SIDE BY SIDE: A Digital Humanities Analysis of the Broadway Musical

By Alison Silldorff

Advised by Professor Brian Kernighan

A Senior Thesis submitted to the Department of Computer Science in partial fulfillment of the requirements for the degree of Bachelor of the Arts

Princeton University

May 2025

I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Alison Silldorff

Abstract

How can data deepen our understanding of Broadway musicals, their history, their actors, and their writers? This project is a data-driven digital humanities approach to understanding the Broadway musical and its related genres on Broadway. Through data collected from the Internet Broadway Database (IBDB), this project reveals the current limitations of digital theatre records and data, which in turn demonstrate the definitional challenges of this form that make it both difficult to work with and fascinating to parse. The data are analyzed to explore the development of the musical genre, the history of screen musicals, and historical demographic and career information about Broadway musical actors and writers. These analyses employ quantitative reasoning to understand theatre narratives in terms of the writers that defined the genre, examine structural gender inequality in the theatre industry, and support and question historical accounts of the Broadway musical. Furthermore, these analyses emphasize the need for further data-driven theatre research and interdisciplinary approaches to musical theatre scholarship.

Acknowledgements

I would like to first thank my advisor, Professor Brian Kernighan, for supporting this research over the past year with sound wisdom, challenging questions, and encouragement to follow the lead of interesting data in any form. I would also like to thank Professor Stacy Wolf, whose advising on this thesis and support throughout my Princeton career have broadened my understanding of and love for musical theatre.

Thank you to my friends Faith Wangermann and Cooper Kofron, who have been my resident sounding boards for odd theatre questions and advice. Thank you for listening to my rants, talking through edge cases, and for making your amazing theatre knowledge and curiosity so easily available to me.

Lastly, to my family: Thank you for your love, support, and laughter, without which I would never have reached this point in my academic journey. Thank you for teaching me the joy of music, for supporting my simultaneous academic and artistic endeavors, and for always encouraging me to find my own way and embrace adventure wherever it presents itself.

Contents

1. Introduction

The term “Broadway musical” calls to mind flashy costumes, bright lights, expensive tickets, loud voices, and catchy tunes. It is an artistic medium which has been thoroughly studied by historians and critics, both in terms of the art form itself and as its function as a pillar of American popular culture. However, musical theatre has rarely been studied quantitatively and in terms of data.

With the rise of digital humanities, a field that applies computational and quantitative methods to humanities fields, theatre has been one of the more recent disciplines to be interpreted with this approach. As such, the existing body of research has only scratched the surface of the potential sources and uses of musical theatre data, especially in terms of theatre as an event rather than a text. Pieces of data we could collect about musicals as events include the performance location, runtime, ticket sales, crew members on staff, the colors of the set, the shoes of the actors, or the lumens of the spotlights. If we disregard the feasibility of collecting some of these data, each of them could tell a fascinating story about that event or its historical context.

Thinking more practically, some of these data on Broadway musicals are currently available to us, most concisely by The Broadway League’s Internet Broadway Database (IBDB). This database keeps a comprehensive record of all Broadway productions, theatres, actors, crew members, and more. Using this information, alongside other data sources, this paper analyzes the Broadway musical from a digital humanities lens, exploring how information regarding genre, form, actors, and writers can support and challenge the existing narratives of theatre history. This paper also details the difficulties of treating theatre events as data to begin with, and explores what the definitional tensions themselves say about the ambiguous bounds of musical theatre.

In analyzing my data, I explore genre features of musical theatre history, overlap between stage and screen musicals, and the career trajectories of Broadway actors and writers, among other categories. Each of these explorations highlights the need for more data-driven musical theatre studies, while also emphasizing the need to keep these studies interdisciplinary and put them in conversation with humanistic research. Furthermore, this process highlights the current limitations of digital theatre data resources, and the challenges these pose to both computer scientists and theatre scholars.

1.1. What counts as a Broadway musical?

Among the first challenges when approaching the Broadway musical is answering the question:What counts as a Broadway musical? A Broadway show must be performed in a Broadway theatre, and a Broadway theatre is distinguished by its seat capacity and geographic location. A Broadway theatre must have 500 or more seats, and must be located in Manhattan, New York, primarily but not exclusively in the Theatre District [1]. There are currently 41 Broadway theatres, but the set of Broadway theatres has changed over time due to destruction and construction of theatres, as well as changing restrictions on what counts in this category.

A Broadway show is not, therefore, defined at all by the genre of performance (e.g. play, musical, revue, musical comedy), success of the show, or other qualities that might be generally associated with the word "Broadway." In *Showtime: A History of the Broadway Musical Theater,* Larry Stempel describes such associations with the term: "It goes a long way dispelling ambiguities. Be it as an indicator of geography or of character, 'Broadway' immediately puts any discussion of musical theater into a modern American framework that distinguishes it" [18]. As such, the term "Broadway musical" is evocative of specific aesthetics, genres, styles, and

individuals. But for the purposes of a digital humanities project, this definition must be as clear as possible.

The second component of this definitional question, what is a "musical" in terms of form, is difficult to answer both historically and currently and is an ongoing topic of debate among scholars and popular audiences alike. Historically, many forms of music theatre outside of the "Broadway musical" we think of today have occupied Broadway stages, including opera, operetta, ballet, burlesque, and minstrelsy. These forms all led to the development of musical theatre, which itself is usually cited as beginning in either 1866 with *The Black Crook* or 1927 with *Show Boat* [2].

For the purposes of this paper, I will use the term "musical" to describe musical theatre as we know it today. A musical, then, consists of song, dance, and scenes, which crucially all work towards a central plot. By this definition, we can consider the first musical to be *Show Boat* (1927), because it was the first work to tie the song, dance, and scenes of the whole show together by a plot. In this paper, I will then use "music theatre" to describe the works that are not musicals, but fit into one of the surrounding genres mentioned above, such as opera, operetta, burlesque, and extravaganza. A musical is a piece of music theatre, but not all music theatre works are musicals. Furthermore, in this paper I sometimes refer to this definition of "musicals" as my "strict definition" of musicals.

I would like to briefly explain some of the more ambiguously defined music theatre forms, which will come up later in this paper, especially extravaganza, revue, vaudeville, and musical comedy. In his book, Stempel charts the various forms of music theatre that led to the musical, noting that overall, "minstrel shows, burlesque, variety, vaudeville, and revues make up some of the major types of musicals without stories, though they differ from each other" [17].

This speaks to this paper's distinction of the "musical" being a piece of music theatre *with* a story (among other identifiers).

In the first chapter of *The Cambridge Companion to the Musical*, Katherine K. Preston describes the difficulty of distinguishing between these forms: "the same show might be called a 'farce-comedy', a 'revue' or an 'extravaganza'; many shows exhibited characteristics of numerous categories" [2]. That being said, Preston, defines **extravaganza** as a form evolved from pantomime that often included acrobatics and stage machinery. This term was often used interchangeably with "spectacle." **Vaudeville** shows were a slight re-imaging of variety shows, and included "instrumental solos, comic skits, dancing, juggling and acrobatics of various sorts." Stempel defines **musical comedy** as "a spoken play with musical numbers inserted […] more relaxed than light opera […] its spirit lightweight altogether" [17]. Preston defines the earliest forms of **revue** as "combined burlesque, satire, specialty acts, minstrelsy, dance […] in essence it was a variety show in the best of the American tradition" [2].

Beyond these outdated forms of music theatre, modern Broadway productions are similarly difficult to categorize, including in the context of the Tony Awards. The Tony Awards must each year decide the award eligibility of different Broadway shows, including making the seemingly clear distinction between "Best Play" and "Best Musical" eligibility. Even this, however, contains many edge cases, such as the recent *Stereophonic*, which features enough music that a cast album was released but was categorized as a play by the Tony Awards. The strong presence of music challenges our definition of what constitutes a play versus a musical versus a "play with music," though this example supports this paper's definition of a musical requiring the songs to work towards a plot.

Another subject of Tony Awards debate is the term "revival," which usually describes a new Broadway production of a show that has already been on Broadway. But what about a show that was previously off-Broadway, and many years later is revived on-Broadway? Take *Gutenberg! The Musical,* which had an off-Broadway run in 2006, but came to Broadway for the first time in 2023. At the 2024 Tony Awards, *Gutenberg! The Musical!* was nominated for "Best Revival of a Musical." These are some of the challenges, even at the broadest level, of clearly defining the musical and its core genre features.

Given these definitions of "musicals" and "music theatre," my database (DB) contains a combination of both. To be precise, my DB contains shows **from Broadway theatres** that were **given the "musical" tag on the Internet Broadway Database (IBDB)**. This encompasses both musicals as we know them today as well as the related music theatre forms such as opera, operetta, vaudeville, and several other forms that were performed in Broadway theatres.

2. Related Work

Data-based studies of musical theatre remain limited, but the scholarship that does exist unearths fascinating trends, and affirms qualitative observations in humanistic musical theatre scholarship. Most comprehensive is "Average Broadway" by Derek Miller, which examines data from all Broadway shows. In particular, he examines "average" cast sizes, production staff billing, and genre, making use of the Internet Broadway Database (IBDB) and Playbill Vault as a data source [3]. Because historical Broadway data remain mostly unexplored, Miller's guiding question regarding what constitutes the "average" Broadway show challenges what historical accounts depict as the norm.

He discusses the importance of the "forgotten middle" of theatre history, explaining how most theatre scholarship relates to extreme and outstanding cases. In fact, by quantitatively analyzing several works of theatre history, he finds that they drastically underrepresent shows that ran on Broadway for a short amount of time, and overrepresent shows with long run lengths. Here, what lies in the "forgotten middle" are shows that ran for an average but unremarkable amount of time. Although it is somewhat logical for historical works to examine the outstanding cases, he notes that "a proper historiography of the popular theatre must also be a historiography of unpopular plays," challenging scholars to think beyond the most popular works and consider the average. This work serves as a main inspiration for this paper.

In *Theater as Data*, Escobar Varela presents a variety of data-centered theatre research where theatre is treated as an event rather than as text [5]. This paper follows a similar distinction, looking at occurrences of performances rather than the textual content of shows. He details what it means to treat theatre as data within the digital humanities, and formalizes what a "computational" approach to theatre data entails. He describes data-assisted theatre research as

employing a “computational defamiliarization strategy,” which certainly speaks to Miller’s storytelling tactics in “Average Broadway” as well. Varela then moves into specific examples of theatre research, with some of the most relevant categories to this project being “Relationships as Data” and “Location as Data.” The former provides examples of network theory as it relates or could relate to theatre research, and the latter looks at the geographic distribution of theatre performances.

“Data Models for Theatre Research: People, Places, and Performance” by Jonathan Bollen presents exciting uses of digital humanities methods within theatre research. One such example models the transmission of theatre via a network of regional theatres in Australia [22]. Bollen emphasizes the novel ways in which data can be interpreted and visualized, and details many of the limitations of digital theatre records including their access and interpretability.

In “Reassessing Obscurity: The Case for Big Data in Theatre History,” Debra Kaplan advocates for theatre data analysis as a means to contextualize forgotten components of theatre history. Kaplan collects and analyzes data on members of the Yiddish theatre troupe The Vilna Troupe, and in turn is able to demonstrate the cultural centrality within the entertainment industry of these individuals. Through network analyses, Kaplan is able to challenge narratives of theatre history that put Yiddish theatre on the periphery [4].

“An Architectural History of NYC Theater” is a digital interactive map of past and present New York City performance spaces [23]. This geographical lens to understanding theater history demonstrates another creative way of viewing theatre through data. It underscores the wide-reaching capabilities of digital humanities applications.

3. Methods

For this project, I collected data from the Internet Broadway Database (IBDB) and The Movie Database (TMDB), and organized these data in a MySQL database. Below I will detail the data collection, data labeling, and database design processes.

In these sections, I will often refer to "works" and "properties." By "work" I refer to a particular stage production, movie, proshot, etc. A "property" on the other hand, refers to the intellectual property itself. A property can have several works, but a work must belong to exactly one property.

3.1. Data collection and cleaning – IBDB

According to their website**,** IBDB "provides a comprehensive database of shows produced on Broadway". The database is run by The Broadway League, the national trade association for the Broadway theatre industry, and their main data source is theatre programs [6] [24].

I set out with a goal of collecting a comprehensive list of all Broadway musicals. My first approach to scraping IBDB was to use their search function, and search for an empty string with the "musical" filter applied. According to their help page, this approach should return all musicals, but the 807 results returned oddly do *not* encompass all musicals nor is it some particular subset of all musicals. Therefore, this approach to scraping IBDB was not functional [7].

IBDB also houses separate pages for each Broadway theatre, and each of these pages records every show that has been performed at that theatre. As mentioned, Broadway musicals must be performed in Broadway theatres, so my next approach was to scrape the pages of every

current and former Broadway theatre, keeping only the shows with the "musical" tag attached to them. This approach yielded a near-comprehensive list of the musicals that IBDB has a record of, with some error on account of the fact that it is difficult to find a fully comprehensive list of all Broadway theatres.

To scrape these data, I used the Selenium library in Python, alongside BeautifulSoup. From IBDB, I scraped the following information about each production: the title; opening, closing, and first preview dates; credits for writers of the production; the opening night cast of the production; and the genre tags given by IBDB.

Most of the data about the production itself was fairly standardized and needed little cleaning before it could be entered into the database. The most difficult data to work with were those involving people, either on the production/creative team or actors.

With actor credits where doubling was employed in the production, I chose to allow my database to mirror the structure of IBDB, which is not standard. Doubling refers to when an actor plays multiple characters within a single performance of a show[1]. In some cases, IBDB lists an actor's various roles in a production in one line, delineated by slashes. This can be seen, for example, on *Hamilton*'s IBDB page, where Daveed Diggs' role is listed as "Marquis de Lafayette/Thomas Jefferson." On the other hand, the 2022 *Into the Woods* revival lists David Patrick Henry's doubled roles of "Narrator" and "Mysterious Man" on separate lines. I chose to keep these data as they were represented in IBDB, with slashes where they were originally listed, and separate DB entries for roles on separate lines.

For members of the production/creative team, it was necessary to scrape a large chunk of text in order to get specific credits. This is largely due to the fact that IBDB did not use a

[1] In "Average Broadway," Miller uses the employment of doubling to examine aesthetic trends and financial restrictions of Broadway shows [3].

standard separator for these credits, nor was there a standard way of listing a person in relation to their role. For example, a role could be listed as "role: name," "role by name," or many other formats. When people shared a role, it was also not standard whether IBDB would list them as "role: name1 and name2," "role: name1, role: name2," etc. As such, a portion of these credits had to be parsed manually.

In this cleaning process, I also discovered that some people were credited not as writing the music or lyrics of the whole show, but rather with just the music and/or lyrics of just a few songs. I decided to credit such individuals with the role of "song," and keep track of which song they wrote in a separate DB column. For individuals who wrote multiple songs, they would have a row for each song. In this process, I did not distinguish between music and lyrics, though a distinction was made when someone wrote music or lyrics for an entire show. On IBDB, and consequently in my DB, these song credits appear across many different kinds of productions and fail to appear in places that might make sense. For example, *& Juliet*, a musical currently on Broadway, has a score comprised of existing pop songs by a variety of artists. The writing credit is listed as:

> Music by Max Martin; Lyrics by Max Martin; Music and Lyrics by Max Martin and Friends: Klas Åhlund, Dido Armstrong, Jon Bon Jovi, Andreas Carlsson, Robyn Carlsson, Jessica Cornish, Cathy Dennis…

This list of singers/writers goes on for several more lines. In this example, IBDB does not indicate who wrote which song in the score, and instead groups many artists together, even though their contributions were very distinct. These inconsistencies, for better or worse, mirror the historic billing quirks and inconsistencies in Broadway playbills.

After properly separating the roles, I then moved into OpenRefine to standardize which roles I was keeping and which I was discarding. In this process, I decided to discard all credits related to the source material, and keep only writing/arranging credits related to this production. I also worked to standardize the naming of all roles, without losing distinct meanings. This includes standardizing “lyricist” and “lyrics” to both be listed as “lyrics.” Through this process I was able to significantly pare down the number of distinct roles, which later made analysis easier.

Another inconsistency I found in IBDB was their tagging mechanism for “revival” versus “original.” Typically, you would expect each property to have one original stage production, and 0 or more revival productions, however I found several cases where one property had several original productions. Some of these were revivals erroneously marked as original productions, but others fell into a third category called a “return engagement.” As with many categorical theatre terms, “return engagement” lacks a formal definition. It is different from a revival, which indicates a fresh production of the same intellectual property, in that a return engagement generally is a particular production that returns to Broadway, often because of a show’s commercial success. A return engagement keeps the production constant in some way other than the show content (eg. same cast, choreography, set, or costume design.) I ask a few questions that challenge the clarity of this distinction:

1. Let us say a musical is on Broadway and closes. Shortly after, this musical returns to Broadway with the same creative team and design, but different actors. Would these be considered the same production, and would the second be considered a return engagement?

2. Let us say a musical is on Broadway and closes. Shortly after, the musical returns to Broadway with the same creative team, design, and actors, but in a new theatre. Would the second be considered a return engagement?
    a. The *Chicago* revival is an interesting related case that answers this question "no." In January 2003, the production transferred from the Shubert Theater to the Ambassador Theater, yet IBDB has only one page for the *Chicago* revival [11]. Though in this case the production did not take a break, it gives some insight into how "different productions" are understood at large and in IBDB.
3. Let us say a musical is on Broadway and closes. Some time later, the show is produced again by the same producer, but with a different director and different actors. Is the second production a return engagement of the other if the production seems similar?

In fact, the most common occurrence of the term "return engagement" in IBDB was for shows that re-opened after the covid-19 pandemic, such as *Beetlejuice*. When shows re-opened, there were a mix of similarities and differences in the production/creative team, design, and cast, but they were thought of as the same "production," just with two separate IBDB pages and with the second production indicating a return engagement. Other musicals, however, have only one IBDB page despite not running during the pandemic. For example, *Phantom of the Opera* has only one IBDB page. This example illustrates the resistance of theatre terms to formalization because of their inherent ambiguity and case-specific definitions.

3.2. Data collection and labeling – TMDB and movie musicals

In addition to stage musicals, which are the main component of my data, I also have gathered a body of movie musicals, proshots, and other filmed musical works. A proshot refers to a professional recording of a stage production of a musical. To collect this data, I used The

Movie Database's (TMDB) free API, which gives extensive access to the database, containing thorough information on 919,353 movies [20].

Because I wanted stage musicals to be at the center of my data, I only collected screen musicals that had a stage musical equivalent. I use Table 3.2.1 to explain this distinction, as well as how I categorized each included value in my DB.

As mentioned in section 2.1, credits relating to source material were *not* included in my database for stage musicals. Similarly, non-musical source material in the form of a movie was not included in my database. This is in part because many adaptations from screen to stage can greatly alter the source material. Additionally, for many properties there is a chain of source material. For example, the musical movie *West Side Story* (1961) is based on William Shakespeare's *Romeo and Juliet* (1597), but there have been numerous different adaptations of this story, and what would count as "related to the property of the musical" is very difficult to delineate. It could be argued that Baz Luhrman's 1996 film adaptation of *Romeo and Juliet*, which uses original text from the play, could be related to *West Side Story*, while something like the animated film *Gnomeo and Juliet* (2011) would not count because it is a less pure adaptation of Shakespeare's play. It could alternatively be argued that neither of these films are related to *West Side Story* at all because both were produced after the musical was already written. It could also be argued that both are related to *West Side Story*, because these films give greater notoriety to *Romeo and Juliet* and in turn could increase ticket sales for a production of *West Side Story*. Thus, I have decided to simply exclude all non-musical source material from my database.

Table 3.2.1. Movie musicals included in my database along with their "type"

Type of screen work	Included in my database?	type	Examples
Movie musical turned into a Broadway musical	yes	movie	*An American in Paris*, *Beauty and the Beast*
Movie musical *not* turned into a Broadway musical	no		*The Wizard of Oz* (1939), *Everybody's Talking About Jamie* (2021) *
Non-musical movie adapted into a Broadway musical	**no**		***Back to the Future*, *Mean Girls* (2004)**
Movie musical adapted from a Broadway musical	yes	movie	*Mamma Mia!*, *Mean Girls* (2024)
A proshot (professional recording) of a musical that has been on Broadway	yes	proshot (Broadway) or proshot (other) **	*Newsies* (2017), *Sunday in the Park with George* (1986)
A bootleg (illegal recording) of a Broadway musical	no		Many on the internet, some of which are listed on TMDB.
A live TV version of a musical that has been on Broadway	yes	other	*The Sound of Music Live!*, *Hairspray Live!*
A recording of a concert version of a Broadway musical	yes	other	*Chess in Concert* (1989)

* *Everybody's Talking About Jamie* (2021) is a film adaptation of the stage musical that has run on London's West End, but this musical has never been produced on Broadway

** proshot (Broadway) refers to a professional recording of a Broadway production. proshot (other) refers to a professional recording of a stage production *not* on Broadway

In order to scrape this particular subset of movie musicals, I queried TMDB's API with each distinct Broadway musical title in my DB and kept only the films that contained the

"music" genre tag. Because of the particularities of what I was including in my database, I needed to manually label all of the films. If I could not tell by the writing credits whether something was the same property as a stage musical, I would do further research to fill in these gaps. In this labeling, I found many notable movie musicals that were missing, such as the recent *Wicked* film. This is due to the inconsistency of which movie musicals are given the "music" tag by TMDB.

To remedy these important missing data points, I used user-generated lists from the platform Letterboxd, a social media app for movie reviewing. Letterboxd's source database is TMDB, so each entry in the app can be easily linked back to an entry in TMDB. Using a range of keyword searches, I found ~64 such user-generated lists of screen musicals, scraped the contents of each list, and labeled these manually in the same way as with the TMDB search results.

More specifically, the labeling process involved deciding whether an entry should be included in my database, determining what kind of screen work it was, and determining which property in my database it was attached to. Fig 3.3.1 above details the first two steps of this process, and gives examples of how different entries were categorized into one of four categories: movie, proshot (Broadway), proshot (other), or other.

This process yielded 622 screen musicals that I added to my database, along with their casts and crew members. For the casts, I have decided to include the entirety of TMDB's cast credits for each film in my DB, all the way down to uncredited extras. I made this decision so as to afford the possibility of looking more granularly at actors' careers between the Broadway and film industries. For crew members, I took a similar approach to the production/creative teams of Broadway musicals, and chose to only keep "above the line" crew members.

For my junior independent work, I used this description: Above the Line (ATL) crew positions is a term that literally comes from film budgets, and which individuals are listed “above the line” versus “below the line” on those budgets. ATL positions control the creative vision for a film, and are also generally present for the entire film’s process. Which positions fall into this category is slightly different from film to film, but for my study the following positions are considered ATL: Producer, Executive Producer, Director, Director of Photography, Screenwriter, Music writer, Lyricist. Below the Line (BTL) includes all other positions [8] [9].

It must be emphasized that this list of 622 screen musicals is by no means exhaustive. As illustrated above, I used a few different approaches to get as close to a complete list as possible, but there are surely adaptations, proshots, and other eligible entries that have gone undetected by my scraping mechanisms. That being said, I feel that the second approach of using user-generated Letterboxd lists likely yielded all of the musicals that have been most commercially successful or most ingrained into popular culture. By a metric of determining the most relevant and impactful movie musicals, rather than compiling an exhaustive list of every screen representation of a musical ever made, I would deem this process successful.

3.3. MySQL Database Design

Upon scraping and cleaning all of my stage and movie musical data, I organized this data in a MySQL database, represented in fig. 3.3.1 in a MySQL Enhanced Entity Relationship (EER) Diagram. The database is primarily organized around representing people and works, and I will explain each below.

Works

Each work has a distinct 8-digit work_id. This work_id is comprised of a 4-digit property_id, 2-digit sequence_id, and 2-digit type_id. Works refer to a particular stage production, movie, proshot, etc, while a property refers to the intellectual property. A “type” is the kind of work it is, with the options being:

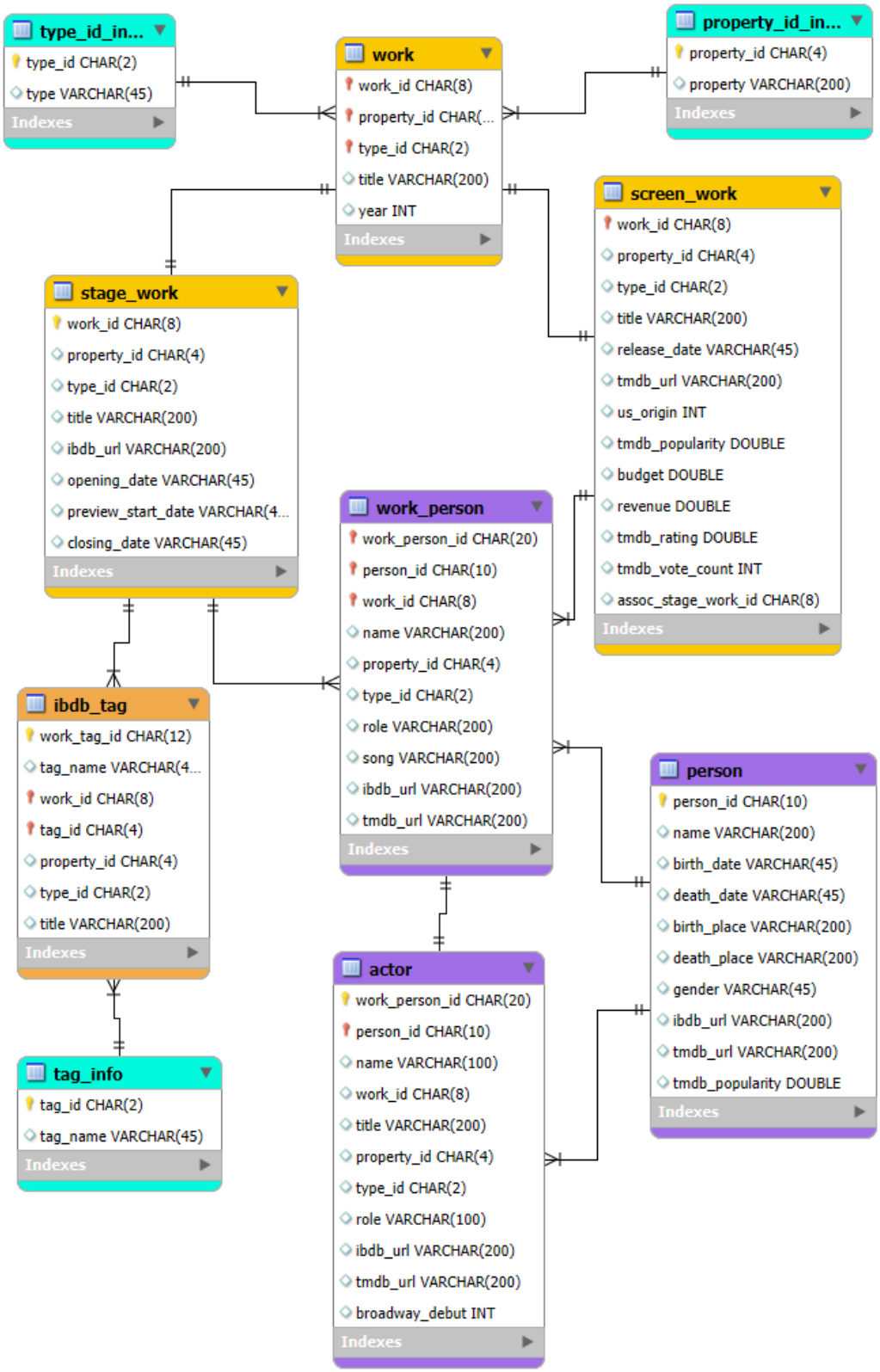
stage musical, original	screen musical, movie
stage musical, revival	screen musical, Broadway proshot
stage musical, other	screen musical, other proshot
stage musical, return engagement	screen musical, other

The two sequence digits serve to distinguish between works with the same type and property.

The tables property_info and type_info contain the name associated with each property_id or type_id, respectively.

From the “work” table, there are two sub-tables: “stage_work” and “screen_work.” The stage_work table contains all Broadway productions scraped from IBDB. The screen_work table contains all screen musicals scraped from TMDB. Entries in these tables are still distinguished by their work_id. There is a one-to-one relationship between both work and stage_work and work and screen_work.

Fig 3.3.1. MySQL EER Diagram



The stage_work and screen_work tables have distinct traits that make this separation helpful. The stage_work table, for example, has the columns ibdb_url, opening_date, closing_date, and preview_start_date, while the screen_work table has the columns tmdb_url and release_date. Because of the robustness of TMDB, I was able to include some extra metrics in the screen_work table. Here are the definitions of these columns:

us_origin: 1 if a screen work was made (either wholly or partially) in the United States
tmdb_popularity: a metric created by TMDB that is calculated using several metrics within the platform [10]. These scores are calculated each day, so all popularity scores are from the day they were scraped. This date is uniform across all entries.
budget: the film's budget, as reported by TMDB
revenue: the film's revenue, as reported by TMDB
tmdb_rating: the rating users have given the film on TMDB
tmdb_vote_count: the number of users who have rated the film on TMDB
assoc_stage_work_id: Unrelated to TMDB, this gives the associated work_id for screen works of type proshot (Broadway).

The stage_work table also has a one-to-many relationship with the ibdb_tag table. A row in this table represents one IBDB tag on one production. Examples of these tags are "Broadway," "Opera," "Comedy," "Revue," etc. Therefore, there will be several rows for the same work, and several rows with the same tag_name. Each tag_name has an associated tag_id, which can be seen in the tag_info table. There are 42 distinct tag names in the database.

People:

Each person, either from the cast or crew of a movie or stage musical, has a distinct person_id and distinct row in the person table. This table contains information regarding the person's birth and death date and place, gender, the links to their pages on IBDB and/or TMDB,

as well as their tmdb_popularity score, which is calculated in a similar way to the popularity scores of individual films on TMDB. For many people, some of these columns remain blank due to incomplete information on their IBDB and TMDB pages.

People are also associated with works in the tables work_person and actor. In work_person, there is a distinct 20-digit work_person_id for each row, consisting of the 8-digit work_id, 10-digit person_id, and 2 sequence digits. The role column describes the role the person had in the work. In this table, the role of all actors is "actor." The actor table comprises all the work_person_ids where the role is "actor," and in this table the role column refers to the character name of each actor. Both the work_person and actor tables contain credits from both screen and stage works. The actor table contains a broadway_debut column that is 1 when the actor made their Broadway debut in the production, 0 if they did not, and empty if the associated work is a screen work rather than a stage work. There is a one-to-one relationship between the actor and work_person tables.

Calling on this database, I analyzed and visualized my data in Python.

4. Data Exploration 1: Overall Metrics, Stage and Screen

To begin demonstrating my analyses of my data, I want to look broadly at traits of Broadway musicals over time.

My database contains 3,097 works of Broadway music theatre, 2,520 of which are distinct properties. Below I explore other features of the distribution of these data on the whole and over time.

4.1. Stage Revivals

In general terms, a Broadway revival is a production of a show that has already been on Broadway. In my methods section, I discussed some of the definitional difficulties of this term as theTony Awards define a revival somewhat differently to IBDB.

For the purposes of these analyses, I have used the default IBDB tags of "original" and "revival," with some changes I made by hand to cases that I found that were incorrectly labeled by IBDB. A surprising trait of this data is that there are about 70 musical properties that have a revival but no original which I cannot entirely explain.

Figure 4.1.1 shows the number of music theatre shows opening each year, and the distribution of originals, revivals, and return engagements. We can see that the number of shows opening per decade spiked from the 1900s-1930s, then decreased to the fairly steady rate it has remained at until today. This can be largely attributed to the much shorter run lengths of musicals in this era. Figure 4.1.2 shows a graph which Miller presents in "Average Broadway," illustrating the average run lengths of plays and musicals each year.

Figure 4.1.1. Broadway Music Theatre Openings by Decade

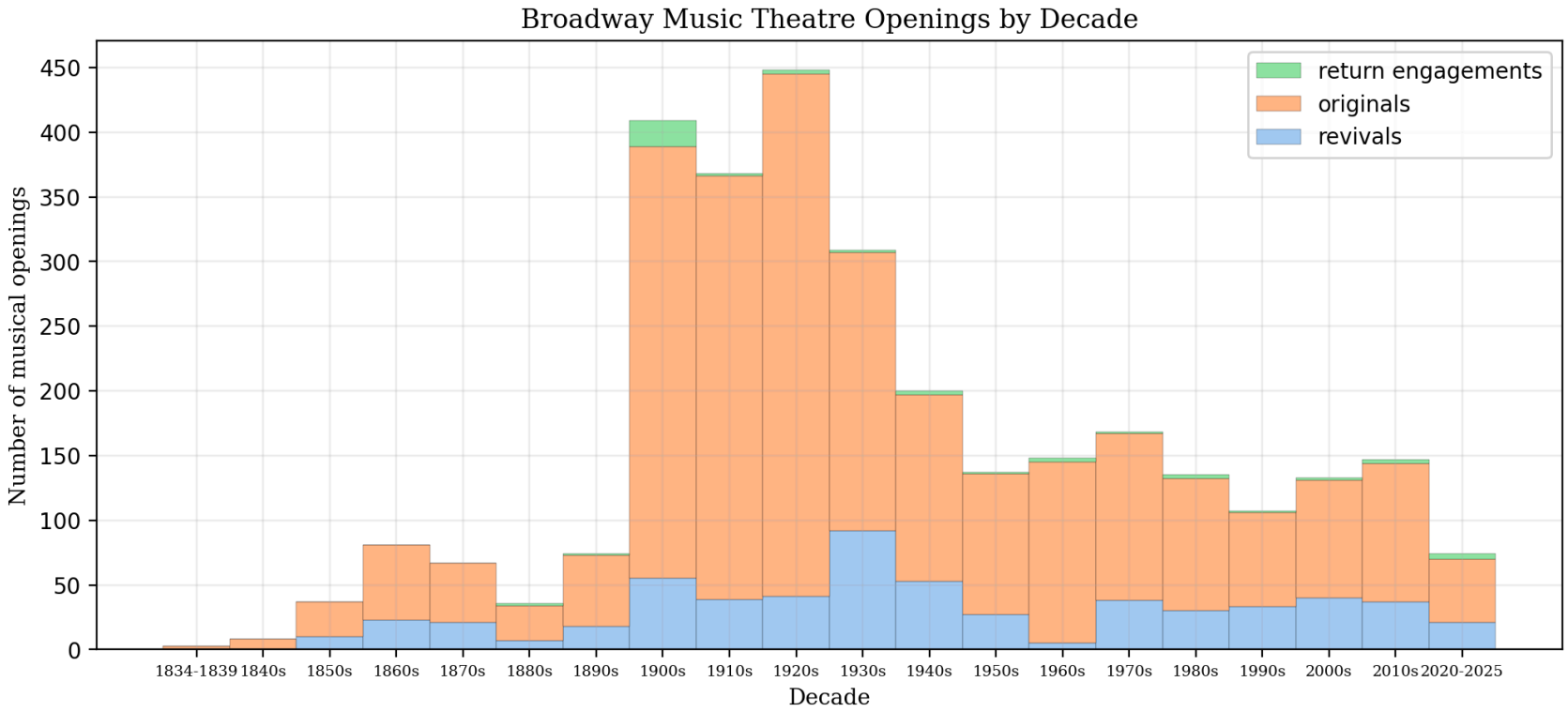


Figure 4.1.2. Mean run lengths of plays and musicals, 1915-2024 [3]

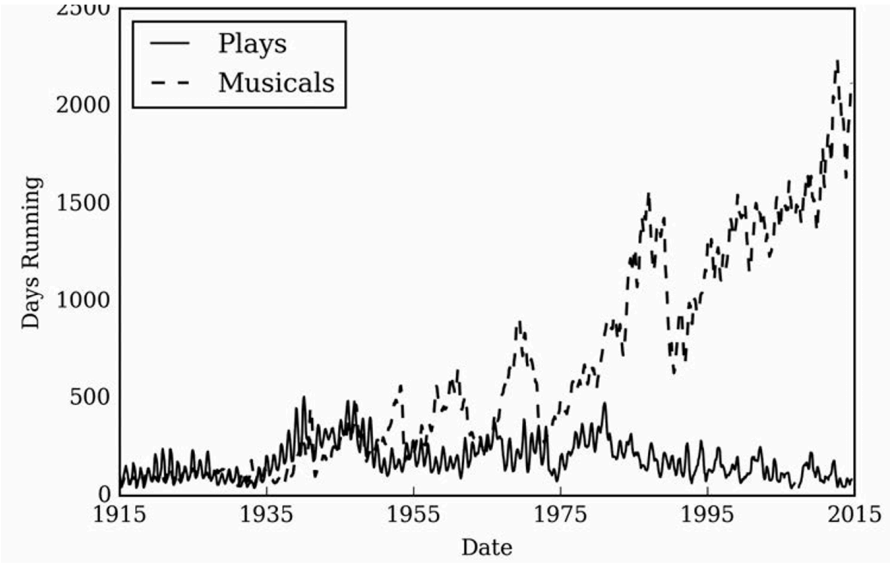
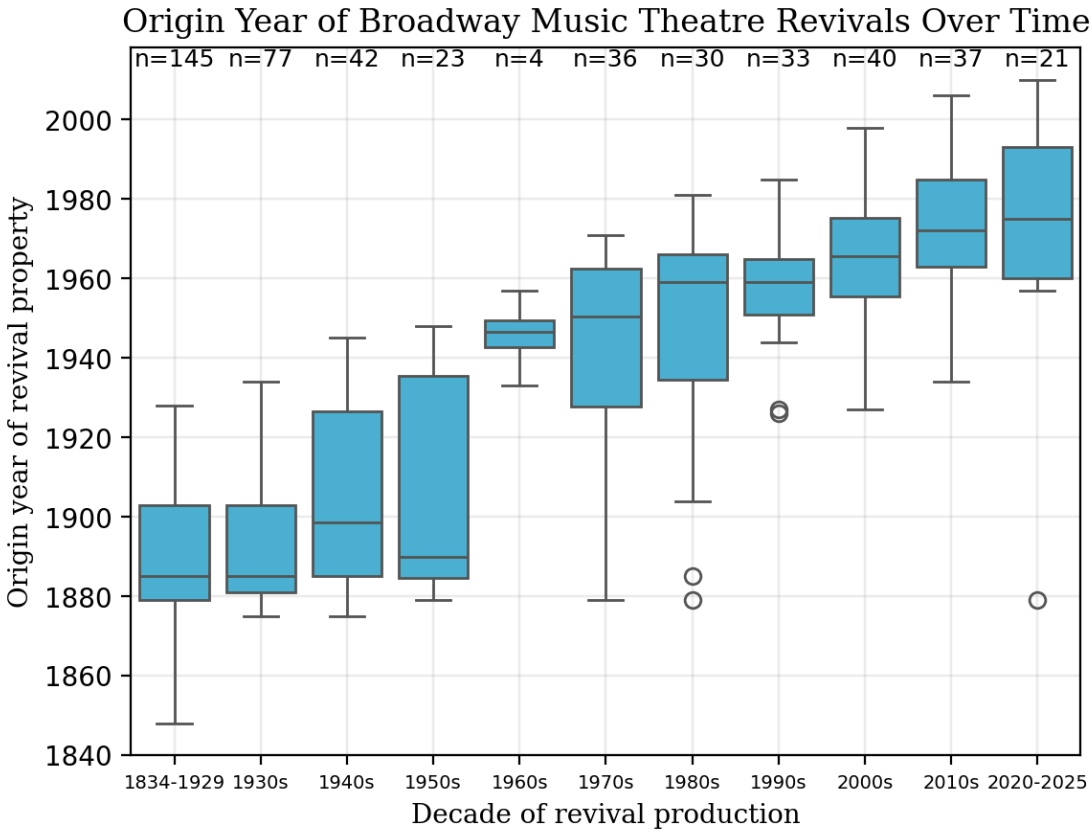


Figure 11. Mean run lengths of plays and musicals, 1915–2014.

In fig. 4.1.1, although the total number of musical openings stays fairly constant from 1940-1979, there are significantly fewer revivals in the 1960s than either the 50s or 70s. This aligns with a similar trend that Miller observed in Broadway play and musical revivals, to which he remarks that "the popularity of revivals, like so many other things on Broadway, is subject to large, cyclical processes that always have local causes" [3]. I emphasize Miller's observation that the data alone cannot answer why this is, including in the case of the drop in revivals in the 1960s. They can, however, prompt further questioning, underlining the importance of interdisciplinary research of music theatre.

Beyond the popularity of revivals on the whole, we can look at what sorts of shows were revived in each decade. Figure 4.1.3 shows distributions, in the form of, boxplots of the origin

Figure 4.1.3. Origin year of Broadway Music Theatre Revivals Over Time



year of the revivals produced in each decade. More clearly, the x-axis shows the year that a revival production was produced, and the y-axis shows the origin year of that property. The x-coordinates of the 2024 *Gypsy* revival, for example, would be (2024, 1959). The boxplots summarize the distribution of these coordinates, binned by the decade of the revival production. The top of the graph shows the total number of revivals produced each year as a measure of *n*.

This graph reveals many interesting trends. First, from the 1960s to now the median origin year of revivals has been (for the most part) increasing. The median year can suggest which older shows were "trendy" in a given decade, while extrema and outliers could indicate which revivals were a surprise to see on Broadway stages at that time. For example, this graph suggests that musicals from the 70s were trendy in the 2010s. The interquartile range of each box

poses many interesting questions. What might account for the small interquartile range in the 1990s? Why were the revivals of that decade so concentrated around a particular era of origin, especially after the 1980s showed a wider range of revivals in terms of origin year?

We can also observe that, since the 1970s until the 2020s, the oldest revival produced in the decade (or the minimum at each decade) has been increasing. This trend is broken with the outlier seen in the 2020-2025 bin, which is *Pirates! The Penzance Musical*. This musical is an adaptation of the Gilbert and Sullivan operetta *The Pirates of Penzance*, which was first produced on Broadway in 1879. This is one such example of a surprise to the Broadway season. The series of boxplots from 1970-2025 shows that it is extremely uncommon for a property this old to be seen on Broadway stages.

It is worth noting that I, as the constructor of the database, made a conscious decision to count this as a revival of *The Pirates of Penzance* rather than as a separate property. In this production's IBDB credits, Gilbert and Sullivan are still credited as having written the music and libretto, and there is a separate person credited for this adaptation. Additionally, IBDB still lists it with the tag "operetta" [12].

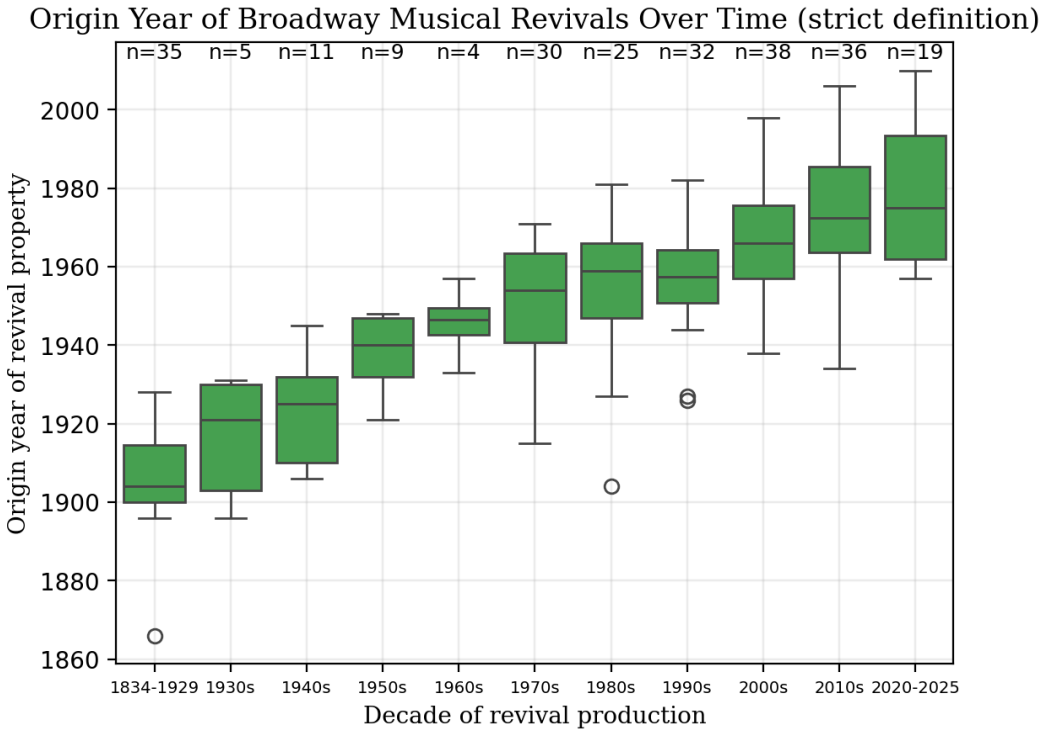
Similarly, *The Gershwins' Porgy and Bess* was a 2012 Broadway production of the opera/musical *Porgy and Bess,* starring Audra McDonald. The music and lyric credits are the same as the original 1935 production, but the 2012 production has an additional credit for adaptation [13]. According to gershwin.com, the official site of Gershwin Enterprises, the adaptor "reinvented DuBose Heyward's 1925 novel… to 'retain all the best-loved elements of the original while crafting a piece that speaks to contemporary audiences'" [14]. This description tests my own distinction of what constitutes a "revival" versus an entirely new production, but I decided to count *The Gershwins' Porgy and Bess* (2012) as the same property as the original. By

being more inclusive with what counts as the same property, we can in this graph see when these revivals of old material occurred.

Both of the outliers I mentioned do not fall under this strict definition of a "musical." *The Pirates of Penzance* is an operetta, which is a genre that paved the way for modern Broadway musicals, and *Porgy and Bess* is an edge case between opera and musicals. What changes if we try to isolate musicals?

Figure 4.1.4 again shows boxplots of the origin year of the revivals produced in each decade, but this time uses our stricter definition of a "musical." For this analysis, I have removed all properties tagged with the following on IBDB: opera, operetta, revue, extravaganza, vaudeville, burlesque, minstrel, variety, and ballet.

Fig. 4.1.4. Origin year of Broadway musical revivals over time– Strict definition of musicals



this does not include data for related genres such as opera, operetta, vaudeville, etc.

This graph, compared to fig. 4.1.3, has a more consistent upward trend of medians that starts at the beginning of the data rather than in the 1960s. Additionally, several of the outliers and minima no longer appear, due to the fact that many of the oldest revived properties were outside of this strict definition of a musical.

In fig. 4.1.4, the interquartile ranges for the 1950s, 70s, and 80s have decreased significantly from those in fig. 4.1.3. Relatedly, *n* has decreased for each of these decades by 14, 6, and 5, respectively. The lack of effect on the 1990s datapoints indicates that, by this point, the identity and form of Broadway musicals had homogenized. This follows from the fact that non-musical works were no longer produced as revivals, for reasons relating either to consumer demand or artistic interest. While in the 50s-80s things like operetta, opera, and vaudeville were still being revived, by the 1990s a stricter definition of a "Broadway musical" manifests in what kinds of revivals were produced. In other words, by the 1990s the idea of operetta, opera, vaudeville and the like on Broadway appears to be defunct.

Within my database, my ability to isolate works within the strict definition of a musical is limited by the amount of information that IBDB tags can provide. As seen, the number of musical revivals pre-1930 has gone down drastically because of this restriction, but it reflects many more musicals than we might consider there to have been in this period: If *Show Boat* (1927) is counted as the first musical, then there should very few musicals before the 1930s. I have observed that IBDB tags seem particularly unable to separate "musical comedy" from modern comedic musicals, for the "comedy" tag is applied to both distinct forms. As such, shows tagged "comedy" cannot be removed in order to isolate strictly defined musicals from broader music theatre, though many of them do not align with my definition of musicals. The challenge in translating this qualitative distinction to something separable within the data speaks to another

instance of both the limitations of digital theatre records themselves and their resistance to taking on clear-cut distinctions.

4.2. Screen Musicals

In this section, I look at a few kinds of screen musicals and their interaction and overlap with stage musicals. These types of screen works are movie musicals, Broadway proshots, other proshots, and other kinds of screen works such as live TV specials. As a reminder, only musical screen works with an equivalent Broadway property are included in this data. See Table 3.2.1 for more details.

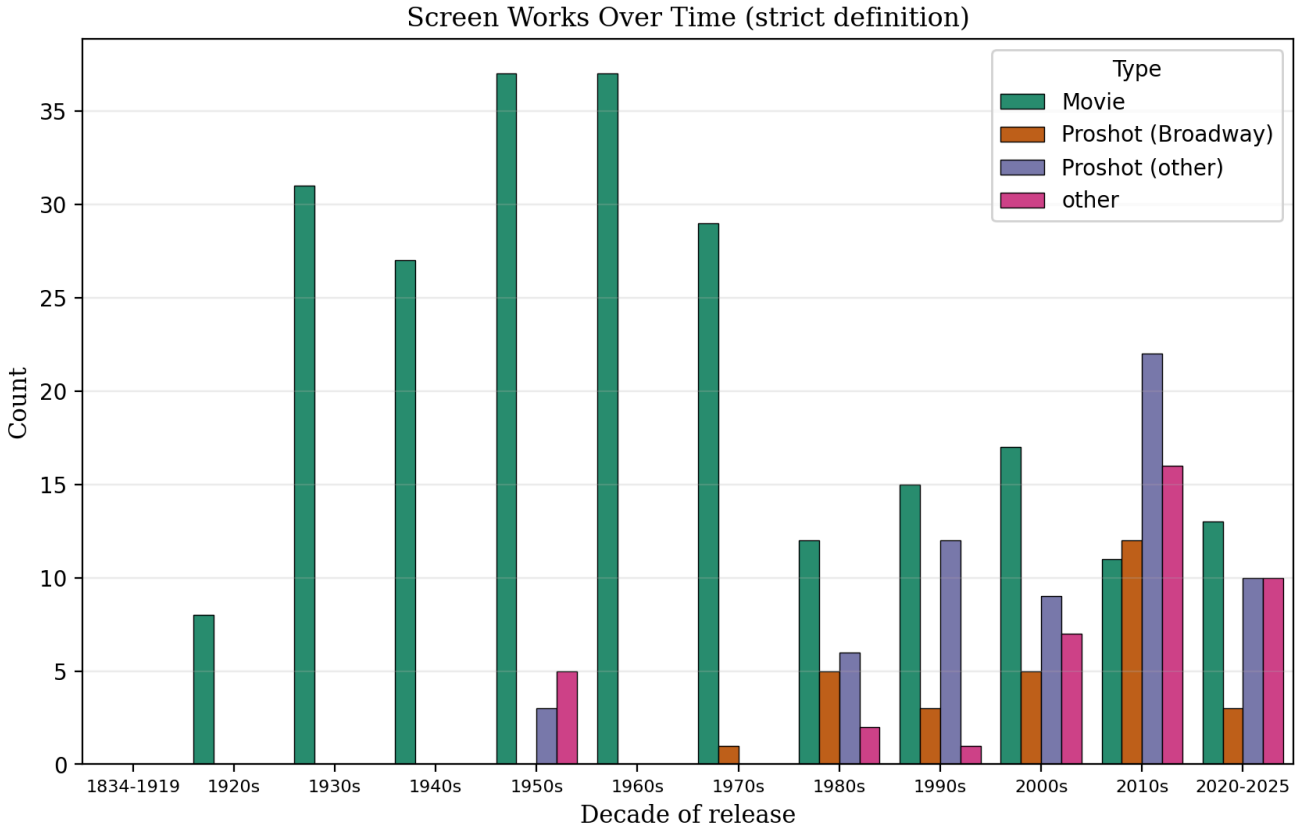
Let us first look at the presence of each of these types of screen musicals over time, as shown in figure 4.2.1. Beginning with movie musicals, shown in green, the graph shows the strong presence of movie musicals from the 30s through the 70s, with a sharp decline starting in the 80s.

I would like to momentarily turn away from data-based findings, and to historical scholarship relating to even these basic statistics on movie musicals. These findings support the more zoomed-in historical claims on the topic, such as in Holley Replogle-Wong's article "The Great Generational Divide: Stage-to-Screen Hollywood Musical Adaptations and the Enactment of Fandom" [15]. Before making her arguments about how fandom plays into the success or failure of a movie musical adaptation, Replogle-Wong gives an overview of this history:

> By the mid-1950s, the Broadway musical was riding a creative tide, spurred by the model of Rodgers and Hammerstein's *Oklahoma!* in 1943, and the musical genre was a key player in American popular culture… Hollywood began looking to existing stage shows for adaptation.
>
> During this time, film musicals were rendered with increasing fidelity to the stage musical, and with a relatively small gap of time between the stage premiere and the adaptation. Films from this period allowed audiences access to stage productions shortly

after their debuts, during their initial period of currency. Through the next two decades, Hollywood enjoyed its fair share of musical box office successes… There were also a few spectacular failures: *Camelot* (1960 [original stage production]/ 1967 [film]), *Hello, Dolly!* (1964/1969), *Man of La Mancha* (1964/1972), and *Mame* (1966/1974). These last four films, all lavish, expensive productions that flopped, ended the trend for Hollywood musical adaptations… The genre experienced a lull in popularity until *The Little Mermaid* (1989), which spurred what has been dubbed the "Disney Renaissance," but the film musical never completely went away, with several enduring successes produced during those decades: *Jesus Christ Superstar* (1973), *Grease* (1978), *Fame* (1980), *Footloose* (1984)[2], *Little Shop of Horrors* (1986) [15].

Fig. 4.2.1. Screen Works Over Time (Strict Definition)



Replogle-Wong later explains that many film studios had purchased the rights to musicals in the 1990s, but "it wasn't until the success of *Chicago* in 2001 that these films were pulled out

[2] The *Footloose* film has not been included in my dataset, based on the distinction that the characters in the film do not sing. The subsequent stage adaptation of the film is, however, included.

of purgatory and placed on the front burner" [15]. This historical study allows the data to take on a more meaningful shape, as well as point out the individual data points that shaped the trends at large.

Beyond movie musicals, fig. 4.2.1 also illustrates the rise of the proshot and other kinds of screen representation. By this figure, Broadway proshots were few and far between from the 1970s to the 2000s, with a dramatic increase in the 2010s. Some examples of early Broadway proshots include PBS broadcasts of shows such as the original Broadway production of *Sunday in the Park with George* (1986). This particular recording is now easily accessible for free on YouTube.

More recent Broadway proshots are often behind a paywall on platforms including BroadwayHD and Disney+. In 2017, Disney released the Broadway proshot of *Newsies* (2012), a property which Disney already owned prior to the stage production. In 2020, the *Hamilton* proshot was released on Disney+, but for this endeavor it cost Disney $75 million to acquire the rights to the proshot. This deal was "the most expensive single-film acquisition in Hollywood history" [16]. This indicates some strong incentive for the acquisition of such proshot rights, which could in turn mark an increasing profitability for proshots.

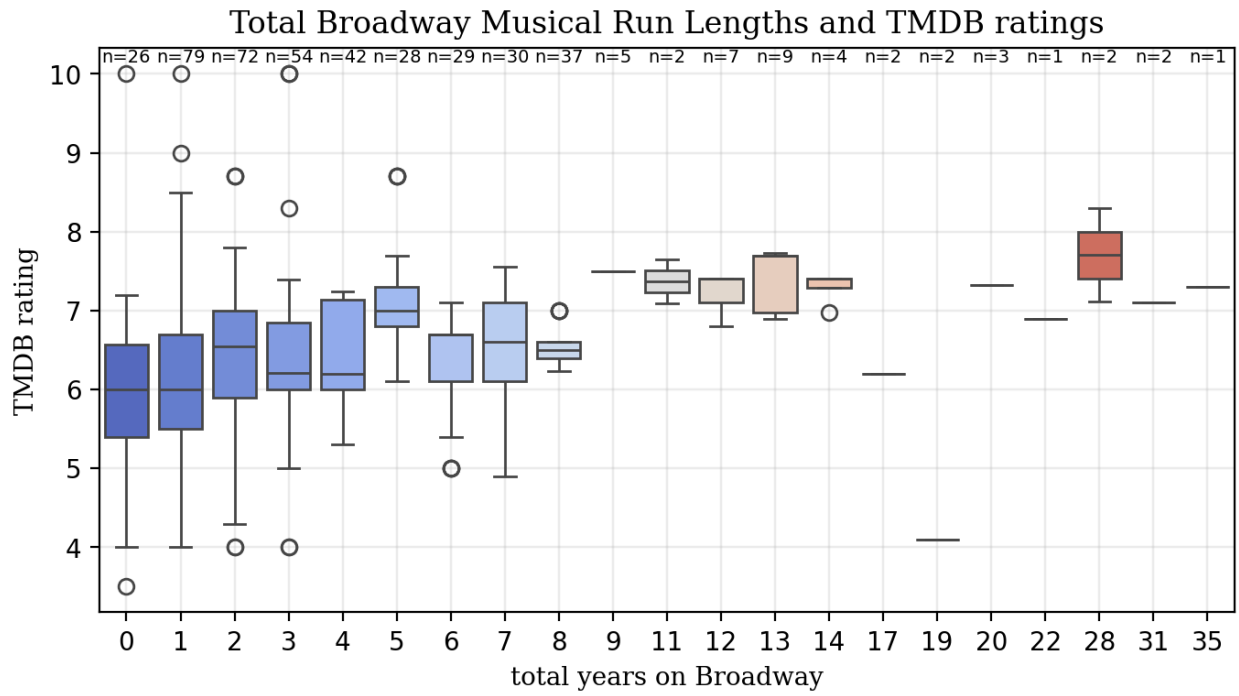
Screen musicals are not isolated events from their stage counterparts, however. Replogle-Wong discusses how the fandoms of stage musicals have influenced not only the reception but the production of movie musical adaptations. Fans' protectiveness over the source material has, in some cases, directly impacted the studio's creative decisions about the film [15].

Screen productions can also have an impact back on subsequent or simultaneous stage productions. The release of the *Hamilton* proshot, rather than making the stage production redundant and irrelevant to consumers, increased demand for stage tickets [17]. A similar

phenomenon has occurred for *Wicked* (2003) on Broadway since the film adaptation's release in the fall of 2024.

Returning to Replogle-Wong's arguments about the influence of musical fandoms, fig. 4.2.2 models a version of this relationship by looking at a musical's total run length on Broadway as compared to the TMDB rating of its movie adaptation (scored by users on a scale of 0 to 10). Note that the "total years on Broadway" represents the years that a property has been on Broadway over all its productions, rounding down by year. The y-axis shows the distribution of TMDB ratings for each run length as a boxplot, and the top of the graph shows the bin sizes for each year count as values of *n*.

Fig. 4.2.2. Total Broadway Musical Run Lengths and TMDB ratings



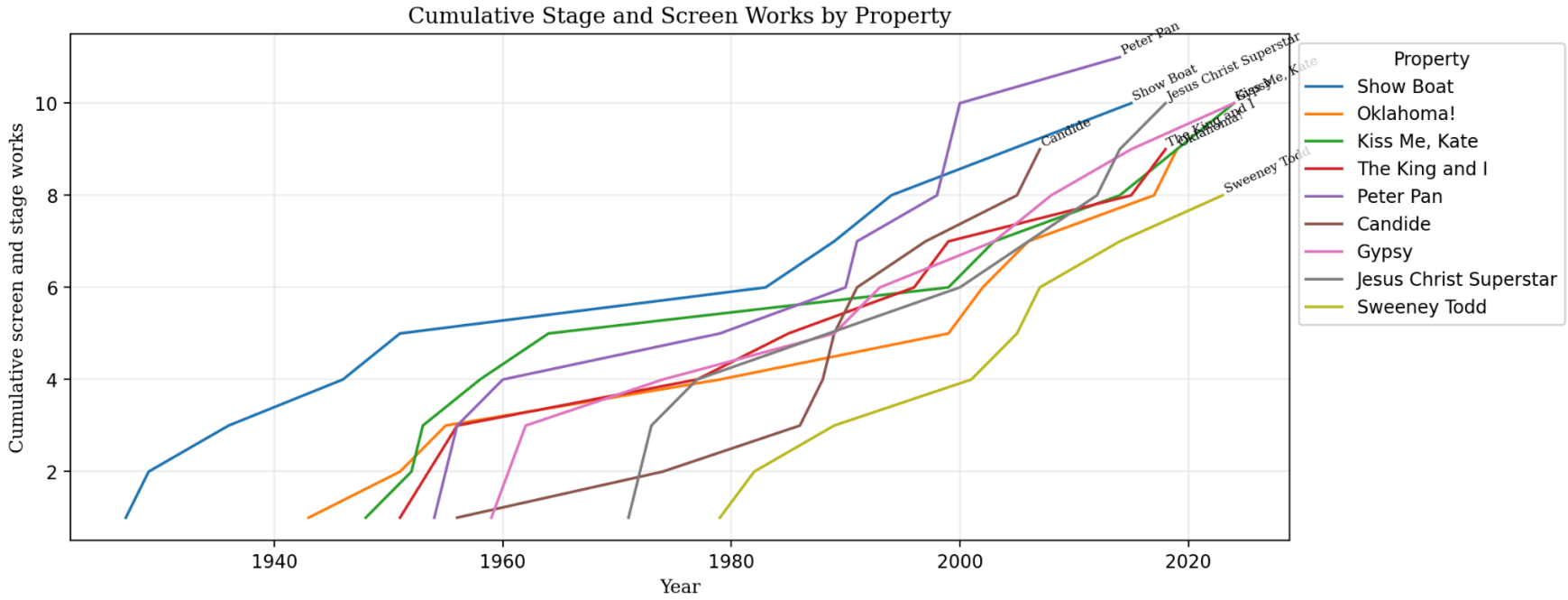
Looking at the median values for each distribution, this visualization shows a slight upward trend in ratings as the run lengths increase. This could draw a connection between large fan followings and movie adaptation success, under the reasoning that a longer Broadway run

corresponds to a larger fanbase. That being said, if we take the run lengths to instead indicate the quality of the show or the intellectual property, the film adaptations might be getting better ratings simply because of the quality of its source material. These data alone are not enough to distinguish between fan preferences and some objective artistic quality, but it is exciting to see how easily Replogle-Wong's observations begin to map to large datasets instead of case studies.

Combining data across screen and stage musicals also allows us to look beyond New York City at the overall cultural significance of musicals. A naive example of this measure can be seen in figure 4.2.3, which counts the cumulative stage and screen works for each property over time for properties with at least 8 works. Among these properties, it is interesting that no particular curve shape stands out. For the most part regardless of the property's origin year, each curve has a similar steadily increasing slope. The grey line representing *Jesus Christ Superstar*, among some other lines, shows an initially steep slope, perhaps representing the rapidity with which the show gained traction.

This is one measure of property significance in my database, though the measure clearly leaves out many factors relating to the relevance of a show. A notably absent data point in these top nine shows is *The Phantom of the Opera*, which is the longest-running show in Broadway history. While a show like *Oklahoma!* has had many revivals and thus many points by this metric, *The Phantom of the Opera* had no need to be revived so many times because the same single production stayed open for 35 years. This measure of database occurrences can display how often artists have returned to the same properties (and relatedly the audience demand to see those properties revived or adapted), but it does not encompass the overall cultural relevance of a property.

Fig. 4.2.3. Cumulative Stage and Screen Works by Property



To better make this estimate, I would posit that one should weight each stage production with the length of the run and its commercial success, and weight each screen work with its commercial success or another measure of its reach. By the data provided by TMDB, creating a proper weight for screen productions is tricky. Although TMDB provides a "popularity score," it is not possible to discern exactly how this score is calculated or how different weights contribute to this value. TMDB also provides data on the budget and revenue for a film, but many screen works did not have a theatrical release or otherwise are missing this data. The *Hamilton* proshot, for example, is lacking these data points, although we know that the proshot is behind a paywall and made a profit.

In short, combining different mediums of theatre consumption is a helpful and interesting tool to understand the cultural reach of this form over time, and warrants further research to discover a quantitative model that can capture the many dimensions of this concept.

4.3. Development of the Musical Genre on Broadway

Given the mentioned restrictions of IBDB tags on determining when the modern Broadway musical solidified in section 4.1, let us look at what information IBDB tags *can* discern about the development of the Broadway musical.

As mentioned, the "comedy" tag seems to be minimally informative in this sense, for it is applied to both "musical comedies" and musicals that are comedic. Thus, the tag cannot show when the "musical comedy" form died out. This is seen in fig. 4.3.1 by the fact that "comedy" continues into the later part of the 20th century rather than dying out with the musical comedy genre. Furthermore, fig. 4.3.1 shows that the ratio of the total musicals to the musicals with the "comedy" tag stays mostly constant.

Fig. 4.3.1. Broadway Music Theatre Openings by Genre (“comedy”)

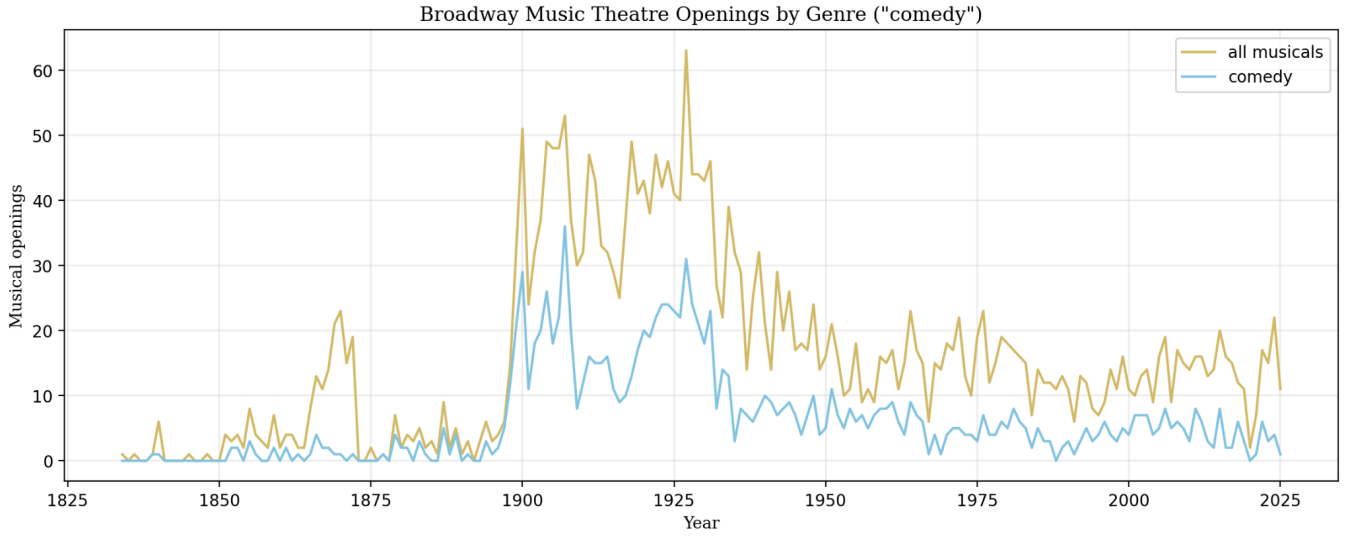


Fig. 4.3.2. Broadway Music Theatre Openings by Genre (opera, operetta)

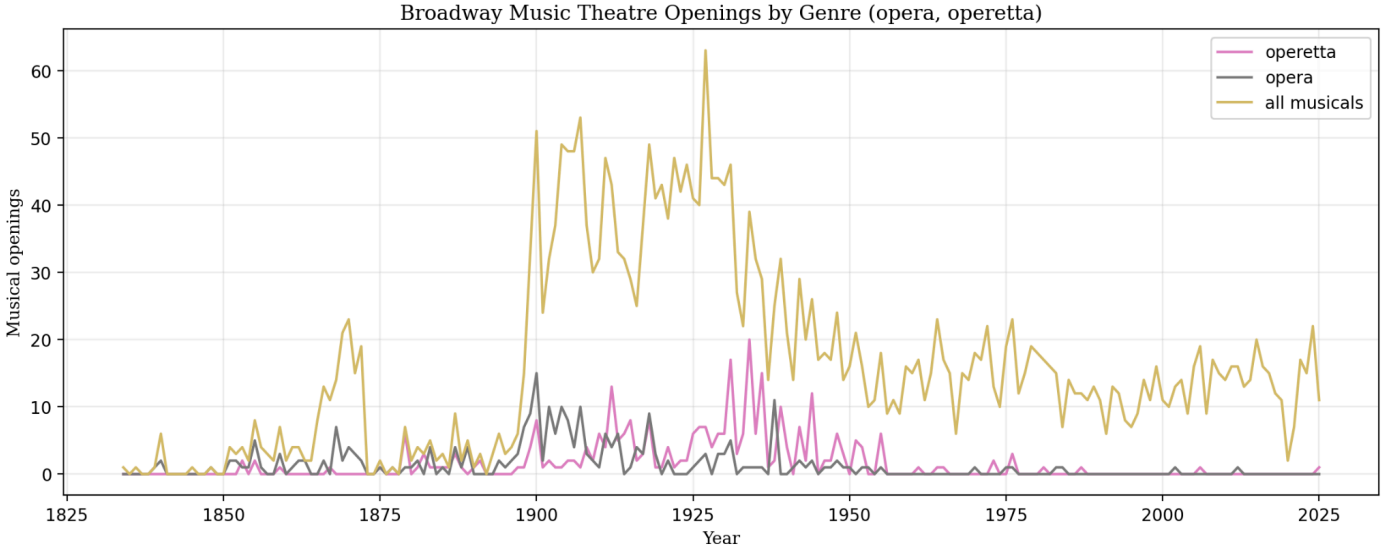


Fig. 4.3.3. Broadway Music Theatre Opening by Genre (burlesque, revue, extravaganza, vaudeville)

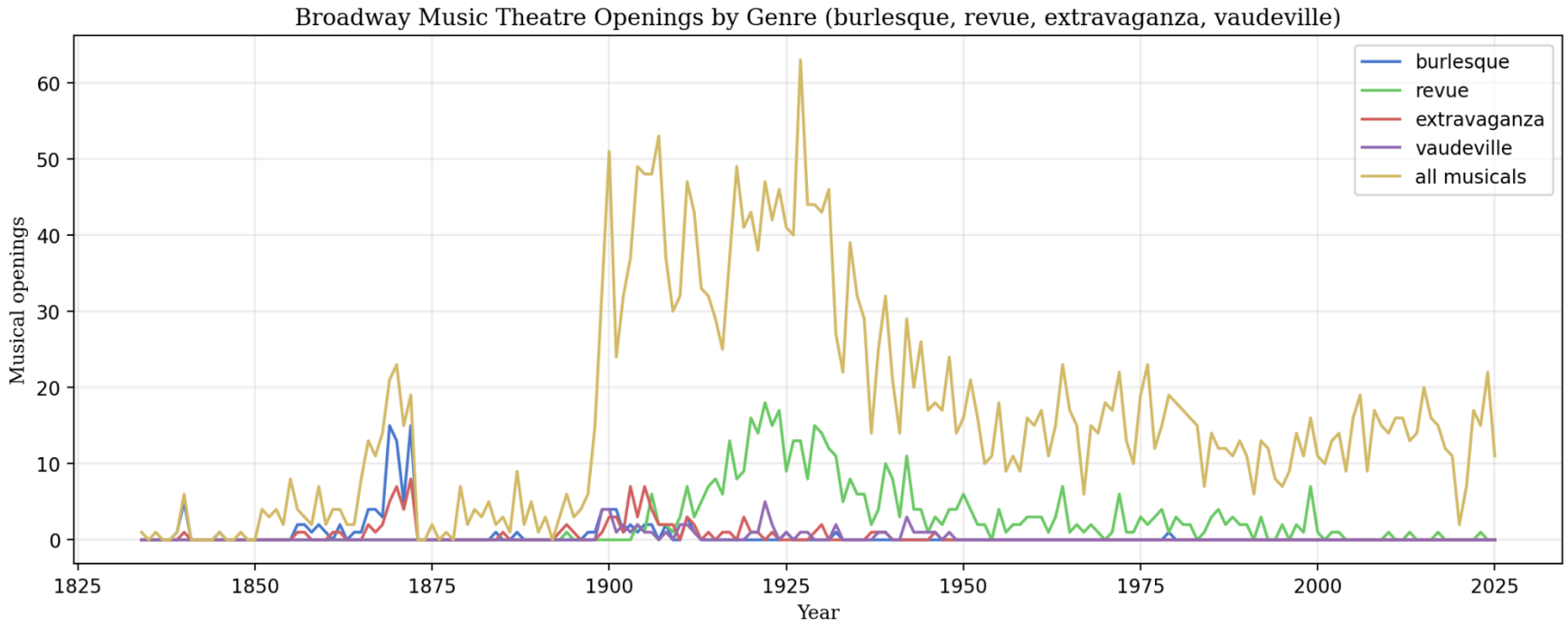


Figure 4.3.2 shows the presence of opera and operetta in the Broadway music theatre scene, according to IBDB tags. Note that in the DB there is no overlap between these two tags, in other words no shows that have both the tag "opera" and "operetta." Note that these genres are *not*, however mutually exclusive with the "comedy" tag shown in Fig 4.3.1.

By fig. 4.3.2 opera and operetta seemed to dominate the Broadway musical scene around 1880-1900. They comprised a significant fraction of musicals through the 1950s, until they appeared on Broadway very infrequently after that point including today.

Fig. 4.3.3 looks at four similar genres: burlesque, revue, extravaganza, and vaudeville. Burlesque, extravaganza, and vaudeville all seem to follow similar trajectories of popularity: burlesque and extravaganza both peak around the early 1870s, and then all three have a moderate presence from around 1900 to the 1940s. Note that, although there is some overlap in the DB between the shows listed with these tags, the overlap is fairly small. This is surprising given the qualitative overlap of these genre distinctions, as discussed in Section 1.1.

Figure 4.3.3 shows revue as having a different trajectory altogether, even though it is very similar to the other three in form. Revue was defined above as some combination of the music theatre sub-genres that came before, which can be seen by the genre's peak around 1925. That being said, unlike the other genres on this chart, the term revue has continued into the 20th century. According to IBDB tags, modern shows like *Ain't Misbehavin'* (1978) and *Dancin'* (1978) also fall into this category. In this case, revue refers more generally to a collection of musical numbers that are not driven by a particular plot. Without diving into the text of each revue show, it is difficult to discern whether, like the "comedy" tag, revue has meant different things over time, or whether this genre has lingered on Broadway while similar genres have died out.

Due to the similarities in these forms, some of what these graphs show is really the changing terminology used to describe theatre. Perhaps it is possible that one or more of the revue shows of the late-1900s bear strong resemblance to an extravaganza or vaudeville show in terms of form, but its cultural context led it to be understood and marked by audiences as "revue." Regardless, the presence or lack thereof of these terms throughout history allows us to understand the findings of theatre historians in a broader way via data.

5. Data Exploration 2: Broadway musical actors

In the following section, I will use the data collected on Broadway actors in the opening night casts of musicals to examine who a Broadway actor is/has been and what their career might look/have looked like. My database contains 99,816 rows of music theatre acting credits, 46,314 of which are associated with distinct stage actors.

5.1. Broadway actors– Age and Gender Demographics

In this section, I am particularly interested in age and gender[3] demographic components of Broadway music theatre actors.

Let us first look at the distribution of actors' ages in each decade in figure 5.1.1. The total number of actors in the shows in that decade are shown along the top of the figure as values of *n*. This plot shows a fairly steady median value across the decades, not far at any point from the

[3] IBDB contains two possible values for "gender": Male and Female. This poses two problems, the first being that "male" and "female" would better describe sex than gender. Second, although there are non-binary actors in the DB, it seems that those actors are either erroneously marked or have no gender listed. For consistency with my data source, I will follow IBDB's use of the word "gender," but note that it is used somewhat inaccurately.

Figure 5.1.1. Age of Broadway Music Theatre Actors Over Time

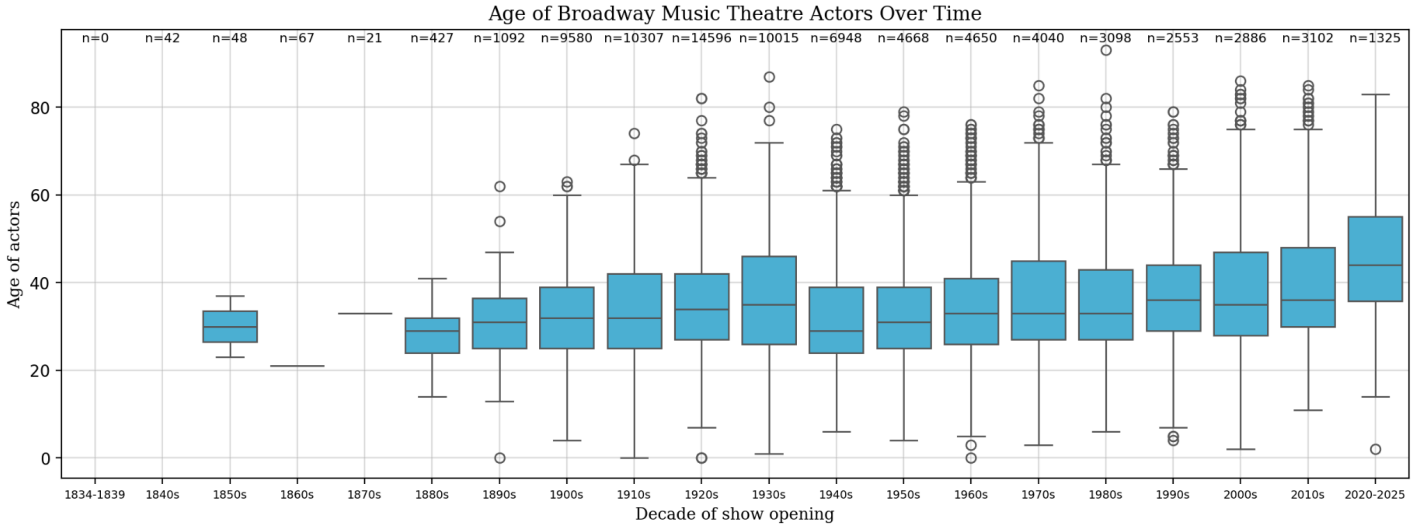
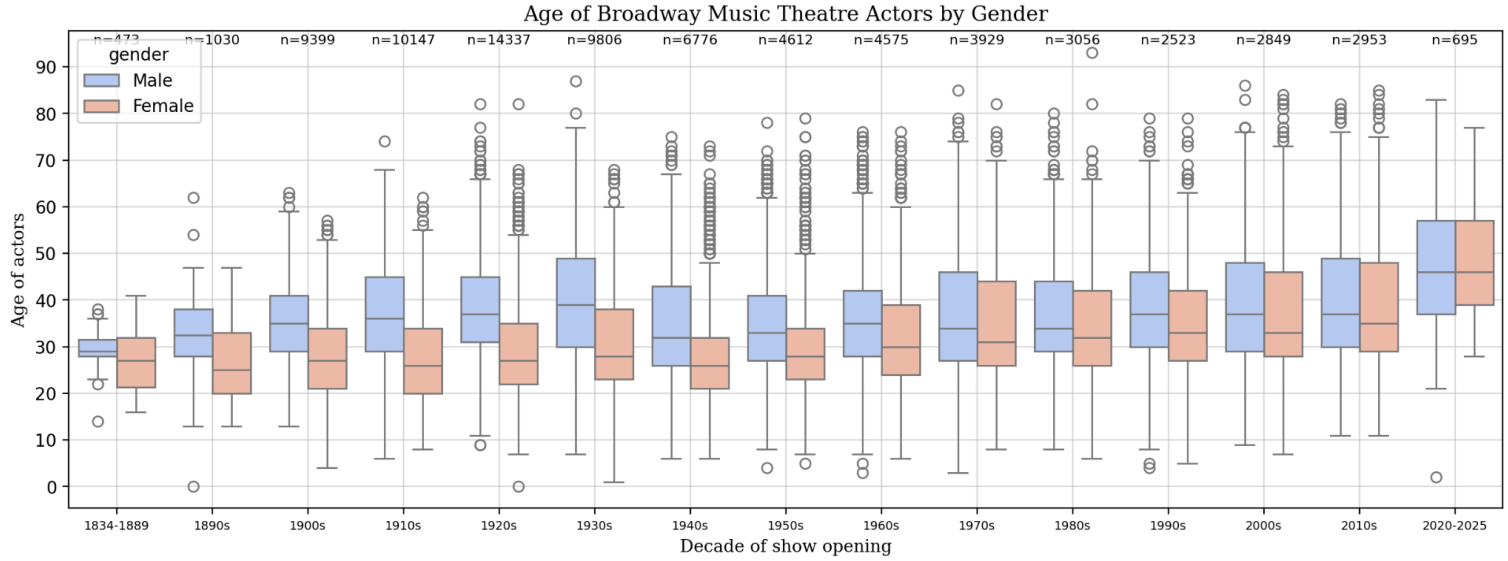


Figure 5.1.2. Age of Broadway Music Theatre Actors by Gender



overall median age of 33. In other words, Broadway actors throughout history have tended to be about 33 years old.

Let us now look at these ages by gender in fig. 5.1.2. Note that actors missing gender data have been removed in this graph. In every decade (except the 2020s), the median age of male actors on Broadway has been higher than female actors. To answer the question of "why," there are many possible explanations: Could it be that the female *characters* are younger than the male characters in musicals? If not, why are women cast at a younger age to play the same ages that men play at an older age? If so, why is there a trend for female characters to be younger than male characters?

Could the explanation, instead, be that the peak skill level of women is younger than that of men, and that casting directors have been exercising this knowledge for decades? Relatedly, might a female singer's sound be in its ideal form when she is younger, while for men the ideal musical theatre sound is at a slightly later point in life? Maybe the relationship here can be related to the history of "chorus girls" on Broadway, and fig. 5.1.2 is confounded by the fact that ensemble members on Broadway might need to be younger than principal actors to perform the physical demands of such a role. Or perhaps the graph below is a symptom of the higher pressures on women than men to stop working if they have a child.

More obviously, these data might indicate the patriarchal context in which Broadway musicals exist. They support the claim that female performers are expected to look and be younger than male performers, which is in turn related to the history of the objectification of women in music theatre. This observation could, however, incorporate some of the questions posed above, including that of the age difference of male and female characters, which is certainly present at least in the heterosexual couples represented in musicals. That being said,

rather than seeing these trends in performer ages as an isolated event within musical theatre, it is more appropriately viewed as being related to gender dynamics in the United States on the whole. Because theatre is commercially driven, the representation onstage must appease the expectations of the audience.

Let us take, for example, the musical *Gigi* (1973), which centers around a girl named Gigi and the eventual romance between her and Gaston, whom she had formerly regarded as something of an older brother or young uncle. In a modern presentation of the musical, the age gap and resulting power dynamic between Gigi and Gaston can cause discomfort for audiences, while it appears to not have caused much alarm when the musical (and Best Picture-winning film) first premiered [21]. This is just one example of how the dual societal contexts of audiences and creatives may have worked together to produce symptoms such as the age gap shown in figure 5.1.2.

5.2 Distribution of Broadway Acting Credits

Let us now look at the Broadway music theatre careers of actors. We can first look at the ages of actors in their first stage role in my database. More specifically, this refers to the first time an actor was in the opening night cast of a musical. Figure 5.2.1 shows these data, again separated by gender.

The overall median debut age is 29, and for men and women respectively the medians are 31 and 26 years of age. The distance between the medians in each decade, as well as the trends of the median values over time seem to parallel the motion of fig. 5.1.2. Furthermore, the median ages in fig. 5.2.1 are always under 10 years away from those in fig. 5.2.1, and are often under 5

Fig. 5.2.1. Age of Actors in Their First Musical Role (opening night cast)

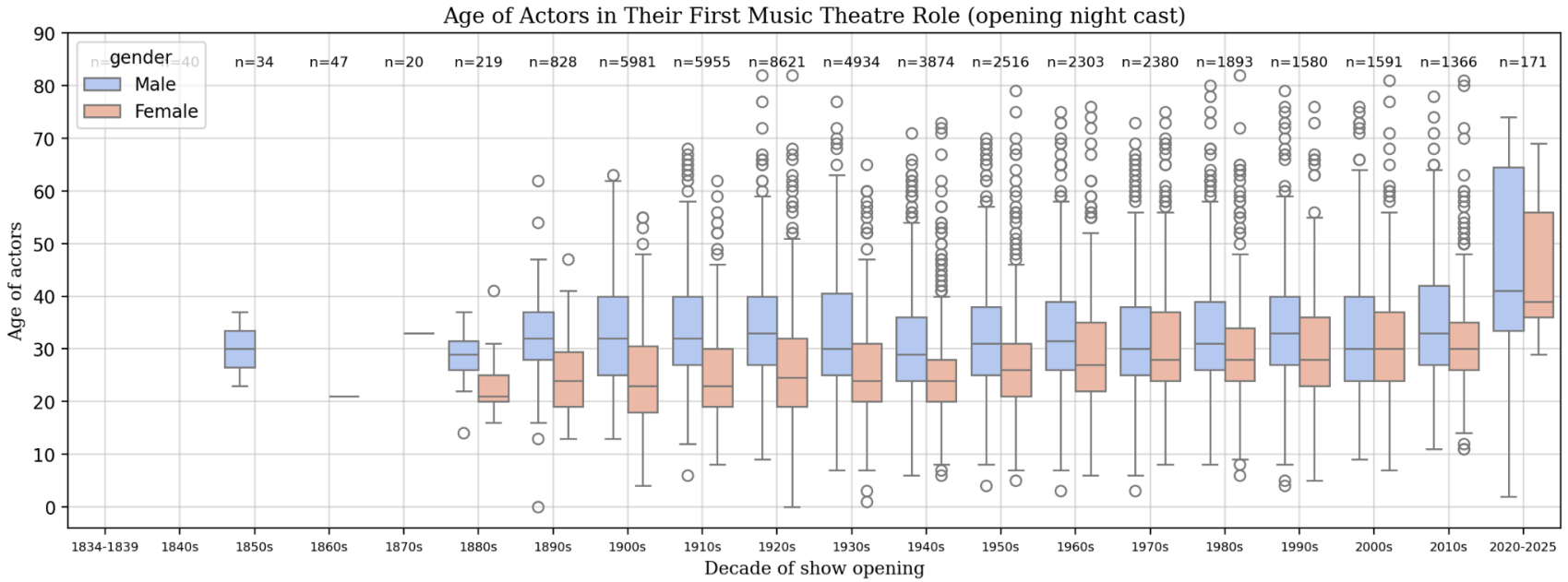
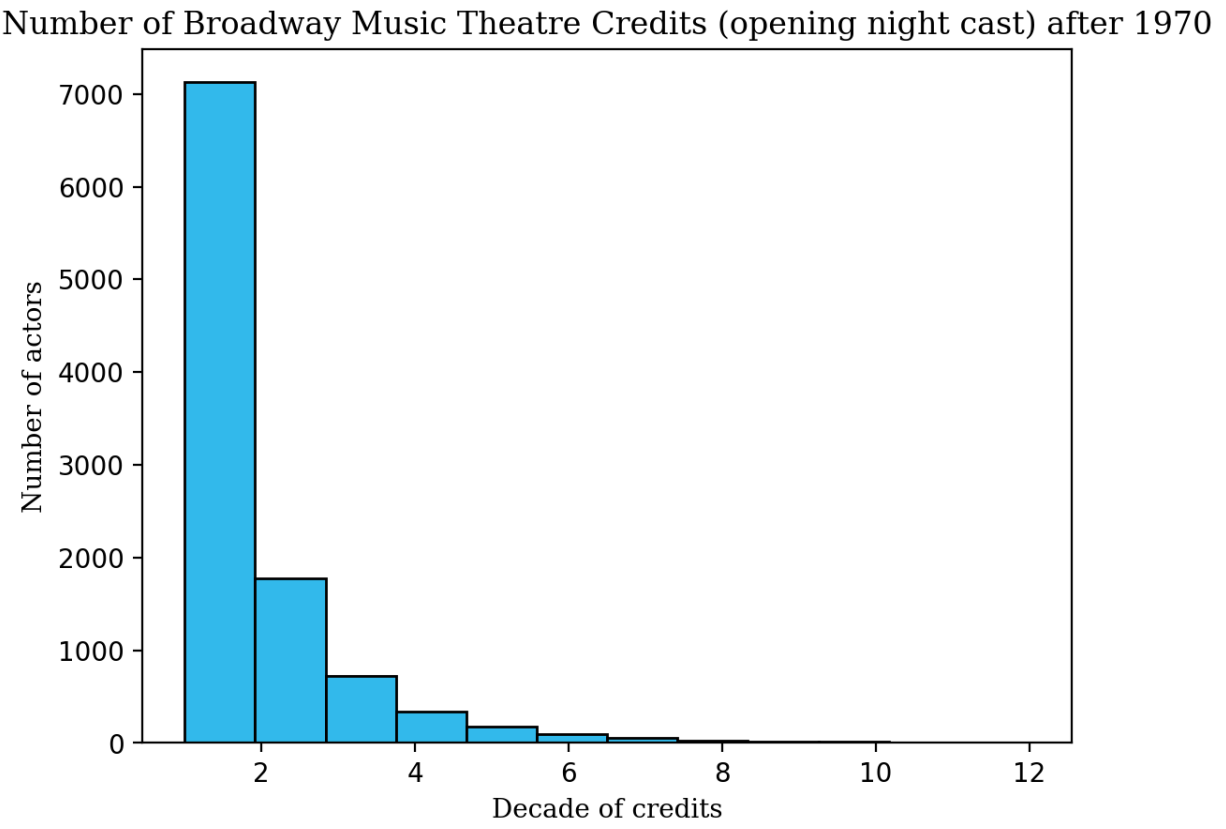


Fig. 5.2.2. Number of Broadway Music Theatre Credits after 1970



years away. This is perhaps explained by fig. 5.2.2, which demonstrates that the large majority of actors in this dataset have only one credit in a Broadway musical opening night cast. This figure looks solely at credits after 1970 to control for missing data in earlier years, but a similar trend is present when looking at the full dataset.

An unsurprising component of this figure is that there are few people with enough credits that we would consider them Broadway "stars," and more people with just a few credits. What is surprising, however, is the quantity of actors with only one credit. Anecdotally, it is known that it is hard to make a career out of acting, and it is hard to get Broadway acting jobs. This shows that this holds even when an actor has made their Broadway debut. It goes against the idea that a Broadway credit will eventually lead to an actor's "big break" and eventual success in the

industry, for the large majority of actors in this dataset were never in another opening night musical cast after their first credit.

Beside a pessimistic look at acting careers, this figure shows that the pool of Broadway performers is quite large. Although the general public may only know a few handful of Broadway stars, this is clearly the tip of the iceberg for the pool of Broadway-level performers.

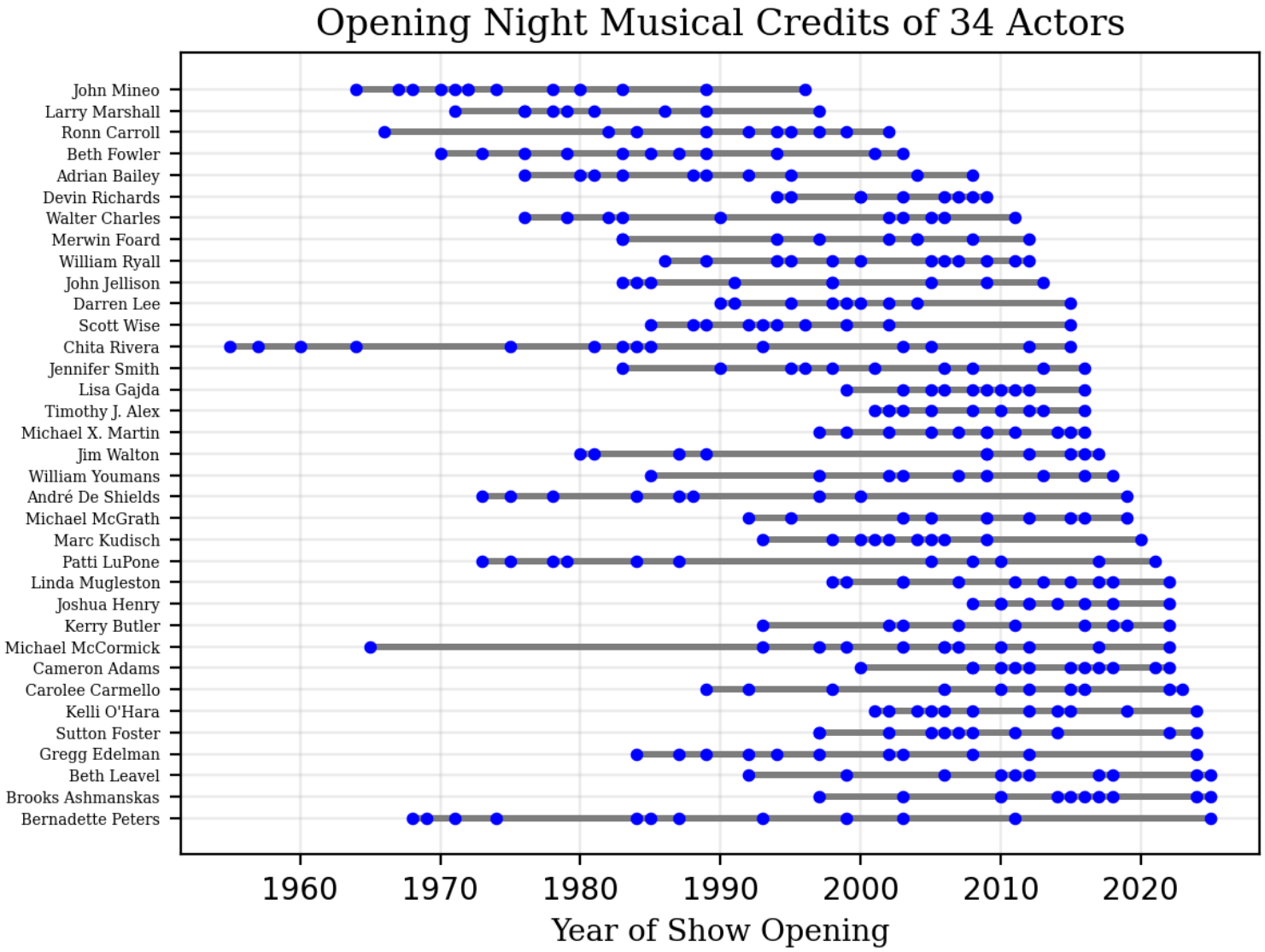
More optimistically, it must be recognized that this figure does not begin to encompass whether it is possible to make a living out of being an actor. First of all, these data do not take into account the length of an actor's run in a musical. Maybe an actor has only one credit, but that one credit is that they have been performing in *The Lion King* for 15 years. Secondly, this does not shows replacement credits, swings, or understudies. An actor would be considered a "replacement" when they replace an actor in the original cast, whereas swings and understudies fill in for actors who are missing on a given day. Further data collection and re-examination of these metrics could give us a better sense of the statistical feasibility of having a career in acting.

With actors, I would last like to look at the career trajectories of the most frequently appearing stage actors in my dataset. In fig. 5.2.3, I have taken the 34 actors with the most Broadway musical credits (again, for opening night casts), and plotted each of their credits in my DB. These 34 actors have between 9 and 14 such credits, with 14 being the maximum credits of anyone in my dataset. The actors are sorted from top to bottom by their most recent role.

Looking at the extremes of these data, Chita Rivera's career started the earliest (*Seventh Heaven*, 1955) and Joshua Henry's the most recently (2008, *In the Heights*), yet they are only a few credits apart by this metric (Rivera with 14 credits and Henry with 9).

This visualization serves to visualize the variety of what "success" as a theatre actor, defined here by frequency of employment, can look like. It raises lots of questions and

Fig. 5.2.3. Opening Night Musical Credits of 34 Most Credited Actors



for further research and visualizations, and allows us to easily compare between several actors who might seem entirely unrelated. For example, out of the most successful actors, how many of them had a role on Broadway in 2021 immediately after the pandemic? Out of the actors in this graph, about 7 of 34 went immediately back to work, while other actors have only recently taken up Broadway performing again.

We can also ask whether the distribution of acting credits could say anything about the importance of stardom in today's casting environment. The median year of these credits is 2003, meaning that most of the top Broadway actors were working in or around 2003. 2003 is obviously much closer to the present year (2025) than to the start of musicals in 1927– does this indicate that stardom has become more important in finding employment as an actor? These data cannot conclude this trend, but further visualizations of this nature could bring us closer to our understanding o this concept.

6. Data Exploration 3: Broadway writers

This section will examine Broadway writers via the three main writing credits for musicals: book, music, and lyrics. “Book” generally refers to the book of the musical, or the unsung scenes in a show. “Music” refers to the songs in a musical, though in getting the music into its final form many productions also have an orchestrator, arranger, vocal arranger, and many other music-related credits. It is occasionally true that, in addition to a music or lyrics credit, there also exists someone else credited for contributing a few songs to the show. This section, however, will use only the main credits for book, music, and lyrics, in addition to a few stray credits for “writing.”

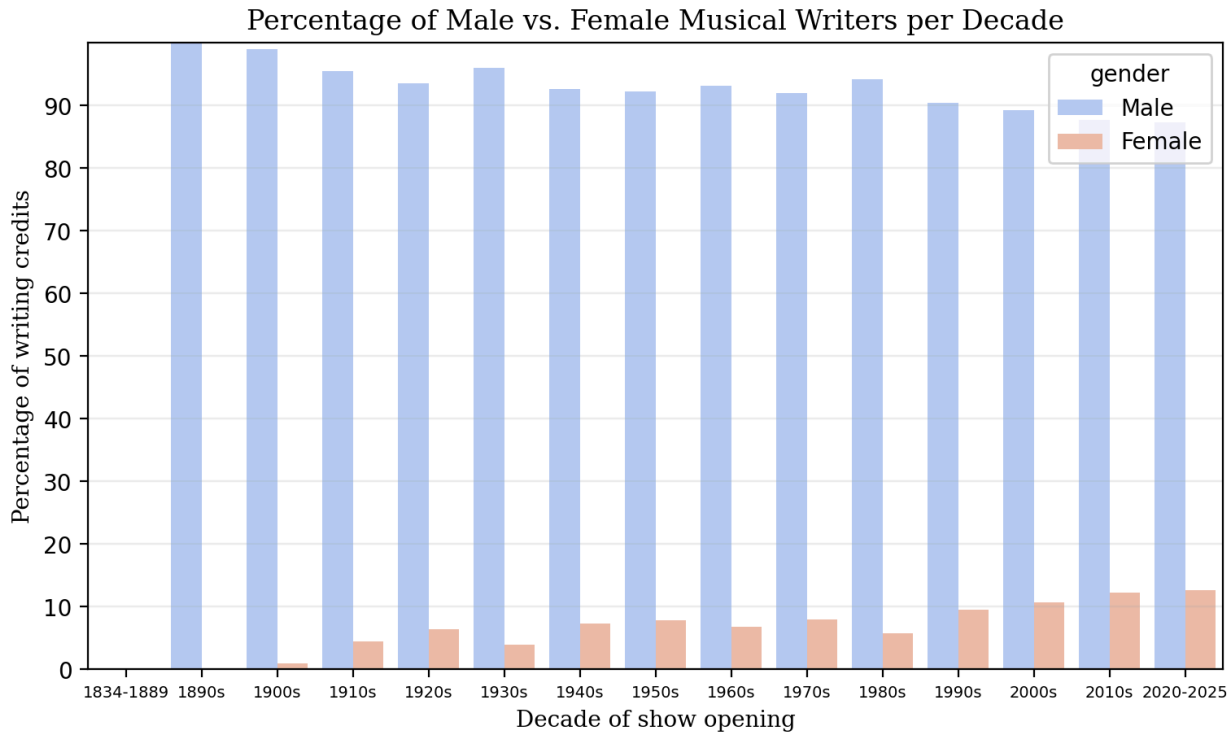
Like many art forms, theatre history tends to be organized not only by decade, but by the influence of certain artists. In the case of musicals, it is writers whom theatre histories tend to refer to when defining changes in the musical form. Richard Rodgers and Oscar Hammerstein II, for example, are known to be instrumental in defining the musical landscape both before and after their death. It is therefore important to quantitatively examine the influence, career trajectories, and demographics of this group.

6.1. Writer Demographics

First, let us examine the gender breakdown of Broadway musical writers. Of the distinct writers in my dataset, 1,514 are men (84.3%), 154 are women (8.6%), and 129 are unlabeled (7.2%).

Figure 6.1.1 shows the percentage of male versus female writers in each decade. For this analysis, I have removed people whose gender is unlabeled, and proceed by assuming that these unlabeled individuals’ genders are divided relatively proportionally to those of the labeled

Fig. 6.1.1. Percentage of Male vs. Female Writers per Decade



individuals. Additionally, the percentages in fig. 6.1.1 are calculated by looking at the individuals on each production that opened in that decade. They do *not*, however, count an individual twice if they were credited in two roles on the same production (e.g. music and lyrics).

As seen, women are severely underrepresented among Broadway writers, and at their maximum, from 2020-2025, have held only 12.6% of writing credits.The fact that men dominate musical writing is likely unsurprising, for the majority of musicals in the musical theatre canon were written by men. However, the degree to which this field has been and continues to be male dominated is quite shocking.

By my data showing theatre employment by gender we know that theatre has not been absent of female representation on stage, at least not in terms of "female representation" defined merely as women present onstage. Many popular and classic musicals have female leads, written

by both men and women. What this graph shows, however, is that those female characters must have been written almost entirely by men.

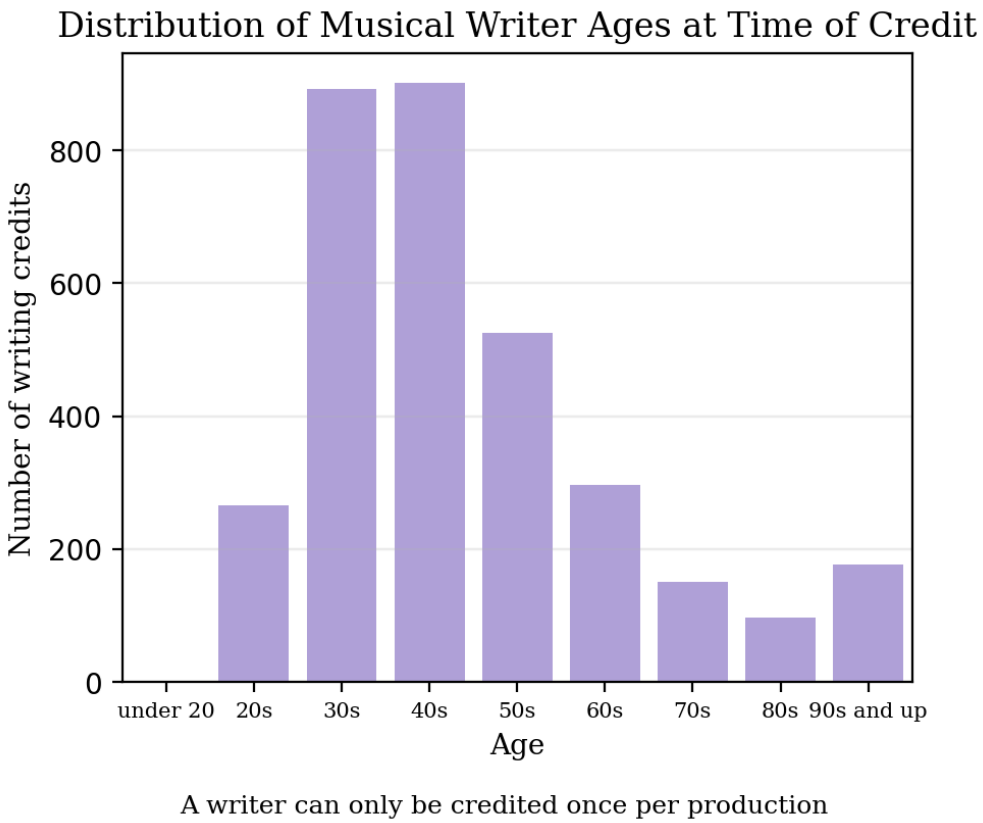
This fact alone bodes poorly for the quality of female representation on Broadway stages, and there indeed exists a body of theatre scholarship approaching this very issue, and approaching theatre history and criticism with a feminist lens. This work becomes all the more necessary once we see, from fig. 6.1.1, that the representation of women on musical writing teams is depressingly low.

Beyond suggesting lower quality female representation than male representation on Broadway, these data also demonstrate that there are likely factors making it more difficult for women to break into this career. Why aren't more women represented here, and why has this representation been so slow to improve? Since the 1920s, just after women were given the right to vote in the United States, the percentage of female writers per decade since the 1920s has increased only from ~6% to ~12%. These numbers greatly challenge our general understanding of how much Broadway has improved for women both on creative teams and in onstage representation.

Lastly, this study unfortunately cannot even begin to cover the representation of gender non-conforming writers, for IBDB makes no "gender" distinction beyond "male" and "female."

In addition to gender, we can explore the age demographics of these individuals. Figure 6.1.2 shows the ages of musical writers at the time they are credited. This figure shows that most Broadway musical writers are credited in their 30s or 40s. The "90s and up" category also encompasses writers credited posthumously, a feature I will examine in section 6.2.

Fig. 6.1.2. Distribution of Musical Writer Ages at the Time of Credit



6.2. Posthumous Writing Credits

Figure 6.1.2 shows that, under a strict definition of a musical, a substantial amount of Broadway shows are produced after one or more of its writers have died. In this section, I define an individual as having a posthumous writing credit when a show they wrote opens on Broadway after the year of their death.

To get a sense of the distribution of these credits, fig. 6.2.1. shows all of the writers who hold more than one posthumous credit. Notably, Oscar Hammerstein II has twice as many posthumous credits as the writer with the third most posthumous credits, Frank Loesser. As a reference, fig. 6.2.2 shows instead the total number of writing credits for each of these writers, both before and after their deaths.

Fig. 6.2.1. Posthumous Writing Credits by Writer

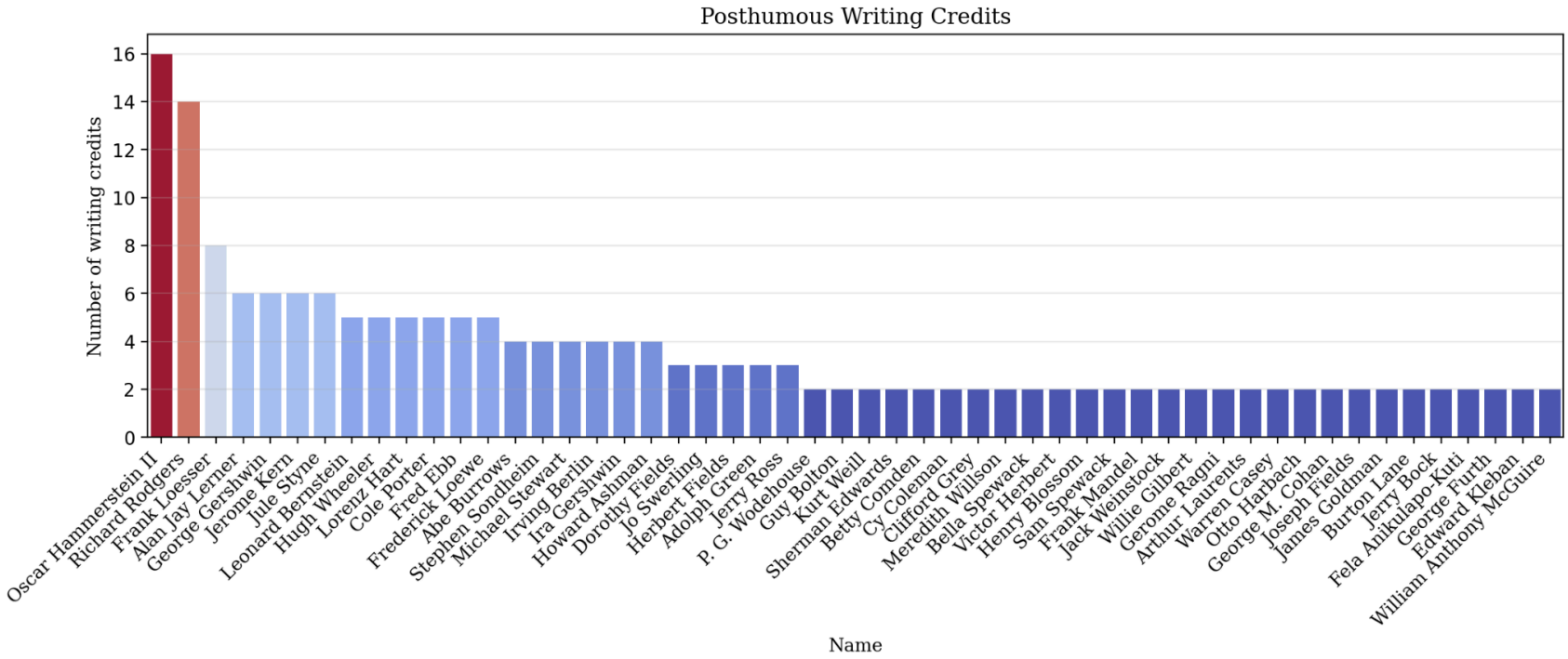
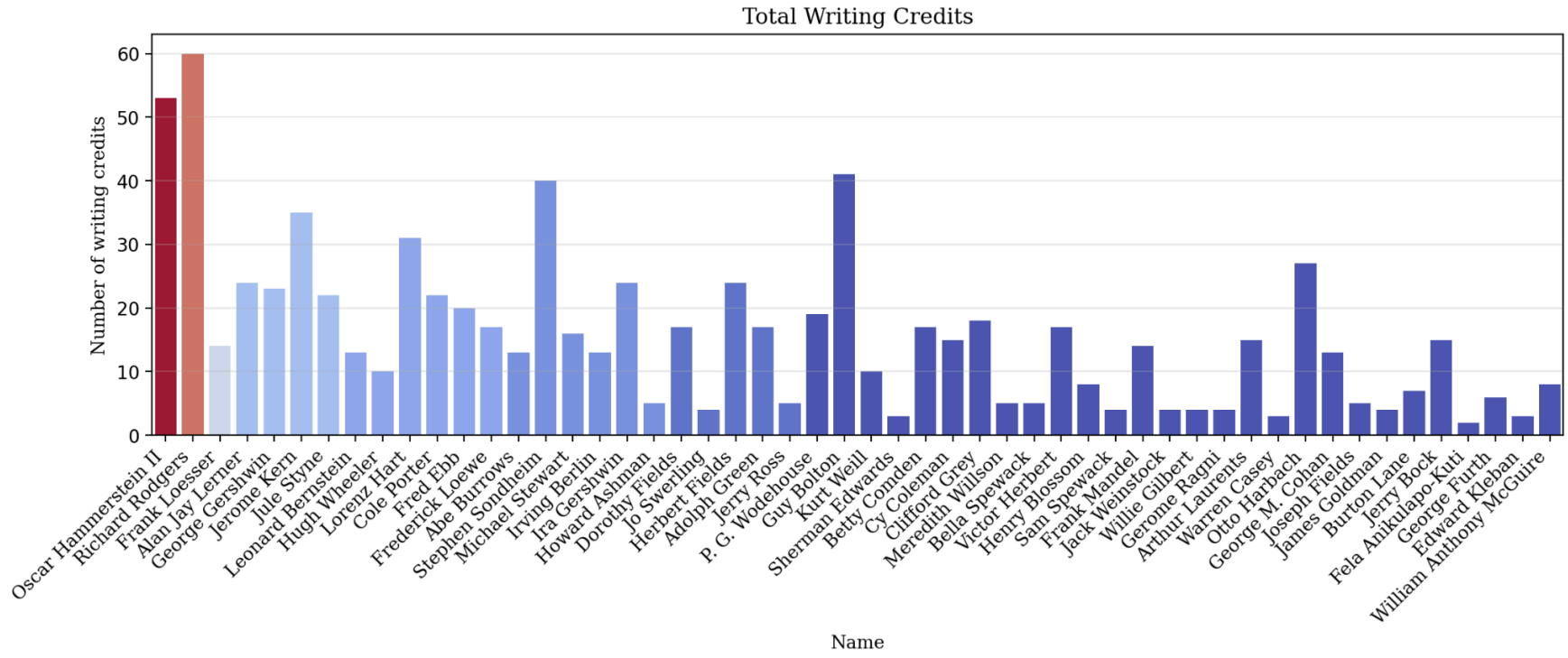


Fig. 6.2.2. Total Writing Credits for Individuals in Fig. 6.2.1



The number of posthumous credits is, of course, skewed by the number of years since the writer's death. To account for this factor, fig. 6.2.3 shows the density of posthumous writing credits, calculated as:

$$density = total\ posthumous\ credits\ /\ (current\ year - death\ year)$$

We see that Richard Rodgers and Oscar Hammerstein are still among the top three most credited by this metric, but Stephen Sondheim has moved strongly ahead of them both with a density value of 1.0. In other words, Sondheim, who died in 2021, has been credited on average every year since his death (in order, *Into the Woods, Sweeney Todd, Merrily we Roll Along,* and *Gypsy*). Could this support the claim that he has become more popular after his death?

Figure 6.2.4 maps each writer's cumulative posthumous credits in each year since they passed. Shown here are all writers who have at least 4 such credits. On the far left, we can see Sondheim's posthumous credit trajectory in the form of a short but steep line. Hammerstein, at the top extreme of the graph, has a steady rate of cumulative posthumous credits that brings him to the current high. Gershwin's curve, on the other hand, shows a fairly gradual increase, where he is credited only occasionally. We can think of flattening curves as indicating a writer losing relevance, and steepening curves as a writer becoming more popular. Additionally, the x-intercept represents how the timing of the first production after their death. How, though, do these trajectories compare to each writers' success while they were alive?

Similarly, fig. 6.2.4 measures cumulative credits for all time, in relation to their death year. Negative years since death are equivalently years before death. If we look now at Sondheim's credit data, the slope steepens at the year of his death (0 on the x-axis). This could support the claim that his work has grown in popularity since his death. This makes sense with

Fig. 6.2.3. Density of posthumous writing credits

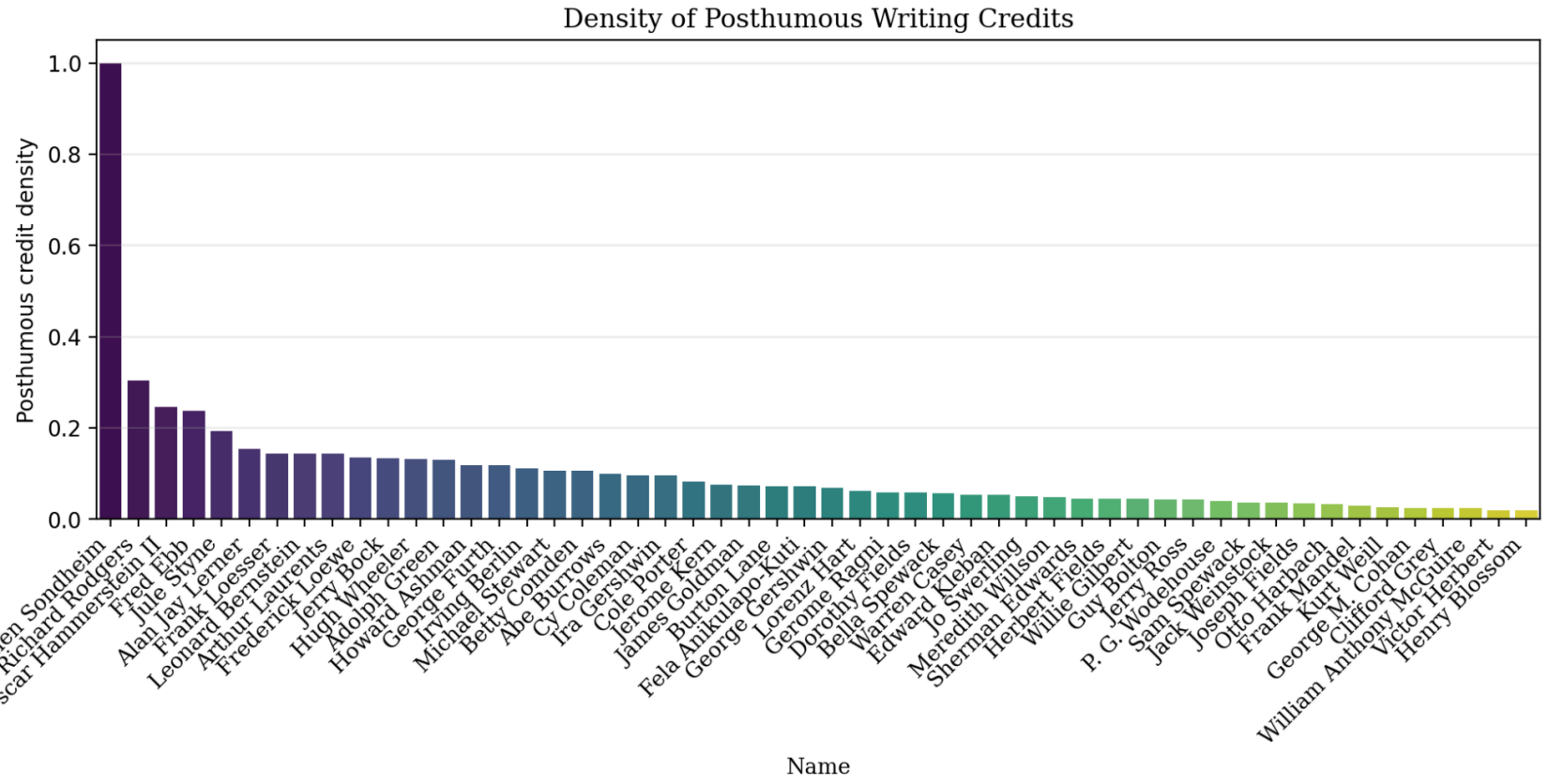
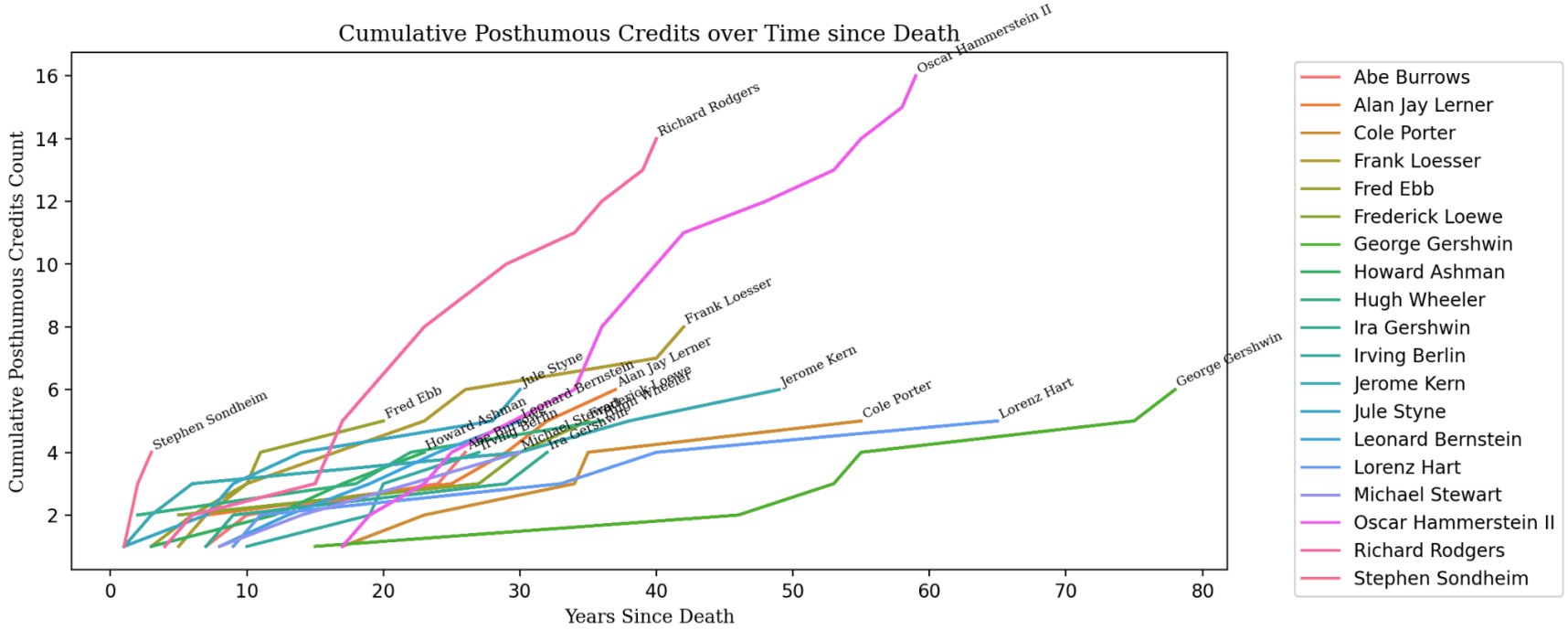


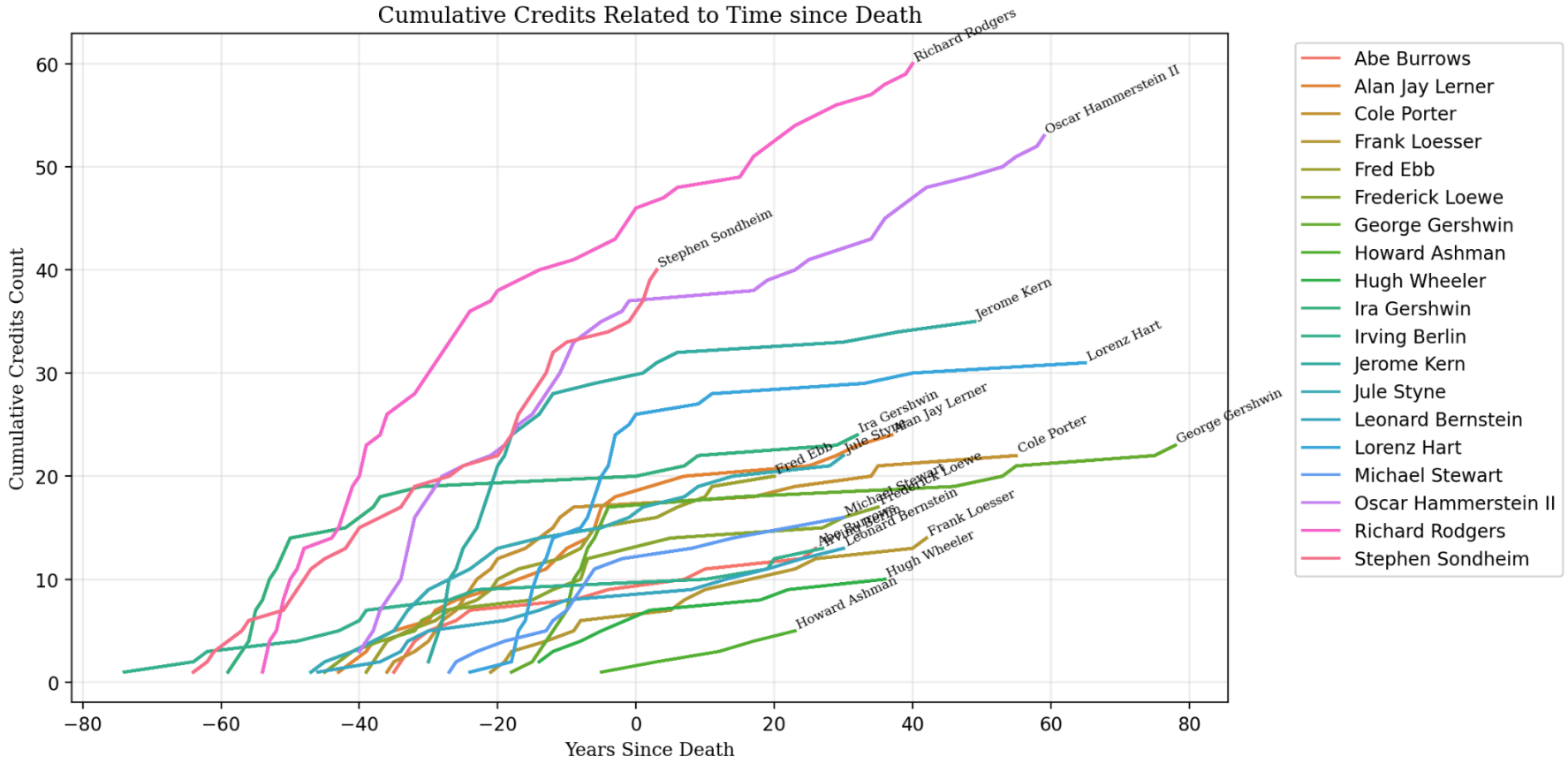
Fig. 6.2.4. Cumulative Posthumous Credits over Time since Death



our qualitative understanding of Sondheim's works, which were not nearly as popular or commercially successful when they were first released as they are now.

On the contrary, this graph shows writers whose work has lost traction after their death. In fig. 6.2.5, Jerome Kern and Lorenz Hart, for example, have curves that flatten out at the year of their deaths. For all writers with smaller bodies of work than, say, Hammerstein, Rodgers, and Sondheim, it must also be noted that this alone gives them a lower likelihood of receiving a posthumous credit. To maintain consumer interest, Broadway revivals must remain varied year to year, thus a writer with a small number of shows might not get multiple posthumous credits within five years while a writer with a large body of work might because it still maintains variety on Broadway.

Fig. 6.2.5. Cumulative Credits Related to Time since Death



6.3. Roles on Musical Writing Teams

In “Average Broadway,” Miller spends some time examining the presence of certain production credits on the first page of Broadway playbills, and using this as a tool to look at how the structure and valuation of industry careers throughout history [3]. For example, he finds that producers have always been listed on this page, while “lighting designer” is a credit whose presence on this page has been increasing from former nonexistence.

With my writer data, I was curious to see how often a writer takes on more than one of the main roles of book, music, and lyrics. Figure 6.3.1 shows that, as an individual credit, it is most common to write the book, and as a multi-credited writer it is most common to write book and lyrics. Most famously, Hammerstein tended to write both the book and lyrics for his collaborations with Rodgers. Only a small number of individuals are credited with writing book, music, and lyrics at once.

Fig. 6.3.1. Frequency of Combined Writing Credits

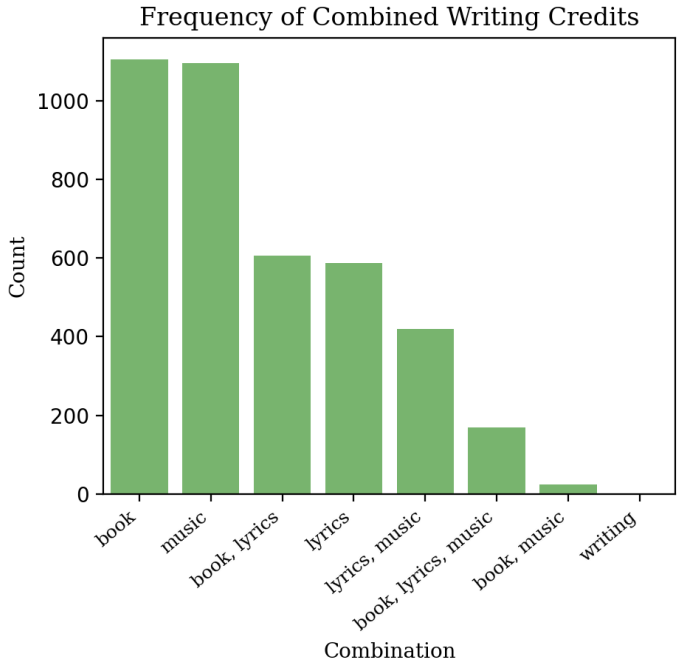
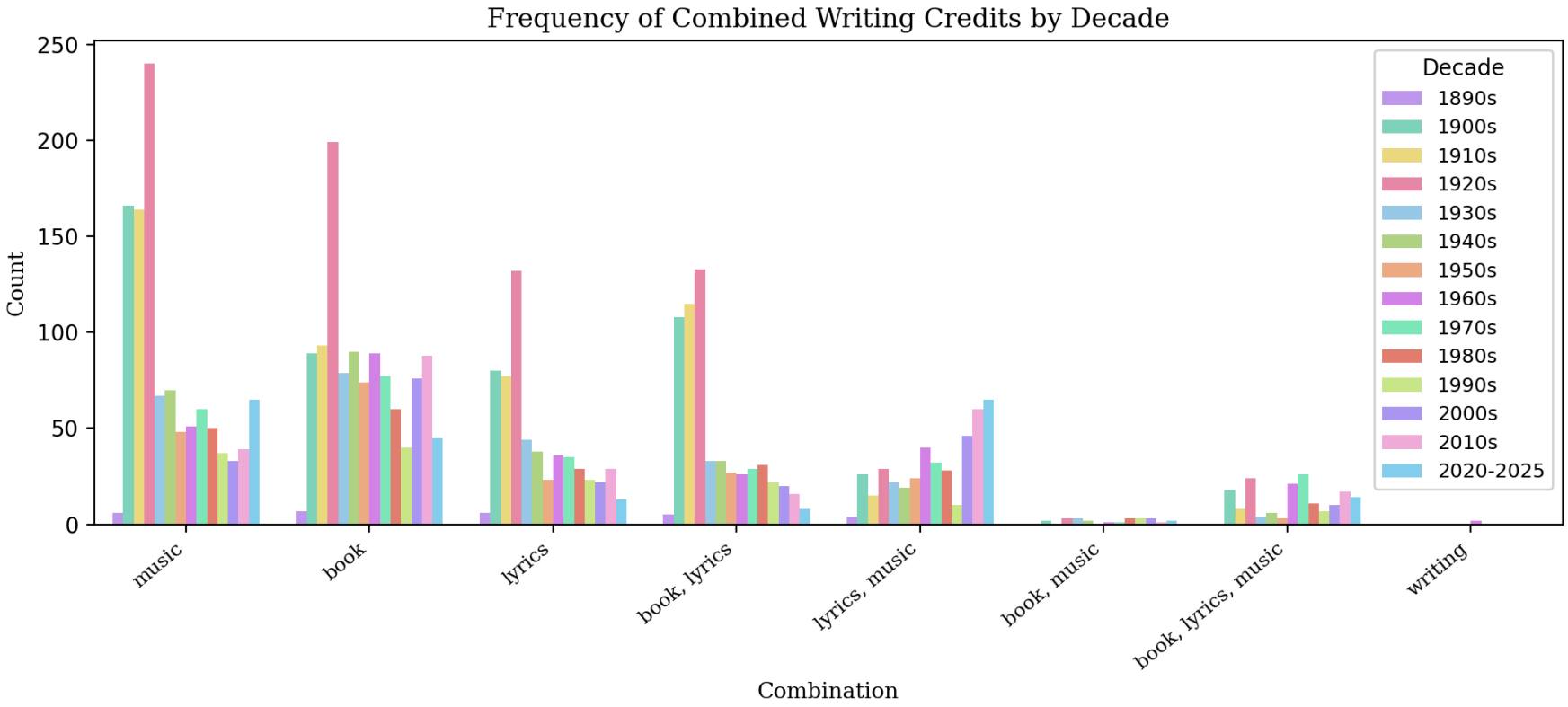


Figure 6.3.2 provides further detail to the prevalence of these writing credits over time. Each category is now broken down by the frequency in each decade, showing, for example, that while it has become less common to see one individual credited with book and lyrics on a Broadway musical, it has become more common to see one individual credited with lyrics and music.

Though such observed trends are influenced by, say, the frequent credits of book/lyric writer Hammerstein, it also might be true that that prevalence is what *sets* trends for writers of that era. Did the success of the way Rodgers and Hammerstein work influence other writers to organize their roles in similar ways? Rather than continuing to focus only on the products that comprise musical theatre history, this understanding of roles begins to interrogate the process as well.

Fig. 6.3.2. Frequency of Combined Writing Credits by Decade



7. Conclusion

Approaching musical theatre history from a digital humanities perspective emphasizes the importance of interdisciplinary work within the field. Though musicals often prove resistant to being flattened into data, this resistance helps us understand what has defined the Broadway musical at its most basic level and provides valuable insights into where tensions lie between the definitions and histories given by scholars and institutions. In fact, these tensions begin at the very definition of a "musical," which with it calls into question the distinction between its preceding sub-genres of music theatre. As seen, the overlap and divergence of these sub-genres reveals both an ever-evolving form and an ever-evolving vocabulary to describe the form.

One of the most obvious advantages of the application of the digital humanities to theatre history is the ability to examine trends via large quantities of data. In addition to supporting the findings of historians, these data-based trends are often able to uncover unexpected or unexamined patterns. For example, the findings in Section 3.1 show a shockingly sharp decrease in revivals in the 1960s, unexplained even when we control for the total number of shows in the season and the kinds of shows represented. On the other hand, this section confirms an intuitive understanding that old musicals after a certain point fall out of style and stop being revived. These results turn the conversation back to historians, who with these quantitative findings might find explanations at a more micro level.

As seen in this paper especially by the gender disparity in musical writing credits, data can also call attention to structural inequalities that have otherwise failed to be formally affirmed. Historic gender data demand that the stubbornly low representation of women receives serious attention from the industry as a whole.

Lastly, this data-based approach to musicals gives context to why the historical narrative is described the way it is: the emphasis on Rodgers and Hammerstein in any theatre history can be understood by the fact that they have more Broadway credits than any other individuals, and their relevance has not decreased since either of their deaths. Without using data to zoom out, it is harder to catch a hold of what was going on at any moment in Broadway history. Data can easily give us several different snapshots of a given moment, in terms of the actors who were performing, the writers who were credited, the locations of the active theatres, and hundreds of other insightful measures.

Above all, this paper demonstrates the massive body of unresearched data-driven questions within Broadway musicals. In a few short explorations, I was barely able to scratch the surface of the ways my data alone might give us a new understanding of theatre history, and nearly every analysis left me with tens of questions about the displayed trends as well as the possible gaps and inconsistencies in the data. This paper also demonstrates the exciting discoveries that improved digital theatre records might uncover, and the various reasons why such an investment would be important.

8. Future Work

As described, there are countless possibilities for future work on this subject, both within the scope of the data of this project and beyond. Starting large, I would be interested in addressing many of the genre inconsistencies in musical history, and going into various forms of archival records to be able to define a better method for turning those genres into standardized data. With this, I would be interested in finding which features actually distinguished these related genres at different points in time, and analyzing based on these findings.

I would also like to use network theory to study the relationships between actors, writers, directors, and producers. In an industry with interpersonal relationships at its core level of functioning, it would be fascinating to see whether data express how established relationships affect future jobs, generate new work, and promote collaboration.

With more time with this dataset, I would like to connect musical film actors and stage actors to examine the overlap between these two groups. Though I was able to automatically connect the TMDB and IBDB pages of actors who had birth dates on both pages, the large majority of these actor pages would have to be evaluated in some other way to confirm where people with the same name on both platforms are, in fact, the same actor. With this information, I could connect the networks of stage and screen actors, expanding the previous piece of future work.

As mentioned in section 4.2, where I illustrated a naive approach to measuring relevance of a property, it would be interesting to try to generate a better quantitative definition via more mathematically sophisticated methods. With this particular question, I would also be interested in factoring in other theatrical entities, such as cast albums and educational theatre productions, to get a holistic view of the way musicals are transmitted beyond New York City.

As a last example, I would be curious to combine these studies of theatre as events with theatre as *text*. In particular, it would be interesting to study different characters in musicals, in terms of some measure of their importance in a given show, the archetypes they fulfill, or some other set of text-based measures. I mean "text" as described by Varela, and indicating the content of the show [5]. Combining these data with the ones used in this project, we might look at questions such a "what vocal range did X actor perform at various points in their career?" or "what percentage of lines on Broadway stages were spoken by women over time?"

These are just a few examples of the future work that could build upon the data and findings of this paper. On the whole, the collaboration between the fields of digital humanities and musical theatre shows promising potential for deepening our understanding of musical theatre history.

Works Cited

[1] nyfa. “Broadway vs Off-Broadway: The Difference Between Common Musical Theatre Terms.” *NYFA*, 12 Dec. 2023, https://www.nyfa.edu/student-resources/difference-broadway-off-off-broadway/.

[2] Preston, Katherine K. “1 - American Musical Theatre before the Twentieth Century.” *The Cambridge Companion to the Musical*.

[3] Miller, Derek. “Average Broadway.” *Theatre Journal*, vol. 68, no. 4, 2016, pp. 529–53.

[4] Caplan, Debra. “Reassessing Obscurity: The Case for Big Data in Theatre History.” *Theatre Journal*, vol. 68, no. 4, 2016, pp. 555–73.

[5] Varela, Escobar. *Theater as Data*.

[6] *About The Internet Broadway Database | IBDB*. https://www.ibdb.com/about/. Accessed 30 Mar. 2025.

[7] *Help | IBDB*. https://www.ibdb.com/help/. Accessed 30 Mar. 2025.

[8] *Above-the-Line vs. Below-the-Line Crew: What's the Difference?* 8 Apr. 2022, https://www.backstage.com/magazine/article/above-the-line-vs-below-the-line-crew-differences-74969/.

[9] Assemble. *The Definitive Film Crew Hierarchy Chart*.

[10] “Popularity & Trending.” *The Movie Database (TMDB)*, https://developer.themoviedb.org/docs/popularity-and-trending. Accessed 31 Mar. 2025.

[11] “Why Don’t We Paint the Town?: Chicago Transfers to Bway’s Ambassador Theatre Jan. 29.” Playbill, https://playbill.com/article/why-dont-we-paint-the-town-chicago-transfers-to-bways-ambassador-theatre-jan-29-com-108928. Accessed 31 Mar. 2025.

[12] *Pirates! The Penzance Musical – Broadway Musical – 2025 Revival | IBDB*. https://www.ibdb.com/broadway-production/pirates-the-penzance-musical-539492. Accessed 2 Apr. 2025.

[13] *The Gershwins' Porgy and Bess – Broadway Musical – 2012 Revival | IBDB*. https://www.ibdb.com/broadway-production/the-gershwins-porgy-and-bess-490541. Accessed 2 Apr. 2025.

[14] “The Gershwins’ Porgy and Bess [Musical].” *Gershwin*, https://gershwin.com/publications/gershwins-porgy-and-bess-musical/. Accessed 2 Apr. 2025.
[15] Replogle-Wong, Holley. *The Great Generational Divide: Stage-to-Screen Hollywood Musical Adaptations and the Enactment of Fandom*.
[16] *Disney’s $75M Hamilton Acquisition Is a Franchise Power Move | Observer*. https://observer.com/2020/02/disney-hamilton-movie-release-date-streaming/. Accessed 9 Apr. 2025.
[17] Whitten, Sarah. “Putting ‘Hamilton’ on Disney+ Amplified Demand for the Broadway Show, Lin-Manuel Miranda Says.” *CNBC*, 24 Nov. 2021, https://www.cnbc.com/2021/11/24/hamilton-film-amplified-ticket-demand-lin-manuel-miranda-says.html.
[18] Stempel, Larry. *Showtime: A History of the Broadway Musical Theater*. 2010.
[19] *Gutenberg! The Musical!* https://www.tonyawards.com/shows/gutenberg-the-musical/. Accessed 9 Apr. 2025.
[20] *About TMDB — The Movie Database (TMDB)*. https://www.themoviedb.org/about?language=en-US. Accessed 19 Apr. 2024.
[21] *Gigi*. Directed by Vincente Minnelli and Charles Walters, Metro-Goldwyn-Mayer (MGM), 1958.
[22] Bollen, Jonathan. "Data Models for Theatre Research: People, Places, and Performance." *Theatre Journal*, vol. 68 no. 4, 2016, p. 615-632. *Project MUSE*, https://dx.doi.org/10.1353/tj.2016.0109.
[23] *Story Map Journal*. https://www.arcgis.com/apps/MapJournal/index.html?appid=dbe468bfd33343dc96c23db1da55f803. Accessed 10 Apr. 2025.
[24] *The Broadway League | The Official Website of the Broadway Industry*. https://www.broadwayleague.com/home/. Accessed 10 Apr. 2025.