Data Insights on Gender Representation: Analyzing the Book and Music Industries

Mariana O. Silva, Gabriel P. Oliveira, Mirella M. Moro

Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{mariana.santos,gabrielpoliveira,mirella}@dcc.ufmg.br

Abstract. *The entertainment industry has been historically dominated by men, which motivates growing recognition and advocacy for improved gender diversity and equality. We present a study on gender representation in the book and music industries by analyzing awarded authors and hit song artists. Through Data Science, we uncover patterns and trends that beg for a more balanced and diverse portrayal of gender in creative expressions and offer insights to foster inclusivity, diversity, and equitable opportunities in such a domain.*

1. Introduction

Achieving greater diversity and gender parity in the entertainment industry has become an increasingly recognized imperative [Brannon Donoghue 2020]. However, the journey towards gender equality remains a multifaceted challenge. Such an important issue has garnered significant attention in academic research and public discourse within different entertainment sectors, including books and music. Indeed, the book and music industries play significant roles in shaping cultural narratives, influencing public opinion, and reflecting societal values [Oliveira et al. 2020, Silva et al. 2021]. Therefore, understanding gender dynamics within these industries is crucial for identifying existing inequalities, exploring the factors contributing to gender disparities, and promoting gender equality.

Given recent advances in data management and web technologies, many sources provide huge volumes of data about such industries and the people within them. Then, distinct studies have shed light on gender representation within both industries, revealing persistent disparities, bias, and the underlying mechanisms that aid such inequalities. For example, within the book industry, researchers have explored the gender gap in authorship and readership patterns [Bucur 2019, Ekstrand and Kluver 2021, Szasz et al. 2022]. Likewise, there is research on the underrepresentation of female artists, gender bias in music production, and the unequal distribution of opportunities within the music industry [Betti et al. 2023, Epps-Darling et al. 2020, Watson 2020].

However, there is still a need for quantitative analyses and empirical evidence to instigate deeper understanding of gender representation within specific book and music categories (genres) and to assess whether progress has been made over time. The underrepresentation of female authors and artists, as well as the bias and barriers they face in achieving recognition and success, are areas that require closer examination. Therefore, this paper aims to enrich the existing body of knowledge by providing a comprehensive analysis of gender representation in the book and music industries. Through data science tools, we address three research questions (RQs):[1]

[1]A previous version of this work is published in [Silva et al. 2024] and focuses on the Social aspects of this research, whereas this paper goes further into technical, data-oriented contributions.

RQ1. *Are book/music genres more likely to be dominated by one gender over the other?*
RQ2. *Have certain categories experienced a shift towards greater gender parity, or have they become more imbalanced in terms of gender representation?*
RQ3. *Are there any gender bias or disparities in the recognition and promotion of male and female authors/artists in award processes or chart rankings?*

Overall, we explore distinct Data Science techniques through a methodology that can be applied (or mapped) and extended to any other cultural context, given the proper dataset or source availability. This paper is then organized as follows. Next, Section 2 presents related work. Then, Section 3 illustrates our Data Science methodology, and Section 4 describes the experimental results to answer our RQs. Finally, Section 5 presents the concluding remarks and future work.

2. Related Work

Gender representation in the entertainment industry has garnered significant attention in academic research and public discourse [Salles and Pappa 2021]. Researchers have explored the extent of gender disparities and bias within different sectors, including books [Bucur 2019, Ekstrand and Kluver 2021, Szasz et al. 2022], music [Betti et al. 2023, Epps-Darling et al. 2020, Watson 2020], film [Istead et al. 2022, Kagan et al. 2020], and video games [Heritage 2020]. Within the book and music industries specifically, the gender gap has been analyzed by focusing on authors, artists, and patterns of consumption by listeners and readers.

In the book industry, which has recently shown significant changes in gender representation [Waldfogel 2023], studies have investigated gender imbalance in authorship and readership. For example, Bucur [2019] measures gender homophily in large-scale online book markets to analyze if book consumption is assortative by gender, revealing that gender homophily leads to skewed consumption patterns in specific literary genres and book communities. Ekstrand and Kluver [2021] measure the gender distribution of book authors in user rating profiles and recommendation lists, whereas Szasz et al. [2022] assess gender representation in both illustrations and photographs in children's books.

Similarly, the music industry, traditionally male-dominated [Smith et al. 2020], has been analyzed for gender representation. Researchers have investigated the underrepresentation of female artists in music genres and gender bias in music production processes [Epps-Darling et al. 2020, Watson 2020]. Additionally, studies have explored gender bias and sexism in song lyrics. Betti et al. [2023] analyze English song lyrics for gender-related language bias and sexism, finding that sexist content has increased over time, especially from male artists and in popular songs appearing on Billboard charts. They also found that male solo artist songs contain more and stronger bias.

While such studies have covered gender representation in entertainment industries independently, there are still many ways of scrutinizing data from both book and music sectors. We explore such a gap through data science lenses by analyzing gender representation patterns, temporal trends, and disparities in success in the book and music industries. By studying these aspects collectively, we seek deeper understanding of gender dynamics within the entertainment industry and show how Data Science uncovers persistent gender dissonances as well as progress toward gender parity.

Table 1. Book/song categories obtained from Goodreads/Billboard year-end lists.

	Categories (Genres)	#/year	Period
Books	Fantasy, Fiction, Historical Fiction, History & Biography, Horror, Humor, Memoir & Autobiography, Mystery & Thriller, Nonfiction, Poetry, Romance, Science & Technology*, Science Fiction, Young Adult Fantasy, Young Adult Fiction	20	2012–2022
Songs	Adult Contemporary, Gospel, R&B*, Rap*, Pop	50	
	All (Hot 100), Christian, Country, Dance/Electronic*, Latin, R&B/Hip-Hop, Rock	100	2012–2022

* Science & Technology (2015–2022) | R&B, Rap, Dance/Electronic (2013–2022)

3. Methodology

This section outlines a data science methodology for collecting and processing success data from the book and music industries. First, we obtain year-end lists of awarded books and hit songs from Goodreads and Billboard (Section 3.1). Next, we preprocess such lists to identify the writers and artists who achieved success in each year and category (Section 3.2). Finally, we propose a heuristic for determining the gender of each individual, which is the key information to perform our analyses (Section 3.3).

3.1. Data Collection

To investigate possible gender bias or disparities within the book and music industries, we select Goodreads and Billboard as our primary data sources. Both platforms provide success data grouped by genre, which is essential to analyze gender differences within such markets. Also, we complement our dataset with information from Wikipedia to help in the gender classification of authors and artists. Table 1 presents categories, number of entries and periods covered by the Goodreads Choice Awards lists and Billboard charts.

Goodreads Choice Awards. Goodreads is a social cataloging website that allows users to track and discover books. We collect data on all award-winning books from its annual Choice Awards[2] from 2012 to 2022. The series begins in 2012, marking the first year with complete and consistent data available for analysis. The Goodreads Choice Awards nominate books across various categories, with Goodreads members voting for their favorites. Each category consists of 20 nominees selected from millions of books that users have added, rated, and reviewed. The analysis focuses on books published in the United States in English, including translations and major rereleases. For each year and category, there is one winning book and author(s). We use the Goodreads API to collect data on all nominated and winning books, along with author bios, which help infer their gender.

Billboard Charts. Billboard, a renowned weekly magazine, is a prominent source of information for music professionals, enthusiasts, and fans. Known for its music charts, Billboard ranks songs and albums based on their popularity and performance in the music market. Besides the weekly Hot 100, the most well-known chart, it also publishes charts for specific genres such as rock, country, rap, R&B, dance, and more. To collect data from Billboard's year-end charts across multiple genres, we use the Python library *billboard.py*.[3] We choose year-end charts over weekly ones to maintain comparability with the book industry data, which also tracks annual achievements.

[2]Goodreads Choice Awards: `https://www.goodreads.com/choiceawards`

[3]*billboard.py*: `https://github.com/guoguo12/billboard-charts`

Wikipedia. Wikipedia is a free online encyclopedia with articles on various topics. We use the *Wikipedia* Python library[4] to collect summaries of articles for music artists and book authors who do not have bios on Goodreads.

3.2. Data Preprocessing

After obtaining the Goodreads Choice Awards lists and the Billboard charts, we preprocess the data to prepare it for gender identification. As handling heterogeneous data from different sources is a complex task [Mangaravite et al. 2022], we perform distinct preprocessing methods for book and music data. By performing specific preprocessing procedures for each data type, we address potential issues arising from the heterogeneity of the sources, enabling a more accurate and reliable gender identification analysis.

For Goodreads data, we consider only authors, excluding other roles such as illustrators and editors. Since authors are the primary creators of books, responsible for crafting the narrative and developing the ideas, they play a central role in our analysis of gender representation. For the Billboard data, although songs can feature multiple artists, we focus on the primary artist to maintain consistency in our analysis.

3.3. Gender Identification Method

Following previous works [Morais and Merschmann 2021, Mukherjee and Bala 2017], we use a heuristic based on linguistic features to identify the gender of book authors and hit song artists. Specifically, we analyze pronouns in the person's description to infer gender, assigning *he/him/his* for males and *she/her* for females. We acknowledge the limitations of this approach, as it does not account for non-binary individuals who may use other pronouns.[5] For book authors, we extract information from Goodreads bios; for music artists and authors without Goodreads bios, we rely on Wikipedia summaries. Our approach comprises two main steps: category assignment and gender identification.

Category assignment. This step applies exclusively to music artists. Since the term *artist* serves to solo singers, bands, groups, and duos, we focus on solo singers. We check the first 40 words of the artist's description for the terms *singer, songwriter, musician, producer, artist, rapper, actress*, and *actor* to categorize the artists as a *person*. If none of these terms are present, we check for terms such as *group, band*, or *duo* to classify the artist as a *group*. If no relevant terms are found, we classify the artist as *unknown*.

Gender identification. This step applies to book authors and individual artists. We compare the frequencies of male and female pronouns in their descriptions to find out the gender. If male pronouns outnumber females, we label the person as *male*. Conversely, if female pronouns exceed male pronouns, we label the person as *female*. In cases in which it is not possible to determine the gender, we use the Python library Gender Guesser[6] to infer gender based on the individual's name. If gender remains indeterminable, we assign the label *unknown*. For music groups, we check for terms such as *boy/girl group* and *boy/girl band* to categorize them.

Our dataset comprises 3,415 individuals in both domains: 2,068 book authors and 1,347 music artists. Specifically, we identified 1,251 male authors, 1,941 female

[4] *Wikipedia* Python library: `https://github.com/goldsmith/Wikipedia`

[5] We discuss this limitation in Section 5 for a comprehensive understanding.

[6] Gender Guesser: `https://github.com/lead-ratings/gender-guesser`

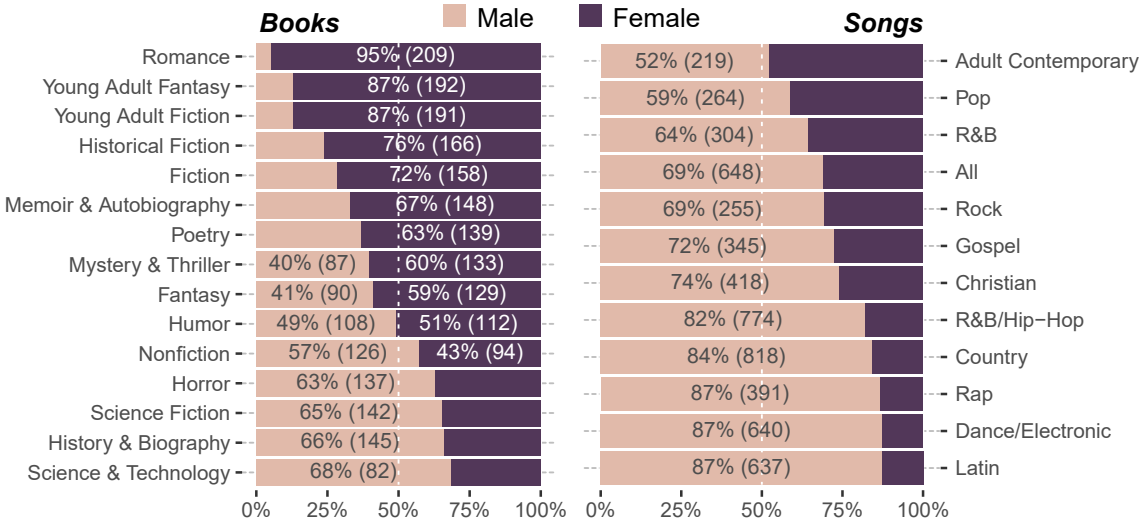

Figure 1. Gender preference across book and music categories.

authors, 824 male artists, and 328 female artists. In addition, there are 195 music artists of unknown gender. This information provides an overview of the gender distribution within our dataset for both book authors and music artists.

4. Experimental Analysis

This section presents the experimental analysis results for the three research questions. In Section 4.1, we review the proportion of male and female authors/artists across various categories in the book and music industries. Then, Section 4.2 investigates changes in gender representation within these categories over time. Finally, in Section 4.3, we analyze the correlation between the gender of authors/artists and their likelihood of winning a Goodreads Choice Award or reaching the first position on a Billboard Chart.[7]

4.1. Gender Representation in Different Categories

To address **RQ1** regarding variations in gender representation across different categories, we calculate the proportions of male and female representation within each category. Figure 1 shows these proportions for both books and music. For books, we observe a general female dominance in several categories, mainly in "Romance", "Young Adult Fantasy", and "Young Adult Fiction". Conversely, categories such as "Science Fiction/Technology" and "Historical & Biography" are predominantly male. Specific categories, including "Fantasy" and "Mystery/Thriller", exhibit a more balanced gender representation.

In contrast, the music industry reveals a predominantly male representation across most categories, as shown in Figure 1(right). Categories such as "Latin", "Dance/Electronic", "Rap", and "Country" are heavily male-dominated, whereas "Adult Contemporary" and "Pop" demonstrate a more balanced gender distribution. These observations align with previous research indicating that many music genres are skewed toward male artists, reflecting longstanding gender disparities within the industry

[7]Note that reaching the first position on a Billboard Chart may limit the scope of the success rate, as it only represents one measure of success. However, we focus on this metric due to its prominence and the significant visibility it provides to artists.

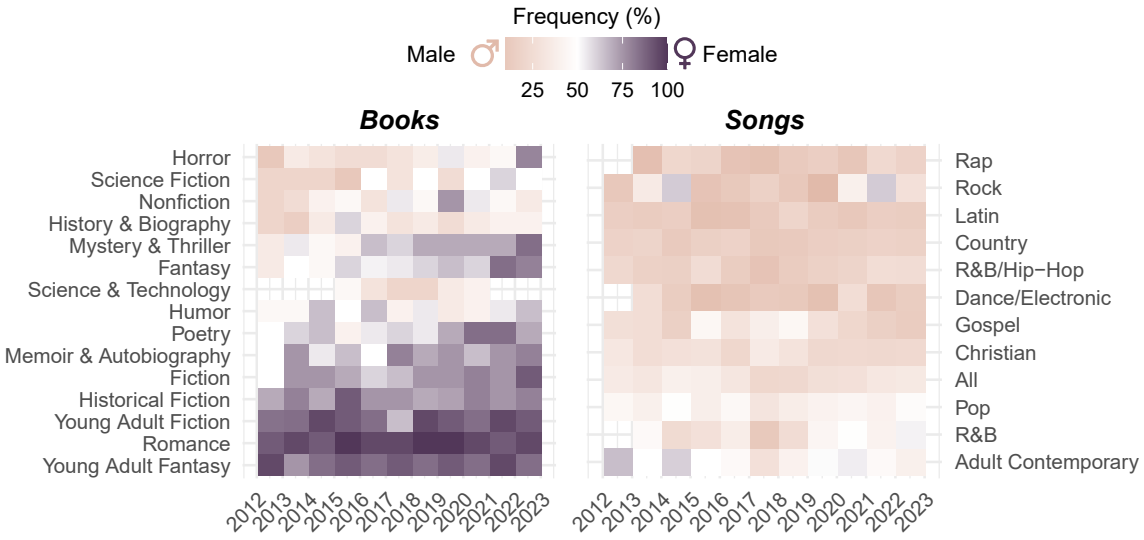

Figure 2. Temporal gender representation trends in the book industry.

[Epps-Darling et al. 2020, Watson 2020]. However, it is important to consider that these results are based on data from Billboard's charts, which primarily reflect trends within the U.S. market. Gender representation in other regions or countries may vary, and the data may not capture all aspects of global music industry dynamics.

These findings answer **RQ1** by revealing a general trend of female dominance in books (over the Goodreads platform) regarding fiction genres and male dominance for nonfiction ones. In other words, there is a greater representation of female voices and perspectives in fictional storytelling, as well as a gender imbalance in factual and informative literature representation. Also, for music, the gender preference analysis unveils a distinct pattern, with most categories being male-dominated and few exhibiting a relatively equal proportion of male and female artists. Such a result suggests a more diverse and inclusive representation within Adult Contemporary and Pop in a lesser degree.

4.2. Temporal Trends in Gender Representation

To answer **RQ2** regarding whether there has been a change in gender representation within different genres over time, we explore historical data and trends to assess any shifts in gender representation patterns within both book and music industries. Overall, Figure 2(left) highlights categories in the book industry that have transitioned towards a more balanced representation, indicating progress toward gender parity. For instance, categories including "Humor", "Poetry", and "Nonfiction" exhibit a noticeable shift towards an equal proportion of male and female authors over time.

On the other hand, Figure 2(left) also reveals categories that have maintained or intensified their gender imbalances, indicating persistent disparities. Categories such as "Romance", "Fantasy", and "Young Adult Fantasy/Fiction" continue to show a higher proportion of female authors, whereas "History & Biography" and "Science Fiction/Techonology" categories show continuous high proportion for male authors. Moreover, there are some categories with an evident shift in gender representation, transitioning from the dominance of one gender to another over time. This shift is observed in "Fantasy", "Horror", "Mystery & Thriller", and "Nonfiction", where there has been a

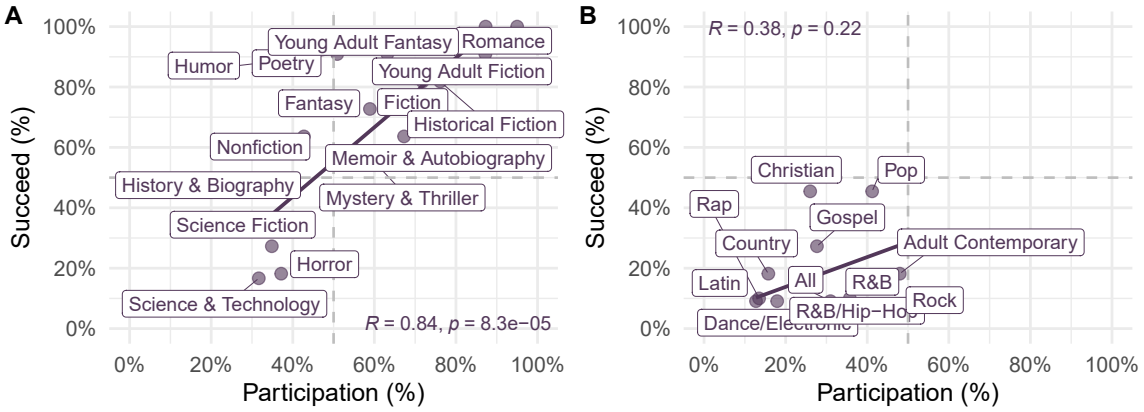

Figure 3. Correlation between female participation and success ratios across (A) book and (B) music categories. Pearson correlation coefficients are depicted in the top left corner (the dashed lines are at 50%).

transition toward a higher proportion of female authors.

In contrast, Figure 2(right) reveals a persistent and intense male dominance within the music industry over time. Unlike the shifting patterns observed in the book industry, the analysis of gender representation in music genres shows a continued imbalance. Most music categories maintain a significant male dominance, indicating limited progress in achieving gender parity within the industry. Categories such as "Latin", "Dance/Electronic", "Rap", and "Country" exhibit a strong skew towards male artists, with a disproportionately lower representation of female artists.

Such result highlights gender bias and structural challenges in the music industry. The underrepresentation of female artists calls for organized efforts to address systemic barriers, promote inclusivity, and create equal chances regardless of gender. The stark contrast between book and music industries asks for targeted interventions within the music industry to promote gender equality. It emphasizes the importance of fostering an environment that supports and uplifts the voices of female artists, promotes diverse representation, and challenges the prevailing norms that perpetuate gender imbalances.

4.3. Gender Participation and Success Ratios

To answer **RQ3**, we investigate if there is a correlation between gender and the likelihood of winning a Goodreads Choice Award or reaching the first position on a Billboard Chart. First, we compute the participation and success ratios for male and female authors/artists in these accolades. The participation ratio represents the percentage of male or female authors/artists participating in the awards or chart rankings. In contrast, the success ratio represents the percentage of male or female authors/artists who actually won an award or reached the first position on a chart.

Correlation Analysis. Figures 3**A** and 3**B** illustrate the correlation between the participation and success ratios in the book and music industries, respectively. To maintain clarity and focus on the gender-specific values, they show only the female data points, as we are dealing with binary data. As expected, both industries have a significant and positive correlation between participation and success ratios. This indicates that a higher participation rate among female authors/artists is associated with a greater likelihood of success

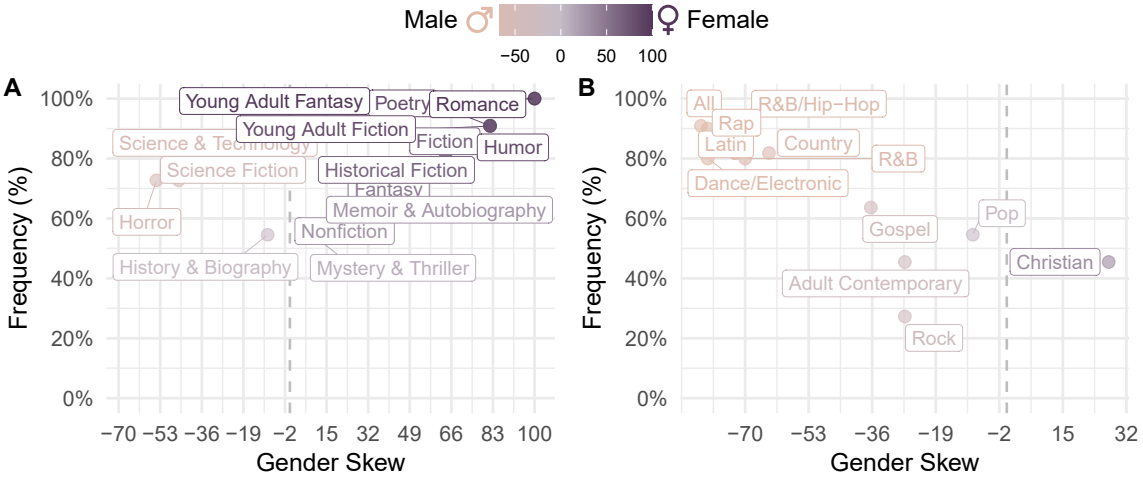

Figure 4. Success disparity analysis, based on the gender skew score of (A) authors and (B) artists (the dashed line is zero).

in winning prestigious awards or reaching top positions on the charts. Conversely, lower participation rates are generally accompanied by lower success rates.

While this result might seem obvious—more opportunities to participate lead to more opportunities for success—it reinforces an important point. Representation matters. The more female authors/artists are included, the more success they can achieve. The more female authors/artists are included, the more success they can achieve. This highlights the ongoing need to break down barriers that limit female participation. Without adequate representation, their chances of recognition and success remain disproportionately low, underscoring the importance of fostering gender inclusivity in these industries.

Gender Skew. To enhance our analysis, we introduce a gender skew score to measure the likelihood of success for male and female authors/artists. Such a score is computed as the difference between the success ratios of female and male authors/artists. A positive score indicates a higher success rate for female authors/artists, a negative score suggests a higher success rate for male authors/artists, while a score close to zero signifies a more balanced representation. Calculating the gender skew score enables to evaluate the magnitude of the gender disparities in terms of success rates. Figures 4**A** and 4**B** show the gender skew analysis for the book and music industries, respectively.

Figure 4**A** shows a more balanced gender skew, with book categories such as "History & Biography", "Mistery & Thriller", and "Nonfiction" exhibiting a skew score close to zero. This indicates a relatively equal likelihood of success for both male and female authors within these categories. On the other hand, categories such as "Romance", "Poetry", and "Young Adult Fantasy/Fiction" display a very high positive skew score, suggesting a higher success rate for female authors, whereas "Science Fiction/Technology" and "Horror" male-dominated categories show a negative skew score.

In the music industry, Figure 4**B** reveals a more pronounced gender skew, with several categories showing a significant negative skew score, indicating a higher likelihood of success for male artists. Categories like "Rap", "Latin", "R&B/Hip-Hop", and "Country" exhibit a notable negative skew, suggesting a clear male dominance in terms of success. Conversely, categories such as "Pop", "Adult Contemporary", and "Rock" show

a skew score closer to zero, indicating a relatively more balanced representation. The only positive skew score observed in the music industry is for the "Christian" category, which indicates a higher likelihood of success for female artists in this specific category.

These findings emphasize the existence of gender disparities and bias within the music industry, where certain genres tend to favour male artists in terms of success. The negative skew scores highlight the need for efforts to address and overcome these disparities, promoting equal opportunities and recognition for artists of all genders. Overall, this final analysis sheds light on the gender dynamics within the book and music industries, answering our **RQ3** and highlighting specific genres or categories where gender disparities in success rates are particularly pronounced.

5. Conclusion

This paper presents a comprehensive study that delves into gender representation within the book and music industries. Our findings show a general trend of female dominance in the book industry regarding fiction genres and male dominance in nonfiction ones. Within the music industry, most categories are male-dominated, and few exhibit a relatively equal proportion of male and female artists. Moreover, the temporal analysis of the book industry reveals a diverse gender representation trend, with some categories transitioning towards a more balanced representation, and others maintaining or intensifying their gender imbalances. This scenario differs in the music industry since all categories maintain a high gender imbalance. Finally, there are specific genres or categories in which gender disparities in success rates are particularly pronounced when compared to others.

By presenting this comprehensive study, we contribute to the ongoing conversation surrounding gender representation in the book and music industries, shedding light on significant challenges within such industries. Addressing these challenges requires concerted efforts from industry stakeholders, policymakers, and society as a whole. It involves fostering inclusive environments, challenging gender stereotypes, promoting equal pay and recognition, amplifying underrepresented voices, implementing comprehensive policies against harassment and discrimination, and providing equitable access to opportunities. Through the promotion of diversity, inclusivity, and gender equality within the book and music industries, a dynamic and inclusive cultural landscape can be cultivated.

Limitations and Future Work. A main limitation of our study is that it relies on data from specific platforms, which may not capture the entire industry landscape. Consequently, our findings may not universally apply to all regions or cultural contexts. Second, our analysis uses a binary gender framework, which overlooks non-binary representation within these industries. Finally, we focus on gender representation and success metrics, but this may not capture other contextual factors, such as race, ethnicity, and socioeconomic background. In future work, we plan to address all such limitations and explore additional factors contributing to gender representation within the entertainment industry.

Acknowledgments. Work partially funded by CAPES, CNPq, and FAPEMIG.

References

Betti, L. et al. (2023). Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science*, 12(1):10.

Brannon Donoghue, C. (2020). Hollywood and gender equity debates in the# metoo time's up era. *Women in the International Film Industry: Policy, Practice and Power*, pages 235–252.

Bucur, D. (2019). Gender homophily in online book networks. *Inf. Sci.*, 481:229–243.

Ekstrand, M. D. and Kluver, D. (2021). Exploring author gender in book rating and recommendation. *User Model. User Adapt. Interact.*, 31(3):377–420.

Epps-Darling, A. et al. (2020). Artist gender representation in music streaming. In *ISMIR*, pages 248–254.

Heritage, F. (2020). Applying corpus linguistics to videogame data: Exploring the representation of gender in videogames at a lexical level. *Game Stud.*, 20(3).

Istead, L. et al. (2022). Evaluating gender bias in film dialogue. In *NLDB*, volume 13286 of *Lecture Notes in Computer Science*, pages 403–410. Springer.

Kagan, D. et al. (2020). Using data science to understand the film industry's gender gap. *Palgrave Communications*, 6(1):92.

Mangaravite, V. et al. (2022). Dedupegov: Uma plataforma para integração de grandes volumes de dados de pessoas físicas e jurídicas em âmbito governamental. In *SBBD*, pages 90–102. SBC.

Morais, J. P. M. and Merschmann, L. H. C. (2021). Uma abordagem híbrida para predição de gênero a partir de textos em português. In *SBBD*, pages 49–60. SBC.

Mukherjee, S. and Bala, P. K. (2017). Gender classification of microblog text based on authorial style. *Inf. Syst. E Bus. Manag.*, 15(1):117–138.

Oliveira, G. P. et al. (2020). Detecting collaboration profiles in success-based music genre networks. In *ISMIR*, Montreal, Canada.

Salles, I. and Pappa, G. (2021). Viés de gênero em biografias da wikipédia em português. In *BraSNAM*, pages 211–216, Porto Alegre, RS, Brasil. SBC.

Silva, M. O. et al. (2021). Exploring brazilian cultural identity through reading preferences. In *BraSNAM*, pages 115–126, Porto Alegre, RS, Brasil. SBC.

Silva, M. O., Oliveira, G. P., and Moro, M. M. (2024). Premiação de Mulheres na Literatura e na Música: Análise de Dados da Billboard e do Goodreads. In Wolff, C. S. and Schmitt, E., editors, *A internet como campo de disputas de gênero*, pages 185–197. Cultura e Barbárie, Florianópolis.

Smith, S. L. et al. (2020). Inclusion in the recording studio? *RATIO*, 25(16.8).

Szasz, T. et al. (2022). Measuring representation of race, gender, and age in children's books: Face detection and feature classification in illustrated images. In *WACV*, pages 3371–3380. IEEE.

Waldfogel, J. (2023). The welfare effect of gender-inclusive intellectual property creation: Evidence from books. Technical report, National Bureau of Economic Research.

Watson, J. E. (2020). Programming inequality: Gender representation on canadian country radio (2005-2019). In *ISMIR*, pages 392–399.