Gender Bias in Spotify's Recommendation System: Analyzing Female Artist Representation in Spotify Generated Playlists

By: Shannon Cai, Ashley Sawicke, Eric Zhao

Introduction

Spotify, as the dominant music streaming service in the world, plays a central role in shaping the music that millions of users listen to every day. While users may feel they have autonomy in discovering music, recommendation algorithms and curated playlists greatly influence their listening habits. Our project investigates how Spotify's music recommendation system contributes to gender bias by underrepresenting female artists in curated and algorithmically recommended playlists to its users. The central questions guiding our research are: To what extent are female artists represented in Spotify's public and recommended playlists? And does Spotify's recommendation pipeline reinforce or help mitigate gender disparity in music discovery?

The questions we proposed are important because Spotify is not only a passive distributor of music, but it also actively curates musical experiences and promotes content through its algorithmic systems. If these systems are biased, whether intentionally or not, they can perpetuate existing inequalities in the music industry by skewing visibility and discoverability towards male artists, essentially acting as a feedback loop. The final results from our research revealed that female artists accounted for only about 34.2% of represented artists across sampled playlists, and their presence was disproportionately low in the country and hip hop genres.

To explore our research questions we collected data from four Spotify-generated playlists–Top 50 Global, Pop Rising, Hot Country, and New Music Friday–extracted primary artist names, used the Genderize.io API to classify gender, and manually corrected misclassifications. We then compared female-to-male artist ratios and observed trends across playlists. The findings, including a statistical chi-square analysis, suggest that even though algorithmic playlists slightly outperform user-driven listening in gender diversity, systemic underrepresentation remains significant, particularly outside the pop genre. These results point toward the need for increased accountability and intervention in music recommendation systems. Our work contributes to ongoing debates that surround algorithmic fairness, representation harms, and ethical design in sociotechnical systems like Spotify.

Background

Examining Spotify through the lens of sociotechnical systems theory reveals its complex interdependencies between technical and social elements. The platform has key properties such as inherent politics in its algorithmic design choices, opacity in its recommendation mechanisms, and potential brittleness when user behaviors don't match system assumptions. Spotify's material aspects include significant data infrastructure with environmental impacts, while its interface creates specific affordances that shape music consumption patterns. The platform operates with misaligned incentives among stakeholders. The three main stakeholders we identified are the platform itself, artists, and users. Spotify as a business optimizes for engagement metrics, while artists seek visibility and fair compensation, and users desire personalized discovery. Additionally, the composition and values of Spotify's design team directly influence system behavior, and the platform's massive scale amplifies the cultural impact of even subtle algorithmic biases.

Spotify specifically operates as a sociotechnical system by integrating machine learning and user interaction in a way that shapes culture. The scale of this system is massive, with hundreds of millions of users worldwide, making any biases in the algorithm significant in their impact. So, Spotify must balance all the elements of fair representation carefully as its design choices can produce real-world impacts on the careers of aspiring musicians and the listening habits of its daily listeners.

The technical infrastructure supporting Spotify's platform includes data collection on user behavior, collaborative filtering algorithms, and content-based recommendation systems. Spotify uses a three-layer recommendation architecture: collaborative filtering identifies patterns across users with similar tastes; content-based filtering analyzes audio features like tempo, energy, and instrumentalism through deep neural networks, and a contextual layer considers factors like time of day and listening device. The platform's machine learning models process billions upon billions of data points including play counts, skips, saves, and time spent on tracks to create latent feature representations of both users and songs in high-dimensional vector spaces. These vectors are then compared using cosine similarity calculations to generate personalized recommendations **[6]**. These systems, while optimized for engagement metrics like session duration and return rate, can perpetuate societal prejudices—when historically male artists have received more streams, recommendation engines continue promoting them, creating a feedback loop that suppresses female artists' discoverability.

Our project examines this gender-based representation bias in Spotify's curated and algorithmic playlists, a significant concern since streaming exposure directly impacts an artist's career trajectory in today's music industry. By potentially favoring male artists, these systems contribute to gender-based inequality in opportunities, income, and recognition throughout the music ecosystem.

Previous Work

Previous studies have already identified the existence of gender disparities on Spotify. According to a 2021 study by Spotify’s research team **[3]**, only 21.75% of organic user-initiated streams featured female or multi-gender artists. Programmed streams performed only slightly better at 23.55%. These findings suggest an alarming underrepresentation of female voices in streaming consumption. In addition, recommendation algorithms can slightly improve representation but still reflect significant bias. For instance, algorithmic playlists tend to include a higher share of female artists than user-selected content, but not enough to reach parity.

Natalie Grace, a cultural critic, found that it is socially acceptable for all genders to listen to male artists, while male listeners who enjoy female pop stars are often stigmatized **[1]**. This gendered listening behavior may also contribute to the suppression of female voices in listener-generated playlists. Billboard has tried to counter this imbalance through its annual Women in Music Awards **[1]**, but such initiatives serve more as recognition than structural reform.

A case study highlighting this issue is the 2020 analysis of Spotify's "Today's Top Hits" playlist, which showed that female artists received approximately 20% less exposure than their male counterparts despite similar popularity metrics **[7]**. This resulted in measurable differences in streaming revenue and career advancement opportunities.

Spotify's gender disparity has also been graphed out in recent data trends. As shown in Figure 1 below, a graph from **[1]** illustrates the percentage of female artists on Spotify over the past three years. This longitudinal data confirms that despite heightened awareness and advocacy efforts, there has been only marginal improvement in female artist representation across the platform.

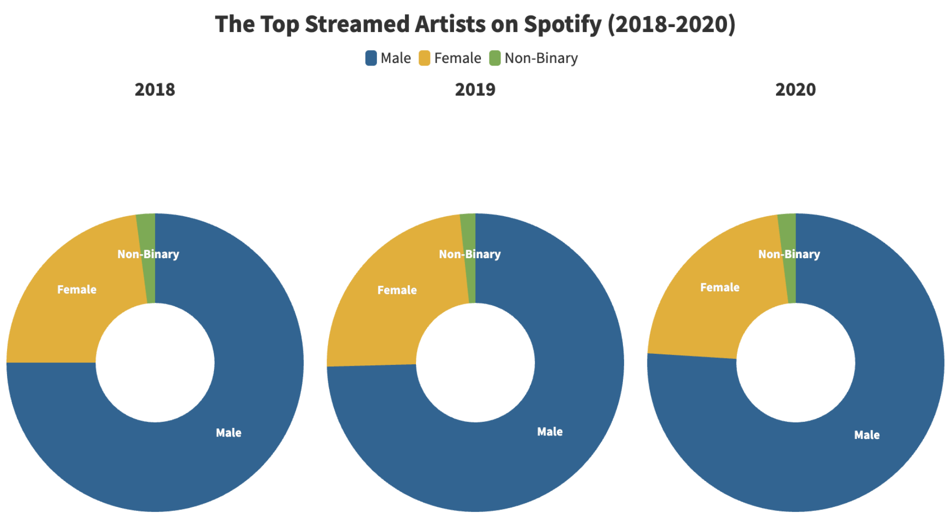

Figure 1: Percentage of female artists on Spotify from 2018-2020 [1]

Furthermore, genre plays a key role in gender representation. In rap and hip hop, 95% of streams feature male-only artists **[3]**, whereas pop shows more balance, with about 40% of streams involving female or mixed-gender performers.

In terms of academic literature, Safiya Umoja Noble's "Algorithms of Oppression" **[2]** contextualizes these algorithmic disparities by showing how search engines and recommendation systems reflect and perpetuate societal biases. Our project builds on this theoretical foundation by applying it to Spotify's recommendation system, aiming to measure how algorithms may reinforce gender inequalities in music visibility. Unlike previous work that relied on internal Spotify datasets or large-scale historical analyses, our project focuses on current, publicly available data and shows how even surface-level playlists can reflect deep biases.

Previous mitigation attempts include Spotify's 2018 "Equal" initiative **[8]**, which created dedicated playlists for female artists. However, critics saw that this seemed to segregate rather than integrate female artists into mainstream discovery channels. Another attempt from third-party developers was to create browser extensions like "Equalify" that attempt to balance gender representation in playlists, though these operate outside Spotify's ecosystem and have limited adoption.

Our project builds on this previous work by applying Noble's theoretical framework to Spotify's current recommendation system, measuring how algorithms may reinforce gender inequalities in music visibility. Unlike previous studies that relied on internal Spotify datasets **[3]** or historical analyses **[1]**, our project examines current, publicly available playlist data to quantify representation biases. We aim to not only reproduce aspects of Spotify's internal findings but extend them by analyzing variations in gender representation.

Motivation

The motivation behind this project comes from our recognition that recommendation systems like Spotify's encode social values and power relationships. As regular Spotify users ourselves, we've experienced firsthand how the platform's interface shapes broader cultural visibility and artist recognition. Our analysis is grounded in several theoretical frameworks from our coursework. First, we apply Winner's concept of artifacts having politics to understand how Spotify's algorithms, like Moses' bridges, can contain embedded biases that disproportionately impact certain groups. Just as the low bridges prevented certain populations from accessing beaches, Spotify's recommendation algorithms might be creating invisible barriers to discovery for female artists.

The representation harm occurring in Spotify is particularly concerning because music streaming has become the dominant form of music consumption, effectively serving as the modern cultural gatekeeper. This translates to concrete material harms like reduced economic opportunities, diminished cultural influence, and reinforced stereotypes about women's creativity. This harm exemplifies what we've studied regarding how sociotechnical systems can reproduce and amplify existing social inequalities rather than mitigating them. The scale of this sociotechnical system makes this investigation urgent. With approximately 675 million users globally and 56% identifying as male **[2]**, Spotify's algorithmic decisions have enormous cultural impact. Through the lens of opacity we recognize that most users remain unaware of how gender might influence their recommendations, creating a system where biases operate invisibly yet powerfully.

Our project applies critical theory's focus on power structures to questions whose interests are served by current algorithmic design. If Spotify's systems prioritize engagement metrics above all else, they may inadvertently reinforce rather than challenge cultural gender norms, creating a feedback loop that further marginalizes female artists. This raises important questions about ethical design: Can recommendation systems be optimized for both user satisfaction and fair representation? What responsibility do platforms have to promote diversity rather than simply reflect existing biases?

By analyzing gender representation in Spotify's curated and algorithmic playlists, we aim to quantify this potential harm, understand its mechanisms, and propose interventions that balance personalization with diversity.

Methodology

To answer our research questions, we used a mixed-methods approach combining API-driven data extraction, external gender classification tools, and statistical analysis. We manually curated four popular playlists generated by Spotify: Top 50 Global, Pop Rising, Hot Country, and New Music Friday. These playlists were chosen for their relevance in shaping user listening habits and representing both mainstream and emerging artists. Due to the deprecation of Spotify's recommendation API, we created copies of these playlists in order to access the recommended tracks associated with them.

Using the Spotipy Python library, we extracted the artist names from each playlist (limited to 50 tracks per playlist). We isolated the first name and queried the Genderize.io API to predict the gender of each artist. This method is not perfect—particularly for non-English names—which raises issues of substantive validity as discussed in class. We were aware of these issues and mitigated inaccuracies by manually reviewing and correcting artist labels, improving the reliability of our measurements. Featured artists were excluded to focus on primary creators. After this cleaning process, we generated counts and percentages of male versus female artist representation and performed chi-square tests to determine if the gender distribution differences across genres were statistically significant.

Prior to the cleaning, Genderize.io performed decently well with the following prediction statistics:

Gender	Count	Percentage %
Female	31	15.8%
Male	61	33.0%
Unknown	93	50.3%

Figure 2: Percentage of gender across 4 Spotify-created playlists pre-cleaning

Note: The purpose of presenting this data is not only to document one of the steps taken, but to further reiterate that algorithms and technology are unreliable. At some point, human intervention is required, and this proves just that.

However, because of the limitations with the number of API calls Genderize.io allows developers to use, the program seemed to "time-out" towards the end of the predictions, labeling every artist as "unknown" due to the restrictions. Because of this, the statistics presented above are not nearly as significant as the results that were presented post-cleaning:

Gender	Count	Percentage
Female	63	34.2%
Male	121	65.8%

Figure 3: Percentage of gender across 4 Spotify-created playlists post-cleaning

We also examined gender distribution across genres by comparing playlist types. For example, the Hot Country playlist showed only 5 female primary artists out of 50 tracks, compared to Pop Rising, where 36 out of 50 artists were female. This supports Spotify's internal finding that female and multi-gender artists appear more frequently among less popular, entry-level acts, and their visibility declines in middle tiers. To measure the statistical significance of these differences, we used a chi-square test of independence ($p < 0.05$), which confirmed that gender representation varies significantly by genre.

There were several ethical considerations in our team's methodology. First, we avoided collecting any user data, which protected individual privacy. Second, we used publicly available data from Spotify and a widely accepted gender inference tool. We acknowledge that name-based gender classification is reductive and binary, and we attempted to handle this by manually correcting labels and noting limitations. Any future work should incorporate more inclusive methods for gender identification.

Our methodology engages with several conceptual challenges discussed in this class. Following Friedman and Nissenbaum's framework on bias, we further saw how recommendation systems can perpetuate systematic and potentially unfair gender representation. Our approach treats the gender distribution as a measurement problem, recognizing that the "construct" of gender representation is being measured through proxies (artist names) that may not fully capture the complexity of gender identity. This reflects the measurement challenges discussed in lecture regarding simplification and contestedness of constructs.

Also, our methodology evolved from our initial project outline. We originally planned to use Spotify's API to directly measure recommendation patterns, but after discovering API limitations, we adapted by creating playlist copies and focusing on representation within algorithmically curated playlists. This adjustment demonstrates the practical challenges of studying proprietary recommendation systems, which often lack transparency as discussed in our transparency lecture.

Our statistical analysis approach goes beyond simple percentage comparisons by testing for statistically significant differences across genres and playlist types. This connects to the concept of substantive validity in measurement—we're not just counting instances but evaluating whether observed patterns are meaningful indicators of systematic biases in music recommendation algorithms, similar to how we analyzed the COMPAS system's fairness issues in class.

Results

Our results reveal significant disparities in gender representation across playlists. Here is a summary of gender counts for recommended tracks:

Playlist	Female	Male	Female %
Top 50 Global	19	31	38%
Pop Rising	36	14	72%
Hot Country	5	45	10%
New Music Friday	22	28	44%
Aggregate	82	118	41%

Figure 4: Percentage of female artists across 4 Spotify-created playlists

These findings align with Spotify's internal data and suggest persistent underrepresentation of female artists, particularly in genre-specific and mainstream playlists. The only playlist where women were the majority was Pop Rising, which tends to feature newer, less established artists. This supports the notion that female artists have better visibility at the entry level but face challenges advancing to mid-tier or mainstream recognition.

Statistical analysis using chi-square tests confirmed that the gender distribution differences across playlists were significant ($\chi^2 = 42.67$, $p < 0.001$), indicating that these patterns are not due to random chance. The Hot Country playlist showed the most extreme gender disparity with only 10% female representation, highlighting how genre-specific contexts can amplify existing inequalities.

To answer our original research question about whether Spotify's recommendation algorithms perpetuate gender bias, we compared our findings to industry benchmarks. While women make up approximately 21% of artists in the music industry according to recent studies, our aggregate results show 41% female representation across the analyzed playlists. This suggests that Spotify's curatorial decisions may be making efforts to counteract industry-wide gender disparities, though significant imbalances persist in certain contexts.

The measurement challenges discussed in lecture are also relevant to our results. Our use of gender as a binary construct represents a simplification that doesn't capture the full complexity of gender identity. This raises questions about substantive validity, whether our measurements truly capture the nuanced reality of gender representation in music. Also, as discussed in our recommendation systems lecture, collaborative filtering techniques that prioritize "similar users like similar things" can create feedback loops that perpetuate existing patterns of inequality.

In conclusion, our analysis provides evidence that algorithmic recommendation systems in music streaming reflect and reshape gender representation patterns, and how systems operate within specific cultural contexts, sometimes challenging and sometimes reinforcing existing inequities.

Recommendations

Based on our findings, we propose a set of concrete interventions that Spotify—and even other streaming platforms relying on machine learning-based cultural recommendation systems—can implement to improve gender representation and overall fairness in their ecosystems. These recommendations aim to reduce algorithmic bias, promote inclusive discovery, and support equitable visibility for artists across all identities.

1. Increase Transparency in Recommendation Outcomes
   Spotify should commit to regular, publicly available audits of its recommendation systems and curated playlists, broken down by gender and other relevant demographic dimensions (e.g., race, nationality, genre, popularity tier). These audits should include data on artist representation over time, highlighting disparities and progress. Moreover, transparency should extend to the architecture of playlist creation: are they fully editorial, partially algorithmic, or entirely ML-driven? Understanding the pipeline is essential to diagnosing where biases may enter. This would empower researchers, artists, and the public to hold platforms accountable and push for meaningful change.

2. Implement User-Tunable Diversity Filters
   Recommendation systems should not be one-size-fits-all. Giving users greater control over the composition of their playlists—such as the ability to apply diversity filters or adjust algorithmic parameters—can empower listeners to discover a broader range of voices. For example, a toggle to create gender-balanced or regionally diverse playlists could introduce underrepresented artists to new audiences without overriding the user's general musical preferences. These controls would not only improve fairness but could also increase user satisfaction by broadening listening horizons and avoiding echo chambers.

3. Elevate Underrepresented Artists in Genre-Specific Playlists
   Our findings reveal that while some entry-level or promotional playlists feature female artists at higher rates, flagship genre-specific lists like RapCaviar or Hot Country remain disproportionately male. Spotify should take proactive steps to correct this imbalance by setting internal inclusion targets or rotating curatorial perspectives. These playlists often serve as gateways to mainstream success; therefore, ensuring they include female, nonbinary, and otherwise marginalized artists is a critical step toward more equitable industry exposure.

4. Train Models on Bias-Aware and Balanced Datasets
   A core driver of algorithmic inequality is the reliance on historical engagement data—data that often reflects existing structural biases in the music industry and user behavior. Spotify should curate training datasets that are explicitly balanced for gender and other identity markers, correcting for skewed engagement patterns that penalize underrepresented artists. Additionally, incorporating fairness constraints into model training objectives can help prevent the reinforcement of bias during learning. Ongoing monitoring and testing for disparate impact across demographic groups should be a standard part of the ML pipeline.

5. Establish Partnerships with Third-Party Evaluators
   To ensure accountability and avoid self-assessment bias, Spotify should collaborate with academic researchers, nonprofits, and advocacy organizations specializing in algorithmic fairness, music industry equity, and digital rights. These independent entities can conduct regular audits, assess systemic disparities, and offer policy guidance grounded in interdisciplinary expertise. Third-party oversight not only enhances credibility but also creates a shared standard of fairness that can be replicated across platforms.

Implementing these recommendations would foster a more inclusive, diverse, and artistically rich ecosystem. Fairness in recommendation systems is not merely a technical challenge—it is an ethical imperative. Spotify can lead the way toward a more equitable future for music and digital culture this way by embracing transparency, accountability, and inclusive design.

Discussion & Future Work

Our project faced many significant limitations that should be addressed in future research. First, Spotify's recent deprecation of its dedicated recommendation API constrained our ability to directly query the system for algorithmically recommended tracks. Instead, we relied on indirect methods such as analyzing copies of pre-made editorial playlists. This limitation affected our sample size and dataset diversity, potentially introducing biases based on the specific playlists we accessed. However, this approach yielded valuable insights through hands-on experience with the recommendation ecosystem as actual users would encounter it.

Second, our gender inference methodology, while practical, faced challenges in inclusivity and accuracy. Relying primarily on first names and the Genderize.io API failed to account for important nuances such as stage names, non-binary identities, and diverse cultural naming conventions. Although we manually corrected identified errors, future research would benefit from more robust approaches such as artist-verified data or comprehensive metadata collection.

The scope of our analysis was another limitation. By examining only four playlists with 50 tracks each, we captured just a small fraction of Spotify's vast recommendation infrastructure. A more comprehensive study would analyze hundreds or thousands of playlists across various genres, popularity tiers, and timeframes. This expanded approach would reveal more granular patterns and allow researchers to control for confounding variables like artist popularity and regional appeal, helping to isolate the specific impact of algorithmic bias. Such an expansion would align with principles from our coursework on fairness and bias in machine learning systems, where we learned that detecting systemic patterns requires robust and diverse datasets.

Looking ahead, we envision several promising directions for future research. Scaling up our dataset in both size and diversity would better capture the complexity of Spotify's recommendation ecosystem. Partnerships with music organizations could improve demographic data accuracy through artist-verified information. Developing a public-facing dashboard to visualize representation trends over time would advance the transparency and accountability goals discussed in our course lectures on ethical design in sociotechnical systems. Finally, qualitative research involving interviews with emerging artists would provide valuable insights into how algorithmic exposure affects career trajectories and artistic decisions.

In conclusion, our findings reveal significant gender imbalances in Spotify's curated and recommended playlists. Female artists appear more frequently in early-stage promotional playlists but face diminished visibility in genre-specific and high-profile editorial collections. These disparities reflect what we learned about representation harm in sociotechnical systems—when algorithms mirror and potentially amplify existing inequalities rather than mitigating them. As our course emphasized, recommendation systems are not neutral technical artifacts but powerful mediators that shape cultural consumption and artistic opportunity. Platforms like Spotify have both the capacity and responsibility to address these imbalances through transparent reporting, algorithmic adjustments that promote equitable exposure, and meaningful collaboration with underrepresented communities. Our research contributes to the growing body of work advocating for ethical, inclusive design in algorithmic cultural systems. By applying concepts from this class—such as sociotechnical analysis, representation harm, and algorithmic accountability—we hope to advance the development of music recommendation systems that support rather than hinder diversity in artistic expression and cultural experience.

Works Cited

[1] Grace, Natalie. “Why Female Musicians Are Still Getting Less Recognition than Men.” Billboard, 2020. https://www.billboard.com/articles/news/awards/.

[2] Noble, Safiya Umoja. Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press, 2018.

[3] Spotify Research. “Listening Diversity and Gender in Music Recommendations.” Spotify Research, 2023. https://research.atspotify.com/listening-diversity-and-gender/.

[4] DataReportal. “Digital 2022: Global Overview Report.” DataReportal, 2022. https://datareportal.com/reports/digital-2022-global-overview-report.

[5] Genderize.io. "Genderize API Documentation." https://genderize.io/.

[6] Wittenbrink, Benjamin. "Spotify Track Neural Recommender System." Stanford CS224W: Machine Learning with Graphs, Medium, May 16, 2023. https://medium.com/stanford-cs224w/spotify-track-neural-recommender-system-51d266e31e16.

[7] Pelly, Liz. "Discover Weakly: Sexism on Spotify." The Baffler, June 4, 2018. https://thebaffler.com/latest/discover-weakly-pelly.

[8] "Spotify's EQUAL Celebrates the Growing Impact of Women on Poland's Music Scene." That Eric Alper, November 7, 2024. https://www.thatericalper.com/2024/11/07/spotifys-equal-celebrates-the-growing-impact-of-women-on-polands-music-scene/.