Data Wrangling 2024-2025
Exploring music evolution over the last century

Radu Nicolae
Computer Science BSc
Vrije Universiteit Amsterdam
r.nicolae@student.vu.nl

Ayush Khadka
Computer Science BSc
Vrije Universiteit Amsterdam
a.khadka@student.vu.nl

Alara Karadeniz
Computer Science BSc
Vrije Universiteit Amsterdam
a.karadeniz@student.vu.nl

Leonid Ivanov
Computer Science BSc
Vrije Universiteit Amsterdam
l.ivanov@student.vu.nl

***Abstract*— In response to the industrial and technological advancement of the past century, our society has become increasingly dependent on music, which is prevalent in most of the activities of our daily lives. With societal, cultural, and technological changes, musical culture is also expected to change, raising the main research question *"How did music evolve over the last century?"* To address the open challenge, we investigate quantifiable metrics and select four representative metrics, used to measure music: acousticness, energy, danceability, and duration. We conduct extensive data-wrangling processes; we conduct the data collection step and select a dataset from Kaggle (the world's largest data science online community) synthesized from Spotify Web API; we conduct the data extraction and processing to bring data to a usable form, thus enhancing readability and usability; then, using the existent data, we create ratios in the data creation step; lastly, in the data analysis process, we investigate the obtained data, conduct relevant statistical analysis and visually depict data. Through the analysis process, we observe that: (i) the energy levels almost doubled throughout the time, while the acousticness dropped by 65%; (ii) energy-danceability ratio almost halved since 1921; (iii) surprisingly, despite cultural changes, the average song duration has remained relatively steady.**

***Index Terms*—music, evolution analysis, energy, acousticness, danceability, duration**

I. Introduction

Over the past century, our society has become increasingly digitalized due to modern advancements in industry and technology [1], [2]. Music became increasingly more accessible and slowly turned into a (sometimes) indispensable component of our daily lives. Albeit apparently indispensable, music can significantly impact our decisions, productivity, and health from a psychological perspective [3]–[5]. It is critical, yet nontrivial, to analyze the evolution of music, correlate it with societal changes, and obtain insights into the impact of music on the future society. This research represents only a starting point for analyzing the evolution of music over the past century and opens potential future research into the impact of societal change into the music types.

Music is *a pattern of sounds made by musical instruments, voices, or computers, or a combination of these, intended to give pleasure to people listening to it [6].* Music reflects diverse cultural backgrounds, allowing listeners to explore cultures through the art of music, independent of the geographic location [7], [8]. Moreover, music is a unique type of art that expresses emotions, thoughts, and visions [9]. From a societal perspective, various genres of music help people concentrate, integrate into groups, and strengthen interpersonal relationships [10]. From a medical standpoint, various music genres have been scientifically proven to help soften or mitigate health issues at all ages [11]–[13].

Music has faced significant changes throughout the last century, tailored to societal changes. Although it is difficult to quantify changes in art fields (e.g., poetry, music, painting), the music contains quantifiable metrics that can determine patterns, e.g., acousticness, energy, danceability, duration, instrumentals, liveness, loudness, speechless, tempo, valence, instrumentalness [14]. In this work, we focus on analyzing changes in music across time using a representative group of metrics: acousticness, energy, danceability, and duration. This raises the main research question (MRQ):

(MRQ) How did music evolve over the last century?

To answer MRQ, we formulate three sub-research questions (RQ) and centralize our research towards answering:

(RQ1) How did the acousticness-energy ratio in songs change over the last century?

(RQ2) How did the energy-danceability ratio in songs change over the last century?

(RQ3) How did the duration of songs evolve over the last century?

Approach and main contribution: In this work, we address the open challenge with a four-fold contribution:

- **C1** (acousticness:energy ratio) We analyze the evolution of acousticness and energy and determine the ratio between these two metrics over the years. We analyze patterns and trends and present the results in Section V-A.
- **C2** (energy:danceability ratio) We analyze the impact of energy on the dancebility factor of the songs throught the years, and determine the ratio. We analyze patterns and trends, and present the results in Section V-B.
- **C3** (duration over time) Throughout the years, notable changes might occur on various properties of songs, and the musical industry overall. We analyze the average duration of songs and the time influence in Section V-C.
- **C4** (open-science) **[to be discussed]** On the written approval of the course coordinators, we provide the datasets, used tools, (data wrangling) methods, and results as publicly available, as an open-science contribution to the scientific community.

II. Data sources

In this work, we used data from the Spotify Web API, provided as an open source dataset[1], which contains data leveraged from 160,000 tracks, collected between 1921 and 2020. The dataset presents a wide variety of attributes, from which, in this work, we select acousticness, energy, danceability, and duration. We present more information on the data wrangling methods (data extraction, processing, analysis, delivery) in Section IV.

III. Background

In this section, we present background on relevant metrics used in this research and widely used in the community.

Acousticness: "Indicates whether a song was played acoustically: where classical instruments are used instead of electric or electronic instruments [15]." Acousticness is quantified with values between 0.0 to 1.0 where the higher the value the more acoustic a song is and is unitless.

Energy: "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. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy [16]." Energy is unitless.

Duration: Duration is a measure that quantifies the length of a track. Duration is measured in international scale units in seconds; thus, in this work, we quantify duration in seconds.

Danceability: "Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable, and 1.0 is most danceable. [16]". Similarly to energy and acousticness, danceability is unitless.

IV. Data wrangling methods

A. Data Collection

We began our journey by formulating the main research question: "How did music evolve over the last century?". Then, we formulated three sub-research questions, which guide our analysis and focus this work towards analyzing four representative metrics in the musical community.

We analyzed numerous sources provided via Canvas, and we chose Kaggle[2], the world's largest data science community, with over 150,000 submissions monthly, offering excellent data searching and filtering options.

In the process of *data collection*, filters played a pivotal role. We established the scale of data and selected only files with sizes ranging between 10 MB to 100 MB. Furthermore, to prevent issues with the data (e.g., missing values, wrong values), we filtered datasets by usability and regarded relevant only datasets marked with usability 10 (highest). Then, we searched the keywords *"Spotify tracks"* and found a dataset labeled *"Spotify Dataset 1921-2020, 160k+ Tracks"* (although Spotify was publicly launched only in 2008). The dataset contains 39.7 MB of data, stored in CSV format, and has a usability of 100%. The data from Spotify Web API, thus ensuring data correctness and validity.

We also observed alternative sources, such as *"Spotify tracks chart dataset[3]"* which, although contains data for one century, overpasses size criteria by 8 times. Although valuable, we chose a lower-sized dataset, which allows for better usability and distribution between regular-user machines.

B. Data Extraction and Processing

Upon selecting the dataset (Section IV-A), we explored the available datasets at a higher depth. The file *data.csv* contains all the relevant data, further filtered by various categories and presented into four additional files *i) data_by_year.csv*, *ii) data_by_artist.csv*, *iii) data_by_genres.csv*, *iv) data_w_genres.csv*. We regard dataset *i)* as the most suitable for this research.

We read the file using Pandas framework (version 2.2.2). To clean unnecessary data for this research, we synthesized relevant columns into a new Pandas dataframe containing five columns: year, energy, acoustics, duration, and danceability. We inspected the selected columns and observed full usability, without non-standard elements (e.g., missing data, corrupted data). We regard this a consequence of the rigorous data collection methodology presented in (Section IV-A).

For readability purposes, we converted duration from milliseconds to seconds (divided the duration column by 10^3). Lastly, we rounded all values to three decimal places to ensure consistency and readability.

C. Data Creation

To answer RQ1, we computed the energy-acousticness ratio by dividing the energy column by the acousticness column, and further storing the data in the data table created in Section IV-A. To answer RQ2, we followed a similar process for energy-danceability. To answer RQ3, we analyzed the duration of the songs, following suitable statistical methods.

D. Data Analysis

Lastly, we conduct the *data analysis* step. To answer MRQ, driven by the time dimension, we regard time-series plots as the most suitable, in which the horizontal axis depicts time over one century, and the vertical axis describes the measured metric (e.g., acousticness, ratio). We summarize our findings in Section V and visually represent them in Figures 1-5.

V. Findings

In this section, we analyze findings and answer the main research question *"How did music evolve over the last century?"*. Figure 1 presents the overall findings and metric evolution between 1921-2020.

[1]https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-1921-2020-160k-tracks

[2]https://www.kaggle.com/

[3]https://www.kaggle.com/datasets/jfreyberg/spotify-chart-data

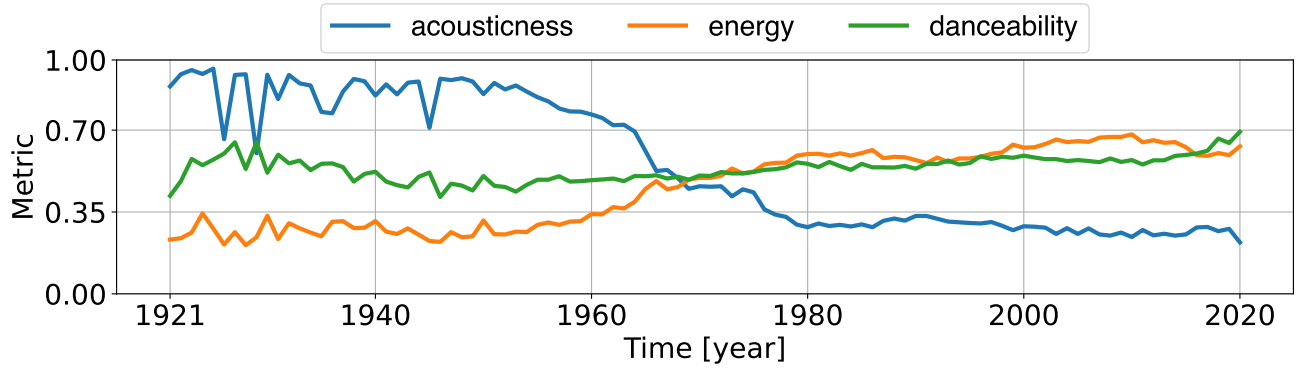

Fig. 1. Changes of acousticness, energy, and danceability factors in songs over the last century.

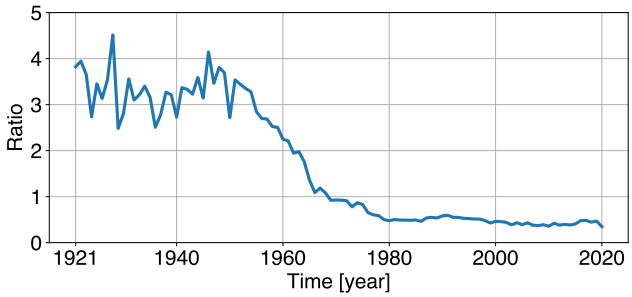

Fig. 2. Energy-acousticness ratio in songs over the last century.

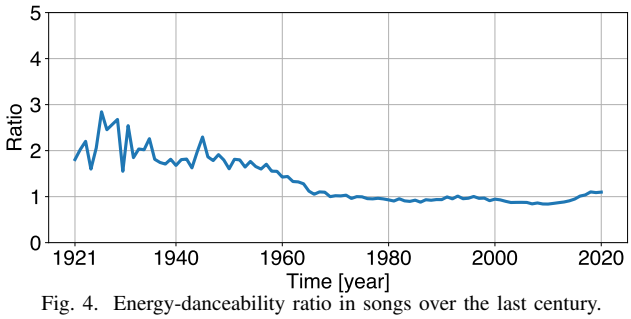

Fig. 4. Energy-danceability ratio in songs over the last century.

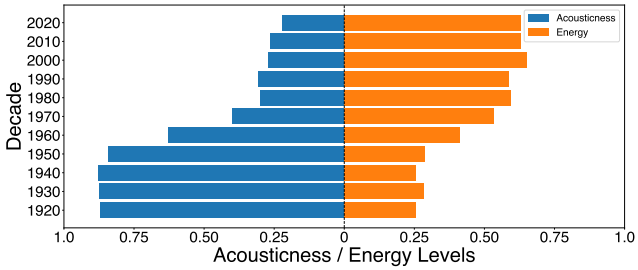

Fig. 3. Evolution of acousticness levels compared to energy levels. The plot, which takes its structure from Tukey's population pyramid, plots on the horizontal axis average measurements: (left) acousticness, (right) energy. On the vertical axis, is the decade in which data was measured.

A. *Effect of acoustics on energy (on RQ1)*

(RQ1) How did acousticness-energy correlation in songs change over the last century?

We answer *RQ1* by analyzing the evolution of the accousticness-energy ratio across time and make contribution **C1**. We visually depict our findings in Figure 1, Figure 2, and Figure 3. Overall, we observe a decrease of 75.1% in acousticness levels between 1921 and 2020, while the energy levels increased by 171.9%.

Figure 3, inspired from Tukey's population pyramid depicts evolutions in acousticness and energy levels throughout the time. We identify a significant drop in acousticness between 1950s-1970s, period which, however, contains a significant increase in energy levels. We also identify that decreases in acousticness levels compensate by increases in energy levels.

Upon determining the ratio energy-accousticness, we identify one order of magnitude decrease, from the peak of 4.51 (in year 1928) to the minimum of 0.34 (in year 2020).

Evaluating all three plots, we observe a key period for music, represented by 1950s-1970s. We argue that this musical trend mirrors the societal changes, and the increased alertness of society, perhaps a reflection of the daily lives intensity into the artistic process of music creation. This change can also be due to the increased accessability to music and the disco (sub)culture, rooted in the disco wave from 1960-1970s [17].

B. *Effect of energy on danceability (on RQ2)*

(RQ2) How did the energy-danceability ratio in songs change over the last century?

We answer *RQ2* by analyzing the evolution of energy-danceability ratio across time and make contribution **C2**. We visually depict our findings in Figure 1 and Figure 4. Overall, we observe similar trend between Figure 2 and Figure 4, yet at a lower magnitude in the latter.

Throughout the last century, although the energy levels increased by 171.9%, the danceability remained almost steady, showing a mean of $\mu = 0.54$ and a standard deviation of only $\sigma = 0.05$. Still, the danceability factor increased between 1921 to 2020 by 65.3%. To answer RQ2, we may conclude that, despite cultural and societal changes, songs remained similarly dance-able throughout the years (perhaps people stayed similarly happy, or sad throughout the time) :D.

C. *Analysis of song duration over time*

(RQ3) How did the duration of songs evolve over the last century?

We answer *RQ3* by analyzing the evolution of song duration throughout the last century and make contribution **C3**. We visually depict our findings in Figure 5.

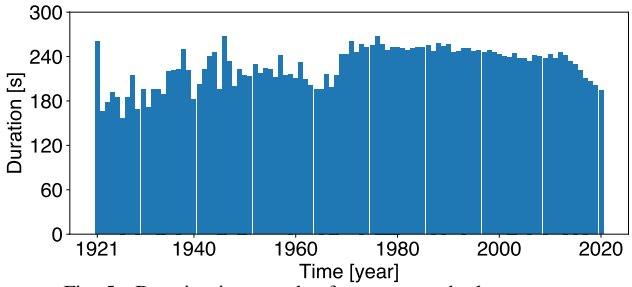

Fig. 5. Duration in seconds of songs over the last century.

We observe a significant fluctuation between 1921-1970 ($\mu_1 = 211.3, \sigma_1 = 24.28$) and 64.5% less fluctuation in duration between 1971-2020 ($\mu_2 = 243.3, \sigma_2 = 14.68$). We argue that the volatility between the 1920s-1970s is mainly due to the different genres that were popular at the time (e.g., jazz, swing, rock'n'roll). In contrast, the rise of pop music from the 1970s onward is reflected in a higher similarity amongst songs, including a higher duration similarity [1].

Surprisingly, duration has remained similar over the past century. However, we observe a 20.2% decay in duration between 2010-2020, potentially attributed to the decreased attention span of the public, also reflected in the music they consume. We expect this decreasing pattern to continue throughout the 2020s and 2030s, yet dependent on governmental regulation on factors that affect attention span (e.g., social media, ads).

VI. Limitations

The limitations of the approach that need to be addressed in future research are (1) dependence on instrumentation/production, (2) limited analyzed metrics, and (3) potential subjectiveness of metrics.

Related to (1), certain instruments and production techniques can impact acoustics and energy, which makes comparison challenging, especially based on the genre. Analyzing and comparing different genres can provide more extensive research with a potentially more comprehensive contribution to the community.

Related to (2), while the small number of analyzed metrics is a breakthrough in the community, future research can analyze other valuable metrics, thus providing a wider overview. The somewhat narrow focus of this research can lead to certain factors being ignored, such as correlation of music with culture and emotions.

Related to (3), metrics analyzed in this work could be prone to subjectiveness (e.g., danceability, energy), due to the potentially challenging quantification process. A comprehensive future analysis of the measurement process could be beneficial for the scientific community.

VII. Conclusion and Future Work

Music is prevalent in most of the activities of our daily lives, sometimes subconsciously influencing decision-making processes. The major cultural, industrial, and technological changes from the last century are reflected also music. In this paper, we address the main research question *"How did music evolve over the last century?"*. To address MRQ, we synthesize four quantifiable metrics and formulate three subsequent research questions. Throughout the entire research process, we applied data-wrangling processes to ensure high-quality information, correctness, and reproducibility. We selected data from Kaggle, the world's largest online community for data science, based on well-defined criteria. Then, we processed data to match this research's needs (Section IV-B) and conducted data creation processes to compute ratios between metrics (Section IV-C). Lastly, we analyzed data (Section IV-D) and summarized our findings in Section V. Overall, we observed a major increase in energy levels and a major decrease in acousticness, over the last century. Still, the danceability and duration of songs have remained similar, although with fluctuations, especially during 1920s-1970s. We envision future research that could empirically analyze and/or predict trends in musical evolution as a bi-directional correlation between music and society at large.

References

[1] United Nations, "The digital transformation of societies." https://www.un.org/en/digital-cooperation-panel. Accessed: January 2025.
[2] Y. Chang, "A century of digital evolution and its cultural impact," *Nature Digital Innovations*, vol. 23, no. 4, pp. 200–210, 2021.
[3] American Psychological Association, "Effects of music on mental processes and well-being." https://www.apa.org/topics/music. Accessed: January 2025.
[4] V. Menon and D. Levitin, "Neural correlates of music perception and cognition," *Nature Neuroscience*, vol. 23, no. 9, pp. 1152–1160, 2020.
[5] World Health Organization, "The role of music in mental health interventions." https://www.who.int/health-topics/mental-health. Accessed: January 2025.
[6] "Music - definition." https://dictionary.cambridge.org/dictionary/english/music. Accessed: 2025-01-20.
[7] Smithsonian Institution, "Music as a reflection of cultural identity." https://music.si.edu/. Accessed: January 2025.
[8] World Bank, "Cultural diversity and global music exchange." https://www.worldbank.org/. Accessed: January 2025.
[9] D. M. Greenberg, "Emotional expression and interpretation through music," *Journal of Personality and Social Psychology*, vol. 114, no. 2, pp. 234–245, 2018.
[10] American Psychological Association, "Group music and social integration research." https://www.apa.org/pubs/highlights/sample/music-therapy. Accessed: January 2025.
[11] "How music helps people heal: The therapeutic power of music." https://www.harmonyandhealing.org/how-music-helps-people-heal/#:~:text=Music%20has%20been%20found%20to%20have%20powerful%20effects%20on%20the,time%20from%20illness%20or%20injury. Accessed: 2025-01-20.
[12] J. Bradt and C. Dileo, "Music therapy for stress and anxiety reduction in adults," *Cochrane Database of Systematic Reviews*, vol. 1, p. CD006577, 2021.
[13] National Institutes of Health, "Music-based interventions for improved well-being." https://www.nih.gov/. Accessed: January 2025.
[14] M. D. Baron and J. Bansal, "Acoustic features influence musical choices across multiple genres." https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2017.00931/full. Accessed: 2025-01-20.
[15] K. Adiasto, M. L. van Hooff, D. G. Beckers, and S. A. Geurts, "The sound of stress recovery: an exploratory study of self-selected music listening after stress," *BMC psychology*, vol. 11, no. 1, p. 40, 2023.
[16] "Get track's audio features." https://developer.spotify.com/documentation/web-api/reference/get-audio-features. Accessed: 2025-01-20.
[17] Wikipedia, "Clubbing (subculture)." https://en.wikipedia.org/wiki/Clubbing_(subculture). Accessed: January 2025.