DATA201/422 Group Project Report

Ryan Moore, Will Durkin, Zheyu Li

Objective

What makes an album successful?

- What era has the most critically acclaimed albums?
- What makes an album critically acclaimed?
- What makes an album sell?

Project Management

For ease of collaboration, we used Git and GitHub version control software. This made it easy to code together during project development.

‘brief.ipbynb’ is the provided project brief.
‘diary.ipynb’ is the group diary.
‘outline.ipynb’ is a summary of the topics to cover in the presentation.
‘project.ipynb’ is the master copy of the notebook, containing functions, graphs, model, etc.

Data Sources

Rate Your Music - https://rateyourmusic.com

Rate Your Music is a website that ranks albums based on a crowdsourced approach. The ratings from users, along with the number of ratings, are aggregated to produce a theoretically 'objective' assessment of an album's quality. We used this dataset extensively, taking Rate Your Music as our benchmark for determining a ‘highly rated’ album.

Best Selling Albums - https://bestsellingalbums.org/

Best Selling Albums is a dataset that showcases historical album sales data, breaking down the units sold by decade. We utilized this dataset to examine if there were any common traits among the best-selling albums that could suggest causation. For instance, we analyzed each album's 'energy' levels to determine if there was a consistent trend in energy levels associated with highly successful albums.

Spotify API - https://developer.spotify.com/documentation/web-api

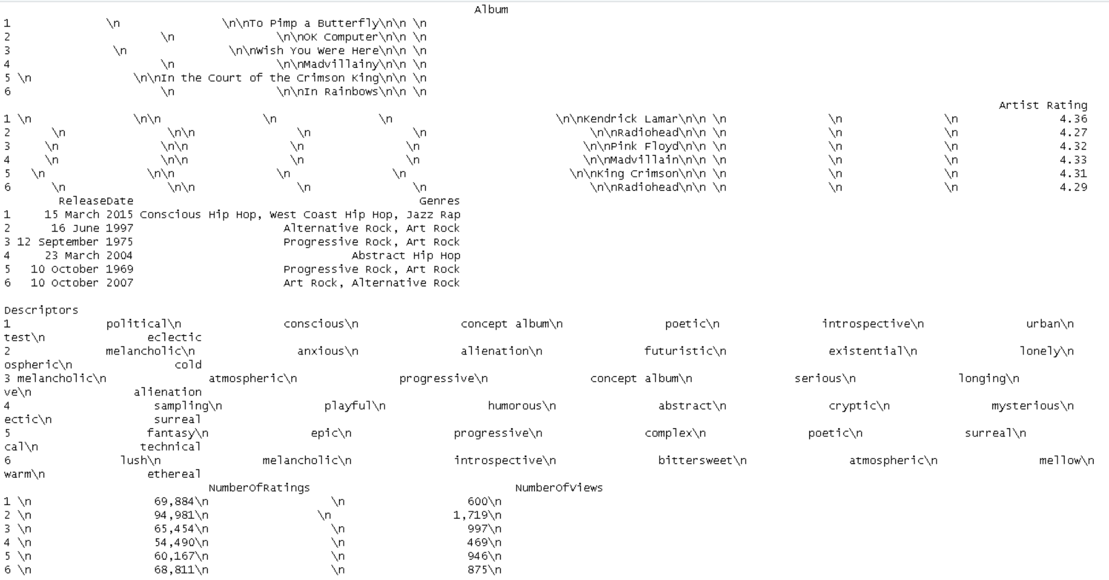
The Spotify API contains many endpoints which make it possible to query almost any data you could think of from within Spotify. This includes artists, tracks, playlists, users, albums and more. Spotify also does an internal audio analysis of all tracks in its library, in order to recommend tracks to users. Conveniently, this audio analysis is free and publicly available through Spotify's API. The audio analysis contains many values that quantify the track, such as acousticness, danceability, loudness and energy. From the Spotify API: "Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy." We used energy as an attribute of an album to see whether there was any correlation between the energy of an album and its sales, for example, or whether there was little correlation. One hypothesis is that there is a 'sweet spot' of energy wherein an album is more likely to sell well. For example, albums with very high energy (heavy metal, noise music) are not appealing to the masses, and neither are albums with very low energy (ambient music, etc.)

Data Collection and Cleaning

Rate Your Music

The rate your music website didn't have an API. That meant we had to scrape the data from the website. First, we scape the first page of data in Rate Your Music, using the rvest package. By targeting specific CSS selectors, we could efficiently extract various details such as album names, artists, ratings, release dates, genres, descriptors, number of ratings, and number of views. All this data is then neatly organized and stored in a data frame, laying the groundwork for subsequent analysis.

After we collected the data, it appeared as follows:

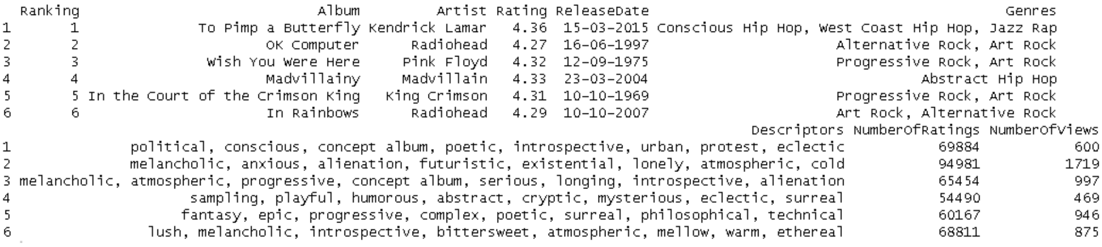


```
                                                                     Album
1              \n               \n\nTo Pimp a Butterfly\n\n \n
2                      \n               \n\nOK Computer\n\n \n
3               \n               \n\nWish You Were Here\n\n \n
4                      \n               \n\nMadvillainy\n\n \n
5 \n               \n\nIn the Court of the Crimson King\n\n \n
6                      \n               \n\nIn Rainbows\n\n \n
                                                                                                                                                 Artist Rating
1 \n               \n\n               \n               \n                        \n\nKendrick Lamar\n\n \n              \n               \n               4.36
2      \n               \n\n               \n               \n                        \n\nRadiohead\n\n \n              \n               \n               4.27
3     \n               \n\n               \n               \n                        \n\nPink Floyd\n\n \n              \n               \n               4.32
4     \n               \n\n               \n               \n                        \n\nMadvillain\n\n \n              \n               \n               4.33
5   \n               \n\n               \n               \n                        \n\nKing Crimson\n\n \n              \n               \n               4.31
6      \n               \n\n               \n               \n                        \n\nRadiohead\n\n \n              \n               \n               4.29
        ReleaseDate                                          Genres
1     15 March 2015 Conscious Hip Hop, West Coast Hip Hop, Jazz Rap
2      16 June 1997                      Alternative Rock, Art Rock
3 12 September 1975                      Progressive Rock, Art Rock
4     23 March 2004                                Abstract Hip Hop
5   10 October 1969                      Progressive Rock, Art Rock
6   10 October 2007                      Art Rock, Alternative Rock

Descriptors
1              political\n               conscious\n               concept album\n               poetic\n               introspective\n              urban\n
test\n               eclectic
2              melancholic\n               anxious\n               alienation\n               futuristic\n               existential\n              lonely\n
ospheric\n               cold
3 melancholic\n               atmospheric\n               progressive\n               concept album\n               serious\n              longing\n
ve\n               alienation
4                     sampling\n               playful\n               humorous\n               abstract\n               cryptic\n              mysterious\n
ectic\n               surreal
5                    fantasy\n               epic\n               progressive\n               complex\n               poetic\n              surreal\n
cal\n               technical
6                lush\n               melancholic\n               introspective\n               bittersweet\n               atmospheric\n              mellow\n
warm\n               ethereal
                              NumberOfRatings                              NumberOfViews
1 \n                  69,884\n                  \n                  600\n
2 \n                  94,981\n                \n                  1,719\n
3 \n                  65,454\n                  \n                  997\n
4 \n                  54,490\n                  \n                  469\n
5 \n                  60,167\n                  \n                  946\n
6 \n                  68,811\n                  \n                  875\n
```

Therefore, we needed to clean this data. After scraping various details from the webpage, such as album names, artists, and release dates, several data-cleaning steps are applied. The extracted text undergoes trimming using the str_trim() function to remove leading or trailing white spaces. Ratings are filtered to avoid potential duplicates. Strings representing ratings, the number of ratings, and views are first stripped of commas for accuracy and then converted to a numeric format. Release dates are formatted consistently to ensure that they are handled as date objects in subsequent analyses. Multiple genres for a single album are fetched and merged into a single comma-separated string for streamlined representation. Lastly, descriptors are cleaned to replace newline characters and remove extraneous spaces. These meticulous steps are taken to ensure that the extracted data is accurate, consistent, and primed for further analysis.

Data is cleaned as shown:

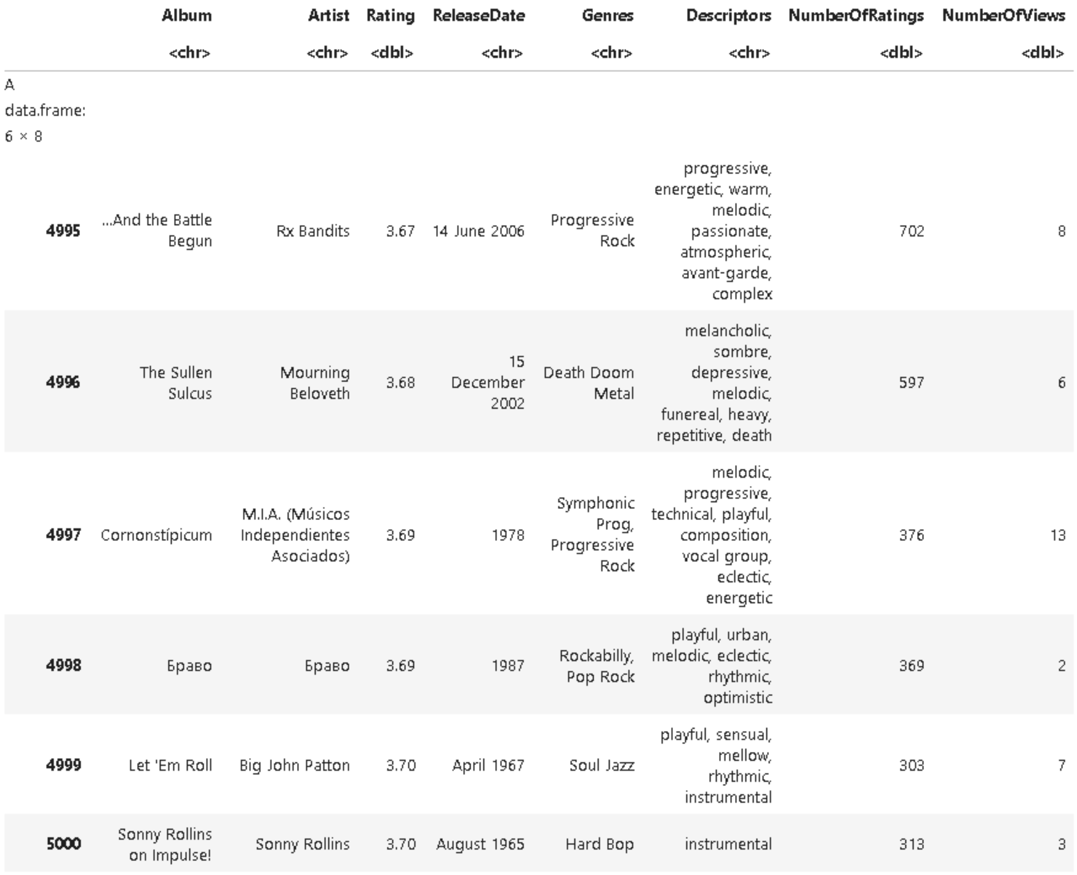


```
  Ranking                            Album         Artist Rating ReleaseDate                                          Genres
1       1              To Pimp a Butterfly Kendrick Lamar   4.36  15-03-2015 Conscious Hip Hop, West Coast Hip Hop, Jazz Rap
2       2                      OK Computer      Radiohead   4.27  16-06-1997                      Alternative Rock, Art Rock
3       3               Wish You Were Here     Pink Floyd   4.32  12-09-1975                       Progressive Rock, Art Rock
4       4                      Madvillainy     Madvillain   4.33  23-03-2004                                Abstract Hip Hop
5       5 In the Court of the Crimson King   King Crimson   4.31  10-10-1969                       Progressive Rock, Art Rock
6       6                      In Rainbows      Radiohead   4.29  10-10-2007                      Art Rock, Alternative Rock
                                                                                        Descriptors NumberOfRatings NumberOfViews
1              political, conscious, concept album, poetic, introspective, urban, protest, eclectic           69884           600
2             melancholic, anxious, alienation, futuristic, existential, lonely, atmospheric, cold           94981          1719
3 melancholic, atmospheric, progressive, concept album, serious, longing, introspective, alienation           65454           997
4                     sampling, playful, humorous, abstract, cryptic, mysterious, eclectic, surreal           54490           469
5                    fantasy, epic, progressive, complex, poetic, surreal, philosophical, technical           60167           946
6                lush, melancholic, introspective, bittersweet, atmospheric, mellow, warm, ethereal           68811           875
```

In the process of web scraping the rest of the data from RYM website, we encountered a challenge related to data consistency. Specifically, the length of the 'descriptors' and 'genres' columns was not aligned with the length of other extracted attributes, like 'album_names'. This discrepancy posed a risk of data misalignment and potential loss of relevant information. To address this, we implemented a safeguard within the code. Using conditional 'if' statements, we checked for inconsistencies in the length of these columns. If a mismatch was detected, we identified the indices where values might be missing. Using the 'map2_chr' function from the 'purrr' library, then inserted 'NA' (Not Available) placeholders for these missing values, ensuring that the resulting data frame maintained the proper structure and alignment. This solution was crucial in preserving the integrity of the scraped data and ensuring that all information was correctly represented in the final dataset.

Another issue was encountered during web scraping, it's vital to be cognizant of the potential for rate-limiting mechanisms on the target website. During our endeavor to scrape 125 pages, we encountered such rate-limiting constraints. These limitations not only risked IP blocking but also substantially increased the duration of our scraping process. To address these challenges and ensure a sustainable data retrieval rate, we integrated a rate-limiting technique within our scraping code using the sys.sleep() function. We introduced deliberate delays between requests to circumvent the server's restrictions. After iterative testing and adjustments, we determined an optimal delay interval that allowed us to successfully extract all the desired data from the 125 pages. However, this approach would have taken a long time.

Here is the data following the RYM scrape:



A data.frame: 6 × 8

	Album	Artist	Rating	ReleaseDate	Genres	Descriptors	NumberOfRatings	NumberOfViews
	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<dbl>
4995	...And the Battle Begun	Rx Bandits	3.67	14 June 2006	Progressive Rock	progressive, energetic, warm, melodic, passionate, atmospheric, avant-garde, complex	702	8
4996	The Sullen Sulcus	Mourning Beloveth	3.68	15 December 2002	Death Doom Metal	melancholic, sombre, depressive, melodic, funereal, heavy, repetitive, death	597	6
4997	Cornonstípicum	M.I.A. (Músicos Independientes Asociados)	3.69	1978	Symphonic Prog, Progressive Rock	melodic, progressive, technical, playful, composition, vocal group, eclectic, energetic	376	13
4998	Браво	Браво	3.69	1987	Rockabilly, Pop Rock	playful, urban, melodic, eclectic, rhythmic, optimistic	369	2
4999	Let 'Em Roll	Big John Patton	3.70	April 1967	Soul Jazz	playful, sensual, mellow, rhythmic, instrumental	303	7
5000	Sonny Rollins on Impulse!	Sonny Rollins	3.70	August 1965	Hard Bop	instrumental	313	3

Best Selling Albums

Best Selling Albums also did not have any public-facing API. However, the website provided a very scraping-friendly interface, and the subpages were organized in a logical way. There was also no rate-limiting, which made scraping this website very convenient and easy. All the charts were stored in a <table /> element at the top of the page, which would easily converted into a dataframe just using the html_table() function from rvest.

Spotify API

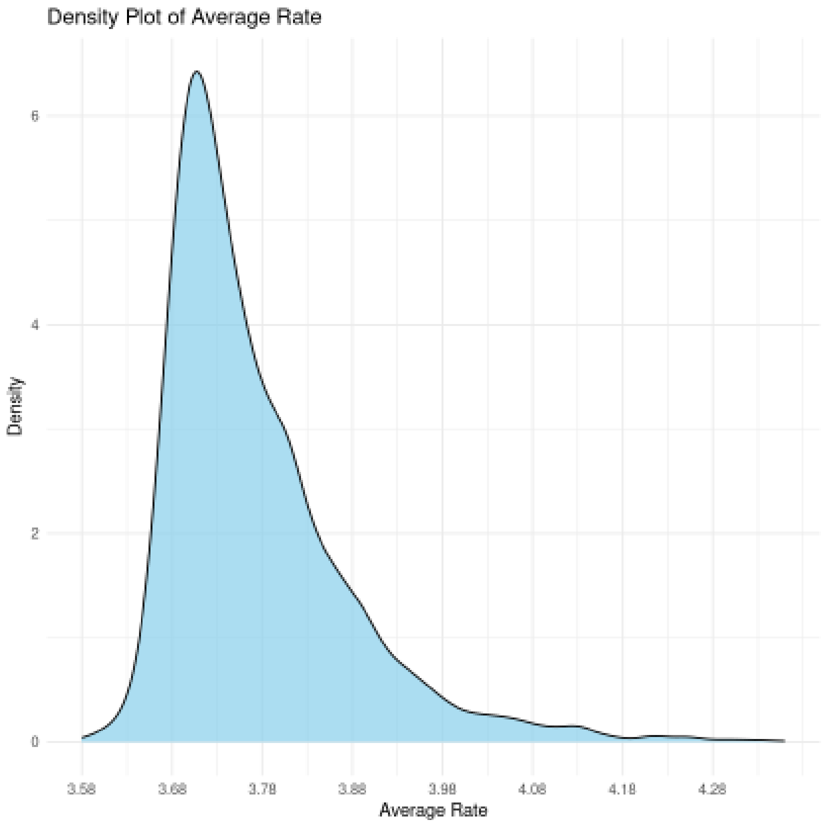
To use the Spotify API, we used the spotifyr package, which conveniently contains R function wrappers for most of the existing endpoints in the API. However, the Spotify API does require the registration of an application on a spotify developer account, and therefore needs a public and private key to be passed to it to start receiving data. However, once this was set up, the data collection was quite easy, due to the convenience of the **spotifyr** package.

Wrangling and Results

How do we do that?

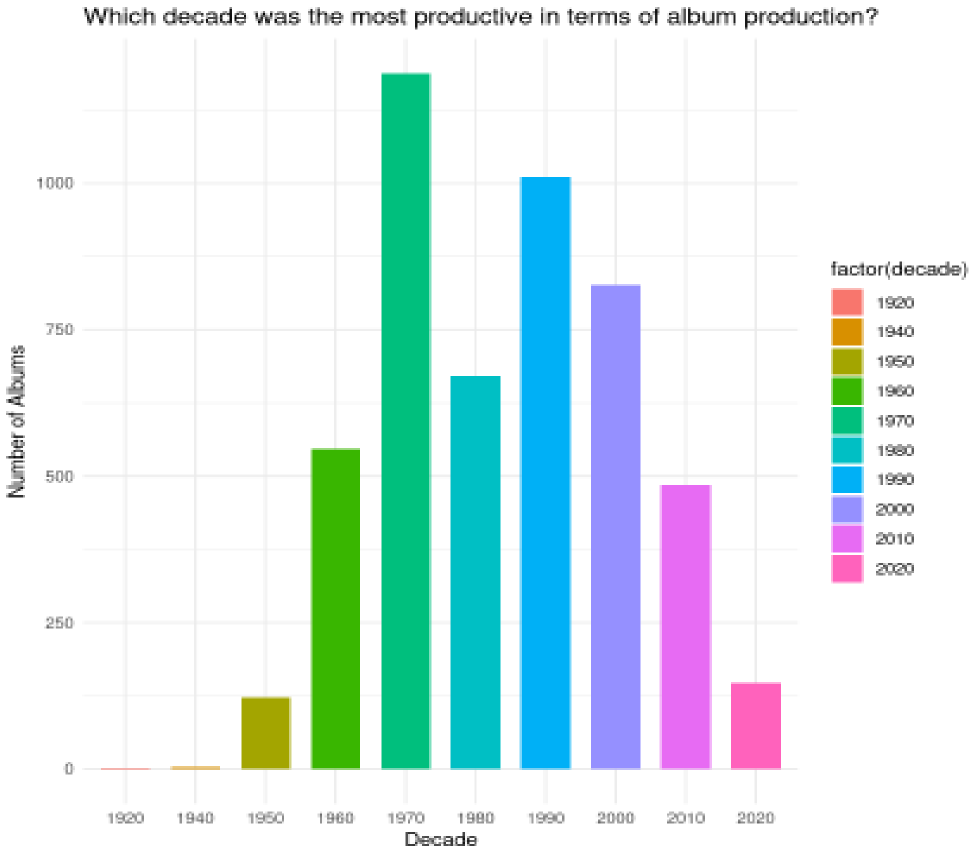
Initially, the code identifies potential date formats and standardizes the ReleaseDate column to ensure consistent date-time representation. This standardized date data is then leveraged to extract the respective year of each album's release, culminating in a bar chart that presents yearly album production counts. To further categorize this data, albums are grouped into their respective decades, and a subsequent bar chart displays the productivity of each decade in terms of album releases. Additionally, a density plot is constructed to showcase the distribution of average ratings given to albums, providing insight into their overall reception.

What era has the most critically acclaimed albums?



To answer this question, we looked at the top 5000 albums on rateyourmusic.com and their ratings. From the "Density Plot of Average Rate" provided, we observe that the majority of album ratings on the "Rate Your Music" website cluster between approximately 3.68 and 3.88. The peak density occurs around a rating of 3.73, suggesting that many albums achieve this favorable score. This rating can potentially be seen as a benchmark for an album's success on the platform. The chart's right tail indicates that while some albums receive ratings above 4.08, they are relatively fewer in number.

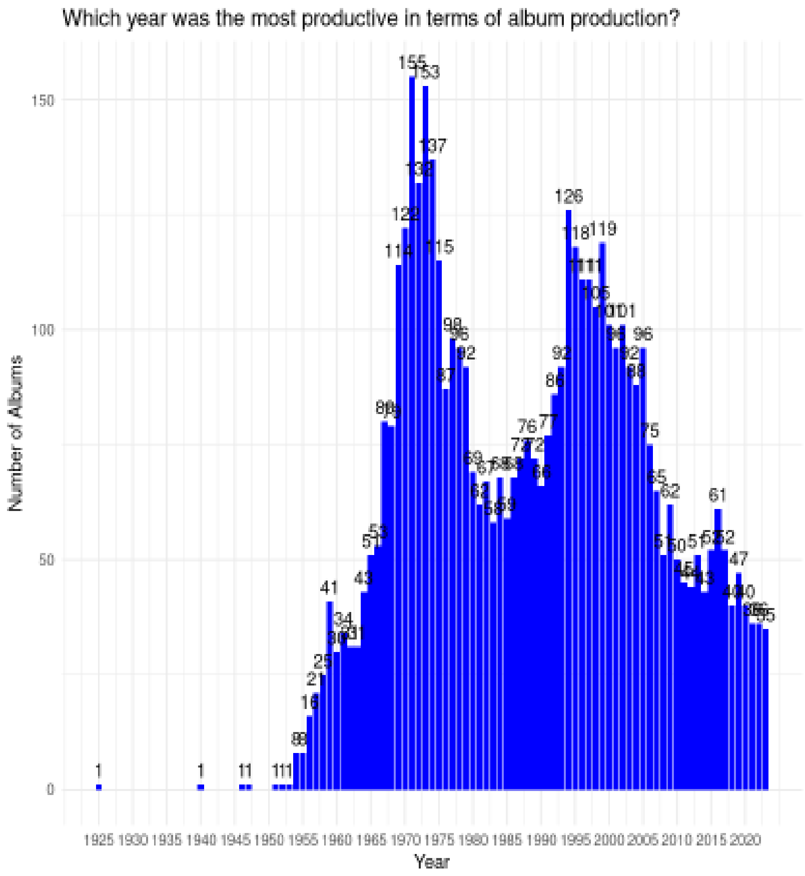
However, we can't directly determine which era's albums were the most critically acclaimed because the chart only shows the density distribution of ratings and has no information about release dates. To answer this question, we provide a graph showing a bar chart of the number of albums vs. year and albums vs decade.



From the bar chart, it's evident that: the 1970s saw the highest number of albums produced, exceeding 1000. This era is often referred to as the "Golden Age of Rock," characterised by the flourishing of the rock genre and the rise of influential music movements like punk, disco, and others, which might explain the surge in album releases.

The 2010s saw a significant drop in album production, and the 2020s even more so. While data for the 2020s might be incomplete depending on the current year, the decline in the 2010s might be due to the dominance of streaming platforms where singles and playlists became more prevalent than traditional albums.

But... is this due to 1 or 2 years in the 70's? Or did the 70's consistently have the 'best' music across the whole decade?

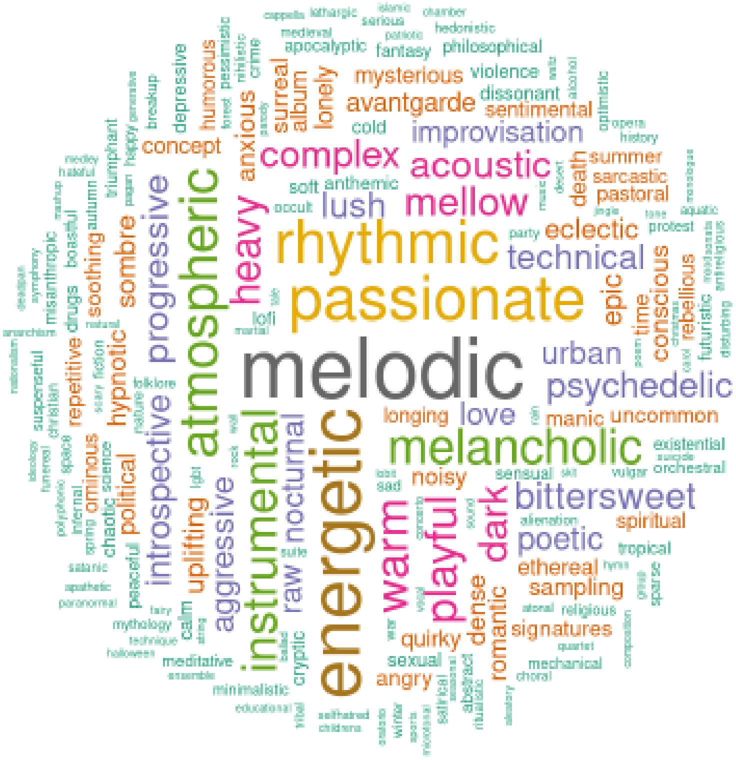


From the histogram above, we observe the annual production of albums spanning several decades. Starting from the 1950s, there was a marked increase in album production, peaking around the 1970s, then experiencing a decline in the 1980s but rising again from the 1990s through the early 2000s. Notably, 1971 emerges as the year with the highest album production, at 155 albums, suggesting a particularly vibrant period in the music industry. We can also see that there is some correlation between years - there is a time-oriented trend and years are not random.

The charts offer a comprehensive overview of album production spanning multiple decades. The first chart breaks down album production by decade, showing that the 1970s was the most prolific era, with subsequent high production in the 1990s, 2000s, and 1980s respectively. The second chart gives a more granular view by year, highlighting specific peaks like 1971, which saw the production of 155 albums. Thus, the prominence of the 1970s in album production might suggest a musically rich and innovative period, possibly due to the cultural revolutions, technological advancements in music recording, and the emergence of diverse genres.

What makes an album critically acclaimed?

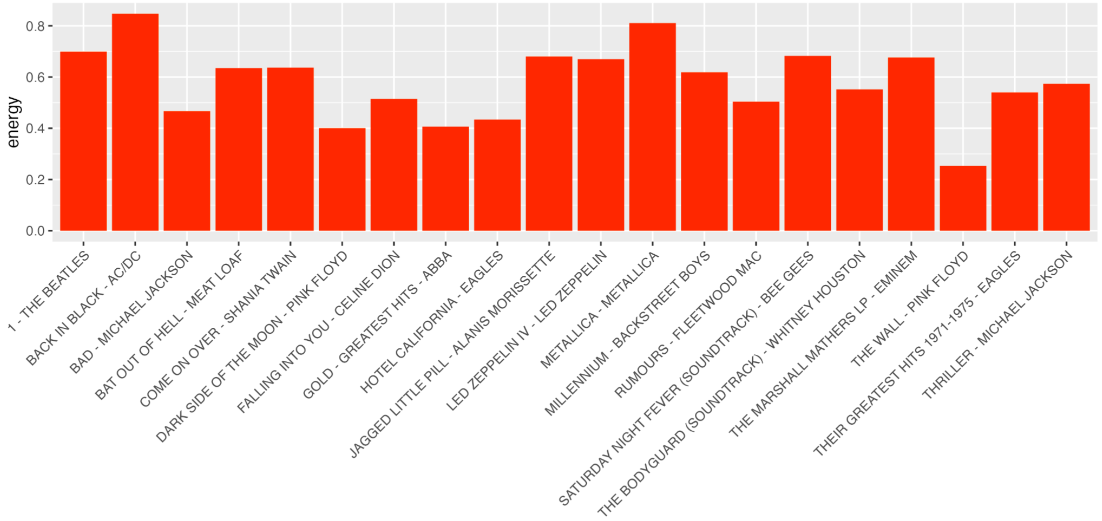
To answer this question, we used text-mining techniques in R. The dataset was cleaned by eliminating common stop words and punctuation marks and converting uppercase to lowercase letters. The data was then structured into a term-document matrix, sorted by word frequency, and visualized as a word cloud.



As this word cloud shows, ‘melodic’ is the most common descriptor among top 5000 albums' descriptors. High-rated albums also usually have music that is energetic, rhythmic, passionate, melancholic, instrumental, or atmospheric.

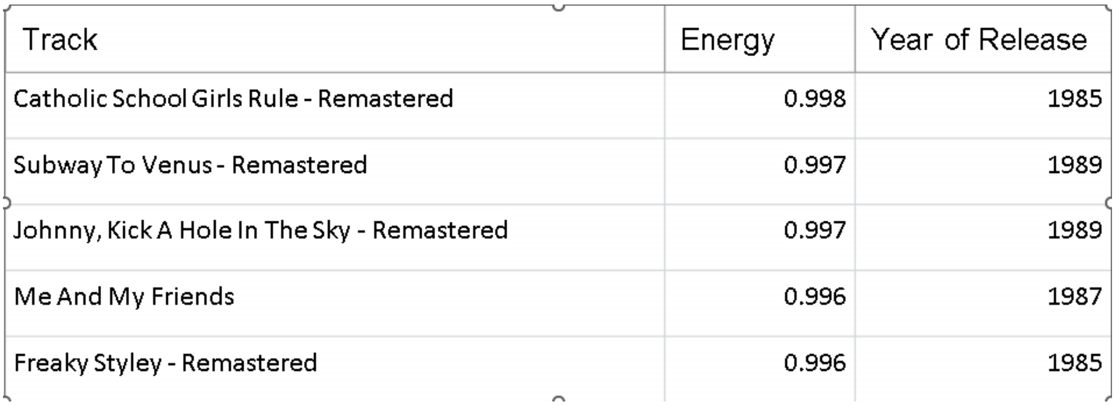
What makes an album sell?

Is it 'energy'? (Top 20 Highest sold albums)



In order to see if there is any trend between 'energy' of an album and how well it sells, we used the best-selling albums dataset to see the average energy of tracks on the top 20 highest-selling albums in the world. Almost all of the albums had moderate energy, between 0.4 and 0.8. Only two albums, by AC/DC and Metallica had energy over 0.8 (due to being metal albums) and only one album, by Pink Floyd, had energy below 0.4.

Is it 'energy'? - Red Hot Chili Peppers case study / deep dive



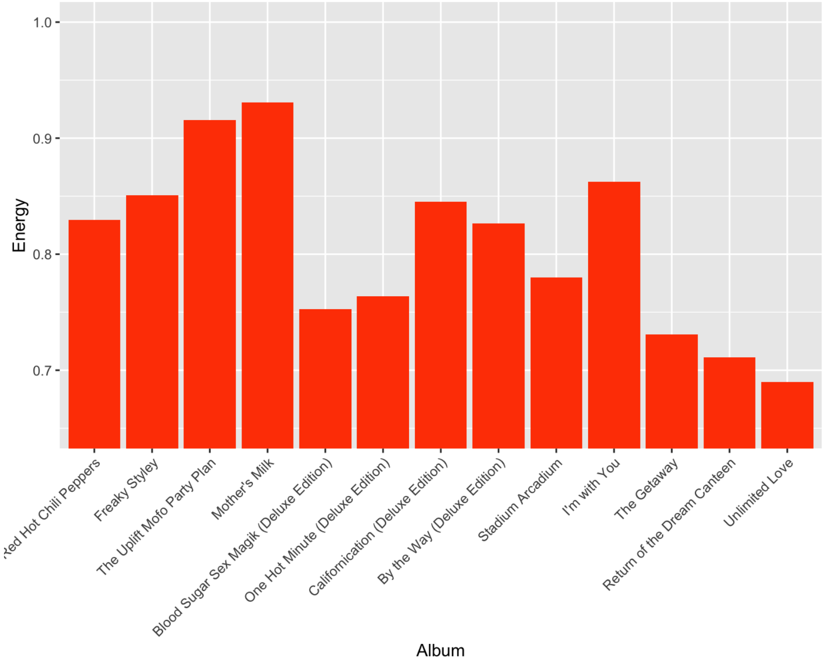
Track	Energy	Year of Release
Catholic School Girls Rule - Remastered	0.998	1985
Subway To Venus - Remastered	0.997	1989
Johnny, Kick A Hole In The Sky - Remastered	0.997	1989
Me And My Friends	0.996	1987
Freaky Styley - Remastered	0.996	1985

Over 75% of the RHCP' album tracks were released AFTER 1990, yet all five of the RHCP' top 5 most energetic tracks were released BEFORE 1990. This implies that the energy of their music may have changed over time, so we investigated this on an album basis.

How does the energy of an album correlate with its sales?

Red Hot Chili Peppers Energy by Album

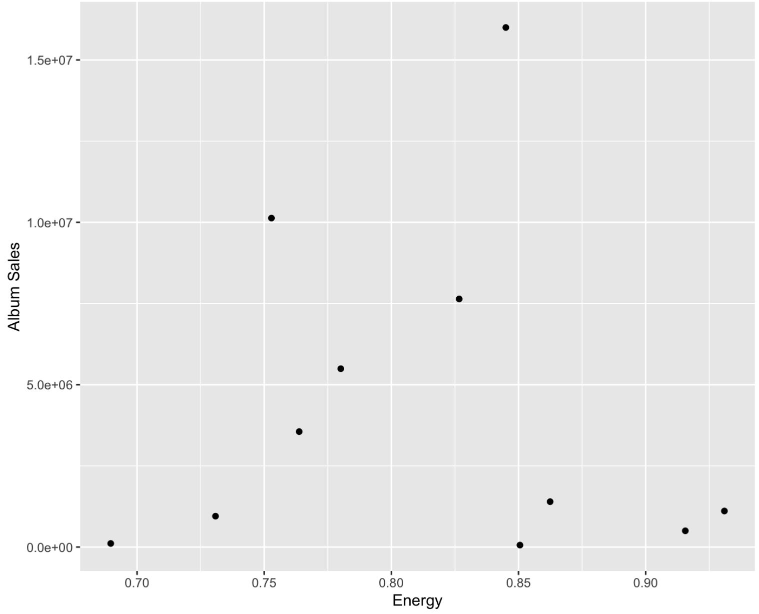
In Order of Release



Prior to 1989’s Mother’s Milk, we can see an increasing trend in energy. During this period, while they were still growing in popularity through the underground scene, the RHCP were experimenting with hard funk rock at high levels of energy. Following Mother’s Milk, we see a decreasing trend - likely due to the success of Under The Bridge from Blood Sugar Sex Magik, a mellow and introspective song. The decrease in energy over time signifies the RHCP’ progression towards a mainstream audience.

Album Energy vs Album Sales

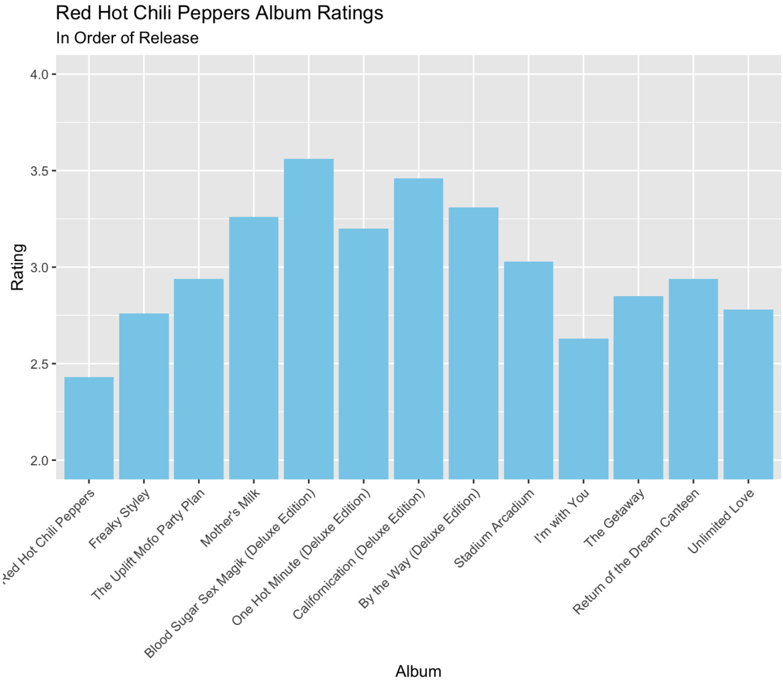
Red Hot Chili Peppers



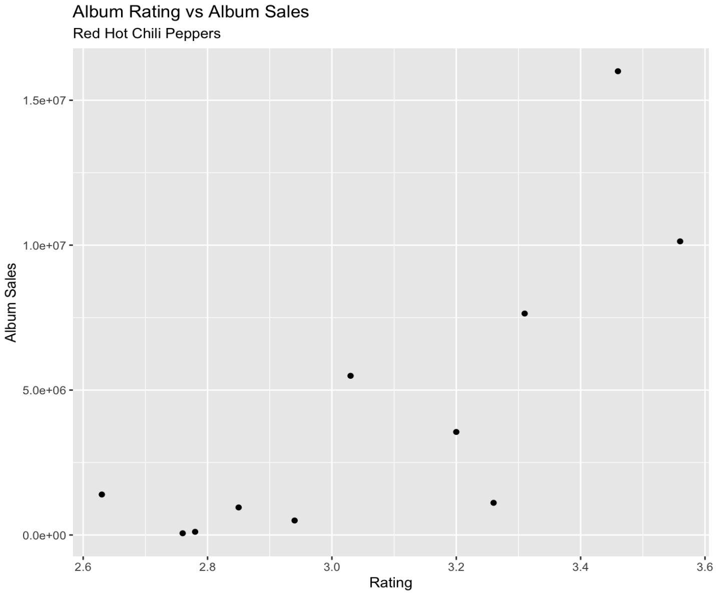
We can see a near-zero R-Squared (0.003) due to the non-linear nature of the relationship (piece-wise function). Below energy levels of 0.85, we see a positive relationship with a huge drop in album sales at very high energy levels, signifying a zero relationship when energy is over 0.85. For low-moderate levels of energy, an increase in energy is associated with an increase in album sales. The nature of the relationship above 0.85 suggests that there is a fine line between producing a “fun” album (higher energy, increase in sales), without being too intense for mainstream (extreme energy, decrease in sales).

Is it high rating? - (Red Hot Chilli Peppers - Deep dive)

How does the rating of an album correlate with its sales?



Here, we observe a similar distribution through time to that of the band’s album sales. Once again, an increasing trend in the funk rock era and a decreasing trend post Under the Bridge (Blood Sugar Sex Magik). The RHCP albums have received low audience ratings since 2006’s Stadium Arcadium.



Here, we have a much higher R-Squared value of 0.63, meaning that audience rating from RYM can explain 63% of the variation in RHCP album sales. We have a moderate and positive relationship between album ratings and album sales. This signifies that an increase in album ratings is associated with an increase in album sales.

Conclusion

The objective of this report was to discover the characteristics which make an album successful. Through web scraping and subsequent analysis of data from sources such as Rate Your Music, Album Sales and the Spotify API, we found a number of factors that contribute towards a successful album.

We found that an indicator of a critically acclaimed album is its rating on Rate Your Music. Albums with an average rating of 3.73 tend to be considered highly successful on the platform. This suggests that favourable reviews and ratings play a significant role in an album's ‘success’.

The critical acclaim of albums varied heavily throughout the decades. Albums from the 1970s and 1990s stood out with record-high ratings. The 1970s, often referred to as the "Golden Age of Rock," witnessed the strongest period of musical innovation, while the 1990s marked the rise of alternative and grunge music, both contributing to the high ratings and critical acclaim of albums from these times.

The word cloud analysis of album descriptors highlighted that terms like 'melodic,' 'energetic,' 'passionate,' 'rhythmic,' 'melancholic,' 'instrumental,' and 'atmospheric' are common descriptors of highly-rated albums. This indicates that the musical characteristics and emotional depth of an album play a crucial role in its critical acclaim.

Analysing the energy levels of high-selling albums revealed that most of these albums have tracks with moderate-high energy levels, typically falling between 0.5 and 0.85 on the Spotify 'energy' scale. Extremely high-energy tracks (e.g., heavy metal) and very low-energy tracks (e.g., ambient music) are less appealing to the masses. This suggests a 'sweet spot' for energy levels in music that is likely to sell well.

A deep dive into the music of the Red Hot Chili Peppers demonstrated the evolution of their sound over time. Albums released before 1990 exhibited higher energy levels, reflecting their experimentation with hard funk rock. However, after the release of more mellow tracks like "Under the Bridge," their energy levels decreased, aligning with their shift towards a mainstream audience.

The correlation between album ratings and sales, (investigated in the context of Red Hot Chili Peppers' discography), revealed that album ratings from platforms like Rate Your Music can explain a significant portion of the variation in album sales. Higher album ratings are associated with increased sales, highlighting the influence of critical acclaim on commercial success.

We can conclude that a successful album can be achieved through the fusion of receiving high critical acclaim, being released in an era marked by musical experimentation, containing specific musical attributes such as melodic and passionate tracks, and boasting a high but balanced level of energy. While there is no definitive formula for success in the music industry, understanding these key factors can significantly improve an album's chances of achieving critical acclaim and commercial prosperity. Artists and music industry professionals can use this knowledge to guide their creative and marketing efforts and create albums that stand the test of time.

Although we have achieved a great depth of analysis and made some interesting conclusions, we failed in generalising our results for the music industry. We selected a handful of convenient (potentially biased) websites to scrape data from and used the results of a case study developed around one band to make comments about the music industry as a whole. If we were to continue this research, we would investigate the websites’ reputability and use data from many different sources. We would also take random samples of industry data before making generalisations, as opposed to selecting one band and assuming the results of that band will apply industry-wide. Although professionals can use this report to guide their efforts, it must be used with caution and in conjunction with their own research from reputable sources.