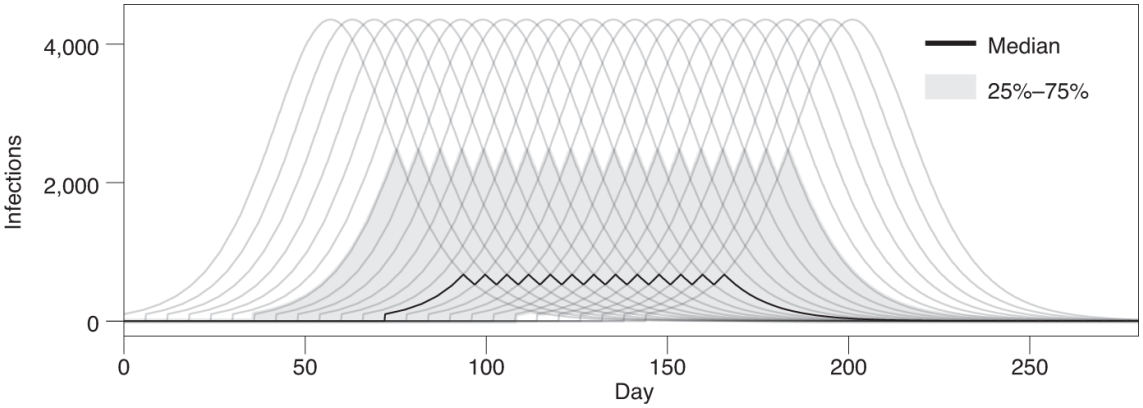


Figure 4.2: **Example of pitfalls of using fixed-time descriptive statistics to summarize curves.**
The illustration shows that using statistics, such as median, does not always capture the features of overlapping trajectories. The data presented in the plots are artificially created time series of possible infection outbreaks. The black lines indicate the median, and the grey area is drawn between the 25th and the 75th percentiles. The actual infection curves are plotted as grey lines. In the example, the infection peaks depicted by median and percentiles are underestimated compared to the data values. (Image taken from Juul et al. [54])

where k is the length of the window, x is a n-element time series, and $t = k+1, k+2, ..., n$.

In my thesis, I used a rolling average to smooth some plots, e.g., yearly lifetimes. It helped with exploring trends in the charts' time series.

4.11 Archetypal analysis

Archetypal analysis reveals similar patterns and features in the system [91]. There are many different approaches to discovering archetypes. These include decomposition, component analysis, and clustering [65].

Archetypal discovery played a crucial part in analyzing the changes in the Billboard Hot 100 charts over time. By discovering potential archetypes for song trajectories, I could show the differences in how songs moved over the decades. I used k-means clustering and non-negative matrix factorization to find the groups of similar curves.

4.12 Machine learning

Machine learning provides many algorithms that learn from data and make predictions [60]. The techniques can be divided into two types based on whether the data is labeled or not - respectively supervised and unsupervised. In the following subsections, I describe a few methods I used from the unsupervised learning group.

4.12.1 Clustering

Clustering is a process of grouping unlabeled samples. One of the simplest yet effective in many cases algorithms is k-means.

K-means clustering is a method that divides observations into k clusters based on their proximity to the cluster's mean (cluster centroid). The algorithm aims to choose centroids

that minimize the criterion [82]:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2), \tag{4.20}$$

where n is the number of samples, x_i is the ith sample, C is the set of clusters, and μ_j is the mean of the points in C_j.

I used the k-means algorithm to group the songs based on the properties of their trajectories (e.g., first position, last position, lifetime, etc.). The method successfully identified some underlying archetypes in the songs, shown in the Results chapter.

4.12.2 Dimensionality reduction

A process of transforming the data into lower space while maintaining its most important features is called dimensionality reduction. It is widely used in machine learning and cluster analysis. Moreover, it simplifies working with sparse matrices by making them easier to inspect.

One of the algorithms used is non-negative matrix factorization (NMF). The algorithm factorizes a non-negative matrix V into two non-negative matrices, W and H, with reduced dimensions [83]. It is used in feature extraction and can be interpreted that W contains attributes and H their importance. In my thesis, I used the NMF algorithm to find the archetypes of trajectories. It did not capture the crucial features of the songs' curves I was looking for, but I included the results in Figure E.8 in Appendix E.

5 Results

The main focus of my thesis is to answer the hypothesis: “Is it getting harder to make a hit?”. While measuring the difficulty of making a hit is not simple, there are ways to explore it with data analysis. The main factors I have investigated are:

1. The lifetime of the songs.
2. Position change distribution.
3. Songs trajectories.
4. Dynamics of ranking.
5. How new artists and hitmakers perform in the charts over time.

5.1 Lifetime of songs is increasing for the best-performing ones and decreasing for others

I first wanted to explore the duration of songs on the charts. It can be considered one of the success indicators, as hits tend to stay longer in the rankings. I plotted the yearly average and median of songs’ lifetime (maximum weeks on the charts), shown in Figure 5.1.

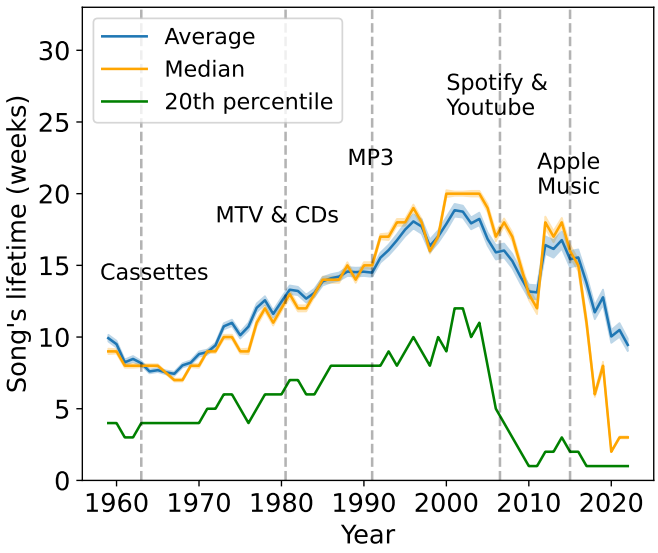
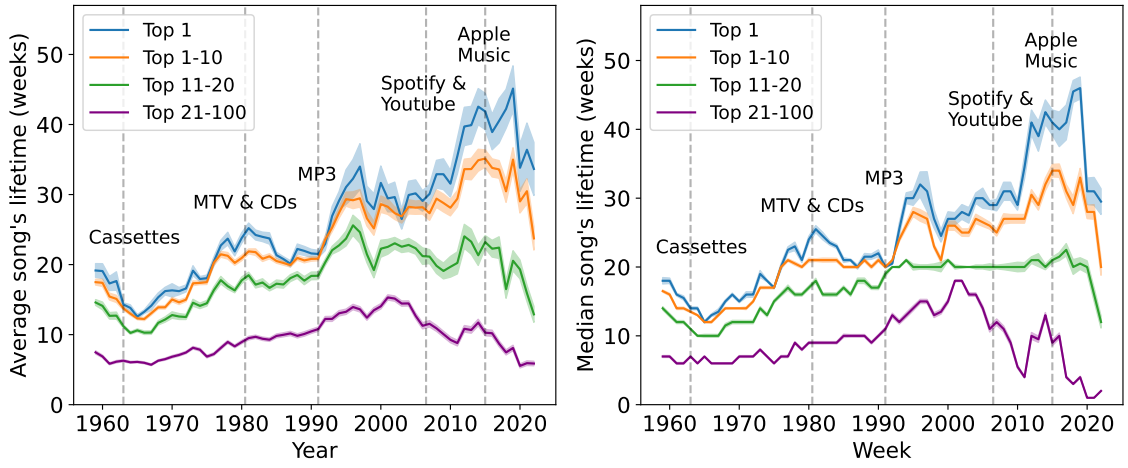


Figure 5.1: **Lifetime of songs on the chart.**
The plot shows the average (blue), median (orange), and the curve indicating the 20th percentile (green) of the lifespans of all songs over time. The shaded areas represent the standard errors. The dashed grey lines indicate some of music history’s most important discoveries/launches. Those include the invention of cassettes, CDs, and MP3s and the launches of MTV, Spotify, Youtube, and Apple Music. Some dates, such as MTV & CDs, were joined and averaged as they were apart by only a few years. All charts have been going up until the 2000s and then started to drop. The median and the 20th percentile have fallen below 5, achieving the lowest values in the chart’s history.

It can be seen that since around 1967, there has been a nearly linear increase in the lifetime of songs in all three categories (average, median, and the 20th percentile). The last three decades show a decrease in the values, especially drastic for the 20th percentile. It is first visible at the beginning of the streaming era (Spotify & Youtube). The 20th percentile and median values since 2020 are the lowest in the chart's history. Compared to the first week of 1959, the average duration has risen to around 20 and dropped back to 10 in the last years.

Similar insights can be spotted in the median. Up until 2015, the values were nearly the same. However, a drop around 2015 has fallen to roughly three weeks in 2022. It is noteworthy that nowadays, around half of the songs on the chart have a considerably shorter lifespan, lasting only a few weeks, whereas, in the early decades, the average and median time on the charts was ten weeks. This means a decrease of more than two times compared to the average.

While the shorter lifetime means more variability, it also means that the songs have difficulty maintaining their positions in the charts. However, it does not necessarily indicate that all the pieces are struggling. To further investigate the changes in the maximum number of weeks on the charts, I looked at the yearly durations for songs that reached top 1, 1-10, 11-20, and 21-100. I present the plots in Figure 5.2.



(a) Average lifetime.

(b) Median lifetime.

Figure 5.2: **Average and median song's duration for different top position ranges.** The figures show average (a) and median (b) lifetimes for songs that reach the top 1 (blue), top 10 (orange), green (top 11-20), and the rest (purple). The shaded areas represent standard errors. The particular ranges have been chosen for readability and to focus on the best-performing songs. The biggest increase in the lifetime occurred for the top 1. For other ranges, it either starts to stabilize or decreases in the later decades. The top 1 has, on average, four times higher lifetime than the bottom 80 (21-100).

The subfigures give a better overview of the actual changes in the average lifetimes of the songs. The average lifespan among the top 10 has increased to approximately 35 weeks in 2015 but has since decreased to slightly above 20. However, the top 1 remains higher than before the 1990s by ten weeks. The same can be spotted in the median. The rise in the duration for the top 11-20 has slowed down around the 2000s. The median stabilized at 20 weeks around the 1990s. There has also been growth for the rest of the songs (top 21-100), but it quickly declined around the 2000s. In the late 2000s, the surge

in popularity of streaming services may have contributed to a shorter lifespan for songs and widened the gap between those at the top and bottom of the charts. I elaborate more on this suggestion in section 6.2.

The statistics plotted in Figures 5.2a and 5.2b provide only some insights into what is actually happening in the dataset. Measures like mean and median are often misleading and highly reductive. I plotted the top 5 songs with the most extended lifetimes to show some of the extreme values appearing in the charts. I demonstrate the data in Table 5.1. From the table, it is evident that some songs stay on the charts for over 90 weeks. The average and the median cannot capture these outliers. Hence, I used the violin plots to show changes in the weeks on chart distribution across different decades. I show this plot in Figure 5.3.

Artist	Song name	Weeks on the chart
Glass Animals	Heat Waves	91
The Weeknd	Blinding Lights	90
Imagine Dragons	Radioactive	87
AWOLNATION	Sail	79
Dua Lipa	Levitating	77

Table 5.1: **Maximum weeks on the chart for the top 5 longest-staying songs.**
The best-performing song in terms of lifetime is “Heat Waves” by Glass Animals, released in 2020. All of the top 5 songs were released after 2010.

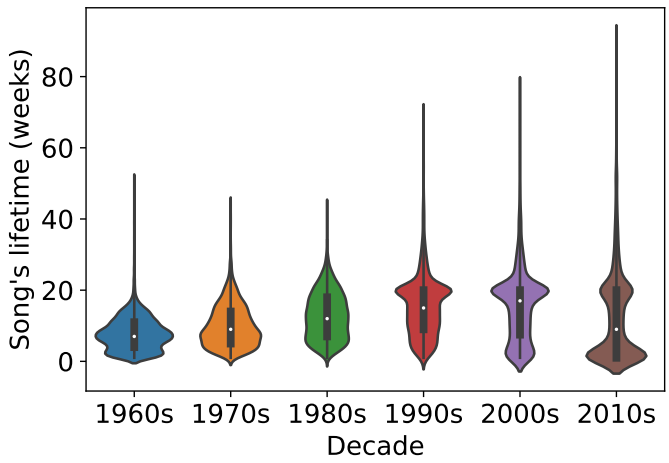


Figure 5.3: **Lifetime of songs on the charts.**
The lifetime distributions are displayed as violin plots for different decades. The white dot represents the median, and the dark rectangle shows the interquartile range (the middle 50% of lifetimes). The colored areas represent the densities of the distributions. As time progresses, the plots become thinner and longer, with the exception of 20-week and 1-week durations on the chart. The long skinny ends of the plots capture the outliers of the data. In the last decades, they have reached the longest lifetimes (four times higher than the upper bound of the interquartile range).

As seen from Figure 5.3, the distributions in the first three decades of the charts are extensive, especially around the first weeks, and very few songs achieve more than 20 weeks. However, the density spreads to the higher regions over the following decades.

The widest part is now around 5 and 20 weeks, but the outlier values are over twice as high as for the first decades. This indicates that songs have a greater chance of having a longer lifespan but also the possibility of disappearing quickly after just a few weeks.

The median, represented by a white dot, shows the same trend as in the previous plot (Figure 5.2). It has been increasing till the 2000s and dropped in the 2010s. On the other hand, the interquartile range (grey rectangle) has been getting wider. It represents the "middle 50%", which can be understood as an increase in the variability of the data.

Lastly, I decided to look at the distribution of songs' lifetimes per decade for different top-position ranges. I show the results in Figure 5.4.

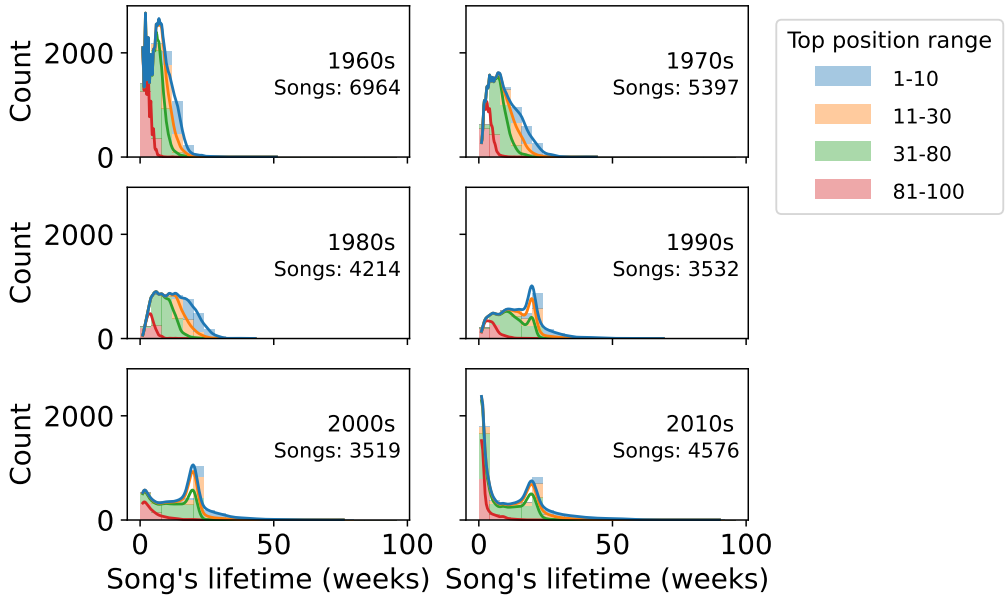


Figure 5.4: **Distribution of songs' lifetimes over decades.**
The data was organized into 33 bins for the top positions ranges: 1-10 (blue), 11-30 (orange), 31-80 (green), and the rest (red). The darker lines indicate the KDE approximations. Some parts of the curves are not visible, as they overlap, due to very small proportions. The text shows the number of songs from each decade. Most songs were released in the 1960s, and the least in the 2000s.

The histograms in Figure 5.4 show the relationship between the top position and the song's lifetime. The densities in Figure 5.3 resemble the presented distributions. The histograms show the same anomalies in the 2010s for the first and 20 weeks. By investigating the rules of the Billboard Hot 100, I found that since the 1990s, they incorporated a new practice for "recurrent" songs: "Descending songs are removed from the Billboard Hot 100 and Radio Songs simultaneously after 20 weeks on the Billboard Hot 100 and if ranking below No. 50, or after 52 weeks if below No. 25" [15]. This citation explains the high number of songs that lasted precisely 20 weeks. Nevertheless, further exploration is required to understand the drastic increase in the appearance of 1-week tunes and the longer durations of the outliers. Though it is out of the scope of this study, I comment on this finding in section 6.4.

In the early decades, most songs that would leave after a few weeks had not reached the top 80. In the 2010s, the number of the top 31-80 songs that disappeared quickly was as high as the bottom 80. There is a visible correlation between the top position and the

lifetime. The first bin was dominated by red (81-100), then the next one by green (31-80), etc. This tendency is not as clear in recent years as in the first decades. Some people suggest that as of late, the music industry has become increasingly competitive, and it is more probable that songs that did not achieve better positions (e.g., top 30) will drop out after a few weeks [33].

Moreover, the top 10 songs are more likely to achieve very long lifetimes, such as more than 20 weeks. Based on the lower counts, the top-performing elements appear to remain in the same position or shift slightly every week, thus occupying the top spots on the charts for extended periods.

After analyzing the data, it is evident that songs with worse peak positions have had a decreasing trend in lifespan in the last decades. Contrarily, the best-performing songs seem to stay longer and longer on the charts. Based on these insights, it appears that the competition in the music industry has intensified, making it more difficult for songs to achieve chart success. In the next steps of the research, I will try to identify the movements in the charts and the groups of artists that occupy the charts for the longest.

5.2 Position change distribution gets more balanced

In the preceding section, I presented the alterations in the lifespan of songs that could signify a rise in the challenge of creating a chart-topping hit. I have shown that some songs generally stay very long on the charts while others disappear rapidly. Still, little is known about the actual movements on the charts. That is why, as my next step, I will look at the position changes in the rankings.

First, I plotted the distributions of position differences over decades. I show the results in Figure 5.5. At first glance, the median shifted to the right for the first three decades. Since the 1990s, it has stabilized at 0. Another difference can be seen in the center area, that is position changes from the range $< -25, 25 >$. For the late decades, the distributions seem more symmetric. Nowadays, songs seem to move more smoothly on the charts, gradually increasing and decreasing their positions.

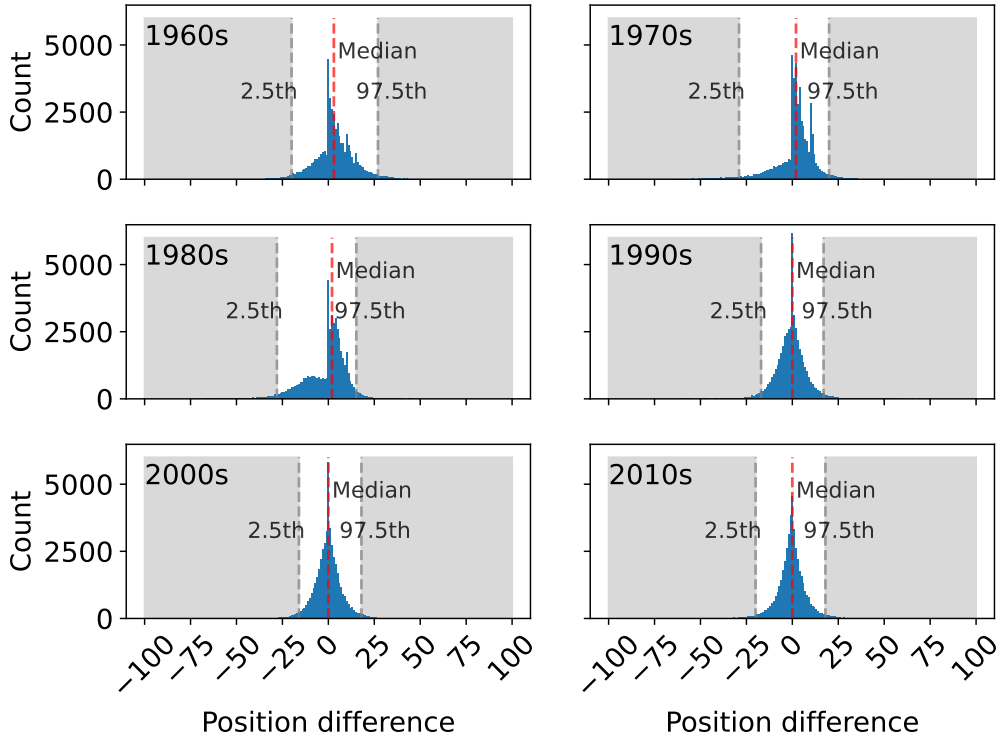
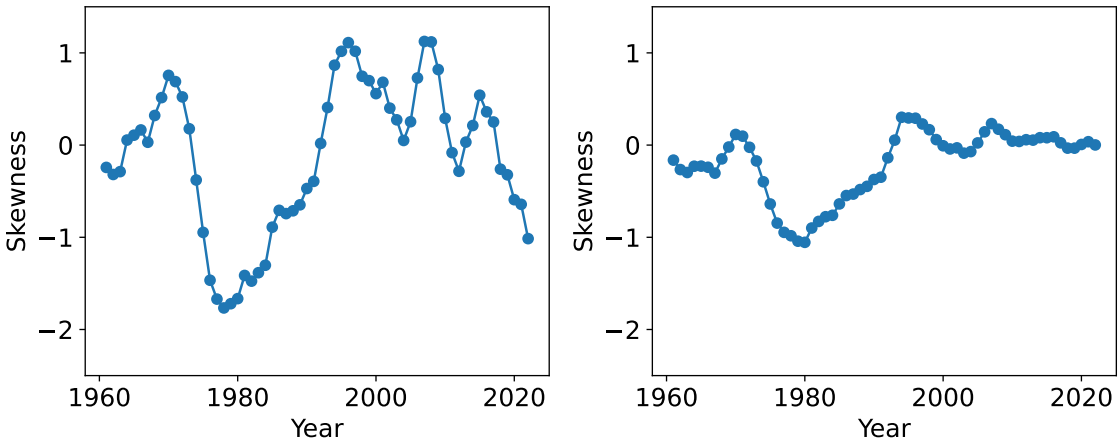


Figure 5.5: **Distribution of weekly position changes over decades.**
Blue bins represent the distributions. The red dashed lines represent the median of the samples. The grey dashed lines represent the samples' 2.5th and 97.5th percentiles, respectively. The shaded grey areas show the elements smaller than 2.5% of the data and greater than 97.5% of the data. During the first three decades, the median skewed towards the right, but it ultimately shifted towards zero in later years. Moreover, the distributions got more symmetrical with time, while before they were concentrated on the right side.

Since it is difficult to guess the skewness from the plots of the distributions, I decided to calculate and plot the yearly asymmetry values for all years. I used the Fisher-Pearson coefficient on the entire data and the center, explained in subsection 4.6.2. I show the plots in Figure 5.6. There are significant differences between the values in different decades. While the centers have become more symmetrical, the overall skewness has become more negative in recent years. Distributions with negative skewness have a heavier right

tail, with more values concentrated on the right side. This indicates that most songs maintain a stable climbing-up flow, with a few experiencing drastic jumps and drops. Streaming services like Youtube speed up the hit-making process, making popular songs jump quickly to the top [37]. However, some songs rapidly drop from the charts, probably due to the industry's growing rivalry [33].



(a) Skewness for the full data.

(b) Skewness for the center.

Figure 5.6: **Skewness of yearly position change distributions.**
The skewness was calculated for the whole dataset (a) and the center (for values from -25 to 25) (b). The plots have been smoothed with a rolling mean of 3 years. The overall skewness has been changing a lot throughout the years. There are fewer fluctuations in the center values. In the first half, the majority of them register values below 0, while in the second half, they record values close to 0.

Next, I calculated the distances between the distributions using Kullback–Leibler divergence measure (explained in subsection 4.6.1) to further analyze the changes between different years. I present the results in Figure 5.7. The heatmap reveals a few patterns in the distributions' similarities. First, there are darker squares along the main diagonal. It shows that groups of consecutive years resemble similar features in their distributions. Next, it can be seen that there is an increase in whitish squares for later years. It shows that there have been many rapid changes in the position change dynamics since the 1990s, overlapping with the rise of digitalization in the music industry (MP3, streaming).

I've created a kernel density estimation (KDE) of weekly position changes, displayed in Figure 5.8, to test the theory that modern songs tend to have consistent diffusive trajectories with occasional significant changes in position. This chart depicts the weekly fluctuations in position volumes. In the first three decades, spikes of higher positive position differences are indicated by darker vertical lines on the right side of the plot. Moreover, there is a notable area of negative values with small densities (light red). This implies that the songs did not experience significant ascents in the rankings but declined rapidly. This would mean that songs quickly disappeared from the charts after starting to fall. The values stabilize around 0 in the 1990s. Nowadays, most songs seem to diffuse around their initial positions, with fewer long jumps. However, due to the smoothing of the KDE plots, there is a possibility that many outliers were erased. Therefore, the contour map gives an overview of the dynamics change rather than the actual shifts.

To further investigate the movements on the charts and discover the types of music pieces

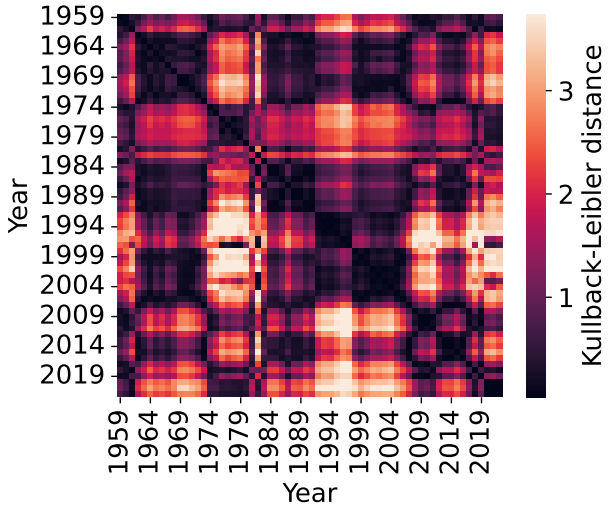


Figure 5.7: **Heatmap of similarities between yearly position change distributions.** The similarities were calculated as Kullback-Leibler divergences between distributions from different years. The colors indicate how well the distribution from one year (x-axis) is approximated by the distribution from another year (y-axis). The darkest colors exhibit great similarity, while the lightest do the contrary. The biggest distribution disparities can be spotted between the years in the last three decades (bottom right corner).

that behave in these particular ways, I plotted the aggregated distributions before and after 1990 for different top-position ranges. I display the results in Figure 5.9. In the early decades, the songs would jump up in the rankings and rarely but rapidly down, quickly disappearing from the charts. This would confirm the findings about the first decades from Figure 5.8. Since the 1990s, the distributions got more balanced, but there are more anomalies in the long jumps. The outliers are significantly visible after 1990, showing that successful songs in better positions are more likely to jump long distances in the rankings. Moreover, it can be seen that the worse the songs perform, the fewer position changes occur in their trajectories. As discovered in the previous section, worse-position songs have a significantly shorter lifetime of only 1-3 weeks. This is a natural outcome of that finding.

Concluding all the results from this section, the following changes in the dynamics have been observed:

1. Before, the songs would climb up, but after starting to decline, they would disappear rapidly.

2. Since the last three decades, the flows have become more balanced, with a few songs jumping long distances - up for the best-performing songs and down for the worst-performing ones.

Two explanations for the behavior changes in the charts over time could be the strengthening of the “long-tail,” and the “winner-take-all” effects [28, 70]. This might suggest that the music industry has expanded to accommodate more “niche” artists, enabling an increasing number of them to enter the charts [28] and therefore growing the competition and the number of position changes. This is a natural cause of the ease of online music access. On the other hand, the most famous artists still maintain great performances on the charts. The dynamics seem to have changed to better accommodate the superstars

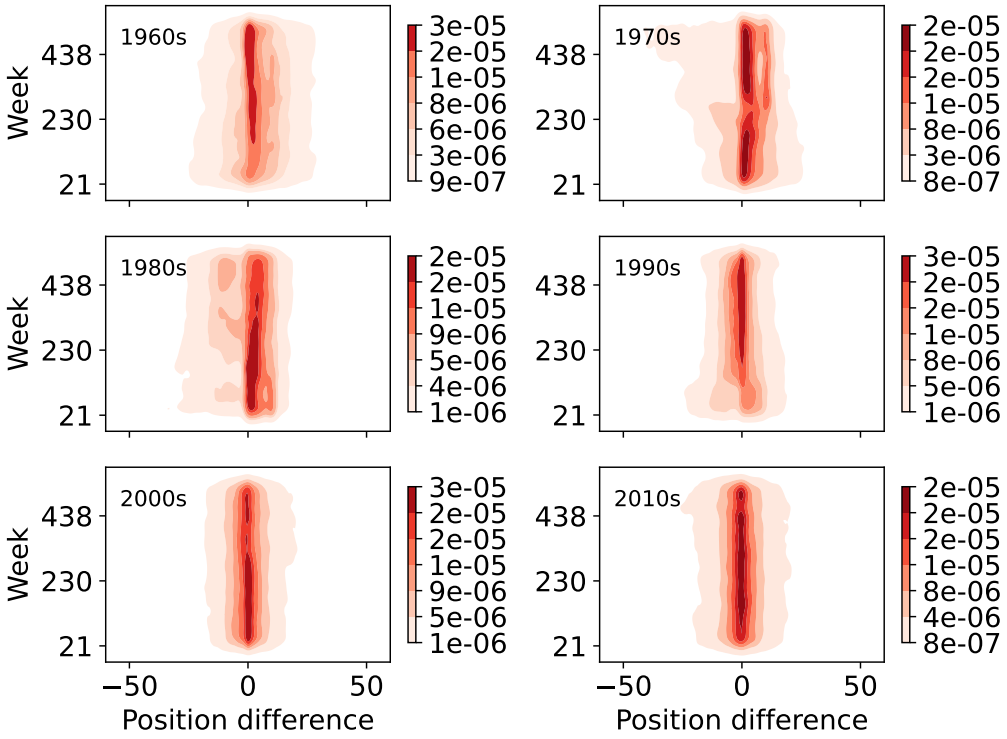


Figure 5.8: **Contour map of weekly position changes over decades.**
The contour map was created for weekly position change distribution and smoothed using KDE. The scale of colors ranges from light red to black, signifying the lowest to highest densities. The contour maps in the first three decades are wider, showing more variety in the moves. Also, the plots were more stretched to the left side, showing that many songs would fall sharply in the rankings. Nowadays, the probability densities are more symmetric. However, the smoothed plot loses outliers that are present in the data.

[70].

The results from this section suggest increasing competition and strengthening of top artists' positions. It might mean that in the last few decades, the industry has become more demanding in terms of creating hit songs. I elaborate on these thoughts in section 6.1.

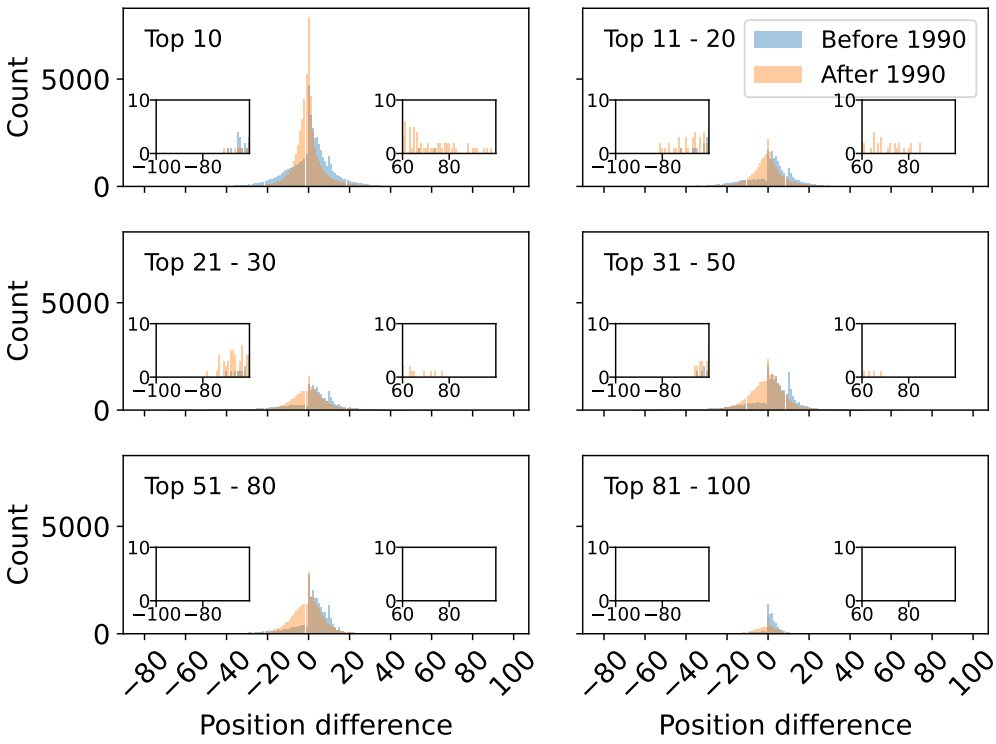


Figure 5.9: **Distribution of weekly position changes over different top-position ranges.**
The distributions were calculated for periods before (blue) and after (orange) 1990. The bottom nested plots zoom in on the outliers, the position differences from -100 to -60 (left) and 60 to 100 (right). The y-axis limits of the inner plots are from 0 to 10. The distribution centers become more symmetric after 1990. Moreover, since the 1990s, there has been an increasing number of outliers - positive for the top positions and negative for the worse positions. This shows that even though the movements became more evenly distributed, there were songs that jumped many positions in the charts.

5.3 Song trajectories get broader and start higher for best-performing pieces

After investigating the changes in the position differences of songs, I decided to look at their trajectories. A trajectory of a track x is defined as trajectory$(\mathsf{x}) = [\mathsf{x}_1, \mathsf{x}_2, ..., \mathsf{x}_n]$ where n is the number of song's weeks on the chart and x_i are positions at week i. The indexes from $1, ..., n$ are discrete observations, where x_1 is the first week of a tune on the chart, x_2 is the second, etc. To show normalized trajectories for sets of songs in decade d I defined trajectory$(\mathsf{x}, \mathsf{d}) = [\mathsf{x}_1, \mathsf{x}_2, ..., \mathsf{x}_n, 101, ..., 101]$. This ensures that all arrays from decade d have equal lengths corresponding to the number of weeks in that decade.

5.3.1 Trajectories analysis

The primary purpose of trajectories in my research is to visualize the individual songs' position differences throughout their lifetime. I created average (normalized) curves as follows:

1. Select songs whose overall top position is in the specified range from currently analyzed decades (e.g., songs that topped at 1-10 positions from the 1990s).
2. Select all songs' positions from the first week of the analyzed decade.
3. Calculate the mean from values in 2.
4. Repeat steps 2-3. for all weeks in the decade.
5. Repeat steps 1-4. for all decades.

I present the plots in Figure 5.10.

The first noticeable thing in Figure 5.10 is that the curves' peaks get smaller and broader over time. It confirms some of my previous findings that earlier, the songs disappeared quickly after reaching their peaks, while now they seem to linger on for longer, slowly decreasing position with time.

The flattening happens very quickly for the worst placements. The average trajectories flatten at 101 after a few weeks. However, it can be seen that all the averaged curves in the 2010s start higher than in the previous decades. Therefore, another insight emerges - on average songs seem to begin their journeys on charts at better positions. The drop in the next few weeks confirms that many worse-performing pieces disappear quickly after entering the chart.

Averaging the curves often loses the important features [54] of the individual trajectories, e.g., peak heights. This is visible for the top 10 in the 2010s, where the averaged curve does not peak in the top 10. I decided to plot the statistics with `curvestat`[1] for the top 1-10 and bottom 50-100 songs over decades to understand the changes in the actual paths. I explain the algorithm in section 4.9 and show the results in Figure 5.11. I also plotted average and median trajectories for these ranges in Figure E.1 and Figure E.2 in Appendix E.

In figure 5.11a, we can see that the trajectories in the first three decades were very alike, ending sharply a few weeks after reaching the top. Since the 1990s, they have started to diverge. The plots have become broader, indicating that a top-10 song can endure for as little as one week or as long as 50 weeks. Moreover, it can be seen that some curves do start at better positions, as the top left corner of the plots is now inside the 95% confidence interval. On the other hand, in the first three decades, the 95% of most central

[1] https://github.com/jonassjuul/curvestat

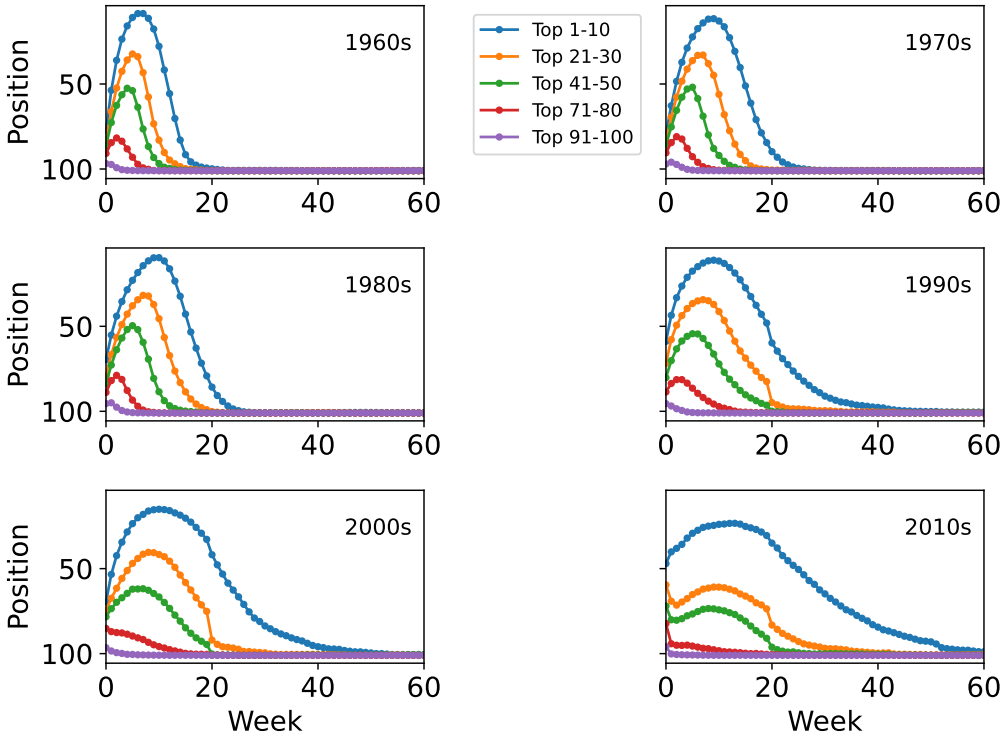
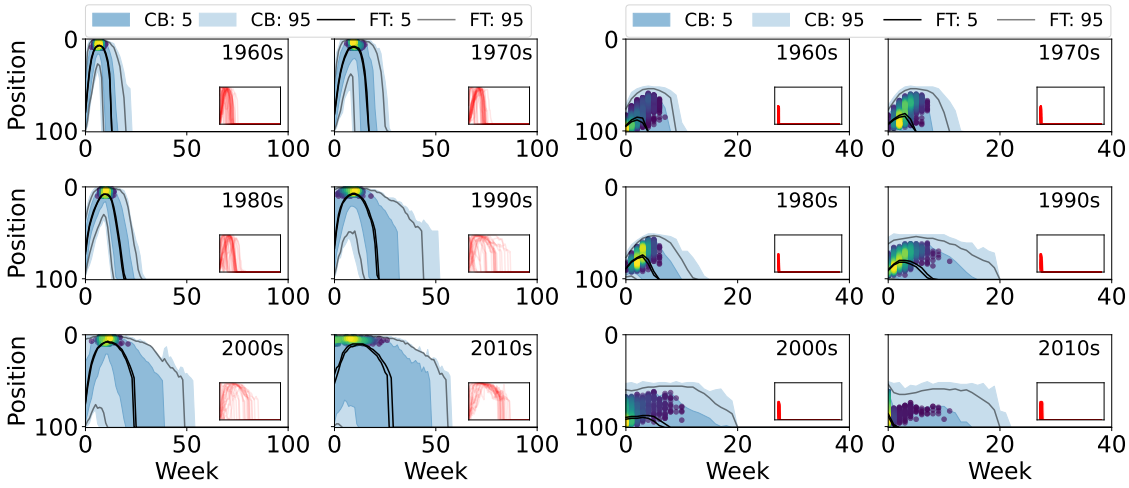


Figure 5.10: **Normalized trajectories of songs from top positions.**
The chosen top positions are top 10 (blue), 21-30 (orange), 41-50 (green), 71-80 (red), and 91-100 (purple). Only a few position ranges were chosen for better readability of the figure. The missing intervals lie between the visible ones (e.g., the 11-20 curve lies between the top 1-10 and the top 21-30). The songs that disappear from the charts have the rest of their trajectories filled with 101. As time passes, the average curves widen and start higher at week zero. There is also a visible drop in the 20 weeks, explained in the previous sections as the “recurrent” status.

songs have lasted at least ten weeks, while in the 2010s, some songs in this range left after 2-3 weeks. This shows a significant variation increase, as some previous plots have hinted. Specific tunes in the top 10 become massive hits and maintain their position on the charts for over a year, whereas others may risk being displaced within a few weeks. Some of these gigantic hits are the outliers, shown in the insets of Figure 5.11a.

The situation looks slightly different for songs that never reached the top 1-50, shown in Figure 5.11b. The variety of possible trajectories also increased. Nowadays, most songs peak in the first 1-2 weeks. Many of them also start at better positions compared to the first decades. Moreover, there is a tendency for the trajectories to gradually decrease after having reached the top. In the previous decades, songs seemed to follow a similar climbing flow as in the top 10. The peaks were more diffused, and the trajectories started at worse positions, slowly climbing to their maxima. The outliers show some of the songs that last for longer times or follow a different flow.

To compare the start and end positions of the songs, I plotted complementary cumulative distribution functions (CCDF), explained in subsection 4.6.3. I show the results in Figure 5.12. I also included the distributions in Figure E.3 in Appendix E. It can be seen from the start positions plot that the CCDF for the 2010s is lower than any other plots till around the position of 90. The gap is evident around position 50, showing that around 20% of



(a) Songs that reached top 1-10.

(b) Songs that never reached top 1-50.

Figure 5.11: **Curve-based statistics for songs' trajectories.**
The shaded areas indicate 5% most central (dark blue) and 95% most central (light blue) song trajectories of the songs that reached the top 10 (left) and bottom 50 (right). The dark lines are percentiles representing the min-max boundaries. For the 5% (black), they are 47.5th and 52.5th, and for the 95% (dark grey), 2.5th and 97.5th. The colored dots show the position and density of the 5% most central songs' peaks, with yellow points indicating high density and violet indicating low density. The insets show 5% least central curves (red). The darker blue areas (5% most central curves) are getting wider each decade, showing an increasing variety in the trajectories. The fixed-time statistics (black lines) do not capture that feature.

songs in the 2010s started at spots better or equal to 50, while for the other decades, the percentage is smaller than 10. The contrary can be observed in the end positions. In the last three decades, around 70% of the songs ended at ranks worse than 70 in the rankings, while in the 1960s, only 40%. These results confirm that the songs tend to start higher (better positions) and end lower (worse positions) in the Billboard Hot 100 compared to before.

Next, I analyzed how long songs take to reach their top positions and how long they have stayed on the charts since their peaks (top positions on the charts). According to Figure 5.11a, the time after reaching the top positions should get longer over decades. For the worse ranks, the songs should achieve their peaks faster. To test these hypotheses, I plotted the average weeks on the chart before and after reaching the top position. I present these plots in Figure 5.13.

The figure shows that the average future weeks on charts have increased over the last three decades for all positions except those that never reached the top 80. The top 10 has experienced an increase of almost three times. During the first three decades, songs took longer to secure their peak position than they did to remain on the chart. It started to change for the top 20 in the 1990s. Nowadays, best-performing songs spend, on average, twice as much time after their peak than before it. However, entries at the worst ranks (bottom 20) peak and disappear faster than in the first decades. These results illustrate the intensifying inequality between the superstars and the other artists.

After seeing all the differences in the dynamics of movements on the charts, I decided

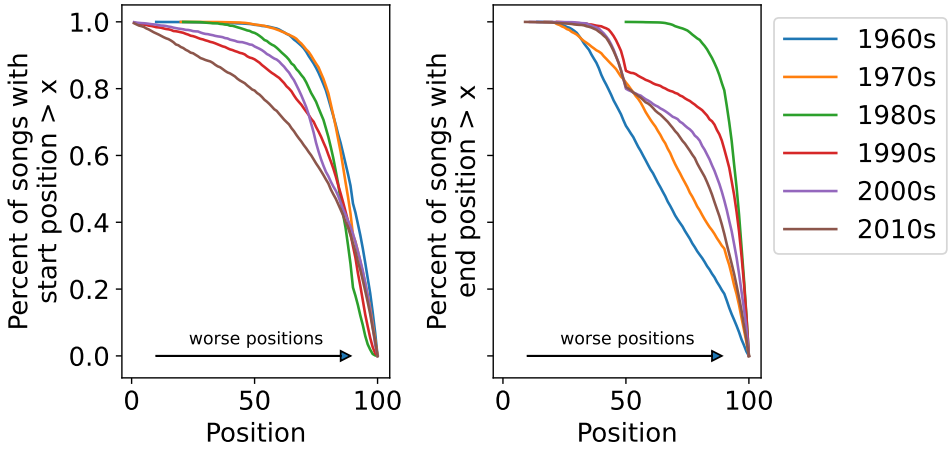


Figure 5.12: **CCDF of distribution of start and end positions.**
The CCDFs were calculated for the first positions (left) and last positions (right) distributions from the 1960s (blue) till the 2010s (brown). The y-axis values represent the percent of songs with start/end positions greater (worse) than x. In the left figure (first positions), most curves start with a plateau at 100%, followed by a sharp drop around position 50. For later decades the plots follow an arc shape, indicating that some trajectories start higher in the ranking. On the contrary, CCDF plots of the last positions (right) in the first decades decrease almost linearly since position 20. The last decades have been stable at 100% till around position 50 and then rapidly drop. This shows that trajectories tend to end at worse positions than before.

to try to measure the similarities between trajectories in different decades. I calculated Manhattan distances between all song curves in each decade. I present the kernel density estimation (KDE) plots of resulting distributions in Figure 5.14. I also show the values using a different metric - Hamming distance - in Figure E.4 and Figure E.5 in Appendix E.

Although it is an artificial measure and the values of some distances are very high (the maximum value is around 50 weeks $* 100$ position difference), the results capture some changes in similarity over time. It can be seen that density at the lower distances (bigger similarities) was higher in the early decades. Later, it shifted to higher lengths, increasing the probability of bigger dissimilarities. For the bottom 50 songs, the 2010s have the most similar trajectories compared to other decades. This results from their lifetimes getting significantly smaller - arrays differ only in the first few elements. These results show that there has been a clear divergence in the songs' path on the chart, except for the worst positions, where the curves are actually more similar compared to other decades.

5.3.2 Identifying trajectory archetypes

All the findings of this and previous chapters show that there has been a change in how songs move on the charts. Before, they started at worse positions, climbed to their top, and quickly disappeared. Nowadays, the best-performing songs start at better ranks, reach their peaks, linger on for longer, and slowly decrease to the bottom of the charts. In contrast, the worst-performing tunes enter the chart at their peak and disappear even more rapidly. During my research, I have identified various song trajectory features. This has led me to question whether it is feasible to categorize songs based on these characteristics. I took into consideration the following features that, in my opinion, best defined the dynamics of the trajectories:

- top position

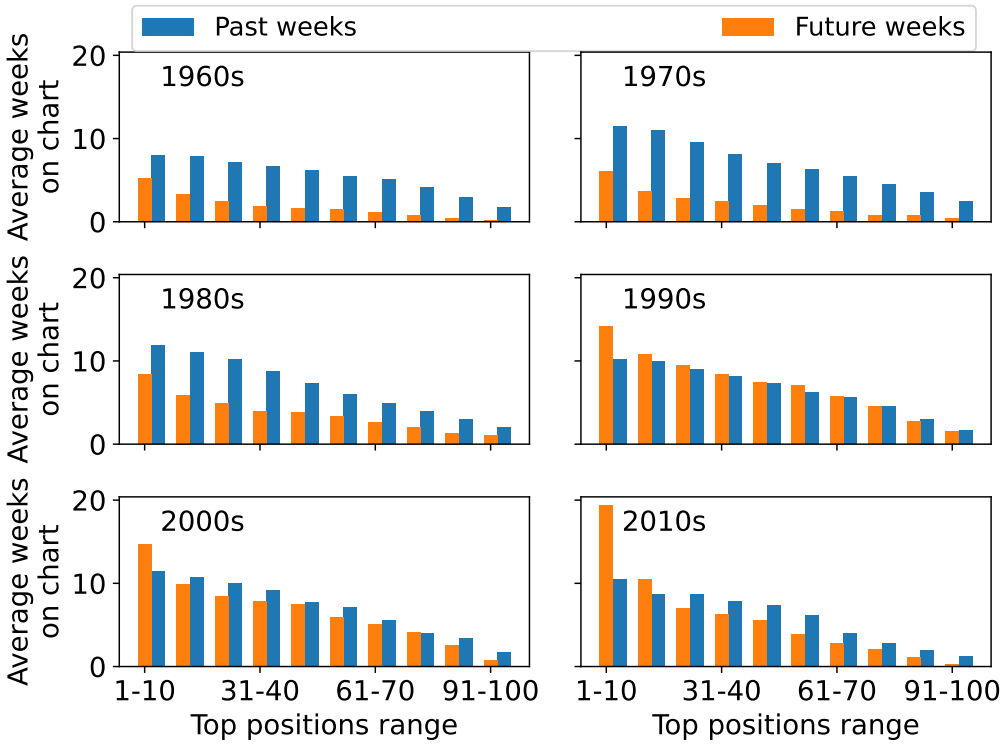
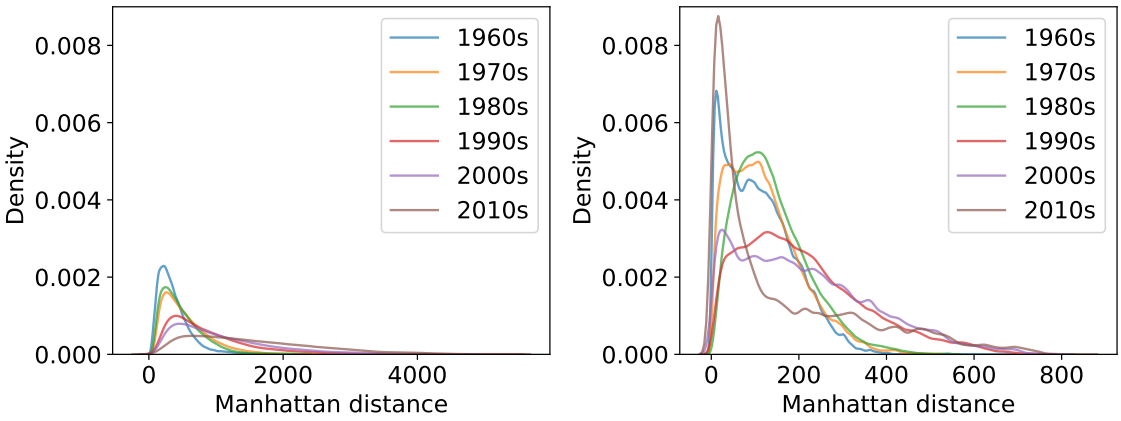


Figure 5.13: **Average weeks on the chart before and after reaching the top position.** The bins represent the average weeks on the chart before the peak (blue) and after the peak (orange). In the first three decades, the blue bars dominated the orange bars, showing that songs quickly disappeared after reaching the peak. Since the 1990s, the bars have been of nearly equal height, except for the top 10, where the average for future weeks is twice higher than for the past weeks. This shows that nowadays, the best-performing songs linger on for much longer.

- max weeks on the chart,
- weeks before the top position,
- weeks after the top position,
- starting position,
- end position.

I chose these features, as they represent some crucial points in the trajectories. One could represent all songs' curves as three points - first position, top position, and end position - connected with two lines. The distances between the points projected onto the x-axis (the lengths of the two lines on the x-axis) would be represented by the weeks before the top position and the weeks after the top position. I visualize my reasoning in Figure 5.15. Such simplified trajectories capture the most important dynamics of the trajectories - rise to the top, the peak, and then the process of dropping out of the chart.

After choosing the features, I created an unsupervised model to cluster the songs that resemble similar chart characteristics. I decided to use the k-means algorithm, which is fast (linear time complexity) and easy to use and interpret. First, I checked which number of clusters would be most suitable based on the clustering quality. I show the silhouette scores for different numbers of groups in Figure E.6 in Appendix E. The highest scores



(a) Songs that reached top 1-10.

(b) Songs that never reached top 1-50.

Figure 5.14: **KDE of Manhattan distances between trajectories.**
I calculated the Manhattan distance for each pair of songs in the decade. Next, I smoothed the distribution of distances using KDE. The values indicate similarities between trajectories from the 1960s (blue) till the 2010s (brown) for songs that reached the top 1-10 (left) and for songs that never reached the top 1-50 (right). For the top 10, the trajectories were the most similar in the first three decades. On the contrary, for the bottom 50, they were the most similar in the last decade

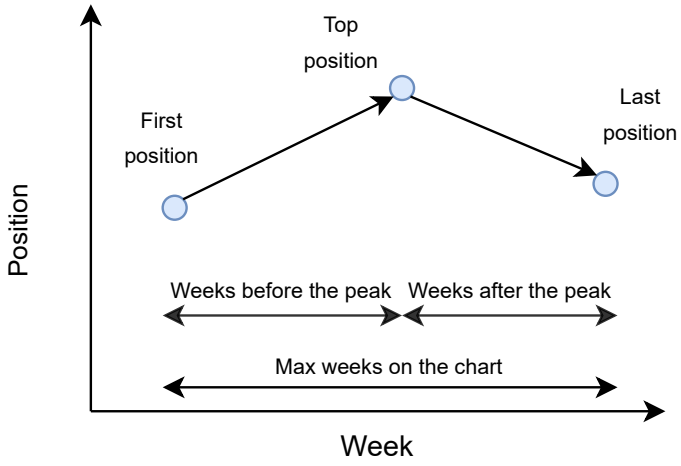


Figure 5.15: **Example of a simplified trajectory.**
The plot shows my reasoning behind choosing the particular measures as crucial trajectories' features. Any song path can be approximated with three points - first position, top position, and end position - connected with two lines. The lengths of the lines projected onto the x-axis are weeks before and after the peak, which sum up to the maximum weeks on the chart.

have been achieved for k values of 5 and 6. I first ran the algorithm on 6 clusters to see if the songs from different clusters resembled similar features. I show some basic statistics in Table E.1 in Appendix E. I chose the 75th percentile to show the relative values of weeks on the chart, first position, and last position. Even though there are some differences in the statistics between different groups, some resemble similar features -

e.g., cluster 1 and cluster 3 both have last position percentiles almost twice lower than for the first position. I also plotted some trajectories from different clusters in Figure E.7 in Appendix E and saw that clusters 1 and 5 were visually similar. That is why, I reran the algorithm for $k = 5$. I present the statistics for 5 clusters in Table 5.2.

Cluster	75th percentile weeks on chart	75th perecntile first position	75th percentile end position
Cluster 0	9	96	97
Cluster 1	20	91	98
Cluster 2	18	93	56
Cluster 3	20	43	95
Cluster 4	40	90	49

Table 5.2: **Statistics for 5 clusters.**
The table shows the 75th percentiles of various features used in identifying clusters. Cluster 4 holds the record for the highest number of weeks on the chart, whereas cluster 1 has the highest end position. Cluster 0 has the smallest weeks on the chart, cluster 3 has the lowest first position. These values give a better overview of the possible archetypes.

It can be seen from Table 5.2 that a few clusters stand out in terms of the statistics. For example, cluster 0 has a significantly lower 75th percentile of weeks on the chart, than the other clusters. The opposite can be spotted for cluster 4, which has the longest lifespan. Taking into consideration the statistics from Table 5.2, I have come up with the following archetypes:

1. Cluster 0 - “Brief songs” that last only a few weeks.
2. Cluster 1 - “Climbing songs” that start at worse positions, climb up to their top positions, and then slowly decay.
3. Cluster 2 - “High-end songs” that leave the charts drastically at better positions.
4. Cluster 3 - “High-start songs” that enter the charts at better positions.
5. Cluster 4 - “Long-lasting songs” that have significantly high lifetimes.

To get a visual impression of the clusters, I plotted some of their trajectories in Figure 5.16. It can be seen from the plots that the trajectories differ for different clusters and they seem to resemble the structures I defined. The brief songs are the songs that last up to 10 weeks. High-end pieces show vertical lines starting at positions better or equal to 50. In the high-start cluster, many songs start at a rank better than 20. The long-lasting songs usually last at least 20 weeks. Lastly, the climbing songs show smooth curves that begin and end at relatively low positions.

To summarize the curves, I also plotted the averaged trajectories of each cluster. I show this plot in Figure 5.17. The figure gives a better overview of the features. The long-lasting curve is broader than for all the other ones. The high-start group starts more elevated than others, and for the high-end songs, it sharply decreases positions after the peak. The brief songs’ trajectory has the lowest height and width. I also show the number of songs in each cluster. The largest group is “Brief songs,” with over 12000 songs, and the smallest is “Long-lasting,” with only 1645.

Lastly, I wanted to see the differences in the proportion of trajectories from each cluster throughout the years. I show the result in Figure 5.18. In the first two decades, almost 80%

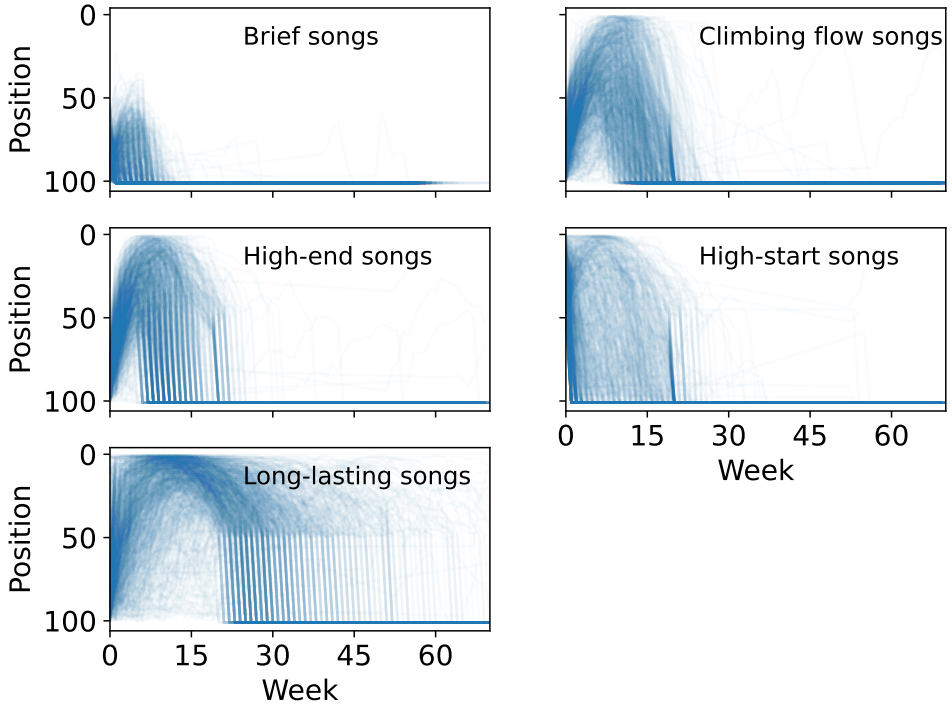


Figure 5.16: **Some of the trajectories of songs from each cluster.**
The trajectories were drawn for 1000 songs from each cluster. The curves were plotted with certain transparency to see the overlap in some areas. The songs from different archetypes seem to have similar features. E.g., most “Brief songs” do not last more than ten weeks. “High-end” songs display many sharp lines from the last positions around 50 to 101 (out of the chart). Many “High-start” trajectories start at positions better than 50, and “Long-lasting songs” last at least 20 weeks. The last group, “Climbing flow songs,” does not depict any significant anomalies.

of the songs are high-end. As I showed in the previous sections, the songs disappeared quickly from the charts in the early years after reaching their peaks. The second largest group is brief songs. Since the 1980s, the high-end trajectories drastically disappeared from the charts, and climbing songs became the new dominating group. In the 1990s, the proportion of climbing songs decreased, and a few new groups rose: the long-lasting and high-start. Moreover, the high-end pieces started to reappear after a break around the 1980s-1990s. The brief songs have gradually decreased since the 1980s but still account for around 10% of all songs. The long-lasting and high-start songs were the most influential in the last years. Combined, they account for approximately 60% of all songs.

Another algorithm I tried for clustering was non-negative matrix factorization (NMF), explained in subsection 4.12.2. However, the groups detected by that algorithm were not as distinguishable as by the k-means. I show the results of the other method in Figure E.8 in Appendix E. Since the trajectories from different archetypes seem to overlap, I decided to abandon the results from that algorithm.

The results from this section show that there have been clear changes in how songs moved on the charts throughout the years. I defined five archetypes of songs that capture different trajectories’ behaviors. These results clearly show the changes in the dynamics over the years. It is evident that songs with extended duration and boosted first positions

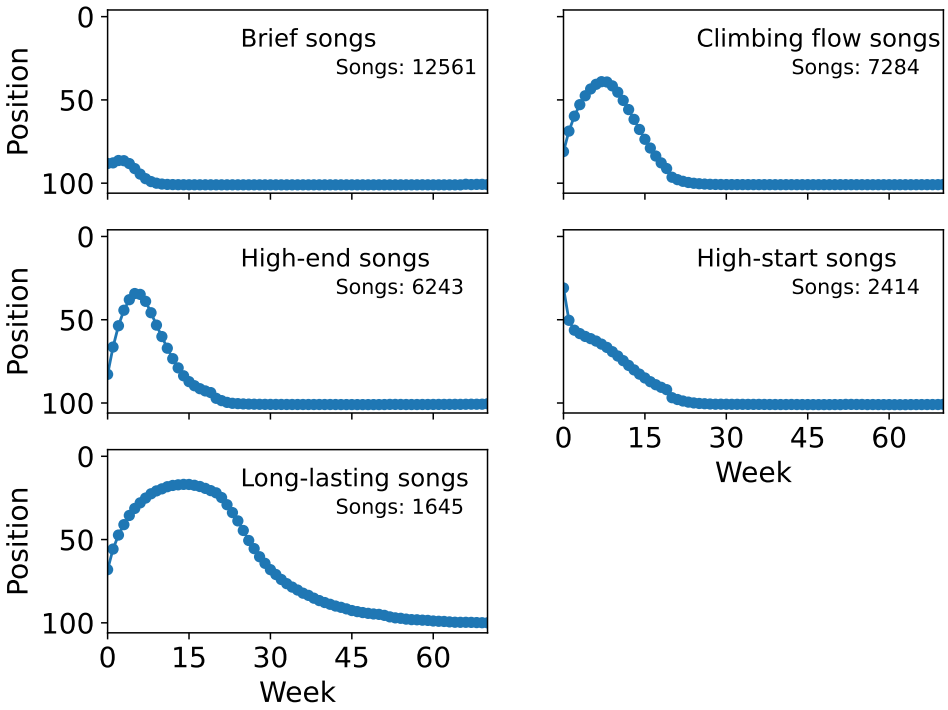


Figure 5.17: **Normalized trajectories of songs from each cluster.**
The plots show normalized trajectories and the number of songs in each cluster. The trajectories were calculated by averaging the positions of all songs in each week. The plots summarize the visual aspects of detected archetypes, described in Figure 5.16. One thing that stands out is that the clusters are not evenly sized. The brief songs group is almost two times bigger than the second-best, climbing flow.

have appeared in the last three decades and dominate the current charts. There are more and more extreme movers in the charts. It could suggest that certain artists who are already popular have an advantage when it comes to creating a hit song. Prestige influences the hierarchies, and big companies strive to strengthen their status by hiring outstanding performers [26, 94]. Lesser-known artists may be facing increasing difficulties due to the success of the superstars.

In the next section, I will categorize the movements on the charts and look at their differences over time. Even though I have shown much evidence of both rapid and mild moves in the Billboard Hot 100, little is known about their quantitative features. I will also test the "openness" of the chart and how it has changed over the years.

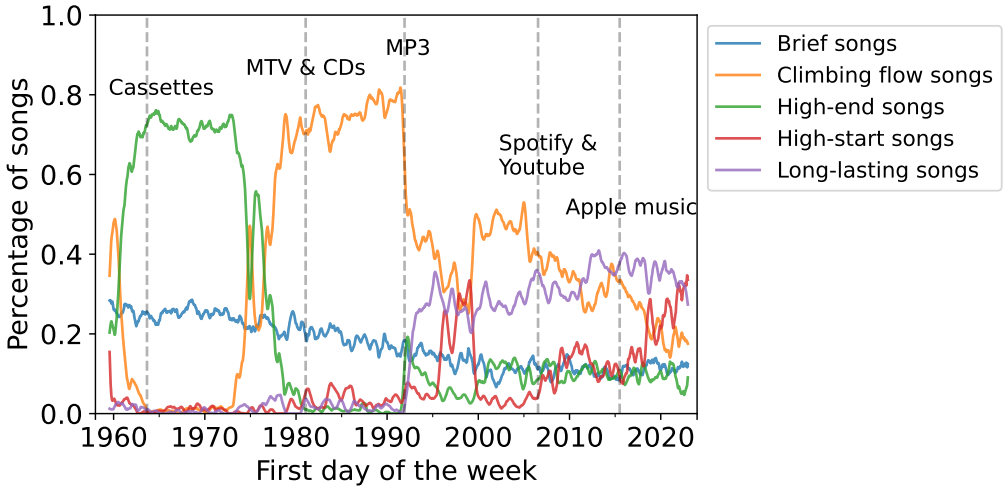


Figure 5.18: **Proportion of songs from each cluster over the years.**
The proportions were calculated by averaging the number of songs in each cluster over the number of songs in the year. The results were normalized with a half-year rolling window. The groups are brief songs (blue), climbing songs (orange), high-end songs (green), high-start songs (red), and long-lasting (purple). The dashed grey lines indicate some of the most important events in music listening history. It can be seen that first, high-end songs dominated until the 1980s when the climbing flow took over. Since the 1990s, the climbing songs have started to decrease, in favor of long-lasting, high-start, and reappearance of high-end songs. Over the course of the chart's history, the prevalence of short songs has steadily declined from 20% to approximately 10%.

5.4 Lower-ranking songs are moving more abruptly on the charts

Iñiguez et al. [48] came up with ways of analyzing ranking dynamics. They divided the charts into open- and closed-oriented. I describe the types of rankings in section 4.8. In this section, I will show that Billboard Hot 100 resembles the features of an open ranking and will show the differences in some of the measures defined in the article over time.

First, I looked at the mean turnover and mean flux over time. Rank turnover indicates how new elements enter the ranking, and rank flux represents the in and out flow of the charts. I show the values in Figure 5.19. Closed rankings have mean turnovers and fluxes of nearly zero. The results for Billboard Hot 100 reveal that the chart has a mean turnover and flux greater than 0 for all decades. Therefore, it is an open ranking from the definitions in subsection 4.8.1 and subsection 4.8.2. This indicates that the chart is relatively stable only in the top positions, meaning they have fewer fluctuations in rank and a higher lifetime than the other elements.

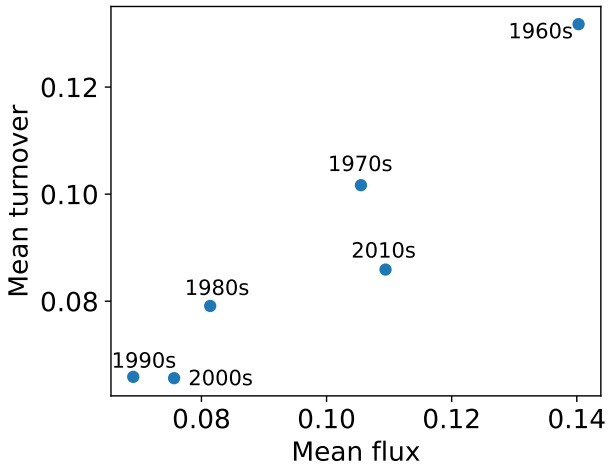


Figure 5.19: **Relationship between mean turnover and mean flux over time.** The x-axis depicts the mean flux, and the y-axis the mean turnover. The dots demonstrate the values for different decades, from the 1960s to the 2010s. The averages are positive for each time range. They slowly decreased till the 1990s, but then they increased again. In the 2010s, the mean turnover was around 0.086, and the mean flux was roughly 0.11.

Next, I calculated displacement probabilities $P_{x,t}$ that an element at rank r will move to position x after time t. I define the measure in subsection 4.8.6. I plotted the values for $t = 1$ week over decades and showed them in Figure 5.20. I also show the results for $t = 4$ in Figure F.1 in Appendix F. Iñiguez et al. [48] define two types of movements in the charts: diffusive moves and Lévy jumps explained in respectively subsection 4.8.4 and subsection 4.8.3. In Figure 5.20, anything outside the dashed hills (diffusion peaks) is a Lévy jump. Closed rankings display symmetry between the heights of the diffusion peaks in the top and bottom initial positions. The probabilities in the Billboard Hot 100 in the first few decades resemble those of the closed ranking. The heights of the lower-position peaks have been getting smaller in the last three decades. The flatness of lower initial position curves suggests that the songs' movements become less diffusive and sharper. The worse-performing songs have higher chances of degrading or jumping many places in the charts.

The definitions of types of movements in the rankings by Iñiguez et al. [48] are complex

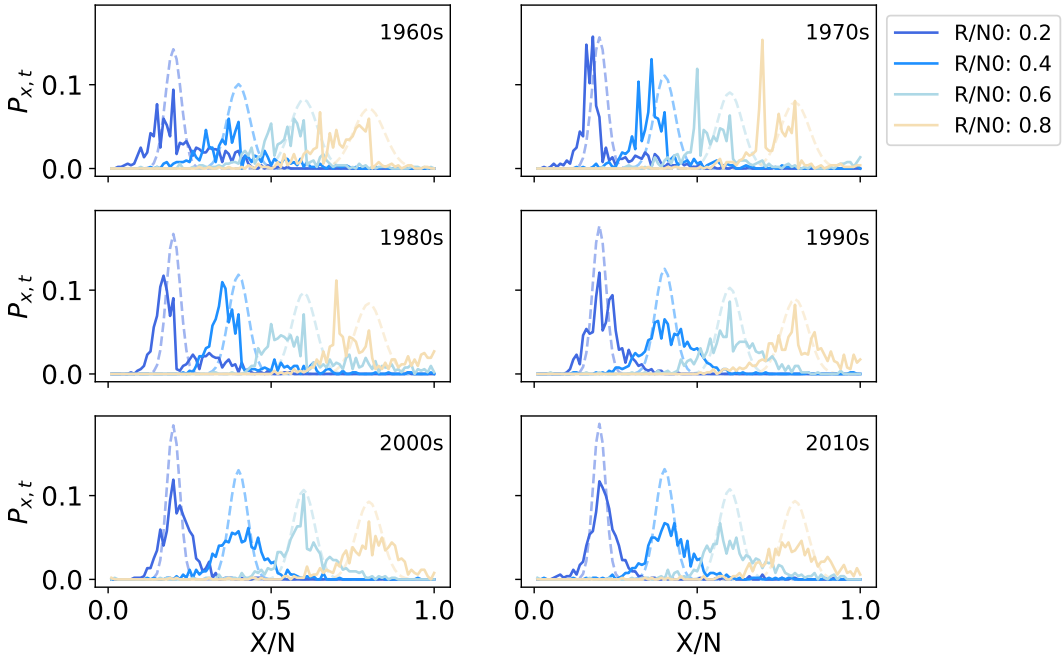
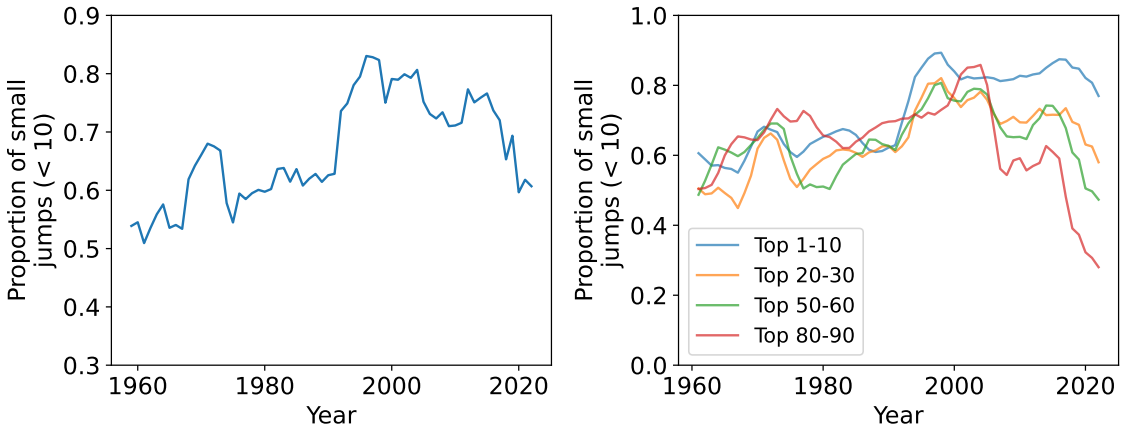


Figure 5.20: **Displacement probability over decades.**
Probabilities were counted for initial ranks 20 (dark blue), 40 (blue), 60 (light blue), and 80 (yellow) for $t = 1$. The dashed lines indicate the $P_{x,t}$ of the model described by Iñiguez et al. [48]. Open systems lose symmetry in the heights of diffusion peaks.

and require a lot of knowledge in statistics and probability theory to understand. Instead of Lévy and diffusion jumps, I propose a simple definition of small/long leaps. A small jump is an absolute position change smaller than 10. Otherwise, a rank difference is considered a long jump. I selected this range because I've grouped the songs based on their consecutive positions in sets of 10. Therefore, a jump of 10 corresponds to a whole range shift. I show the proportions of these measures for all songs and different top position ranges in Figure 5.21.

Figure 5.21a shows that the amount of small jumps has, on average, increased over time but started dropping around 2010. The peak proportion of 80% was achieved in the late 1990s but has fallen to around 60% in the last years. Figure 5.21b shows that the proportion for the top 10 has stabilized around 80% since the 2000s. For all the other ranges, it started to fall at that time. For the worst-performing songs, the proportion has fallen to roughly 30%, compared to 80% of the top 10. It indicates that songs in better positions are more likely to diffuse around their initial ranks. Those not achieving higher rankings will likely make sudden, abrupt moves.

The analysis of the ranking dynamics is a supplement to the work done in the previous sections. The Billboard Hot 100 is an open-oriented ranking with a highly stabilized top and fluctuating middle and bottom. It supports the findings that there have been changes in the movement of songs on the chart and that the top-position songs smoothly climb on the charts, with few experiencing long jumps. New information has been revealed about the lower-ranking songs. It appears that they now feature more sudden changes than they did previously. It supports the theory that competition has also increased for the lower ranks.



(a) All songs.

(b) Songs in different position ranges.

Figure 5.21: **Proportions of small jumps over the years.**
The left figure shows a proportion of small jumps in all songs. The right plot shows the same but divided into the top 1-10 (blue), 20-30 (orange), 50-60 (green), and 80-90 (red) songs. The values have been smoothed with a moving average of 3 years. I have selected these specific ranges for ease of readability and to align with the previous analysis of the best-ranking positions. It can be seen that the overall proportion of small jumps started to drop around the 2000s, but for the top 10, it stabilized at around 80%. For the worst positions, the proportion drastically dropped to around 30%.

5.5 Fewer new artists get to be in the charts and the number of collaborations increases.

So far, I have shown changes in the dynamics of charts that could lead to increased difficulty in making a hit. I have identified types of songs in the charts and their change in prevalence over time. Some findings motivated me to hypothesize that famous artists might perform way better, making it harder for other, less popular, and new artists to succeed in the charts. In this section, I want to focus on the situation of new artists in the charts over the years.

To begin, I plotted the yearly percentage of new artists. This includes instances where the "artist" column displays a name for the first time in a given year. I show the results in Figure 5.22. However, I quickly realized this method is slightly misleading, as it does not include features and collaborations between artists. Hence, the proportions might be overestimated and do not reflect the artists who just made it to the Billboard Hot 100 for the first time. I plot the ratio of new artists (without "featuring" and "&" in the name), as well as the percentage of collaborations in Figure 5.23. Although the method for extracting individuals may not be entirely accurate (it excludes the bands and individuals with "&" characters from the individual-artists group), it approximates the actual values.

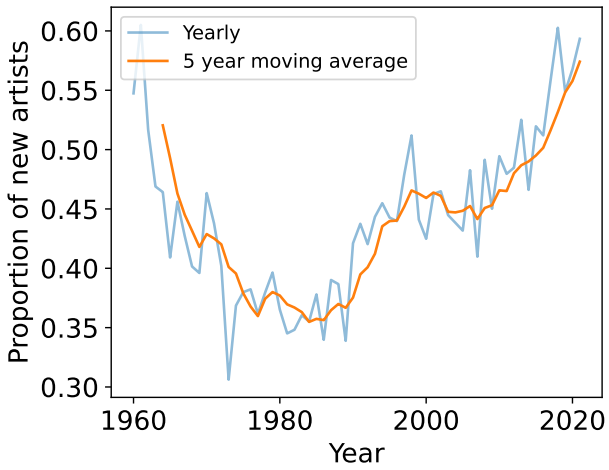
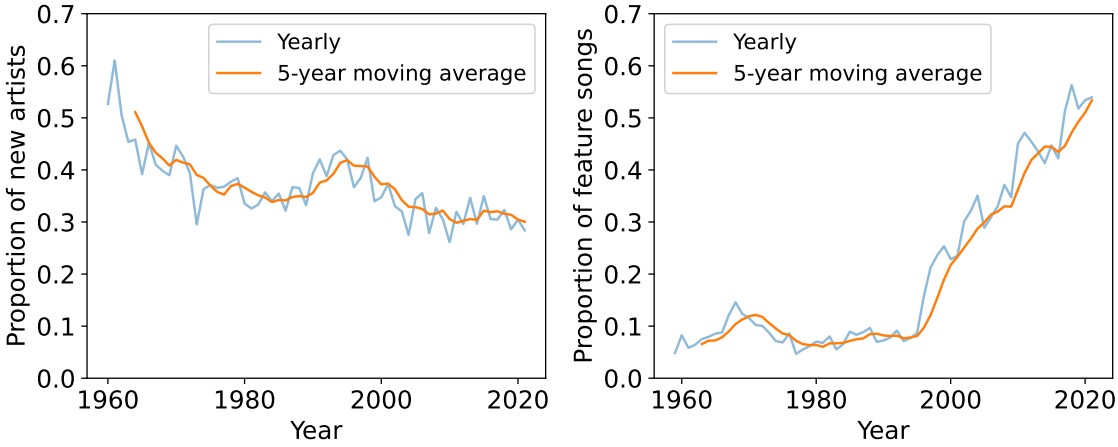


Figure 5.22: **Proportion of new artists.**
The blue lines indicate yearly values, and orange, the 5-year moving average. The proportion has decreased for the first three decades and increased since the 1990s, peaking at around 60%. However, these results also include collaborations and features.

From Figure 5.23a, we can see that there has been a gradual decrease in the proportion of new artists over time, except for a small spike around the 1990s. The value has ultimately fallen from 60% in the 1960s to 30% in the last years. In Figure 5.23b, the proportion of features and collaborations was more or less steady till the 1990s, fluctuating around 10%. However, since the late 1990s, it has rapidly increased, achieving around 50% in 2022. It might mean that more and more artists group together to increase their chances of entering the charts. Collaborating with famous people is a popular way of quickly gaining a vast amount of interest [56].

Next, I looked at the proportion of new artists (both individuals and collaborations) who reached the top 10. I show the results in Figure 5.24. It can be seen that new artists and collaborations do not perform exceptionally well on the charts. Less than 10% makes it to the top 10. The data indicates that most of the top 10 performers are established



(a) New artists (without features).

(b) Features and collaborations.

Figure 5.23: **Proportions of new artists.**
The plots show the proportion of new artists (left) and features/collaborations (right) over time. The blue curve presents the actual values, and the orange is smoothed with a 5-year window. The proportion of new artists without features is decreasing with time, while for collaborations, there has been a drastic decrease since the late 1990s. The ratio of new artists is now at 30% and for the features 50%.

singers, who have appeared numerously on the charts. This might also mean that artists collaborate to be in the charts but not necessarily to achieve the best positions. I elaborate on the phenomenon of features in section 6.3.

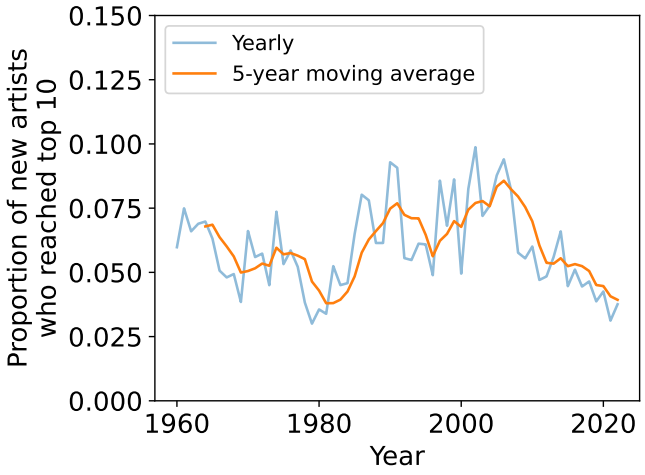


Figure 5.24: **Proportion of new artists and collaborations in top 10.**
The blue lines indicate real values, and the orange is a 5-year moving average. The data includes features and collaborations as well. It can be seen that these groups rarely achieve the top 10. The proportions are below 10% and have been decreasing since the 2000s. Nowadays, only 5% of new artists and features achieve the top 10.

I decided to take this investigation one step further and calculate the proportion of the top 10 artists who have been in the top 10 before. I show the graph in Figure 5.25. Based on the figure, more than 50% of the top 10 artists are well-established, consistently appearing

in the highest positions in previous years.

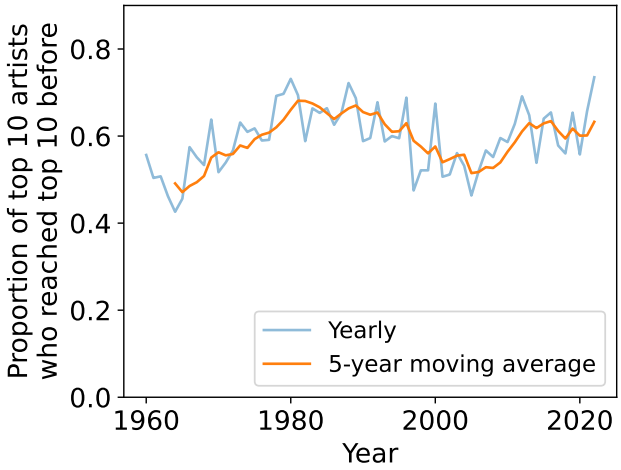


Figure 5.25: **Proportion of artists in top 10 who have achieved it before.** The reappearing artists are those who achieved the top 10 in previous years. The blue lines are real values, and the orange is the 5-year moving average. It can be seen that the ratio of reappearing top artists fluctuates around 60%. In the last years, it peaked at around 70%.

Lastly, I plotted the proportion of artists with different numbers of songs. I show the results in Figure 5.26. The figure shows that the number of artists with more than 30 songs has increased visibly since the 2000s, peaking at around 40%. It is natural for artists to produce more songs over time, which can lead to an increase in the number of chart entries. However, the growth has been very sharp since the 2000s. This might indicate that, nowadays, these artists are consistent superstars who have been on the charts for the longest. However, the percentage of artists with five or fewer songs has been declining and is now less than 20%, which is half as much as those with the most pieces. The gap between the worst-performing artists and the superstars has increased for the last few years. The other values have been fluctuating around 20-30%.

The analysis revealed two radical groups of artists on the charts - new artists and hit-makers. I showed that the situation of emerging artists has been worsening since the 1990s. Fewer and fewer individuals belong to this group, and their likelihood of reaching the highest positions is minimal. In the next section, I will investigate the other group - the hitmakers.

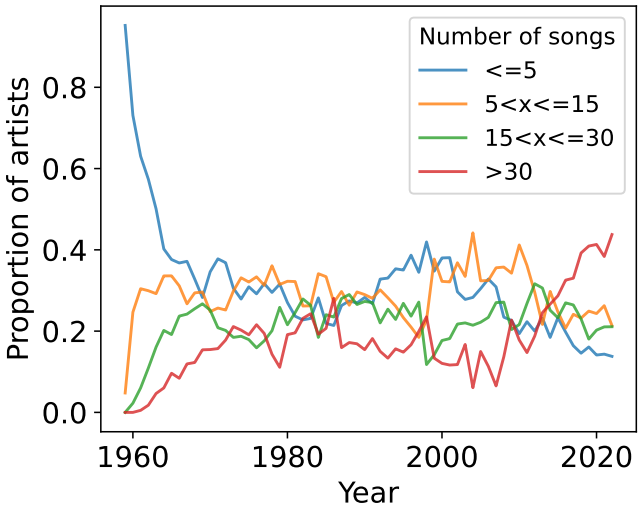


Figure 5.26: **Proportion of individual artists with different numbers of songs.** The different ranges of the number of songs are: <= 5 (blue), 5<x<=15 (orange), 15<x<=30 (green), and more than 30 (red). Plots have been smoothed with a 5-year moving average. The data does not include features and collaborations. The ratio of the smallest number of songs (blue) is naturally very high at the beginning of the chart's history. It has dropped to around 20% in the last years. The medium values (orange and green) have been fluctuating between 20% and 40%. The red curve, which indicates over 30 songs by an artist, has visibly increased in the last two decades.

5.6 Hitmakers dominate the rankings, but their songs tend to perform worse.

So far, I have focused on the situation of new artists in the charts. During the analysis, the second group identified was the hitmakers. I have shown in my results that the most famous artists have been performing exceedingly well in the last few years. In addition, the preceding chapter demonstrated that there has been a rise in the number of artists who have many songs (more than 30) on the charts. That is why, I decided to investigate the group of the most elite artists and whether it dominates current charts.

There are many different ways to classify the best-performing artists. However, I decided to look at the distribution of the number of songs by one artist. I present some key statistics on chart appearances for artists in Table 5.3. Around 80% of all artists in the charts have two or fewer songs.

On the other hand, the percentage of artists with more than ten songs is only 5%. With this in mind, I choose to classify an artist as a hitmaker if they have over ten songs on the charts. I show artists with the most significant number of Billboard entries in Table 5.4. I present the entire song count distribution with the top artists marked in Figure 5.27.

Group	Count	Percentage of all artists
All artists	10416	100%
Artists with 1 song	6829	66%
Artists with 2 songs	1244	12%
Artists with 3 songs	615	6%
Artists with more than 10 songs	552	5%

Table 5.3: **Artists with different numbers of songs.**
Over 10000 distinct artists were on the Billboard Hot 100 from 1959 to 2022. Artists with three or fewer songs account for over 80% of all artists.

Artist	Number of songs
Glee Cast	183
Taylor Swift	163
Drake	124
YoungBoy Never Broke Again	75
The Beatles	65

Table 5.4: **Top 5 artists with the most songs.**
The top 4 artists with the most songs on Billboard Hot 100 debuted in the 2000s. Glee Cast, Taylor Swift and Drake are the only artists with over 100 songs.

After creating a definition of a hitmaker, I looked at their contribution over time. I plotted the hitmaker's average and median number of songs and their proportion over time. I show the result in Figure 5.28. From Figure 5.28a, it can be seen that the average has been increasing till around the 1990s, decreased slightly in a few years, and continued to rise since the 2000s, peaking at around 35 songs per hitmaker. The median follows the mean dynamics, peaking at approximately 30 weeks in 2022.

The proportion of hitmakers in Figure 5.28b goes up and down at similar times as the curves in Figure 5.28a. In the 2000s, only 10% of the artists in the top 10 and 40% overall

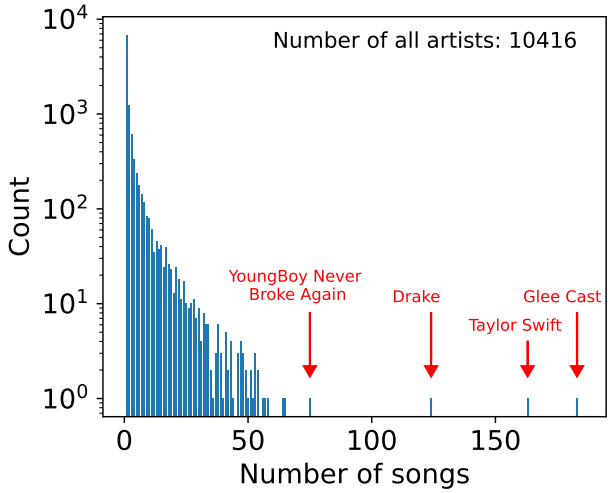
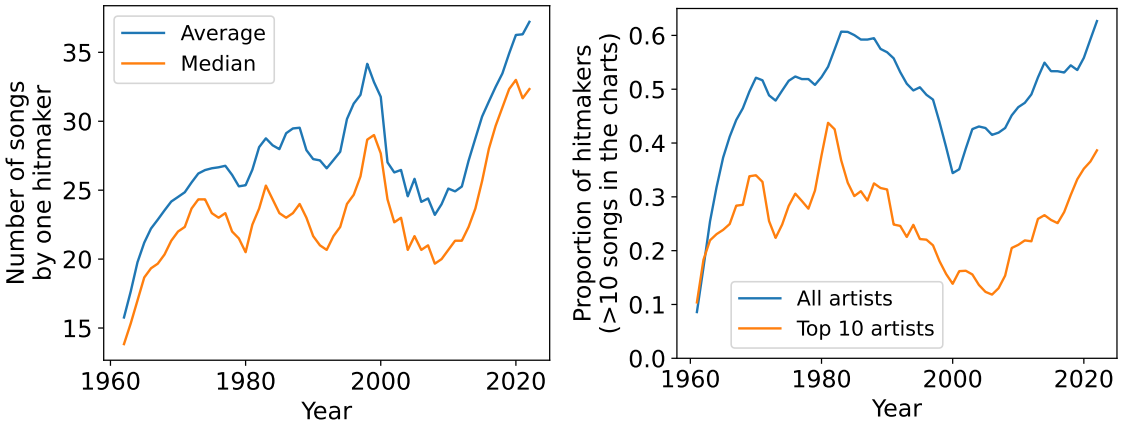


Figure 5.27: **Distribution of the number of songs on the chart per artist.**
The red arrows point to the outliers with significantly the biggest song count. The counts have been presented on a logarithmic scale, as the 1 and 2 values highly dominate the distribution (80% of all artists). It can be seen that the top 3 artists (Glee Cast, Taylor Swift, and Drake) are far from the rest of the distribution. The vast majority of artists have only 1 or 2 songs on the Billboard Hot 100.

had more than ten songs. Today, hitmakers account for 60% of all artists in the charts and over 40% of the top 10. The numbers indicate that consistent and well-established performers occupy most spots on the charts and a considerable fraction of the highest positions.



(a) Number of songs by one hitmaker.

(b) Proportion of hitmakers.

Figure 5.28: **Number of songs and proportion of hitmakers over time.**
The left plot shows the average (blue) and median (orange) number of songs by one hitmaker over time. The right plot displays the proportion of hitmakers in all artists (blue) and those who achieved the top 10 (orange) over the years. Both figures were smoothed with a 3-year moving average.

Next, I decided to compare some statistics between the hitmakers and non-hitmakers regarding the top position, weeks on the chart, etc. I show some values in Table 5.5. It can be seen that the hitmakers tend to get better top positions than non-hitmakers.

However, there are no major differences in the other features.

Group	Median lifetime	Median top position	Median first position	Median last position
Hitmakers	10	37	82	79
Non-hitmakers	9	53	88	87

Table 5.5: **Statistics for hitmakers and non-hitmakers.**
Half the hitmakers achieve a top position higher than 37, compared to 53 for the non-hitmakers. Also, hitmakers start and end slightly higher in the rankings.

Next, I wanted to see whether the hitmakers' trajectories over time. I plot the song paths for all top performers in Figure 5.29. The trajectories have been getting more diversified. Some last only a week, while others last for over a year. In previous decades, top performers were more likely to reach the top 10 whenever they released a new song. In the 1980s, approximately 33% of hitmakers' songs made it to the top 10, whereas in the 2010s, only 15% achieved this level of success. Together with the results from Figure 5.28b, it must indicate that an increasing number of hitmakers end up in lower positions.

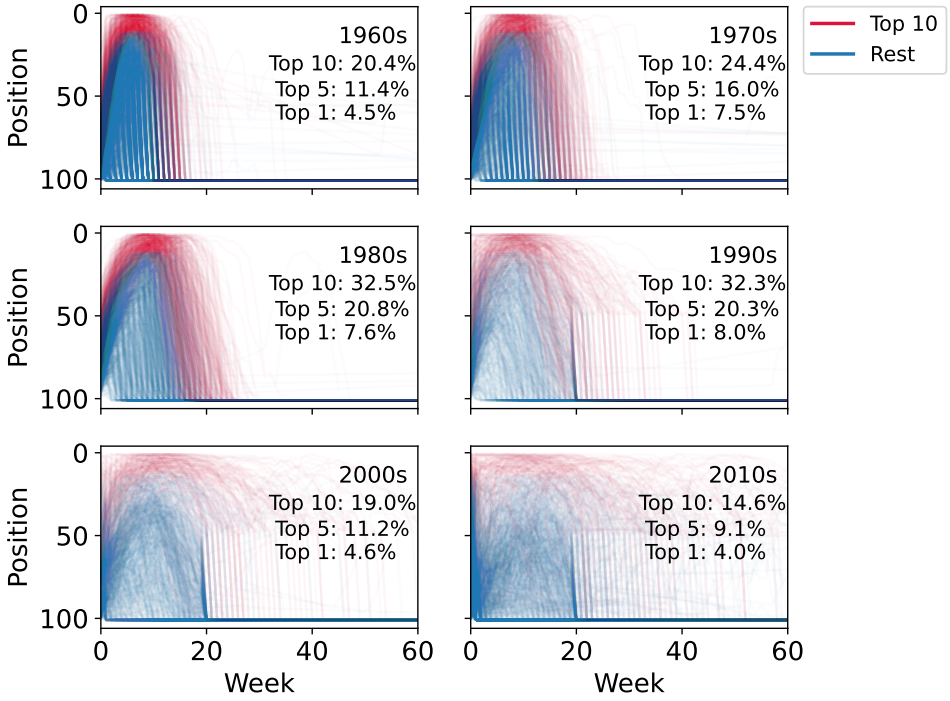


Figure 5.29: **Trajectories of the songs of hitmakers.**
The red lines indicate songs that reached the top 10, and blue the rest. Also shown are percentages of the top 10, top 5, and top 1 songs. In the first decades, the hitmakers were more likely to end up in the top 10, compared to nowadays.

To examine the performance of the hitmakers, I created a graph showcasing the proportion of their songs in various position ranges. I show it in Figure 5.30. In the first three decades, over 50% of the top 20 songs were created by hitmakers. In recent decades, more and more superstars' songs have performed worse in the rankings. However, we can see that there has been an increase in the number of pieces by hitmakers since the 1990s - from 25.6% in the 1990s to 39.5% in the 2010s. I also plotted the top positions of some of the

most famous artists before and after the 1990s in Figure G.1 in Appendix G, to showcase the increase in the variety of positions of hitmakers.

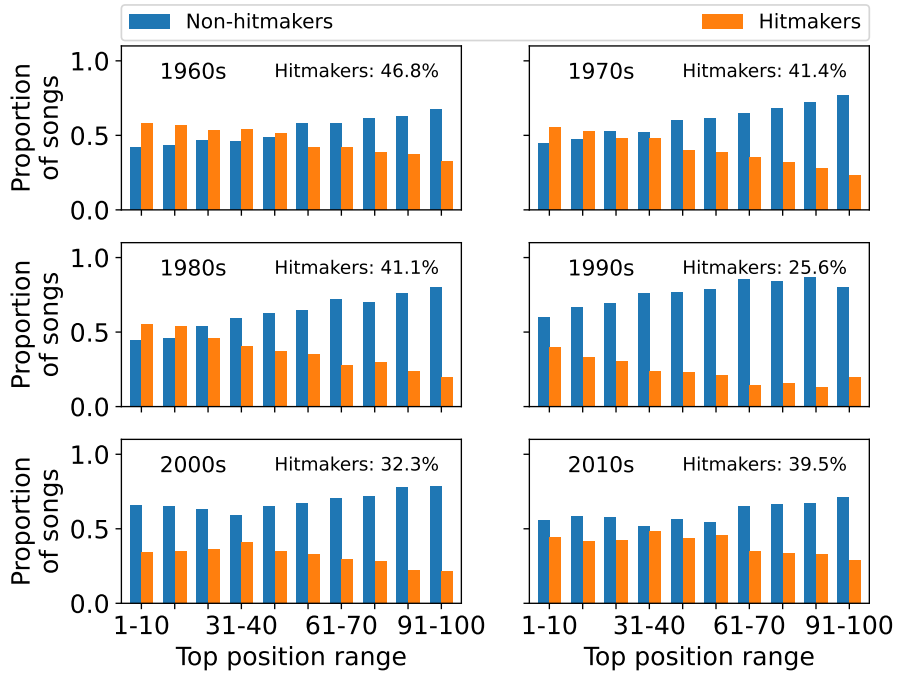


Figure 5.30: **Proportion of songs by hitmakers and non-hitmakers.**
Blue and orange bars represent the non-hitmakers and hitmakers, respectively. The upper right text shows the percentage of songs performed by hitmakers in the decade. In the last decades, the blue bars are higher than the orange in all position ranges. Moreover, the heights of the hitmakers' bars are larger in worse top positions, compared to before.

Concluding all this section's findings, nowadays, hitmakers dominate most charts and have a lot of songs in the top 10. However, many hit songs have performed worse than in the first decades. The results also show that the constant hitmakers might strive to stay relevant on the charts, releasing more and more songs since the 1990s. This could further prove that the music industry has become increasingly competitive, even for the topmost artists.

6 Discussion

The main part of my research was to answer the hypothesis, “Is it getting harder to make a hit?”. I have shown that there are many hints indicating that it has gotten in fact more challenging:

- increasing lifetime of the best-performing songs and shorter duration of worse-performing ones,
- bigger variety in the movements of songs,
- increasing number of long-lasting and high-start songs,
- worsening of performance of new artists and hitmakers.

Some songs occupy the spots for over a year, while others disappear after one week (Figure 5.3, Figure 5.4). There are fewer new artists and more collaborations/features in the charts (Figure 5.23). Moreover, the hitmakers occupy many spots, even in the lower positions (Figure 5.30). All these findings show that there have been changes in the pop music charts, which make it harder to make a hit.

While some of these results were expected, such as the lifetime correlation with the top position, I discovered a few unexpected features. The first surprise was the increasing competition in the lowest places. Based on my analysis, I have found that the lower-ranked chart entries tend to experience much larger jumps in position than in previous decades (Figure 5.20, Figure 5.21). The second unexpected result was the worsening of the performance of hitmakers. Even though they occupy a large portion of the charts, in the earlier decades, they were more likely to achieve the top 10 than nowadays (Figure 5.30).

The culmination of my findings was the research on the archetypes of trajectories over time. I defined five different types of song paths on the charts and showed their change over time (Figure 5.18). The results clearly show the changes in the dynamics over time. Until the 70s, songs would quickly disappear from the charts after reaching their peak. Then, until the 90s, songs would climb from the bottom of the ranking to their peak and again to the bottom positions. The last decades show the emergence of long-lasting songs and songs that start at the top positions of the charts. These results sum up the changes in the dynamics throughout the Billboard chart’s history.

Whereas my thesis hints that the answer to the hypothesis “Is it getting harder to make a hit?” is “yes,” further investigation is required to settle the question completely. Such work might incorporate additional information on, e.g., sales and live performances [64]. The results of my research lay the foundation for future exploration of the music charts and the industry as a whole. There is still a limited understanding of the potential factors and explanations for the fluctuating positions of songs on the music charts. I put forward some possible explanations and theories in the following subsections.

6.1 Increasing rivalry in the music industry

Some of the results of my research show that there might have been an increase in competition in the music industry. Companies introduce new strategies to ensure their artists perform well on the charts [33]. This is especially visible in the increasing number of songs by well-established and highly-paid artists (hitmakers), as shown in Figure 5.30. That the

music industry has changed in recent years is mirrored by findings from other studies. The impact of information spread and advertising is increasing significantly, with the goal of gaining as much attention as possible [23, 41]. The most prestigious artists are more likely to be endorsed and gain the most out of advertisements [26, 94]. In contrast, albums from less famous artists tend to sell below their potential due to a lack of proper promotion [40]. In summary, the industry seems to be more biased toward the superstars.

Two competing and seemingly contradicting theories concerning the skewness of the music market have emerged in recent years: the “winner-take-all” theory and the “long-tail” theory [70]. The supporters of the first theory say that some superstars can capture a significant portion of the market with just a few songs. On the contrary, the “long-tail” effect states that more and more audiences switch to less popular artists, decreasing the concentration around the most famous performers. Interestingly, in my work, there is evidence that both effects are present in modern music. My finding that songs debut at better positions and have longer chart lives may be due to the superstar phenomenon. In my work, the “long-tail” effect can be spotted in the increased traffic in the chart’s bottom parts and shorter durations. This might indicate that the competition between lesser-known artists has also grown. It is difficult to determine which theory will dominate more in the future [19, 28].

Despite the growth in the volume of new music, the market remains highly concentrated around the well-established artists [4, 31]. Fortunately, the digitalization of music has given contemporary artists a greater opportunity to be heard. While the performance of new artists may not be as strong, and they may have a lower presence on the charts, the niche market in the music industry is experiencing an increase in sales [4, 97]. These observations might significantly impact the charts created in the years to come.

6.2 Impact of streaming

In my study, I hinted that streaming may have played a huge role in shaping the Billboard Hot 100 over the last few years. Most drastic changes, e.g., in lifetime and trajectories, occurred in the last dozen years. Some of the most important events in the digitalization of music are:

- launch of iTunes - 9th January 2001,
- launch of Youtube - 14th February 2005,
- launch of Spotify - 7th October 2008.

This gives the idea that streaming may have impacted the dynamics of charts. As shown in Figure 5.2a, the best-performing songs have had a drastic increase in their lifetime since the streaming era [11]. Contrary, the worst-performing pieces lasted even less, sometimes disappearing after one week.

The signs of digital technologies influencing the music industry have also been shown by researchers. Bhattacharjee et al. [12] showed a correlation between the amount of digitally shared music (downloads on Napster[1]) and the Billboard Hot 100 position. They also suggest that the music industry might benefit from monitoring the online sharing activity, as it provides valuable feedback on the interest in songs.

In other work, streaming has been shown to be beneficial for new artists. The ease of access to music helps to increase the sales of lesser-known performers, increasing the

[1] [https://www.napster.com/]

"long-tail" effect [9, 58]. Smaller, independent companies have demonstrated the advantages of creating low-cost music for niche audiences [9].

The emergence of the digital music era may have caused significant shifts in chart dynamics. Thanks to advancements in technology, producing music has become more affordable and accessible, which has opened up opportunities for numerous aspiring artists to enter the industry [9]. This leads to the conclusion that, nowadays, streaming may have a more significant impact on chart performance than any other medium.

However, the results of my study are only suggestions that streaming might have impacted the Billboard Hot 100 dynamics. Calculating the importance of streaming was out of the scope of my study. One could take this investigation further, by studying the charts from different countries before and after the launch of streaming services. Such an experiment might show a correlation between the emergence of streaming and the rankings.

6.3 Collaborations as a way of boosting the performance

My study revealed interesting results about the drastic increase in collaborations and features in the charts. As shown in Figure 5.23b, approximately 50% of Billboard Hot 100 songs in 2022 were features. This might shed light on new strategies in the music market.

Collaboration between creators has been shown to affect performance and popularity positively [34, 56]. In the music world, it might mean one of two things:

- less popular artists collaborate with famous performers to gain a vast amount of interest quickly,
- or well-established artists collaborate with other prestigious artists to maintain their status in the rankings.

In either case, further research on the featuring strategies should be conducted.

6.4 Limitations of this study

The music industry is a complex market - fast-paced, competitive, and highly dependent on consumers [10]. The trends in music have been changing throughout the years. Many researchers have tried to understand the underlying processes, but little is still known about indicators of success [10, 61]. Most of the results of my study are still hypothetical, as I provide only statistics on charts from a complex music market. One might question whether it is enough to answer the hypothesis, "Is it getting harder to make a hit?". However, I provide the results on publicly available historical data, which is relevant to the industry. With all that in mind, my study cannot provide an answer to the question, but instead hints and suggests that the answer could be "yes."

Moreover, the research includes only a tiny proportion of the data on music. While the music charts provide a lot of valuable information on the changes in the industry, there are many different aspects of songs. These include live performance sales, sonic features, social media traffic, etc. There are many ways to improve my results, including different data sources and looking at the charts in a broader context. Only a limited amount of information can be obtained from the charts. Other relevant data could include sonic features from, e.g., Spotify API[2], historical information on listening counts, sales of the singles/albums, etc. In my study, I did not include any of this information, as I wanted to focus purely on the dynamics of charts and whether they provide sufficient knowledge

[2] https://developer.spotify.com/documentation/web-api

about the increasing difficulty of making a hit. Moreover, music rankings are a summary of the raw features, such as album sales, listeners, etc.

Furthermore, I have shown that an increasing number of songs last only one week. However, still little is known about which artists perform so poorly and why. Further investigation into the possible cases of increased tendency to quickly drop out of the Billboard 100 would be interesting.

Lastly, I have briefly shown that podcast and book rankings have different dynamics than pop music charts in Appendix B. However, the data I used was of poor quality. The investigation of the dynamics of other charts would be interesting to carry out if one could find better data sources.

6.5 Future work

My study provides the groundwork for future work on the dynamics of rankings. The first possible direction is to investigate the "diffusion of hits" across different countries. Future work could include creating a simulation of songs' spread worldwide. This could help identify the critical import and export factors in the music markets. For example, `Kworb.net` [59] contains relevant data on charts from different countries and could be scraped and used in the investigation.

Another subject for future work could be developing mathematical models to reflect the challenge of creating a hit. A model could be used to show how the difficulty changes with time. Moreover, it could help compare charts from different industries.

One could also develop predictive models to forecast songs' movements in the charts. However, the Billboard Hot 100 dataset alone contains very few features that will likely provide inaccurate and biased results. That is why much effort should be put into feature engineering and balancing the data. Another model could be developed to forecast songs' lifetime, peak position, etc.

7 Conclusion

Music plays a big part in people's life. It has accompanied humans worldwide throughout their lives since ancient times [20, 55]. This is one motivation why researchers have been trying to understand it for many years. Nowadays, the music industry is a fast-paced market with an increasing number of products to choose from. However, a significant portion of industry profits is dominated by a few highly successful products [40]. Music magazines, like Billboard, have been trying to capture the top performers for many years, constantly changing their rules to reflect the current trends. Digital media, such as streaming platforms, have strongly influenced the perception of music success over the last few years. That is why I chose to investigate the pop music charts changes before and after the digital era, to answer the question, "Is it getting harder to make a hit?".

In this study, I have shown that there has been a vast amount of changes in the dynamics of the songs on the charts. The lifetimes of the top-performing pieces have drastically increased, while those in the lowest rankings have decreased their lifespans to the point where they now last only 1-2 weeks. Moreover, the best entries diffuse around their initial ranks, while the rest of the chart experiences heavy fluctuations. Using a clustering algorithm, I have identified and examined five different archetypes for the song trajectories. I found that certain archetypes were prevalent in the paths of songs on the Billboard Hot 100 at particular points in history. My work showed that up until the 1970s, songs disappeared quickly after reaching their peaks. Then, until the 1990s, they would gradually move from the bottom of the charts to the top and again to the bottom. Nowadays, they tend to start higher in the ranking and last significantly longer. These results provide characteristics of different time periods in the history of Billboard Hot 100. Lastly, I have defined and studied two extreme groups of artists - new artists and hitmakers. The study showed that fewer and fewer new artists get to enter the music charts, which are occupied mainly by constant hitmakers. However, the superstars have also seen a decline in performance - fewer achieve the top 10 of the rankings.

Some explanations for the changes in the dynamics include the winner-take-all and heavy-tail effects [9, 58, 70], more accessible access to music [9, 58], and the race for publicity and attention [41, 56, 78, 94]. The explanation is likely to be a complex mixture of these effects; to settle to what degree each can be attributed is outside the scope of this thesis but would be a valuable research goal to pursue.

All these results suggest that the industry has gotten more competitive and, in some ways, less predictable. While more research is required to answer the hypothesis fully, this study shows that creating a hit has become more challenging. Investigating possible causes of changes in the performance could help songwriters, record labels, and music magazines adjust to the new trends and attract more audiences.

Bibliography

[1] *40+ Fascinating Music Streaming Statistics (2023 Updated)*. 2023. URL: https://headphonesaddict.com/music-streaming-statistics/.

[2] Masato Abe. "Functional advantages of Lévy walks emerging near a critical point". In: *Proceedings of the National Academy of Sciences of the United States of America* 117 (Sept. 2020). DOI: 10.1073/pnas.2001548117.

[3] Luis Aguiar and Joel Waldfogel. "As Streaming Reaches Flood Stage, Does it Stimulate or Depress Music Sales?" In: *International Journal of Industrial Organization* 57 (July 2017). DOI: 10.1016/j.ijindorg.2017.06.004.

[4] Luis Aguiar and Joel Waldfogel. "Even the Losers Get Lucky Sometimes: New Products and the Evolution of Music Quality since Napster". In: *Information Economics and Policy* 34 (Jan. 2016). DOI: 10.1016/j.infoecopol.2015.12.003.

[5] Peter Alexander. "New technology and market structure: Evidence from the music recording industry". In: *Journal of Cultural Economics* 18 (Jan. 1994), pp. 113–123. DOI: 10.1007/BF01078934.

[6] Pantelis Analytis et al. "The Structure of Social Influence in Recommender Networks". In: Feb. 2020. DOI: 10.1145/3366423.3380020.

[7] Pantelis Pipergias Analytis et al. *Sequential choice and self-reinforcing rankings*. WorkingPaper 1819. Universitat Pompeu Fabra, Feb. 2022, p. 46.

[8] Noah Askin and Michael Mauskapf. "What Makes Popular Culture Popular?: Product Features and Optimal Differentiation in Music". In: *American Sociological Review* 82 (Apr. 2017). DOI: 10.1177/0003122417728662.

[9] Mary Benner and Joel Waldfogel. "The Song Remains the Same? Technological Change and Positioning in the Recorded Music Industry". In: *Strategy Science* 1 (Sept. 2016), pp. 129–147. DOI: 10.1287/stsc.2016.0012.

[10] Justin M. Berg. "One-Hit Wonders versus Hit Makers: Sustaining Success in Creative Industries". In: *Administrative Science Quarterly* 67.3 (2022), pp. 630–673. DOI: 10.1177/00018392221083650. eprint: https://doi.org/10.1177/00018392221083650. URL: https://doi.org/10.1177/00018392221083650.

[11] Sudip Bhattacharjee et al. "The Effect of Digital Sharing Technologies on Music Markets: A Survival Analysis of Albums on Ranking Charts". In: *Management* 53 (Sept. 2007). DOI: 10.1287/mnsc.1070.0699.

[12] Sudip Bhattacharjee et al. "Whatever Happened To Payola? An Empirical Analysis of Online Music Sharing". In: *Decision Support Systems* 42 (Oct. 2006), pp. 104–120. DOI: 10.1016/j.dss.2004.11.001.

[13] Wesam Bhaya. "Review of Data Preprocessing Techniques in Data Mining". In: *Journal of Engineering and Applied Sciences* 12 (Sept. 2017), pp. 4102–4107. DOI: 10.3923/jeasci.2017.4102.4107.

[14] Sifeng Bi, Matteo Broggi, and Michael Beer. "The role of the Bhattacharyya distance in stochastic model updating". In: *Mechanical Systems and Signal Processing* 117 (Feb. 2019), pp. 437–452. DOI: 10.1016/j.ymssp.2018.08.017.

[15] *Billboard Charts Legend*. Accessed: 2023-05-07. URL: https://www.billboard.com/billboard-charts-legend/.

[16] *Billboard Frequently Asked Questions*. Accessed: 2023-05-13. URL: https://www.billboard.com/frequently-asked-questions/.

[17] *Billboard Hot 100*. Accessed: 2023-01-03. URL: https://www.billboard.com/charts/hot-100/.

[18] Eric Bradlow and Peter Fader. “A Bayesian Lifetime Model for the ”Hot 100” Billboard Songs”. In: *Journal of the American Statistical Association* 96 (June 2001), pp. 368–381.
[19] Erik Brynjolfsson, Yu Hu, and Michael Smith. “Long Tails vs. Superstars: The Effect of Information Technology on Product Variety and Sales Concentration Patterns”. In: *Information Systems Research* 21 (Dec. 2010), pp. 736–747. DOI: 10.1287/isre.1100.0325.
[20] Andrzej Buda and Andrzej Jarynowski. “Exploring patterns in European singles charts”. In: (Mar. 2015).
[21] Keith Burghardt et al. “Origins of Algorithmic Instabilities in Crowdsourced Ranking”. In: *Proceedings of the ACM on Human-Computer Interaction* 4 (Oct. 2020), pp. 1–20. DOI: 10.1145/3415237.
[22] Cristian Candia-Castro et al. “The universal decay of collective memory and attention”. In: *Nature Human Behaviour* 3 (Jan. 2019). DOI: 10.1038/s41562-018-0474-5.
[23] Yuyu Chen, Hanming Fang, and Hongbin Cai. “Observational Learning: Evidence from a Randomized Natural Field Experiment”. In: *American Economic Review* 99 (May 2009), pp. 864–82. DOI: 10.1257/aer.99.3.864.
[24] SHC Choi, Sung-Hyuk Cha, and Charles Tappert. “A Survey of Binary Similarity and Distance Measures”. In: *J. Syst. Cybern. Inf.* 8 (Nov. 2009).
[25] Johan Chu and James Evans. “Slowed canonical progress in large fields of science”. In: *Proceedings of the National Academy of Sciences* 118 (Oct. 2021), e2021636118. DOI: 10.1073/pnas.2021636118.
[26] Aaron Clauset, Samuel Arbesman, and Daniel Larremore. “Systematic inequality and hierarchy in faculty hiring networks”. In: *Science Advances* 1 (Feb. 2015), e1400005–e1400005. DOI: 10.1126/sciadv.1400005.
[27] Aaron Clauset, Daniel Larremore, and Roberta Sinatra. “Data-driven predictions in the science of science”. In: *Science* 355 (Feb. 2017), pp. 477–480. DOI: 10.1126/science.aal4217.
[28] Manuel Francisco Coelho and José Mendes. “Digital music and the “death of the long tail””. In: *Journal of Business Research* 101 (Aug. 2019). DOI: 10.1016/j.jbusres.2019.01.015.
[29] Aoife Coffey. “The impact that music streaming services such as Spotify, Tidal and Apple Music have had on consumers, artists and the music industry itself”. MA thesis. University of Dublin, 2016.
[30] Alberto Cosimato et al. “The Conundrum of Success in Music: Playing it or Talking About it?” In: *IEEE Access* 7 (Aug. 2019), pp. 1–1. DOI: 10.1109/ACCESS.2019.2937743.
[31] William Crain and Robert Tollison. “Consumer Choice and the Popular Music Industry: A Test of the Superstar Theory”. In: *Empirica* 29 (Feb. 2002), pp. 1–9. DOI: 10.1023/A:1014651130414.
[32] David Doane and Lori Seward. “Measuring Skewness: A Forgotten Statistic?” In: *J. Stat. Educ.* 19 (July 2011). DOI: 10.1080/10691898.2011.11889611.
[33] Christian Essling, Johannes Koenen, and Christian Peukert. “Competition for Attention in the Digital Age: The Case of Single Releases in the Recorded Music Industry”. In: *Information Economics and Policy* (May 2017). DOI: 10.2139/ssrn.2444708.
[34] Santo Fortunato et al. “Science of science”. In: *Science* 359 (Mar. 2018), eaao0185. DOI: 10.1126/science.aao0185.
[35] Samuel Fraiberger et al. “Quantifying reputation and success in art”. In: *Science* 362 (Nov. 2018), eaau7224. DOI: 10.1126/science.aau7224.

[36] Ruth Garcia-Gavilanes et al. “Memory Remains: Understanding Collective Memory in the Digital Age”. In: *Science Advances* 3 (Sept. 2016). DOI: 10.1126/sciadv.1602368.
[37] Lisa George and Christian Peukert. “Youtube Decade: Cultural Convergence in Recorded Music”. In: (Sept. 2014). DOI: 10.5167/uzh-101018.
[38] Ram Gopal and G. Sanders. “Do Artist Benefit from Online Music Sharing?” In: *The Journal of Business* 79 (Feb. 2006), pp. 1503–1534. DOI: 10.1086/500683.
[39] *Hawes Publications: New York Times Adult Hardcover Best Seller List*. Accessed: 2023-04-21. URL: https://hawes.com/pastlist.htm.
[40] Ken Hendricks and Alan Sorensen. “Information and the Skewness of Music Sales”. In: *Journal of Political Economy* 117 (Apr. 2009), pp. 324–369. DOI: 10.1086/599283.
[41] Kenneth Hendricks, Alan Sorensen, and Thomas Wiseman. “Observational Learning and Demand for Search Goods”. In: *American Economic Journal: Microeconomics* 4 (Feb. 2012), pp. 1–31. DOI: 10.1257/mic.4.1.1.
[42] David Hesmondhalgh. “Is Music Streaming Bad for Musicians? Problems of Evidence and Argument”. In: *New Media and Society* 23 (Aug. 2020). DOI: 10.1177/1461444820953541.
[43] *Hitlisten.NU*. Accessed: 2023-01-04. URL: https://www.hitlisten.nu/.
[44] *Hot 100 Impacted by New On-Demand Songs Chart*. 2012. URL: https://www.billboard.com/music/music-news/hot-100-impacted-by-new-on-demand-songs-chart-502020/.
[45] Jiann-wien Hsu and Ding-wei Huang. “Correlation between impact and collaboration”. In: *Scientometrics* 86 (Feb. 2011), pp. 317–324. DOI: 10.1007/s11192-010-0265-x.
[46] Rob Hyndman. “Moving Averages”. In: Jan. 2010, pp. 866–869. DOI: 10.1007/978-3-642-04898-2_380.
[47] Oliver Ibe. “Diffusion Processes”. In: Aug. 2013, pp. 158–174. ISBN: 9781118618097. DOI: 10.1002/9781118618059.ch7.
[48] Gerardo Iñiguez et al. “Dynamics of ranking”. In: *Nature Communications* 13 (Mar. 2022), p. 1646. DOI: 10.1038/s41467-022-29256-x.
[49] Myra Interiano et al. “Musical trends and predictability of success in contemporary songs in and out of the top charts”. In: *Royal Society Open Science* 5 (May 2018), p. 171274. DOI: 10.1098/rsos.171274.
[50] *iTunes Charts*. Accessed: 2023-04-05. URL: http://www.itunescharts.net/us/charts/podcasts/current/.
[51] Milan Janosov, Federico Battiston, and Roberta Sinatra. “Success and luck in creative careers”. In: *EPJ Data Science* 9 (Dec. 2020). DOI: 10.1140/epjds/s13688-020-00227-w.
[52] Jim Jansen et al. “Twitter Power: Tweets as Electronic Word of Mouth”. In: *JASIST* 60 (Nov. 2009), pp. 2169–2188. DOI: 10.1002/asi.21149.
[53] Shuyi Ji et al. “Kullback-Leibler Divergence Metric Learning”. In: *IEEE Transactions on Cybernetics* PP (July 2020), pp. 1–12. DOI: 10.1109/TCYB.2020.3008248.
[54] Jonas L. Juul et al. “Fixed-time descriptive statistics underestimate extremes of epidemic curve ensembles”. In: *Nature Physics* 17.1 (Dec. 2020), pp. 5–8. DOI: 10.1038/s41567-020-01121-y. URL: https://doi.org/10.1038%2Fs41567-020-01121-y.
[55] Anton Killin. “The origins of music: Evidence, theory, and prospects”. In: *Music & Science* 1 (Feb. 2018), p. 205920431775197. DOI: 10.1177/2059204317751971.

[56] Christian Koch et al. “Collaborations on YouTube: From Unsupervised Detection to the Impact on Video and Channel Popularity”. In: *ACM Transactions on Multimedia Computing Communications and Applications* (Oct. 2018). DOI: 10.1145/3241054.
[57] Matthieu Komorowski et al. “Exploratory Data Analysis”. In: Sept. 2016, pp. 185–203. ISBN: 978-3-319-43740-8. DOI: 10.1007/978-3-319-43742-2_15.
[58] Tobias Kretschmer and Christian Peukert. “Video Killed the Radio Star? Online Music Videos and Recorded Music Sales”. In: *Information Systems Research* 31 (June 2020). DOI: 10.1287/isre.2019.0915.
[59] *Kworb.net*. Accessed: 2023-05-28. URL: https://kworb.net/charts/.
[60] Luca Lista. “Machine Learning”. In: *Statistical Methods for Data Analysis: With Applications in Particle Physics*. Cham: Springer International Publishing, 2023, pp. 225–276.
[61] Lu Liu et al. “Hot streaks in artistic, cultural, and scientific careers”. In: *Nature* 559 (July 2018). DOI: 10.1038/s41586-018-0315-8.
[62] Philipp Lorenz-Spreen et al. “Accelerating dynamics of collective attention”. In: *Nature Communications* 10 (Apr. 2019). DOI: 10.1038/s41467-019-09311-w.
[63] *Matplotlib’s Official Website*. URL: https://matplotlib.org/.
[64] Juan Montoro-Pons and Manuel Cuadrado-Garcia. “Live and prerecorded popular music consumption”. In: *Journal of Cultural Economics* 35 (Feb. 2011), pp. 19–48. DOI: 10.1007/s10824-010-9130-2.
[65] Morten Mørup and Lars Hansen. “Archetypal analysis for machine learning and data mining”. In: *Neurocomputing* 80 (Mar. 2012), pp. 54–63. DOI: 10.1016/j.neucom.2011.06.033.
[66] *Music through the Decades*. 2021. URL: https://www.bopdrop.com/post/music-through-the-decades.
[67] Mark Newman. “Power Laws, Pareto Distributions and Zipf’s Law”. In: *Contemporary Physics - CONTEMP PHYS* 46 (Dec. 2004). DOI: 10.1080/00107510500052444.
[68] *NumPy’s Official Website*. URL: https://numpy.org/.
[69] *Official Charts*. Accessed: 2023-01-03. URL: https://www.officialcharts.com/charts/singles-chart/.
[70] Andrea Ordanini and Joseph Nunes. “Fewer blockbusters and more superstars: How technological innovation has impacted convergence on the music charts”. In: *International Journal of Research in Marketing* 33 (Sept. 2015). DOI: 10.1016/j.ijresmar.2015.07.006.
[71] Paul Ormerod et al. “Social network markets: A new definition of the creative industries”. In: *Journal of Cultural Economics* 32 (Feb. 2008), pp. 167–185. DOI: 10.1007/s10824-008-9066-y.
[72] *Pandas’ Official Website*. URL: https://pandas.pydata.org/.
[73] Shraddha Pandit and Suchita Gupta. “A Comparative Study on Distance Measuring Approaches for Clustering”. In: *International Journal of Research in Computer Science* 2 (Dec. 2011), p. 29. DOI: 10.7815/ijorcs.21.2011.011.
[74] Bang-Dang Pham, Minh-Triet Tran, and Hoang-Long Pham. “Hit Song Prediction based on Gradient Boosting Decision Tree”. In: Nov. 2020, pp. 356–361. DOI: 10.1109/NICS51282.2020.9335886.
[75] Má rton Pósfai and Raissa M. D’Souza. “Talent and experience shape competitive social hierarchies”. In: *Physical Review E* 98.2 (Aug. 2018). DOI: 10.1103/physreve.98.020302. URL: https://doi.org/10.1103%2Fphysreve.98.020302.
[76] *PyPDF2’s Official Github*. URL: https://github.com/py-pdf/pypdf.

[77] Filippo Radicchi. “In science “there is no bad publicity”: Papers criticized in comments have high scientific impact”. In: *Scientific reports* 2 (Nov. 2012), p. 815. DOI: 10.1038/srep00815.

[78] Nicolas Ruth and Benedikt Spangardt. “Research trends on music and advertising”. In: *Revista Mediterránea de Comunicación* 8 (July 2017). DOI: 10.14198/MEDCOM2017.8.2.1.

[79] *Ryan Tedder: Classic songs are strangling new music*. 2021. URL: https://www.bbc.com/news/entertainment-arts-58329477.

[80] Matthew Salganik, Peter Dodds, and Duncan Watts. “Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market”. In: *Science (New York, N.Y.)* 311 (Feb. 2006), pp. 854–6. DOI: 10.1126/science.1121066.

[81] Rinaldo Schinazi. *Probability with Statistical Applications*. Jan. 2012. ISBN: 978-0-8176-8249-1. DOI: 10.1007/978-0-8176-8250-7.

[82] *Scikit-learn User Guide: K-means*. Accessed: 2023-05-14. URL: https://scikit-learn.org/stable/modules/clustering.html#k-means.

[83] *Scikit-learn User Guide: Non-negative matrix factorization (NMF or NNMF)*. Accessed: 2023-05-15. URL: https://scikit-learn.org/stable/modules/decomposition.html#nmf.

[84] *Scikit-learn's Official Website*. URL: https://scikit-learn.org/stable/.

[85] *Scipy's Official Website*. URL: https://scipy.org/.

[86] *Scraping Data from Dynamic Websites with Selenium and Python*. 2022. URL: https://medium.com/@traviscroyce/scraping-data-from-dynamic-websites-with-selenium-and-python-f702bb534974.

[87] *Scrapy Documentation: Spiders*. Accesssed: 2023-05-16. URL: https://docs.scrapy.org/en/latest/topics/spiders.html.

[88] *Scrapy's Official Website*. URL: https://scrapy.org/.

[89] *Seaborn's Official Website*. URL: https://seaborn.pydata.org/.

[90] Vedran Sekara et al. “The Chaperone Effect in Scientific Publishing”. In: *Proceedings of the National Academy of Sciences* 115 (Dec. 2018), p. 201800471. DOI: 10.1073/pnas.1800471115.

[91] Diana Sietz et al. “Archetype analysis in sustainability research: Methodological portfolio and analytical frontiers”. In: *ECOLOGY AND SOCIETY* 24 (July 2019), p. 34. DOI: 10.5751/ES-11103-240334.

[92] Roberta Sinatra et al. “Quantifying the evolution of individual scientific impact”. In: *Science* 354 (Nov. 2016), aaf5239–aaf5239. DOI: 10.1126/science.aaf5239.

[93] *The Ledger: Are There Really 100,000 New Songs Uploaded a Day? Maybe More*. 2023. URL: https://www.billboard.com/pro/how-much-music-added-spotify-streaming-services-daily/.

[94] Corliss Thornton and Janée Burkhalter. “Must be the Music: Examining the Placement Effects of Character-Brand Association and Brand Prestige on Consumer Brand Interest within the Music Video Context”. In: *Journal of Promotion Management* 21 (Feb. 2015), pp. 126–141. DOI: 10.1080/10496491.2014.971212.

[95] *Tqdm's Official Github*. URL: https://github.com/tqdm/tqdm.

[96] *US Singles Top 100*. Accessed: 2023-01-03. URL: https://acharts.co/us_singles_top_100.

[97] Joel Waldfogel and Imke Reimers. “Storming the gatekeepers: Digital Disintermediation in the market for books”. In: *Information Economics and Policy* 31 (Mar. 2015). DOI: 10.1016/j.infoecopol.2015.02.001.

[98] Oliver Williams, Lucas Lacasa, and Vito Latora. “Quantifying and predicting success in show business”. In: *Nature Communications* 10 (June 2019). DOI: 10.1038/s41467-019-10213-0.
[99] Li-Chia Yang et al. “Revisiting the problem of audio-based hit song prediction using convolutional neural networks”. In: Mar. 2017, pp. 621–625. DOI: 10.1109/ICASSP.2017.7952230.
[100] Lang-Chi Yu et al. “Hit Song Prediction for Pop Music by Siamese CNN with Ranking Loss”. In: (Oct. 2017).
[101] Burcu Yucesoy et al. “Success in books: a big data approach to bestsellers”. In: *EPJ Data Science* 7 (Dec. 2018). DOI: 10.1140/epjds/s13688-018-0135-y.
[102] Hugh Zehr. “An Economic Analysis of the Effects of Streaming on the Music Industry in Response to Criticism from Taylor Swift”. In: *Major Themes in Economics* 23 (2021), pp. 51–63.
[103] Bo Zhao. “Web Scraping”. In: May 2017, pp. 1–3. ISBN: 978-3-319-32001-4. DOI: 10.1007/978-3-319-32001-4_483-1.

A Datasets

A.1 Billboard Hot 100

week	artist	song name	position	last week's position	peak position	weeks on chart
1959-01-12	The Chipmunks With David Seville	The Chipmunk Song	1	1.0	1	7
1959-01-05	The Chipmunks With David Seville	The Chipmunk Song	1	1.0	1	6
1959-01-12	The Platters	Smoke Gets In Your Eyes	2	2.0	2	9
1959-01-12	Connie Francis	My Happiness	3	6.0	3	6
1959-01-12	Billy Grammer	Gotta Travel On	4	9.0	4	8
1959-01-12	The Teddy Bears	To Know Him, Is To Love Him	5	3.0	1	17
1959-01-12	Fats Domino	Whole Lotta Loving	6	10.0	6	9
1959-01-12	Clyde McPhatter	A Lover's Question	7	8.0	7	14

Table A.1: **First few rows of the Billboard Hot 100 dataset.**

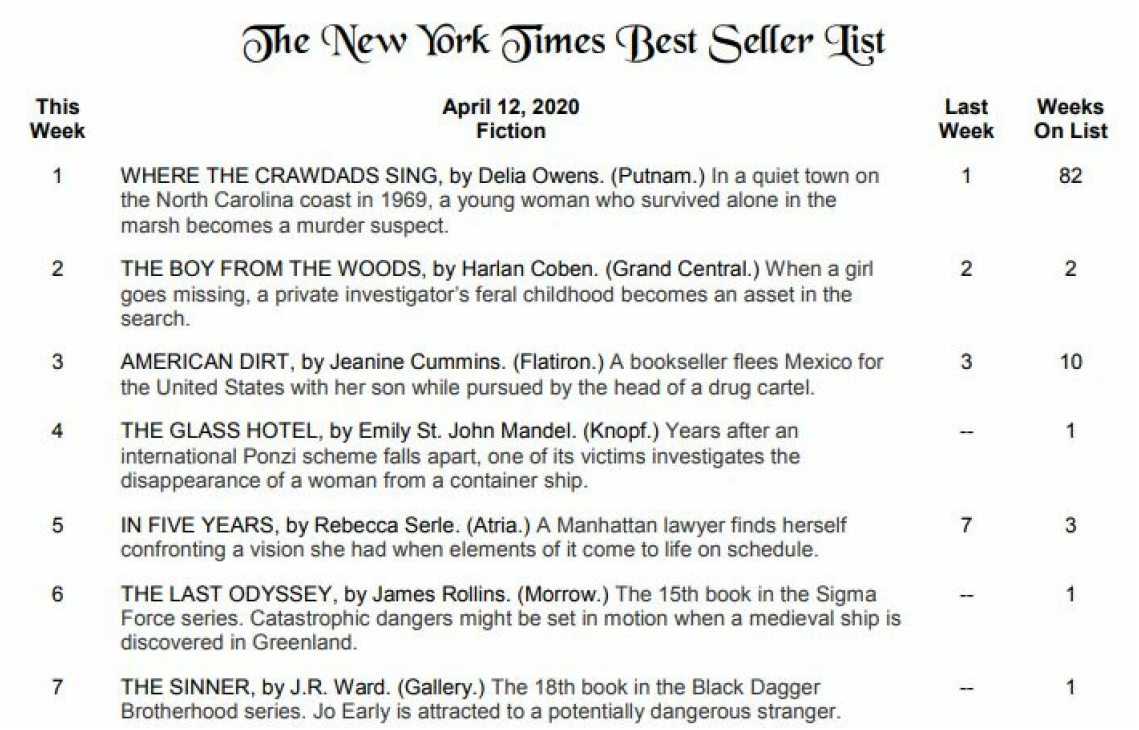
A.2 iTunes Podcasts

day	title	position
2009-11-26	The Twilight Saga: New Moon	1
2009-11-26	This American Life	2
2009-11-26	Celebrity Playlist Podcast	3
2009-11-26	NPR: Wait Wait... Don't Tell Me! Podcast	4
2009-11-26	Stuff They Don't Want You To Know	5
2009-11-26	NPR: Fresh Air Podcast	6

Table A.2: **The first few rows of the iTunes podcasts chart dataset.**

Figure A.1: **Example of the iTunes podcasts chart from 16th May 2023.**
The picture shows an example podcasts chart from the iTunes Charts website. It contains the title, position, days on the chart, and whether a song has changed the position from last week. (Screenshot taken from the iTunes Charts website.)

A.3 New York Times best sellers



The New York Times Best Seller List

This Week	April 12, 2020 Fiction	Last Week	Weeks On List
1	WHERE THE CRAWDADS SING, by Delia Owens. (Putnam.) In a quiet town on the North Carolina coast in 1969, a young woman who survived alone in the marsh becomes a murder suspect.	1	82
2	THE BOY FROM THE WOODS, by Harlan Coben. (Grand Central.) When a girl goes missing, a private investigator's feral childhood becomes an asset in the search.	2	2
3	AMERICAN DIRT, by Jeanine Cummins. (Flatiron.) A bookseller flees Mexico for the United States with her son while pursued by the head of a drug cartel.	3	10
4	THE GLASS HOTEL, by Emily St. John Mandel. (Knopf.) Years after an international Ponzi scheme falls apart, one of its victims investigates the disappearance of a woman from a container ship.	--	1
5	IN FIVE YEARS, by Rebecca Serle. (Atria.) A Manhattan lawyer finds herself confronting a vision she had when elements of it come to life on schedule.	7	3
6	THE LAST ODYSSEY, by James Rollins. (Morrow.) The 15th book in the Sigma Force series. Catastrophic dangers might be set in motion when a medieval ship is discovered in Greenland.	--	1
7	THE SINNER, by J.R. Ward. (Gallery.) The 18th book in the Black Dagger Brotherhood series. Jo Early is attracted to a potentially dangerous stranger.	--	1

Figure A.2: **Example of the New York Times best seller chart from 12th April 2020.** The picture shows an example New York Times best seller chart for fiction. The data is saved in a PDF format and contains the position, last week's position, and weeks on the chart. Moreover, it contains information on the author, title, publishing company, and comments about the book. (Screenshot taken from the Hawes website.)

author_title	week	position	last week's position	weeks on chart
ACT ONE, by Moss Hart. (Random House.)	1960-01-03	1	1.0	14
BARUCH: MY OWN STORY, by Bernard M. Baruch. (Henry Holt and Company.)	1958-01-05	1	1.0	18
PROFILES IN COURAGE, by John F. Kennedy. (Harper and Brothers.)	1964-01-05	1	1.0	114
THE RISE AND FALL OF THE THIRD REICH, by William L. Shirer. (Simon and Schuster.)	1961-01-01	1	1.0	9
THIS IS MY GOD, by Herman Wouk. (Doubleday and Company.)	1960-01-03	2	2.0	13
FOLK MEDICINE, by Deforest Clinton Jarvis. (Henry Holt and Company.)	1960-01-03	3	3.0	38

Table A.3: **The first few rows of The New York Times best seller charts for non-fiction.**

B Results for other datasets

B.1 Podcasts

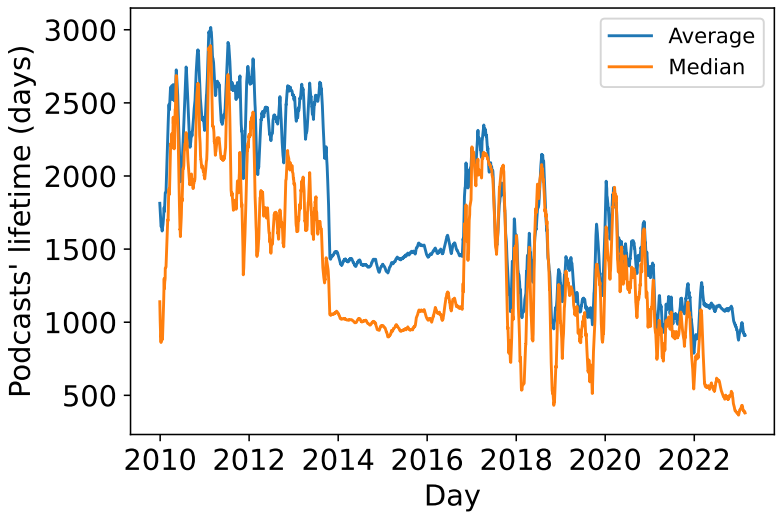


Figure B.1: **Lifetime of podcasts on the chart.**
The plot shows podcasts' average (blue) and median (orange) lifetimes. The values have been smoothed with a 30-day rolling average. On average, podcasts last way longer than songs - sometimes for more than seven years. There is a decreasing trend in the lifetime. Podcasts last, on average, even six times shorter than at the chart's beginning. An anomaly started around late 2013, when the average and median lifespans stabilized around 1500 and 1000, respectively. 25th September 2013 was when the chart changed from top 10 to top 100, which might have contributed to the long stabilization period.

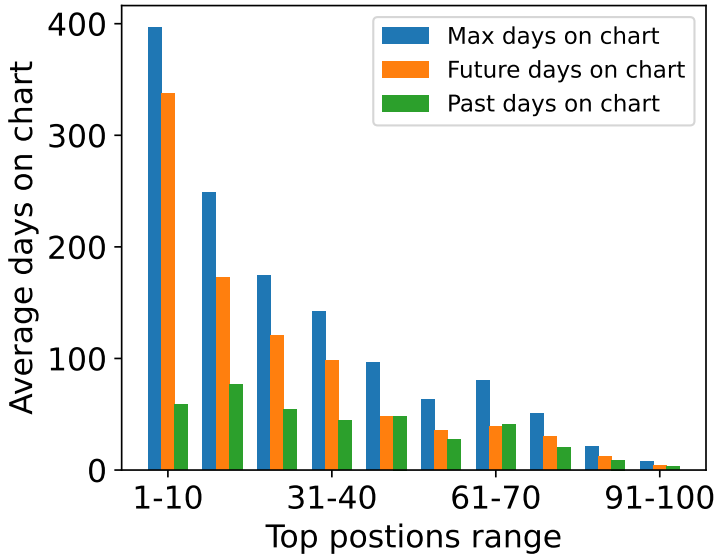


Figure B.2: **Average days on the chart before and after reaching the top position.** The bins represent the average weeks on the chart before the peak (blue) and after the peak (orange). The future days on the chart highly dominate the top positions. Podcasts linger on for a very long time after reaching their peak. This might indicate that the podcast industry is less competitive than the music industry.

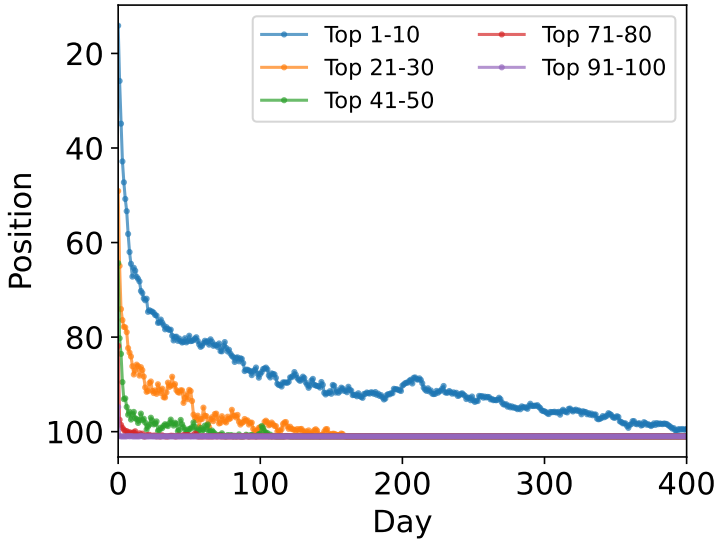


Figure B.3: **Normalized trajectories of podcasts from top positions.** The chosen top positions are top 10 (blue), 21-30 (orange), 41-50 (green), 71-80 (red), and 91-100 (purple) for podcasts after 25th September 2013. Only a few position ranges were chosen for better readability. The podcasts that disappear from the charts have the rest of their trajectories filled with 101. Podcasts tend to start at better positions and then linger on in the charts. There is no climbing up average flow as in the songs.

B.2 Books

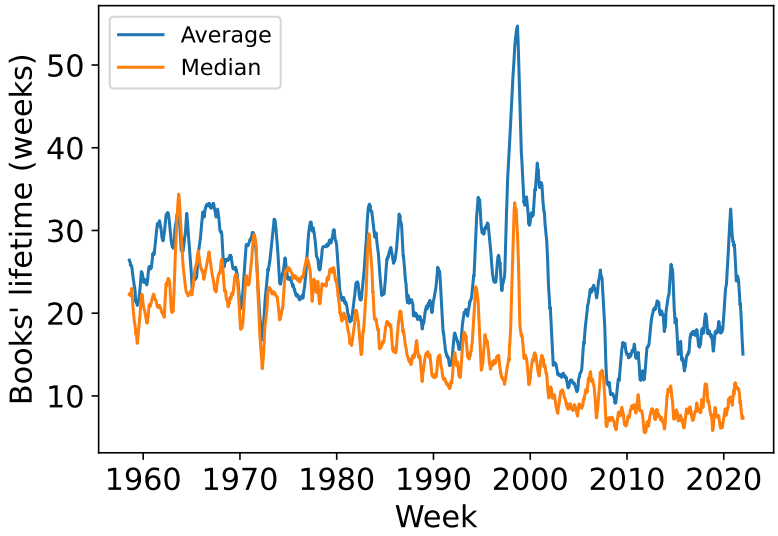


Figure B.4: **Lifetime of books on the chart.**
The plot shows non-fiction New York Times best sellers' average (blue) and median (orange) lifetimes. The values have been smoothed with a half-year rolling average. The average and median lifetimes have decreased over time, except for a sudden peak around 2000. Nowadays, books last, on average, 20 weeks, compared to 50 weeks during the peak. The median lifetime is below ten weeks.

Figure B.5: **Average weeks on the chart before and after reaching the top position.** The bins represent the average weeks on the chart before the peak (blue) and after the peak (orange). In the last decades, books tend to stay way shorter on the charts, even for the top positions. Moreover, books usually take longer to climb to their peak positions. Nowadays, they disappear quickly after reaching their top. This shows that the book industry might be more competitive than podcasts and songs.

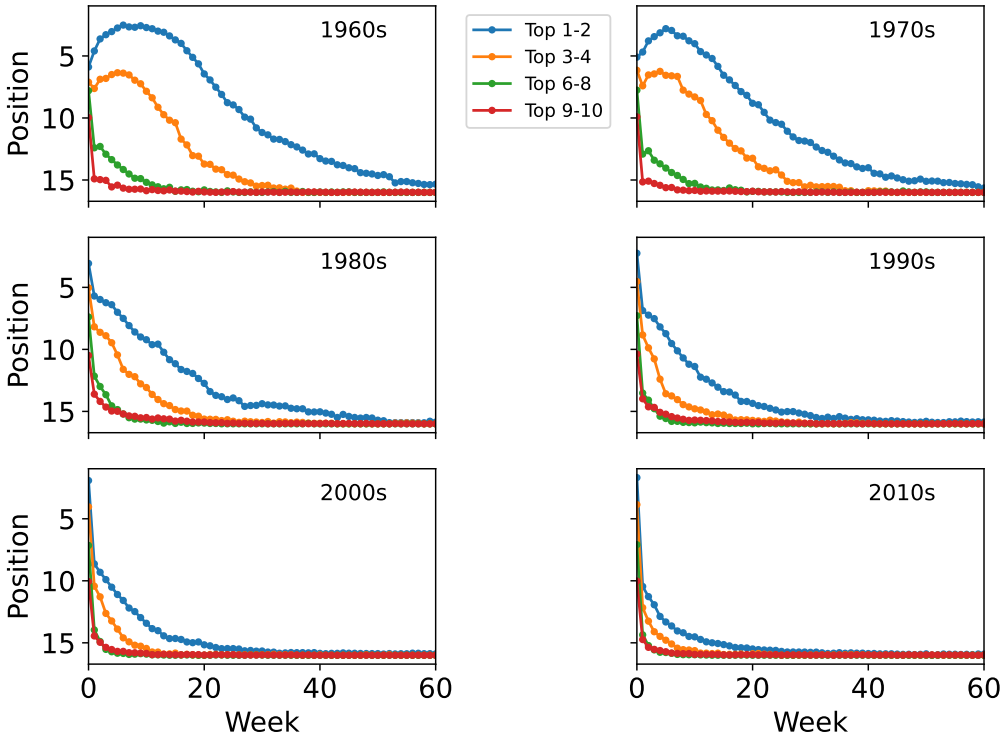


Figure B.6: **Normalized trajectories of books from top positions.**
The chosen top positions are top 1-2 (blue), 3-4 (orange), 6-8 (green), 9-10 (red), and 91-100 (purple) for books in different decades. Before, the books would climb to their top positions and quickly decay. Now, they seem to drop drastically after just a few weeks.

C Code samples

C.1 Billboard Hot 100 spider

```
base_url = "https://www.billboard.com/charts/hot-100/"
first_week = datetime(1958, 8, 4)
last_week = datetime(2023, 1, 6)

class BillboardSpider(scrapy.Spider):
    name = "billboardhot100"

    def generate_weeks(self):
        current_week = first_week

        while current_week < last_week:
            yield current_week.strftime("%Y-%m-%d")
            current_week += timedelta(weeks=1)

    def start_requests(self):
        for week in self.generate_weeks():
            yield scrapy.Request(url=urljoin(base_url, week), callback=self.
                parse)

    def parse(self, response):
        date = response.url.split("/")[-2]

        hits_containers = response.css('div.o-chart-results-list-row-container
            ')
        for hit_container in hits_containers:

            yield {
                'first_day_of_the_week': date,
                'artist': hit_container.css('span.c-label::text').getall()
                    [-7].strip(),
                'song_name': hit_container.css('h3.c-title::text').get().strip
                    (),
                'position': hit_container.css('span.c-label::text').get().
                    strip(),
                'last_week_position': hit_container.css('span.c-label::text').
                    getall()[-3].strip(),
                'peak_position': hit_container.css('span.c-label::text').
                    getall()[-2].strip(),
                'weeks_on_chart': hit_container.css('span.c-label::text').
                    getall()[-1].strip()
            }
```

Listing C.1: **Code for the Billboard Hot 100 spider.**

D Position change distribution

D.1 Pearson's first and second skewness

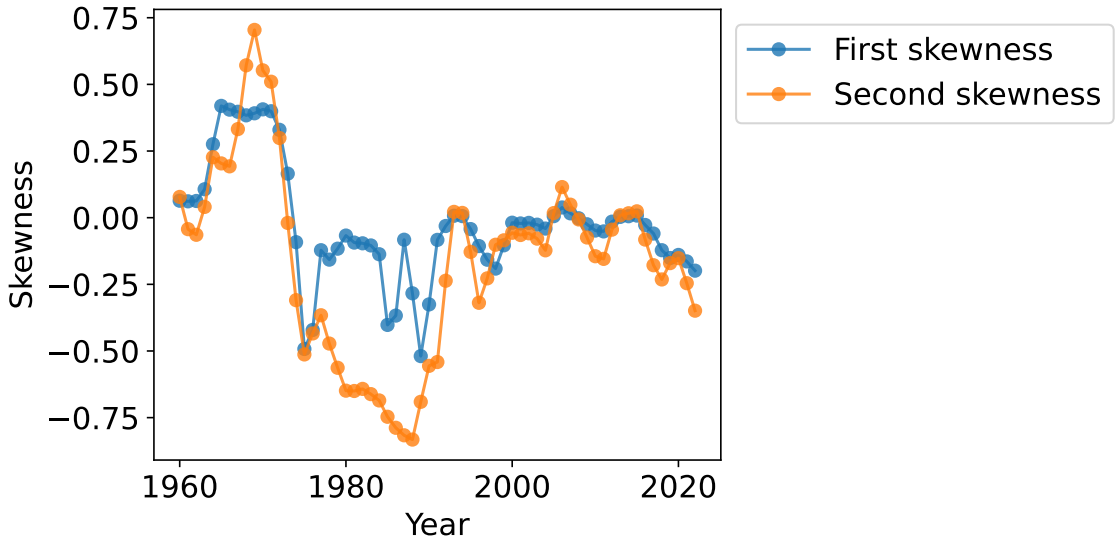


Figure D.1: **Pearson's first and second skewness.**
The first skewness (blue) was positive till around the 1980s and then fluctuated around zero. The second skewness (orange) resembles similar features, but the fluctuations were more sharp.

E Trajectory analysis

E.1 Average and median trajectories

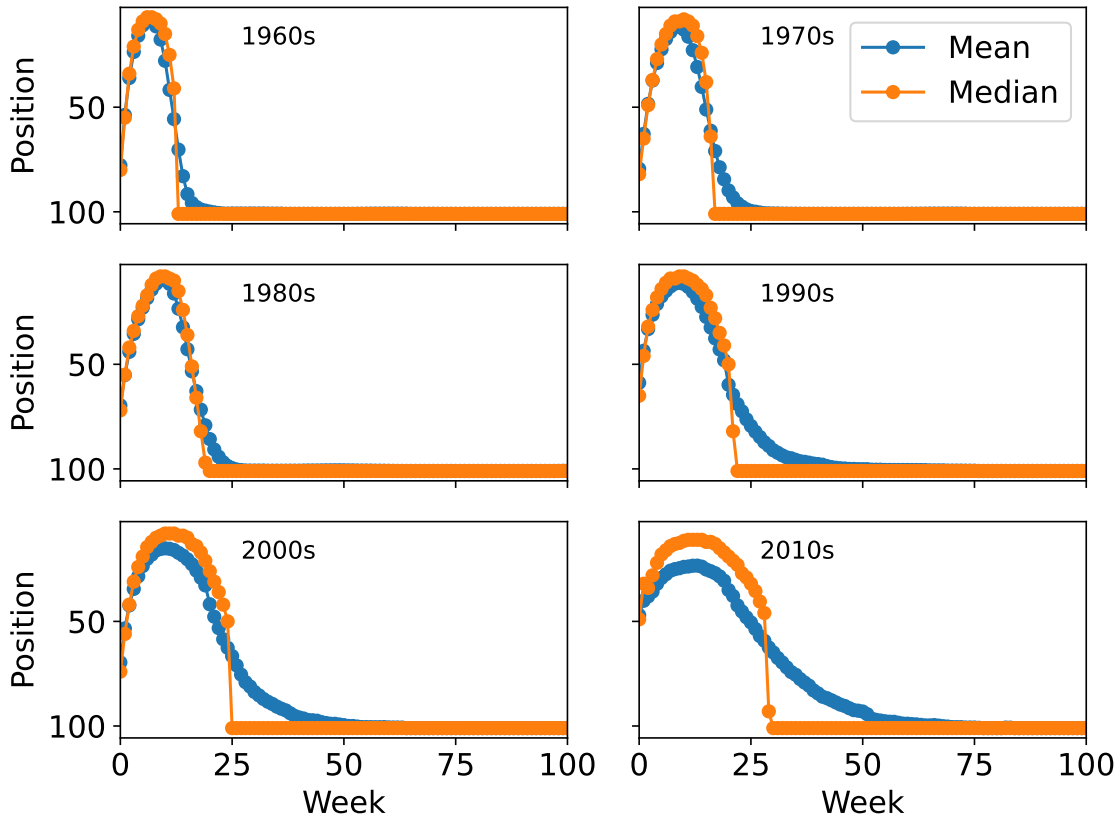


Figure E.1: **Average and median trajectories for top 10 over decades.**
The plots show the average (blue) and median (orange) trajectories of the top 10 songs over decades. The averages get wider and start higher over the years. The same can be spotted for the median, though they started to drop to 101 around 20 weeks drastically.

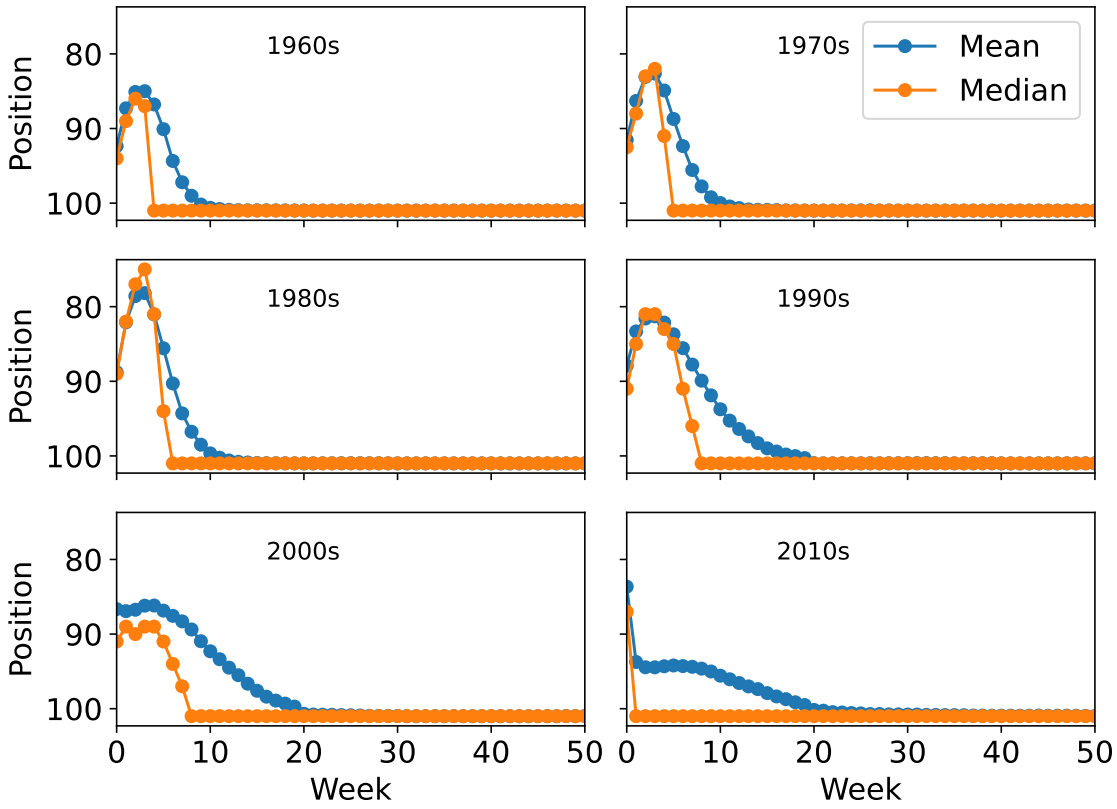


Figure E.2: **Average and median trajectories for sons that never reached top 50 over decades.**
The plots show the average (blue) and median (orange) trajectories of the bottom 50-100 songs over decades. The averages start higher and flatten more quickly over the years. The same can be spotted for the median. In the 2010s, the median dropped to 101 after one week.

E.2 Start and end positions

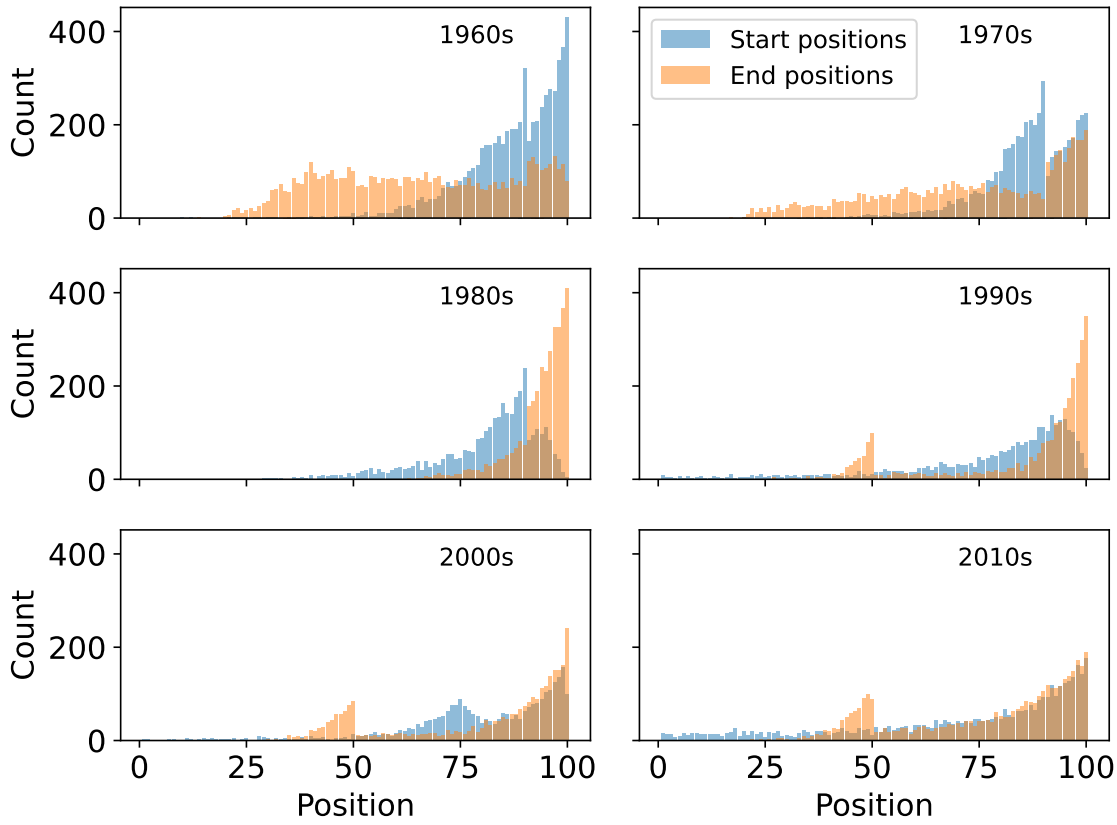


Figure E.3: **Distribution of start and end positions.**
The plots show distributions of start (blue) and end (orange) positions over decades. The heights in the distribution of start positions get shifted to the left. For the end positions, the opposite is observed.

E.3 Hamming distance

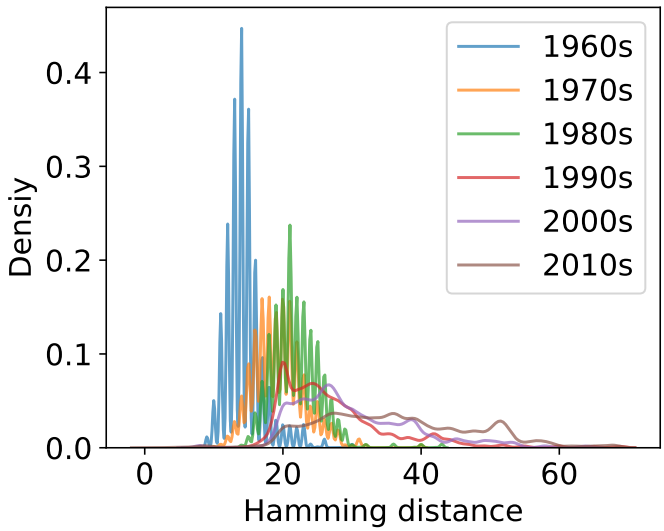


Figure E.4: **KDE of Hamming distances between trajectories for the top 1-10.** The plots are noisier than for the Manhattan distance. The hamming distance was not the right measure in the case of song trajectories, as rarely do two songs have identical paths on the charts.

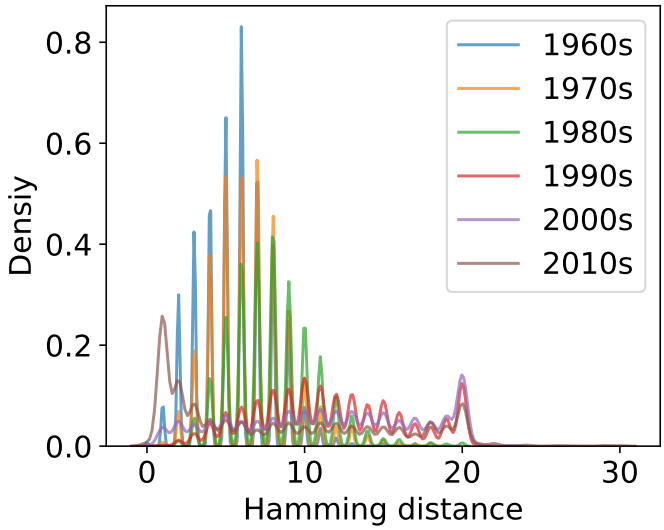


Figure E.5: **KDE of Hamming distances between trajectories for the top 50-100.** The plots are nearly impossible to interpret due to the noise. The hamming distance was not the right measure in the case of song trajectories, as rarely do two songs have identical paths on the charts.

E.4 Clustering

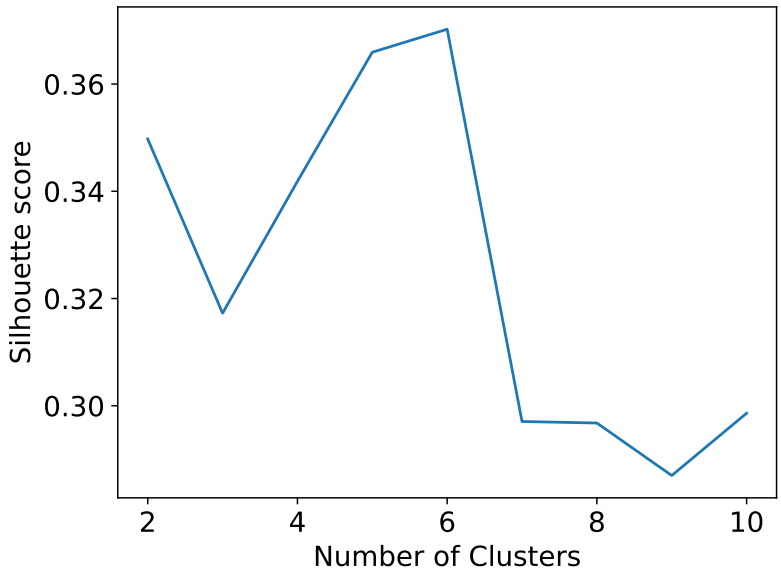


Figure E.6: **Silhouette scores for different numbers of clusters.**
The best score can has been achieved for 6 clusters, but for $k = 5$, it is nearly as high.

Cluster	75th percentile for weeks on chart	75th percentile for first position	75th percentile for last positions
Cluster 0	7	97	97
Cluster 1	29	93	52
Cluster 2	20	90	98
Cluster 3	13	90	58
Cluster 4	6	57	92
Cluster 5	27	43	85

Table E.1: **Statistics for 6 clusters.**
The statistics have been calculated for songs from different clusters. The values differ for various clusters, but some are similar in at least two columns. For example, clusters 1 and 3 both start at worse positions and end up at positions around 50. Songs from clusters 0 and 4 have a timespan of around a few weeks and end in the lowest positions (around 90). That is why, identifying different archetypes is not straightforward from this table.

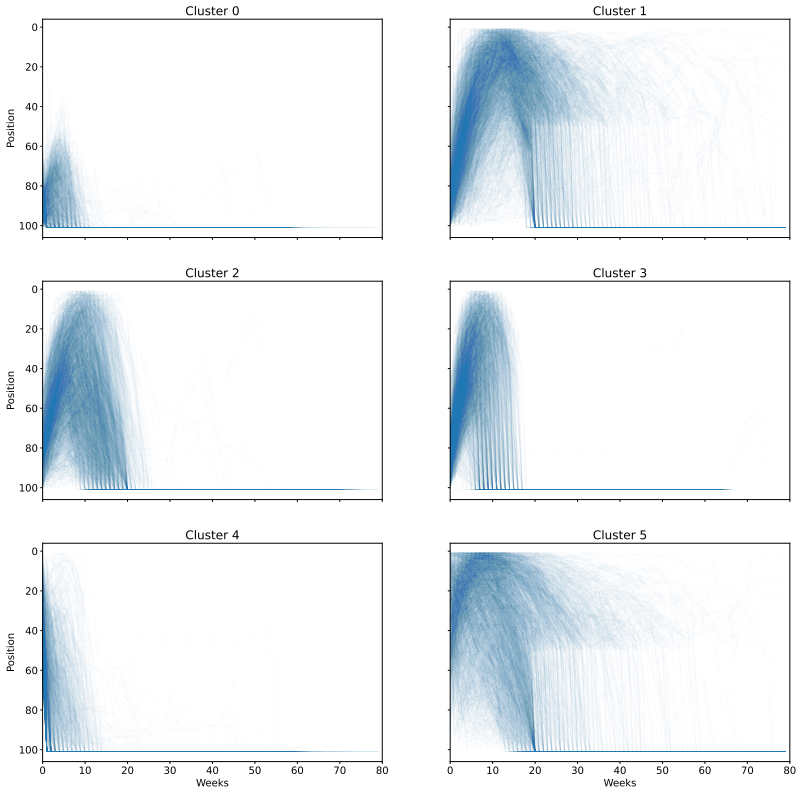


Figure E.7: **Some of the trajectories for 6 clusters.**
The trajectories were drawn for 1000 songs from each cluster. The curves were plotted with certain transparency to see the overlap in some areas. Clusters 1 and 5 seem to have many overlapping trajectories. That is why 5 was chosen for a number of archetypes.

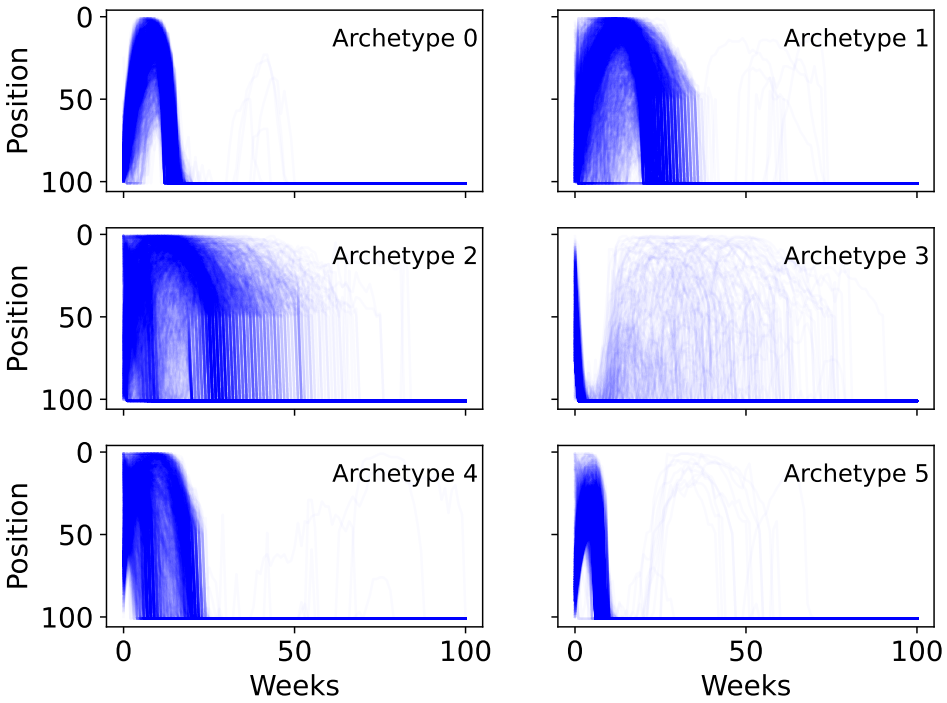


Figure E.8: **Archetypes detected by the NMF algorithm.**
The clusters are not as clear as in the k-means method. For example, archetype 3 seems like a high-start group but contains some long-lasting songs. The NMF method fails to identify similar trajectories.

F Ranking dynamics

F.1 Displacement probability

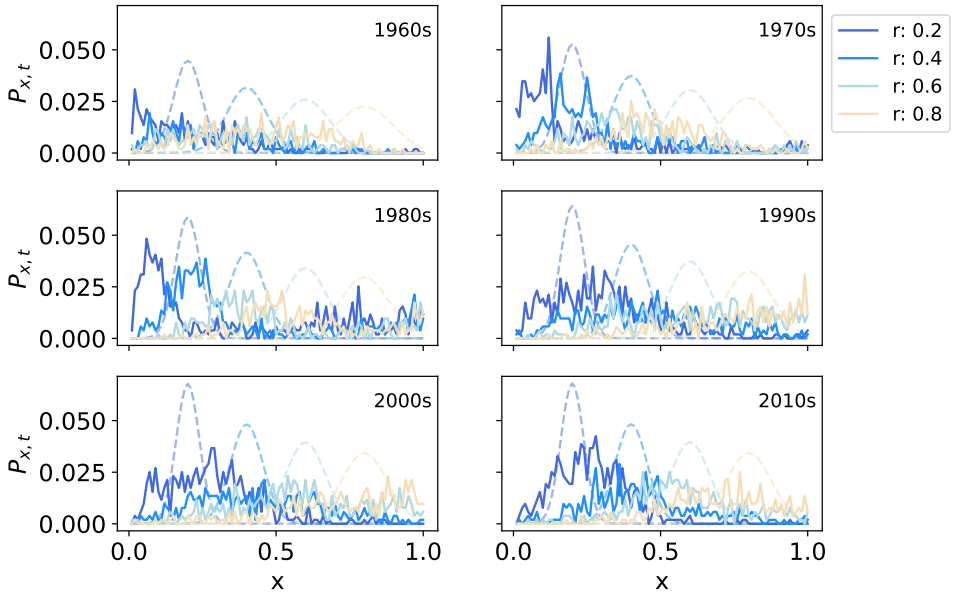


Figure F.1: **Displacement probability over decades.**
Probabilities were counted for initial ranks 20 (dark blue), 40 (blue), 60 (light blue), and 80 (yellow) for $t = 4$. The dashed lines indicate the $P_{x,t}$ of the model described by Iñiguez et al. [48]. For higher time intervals, the probabilities are less densely populated and more difficult to interpret.

G Hitmakers

G.1 Top positions of selected famous performers

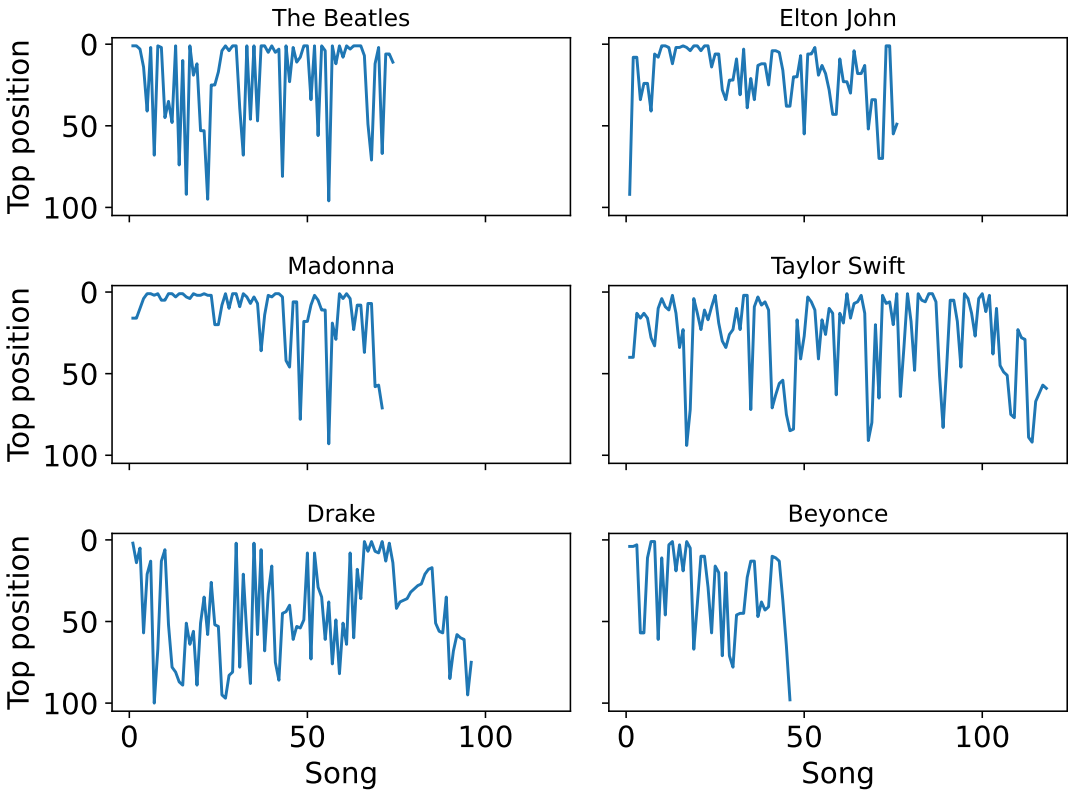


Figure G.1: **Top positions of songs by famous artists.**
The x-axis represents the nth song of the artists, and the y-axis its top position. Some of the most famous performers before the 1990s include The Beatles, Elton John, and Madonna. Many of their songs reached the top 1 and rarely achieved worse positions than 50. For some popular performers after the 1990s - Taylor Swift, Drake, and Beyonce - there are more frequent fluctuations in their top positions. Some of their songs did not even reach top 80.

Technical
University of
Denmark

Richard Petersens Plads, Building 324
2800 Kgs. Lyngby
Tlf. 4525 1700

https://www.compute.dtu.dk/