Black Panther vs The Help, the battle of inclusion versus tokenism:

A comprehensive examination of underrepresentation of ethnic minorities in Hollywood films

A logo with blue and yellow squares

Description automatically generated

**Name:** Bastiaan van Houten

**SNR:** 2082816

**Study program:** MSc Marketing Management/Analytics

**Academic year:** 2023-2024

**Submission date:** January 5, 2024

**Name supervisor** Dr. Arjen van Lin

**Name co-reader** Dr. Kunthi Kusumawardani

Management summary

In recent years, the issue of minority representation in Hollywood films has become increasingly prominent. Countless studies have been published on this topic, and there is a growing demand for clarity about the relationship between representation and its effects.

This paper addresses the pressing issue of diversity and representation within the film industry, specifically focusing on Hollywood films. Films are powerful cultural influencers that shape societal beliefs. The lack of representation in films can perpetuate negative stereotypes and marginalization, which can have a harmful impact on individuals and communities. The study highlights that films with diverse casts tend to perform better at the box office due to their wider appeal. This suggests that diversity is not only a moral imperative, but also a financial necessity. The study highlights the economic significance of the film industry, which generates billions of dollars in annual revenue and produces some of the most profitable products in the world.

Despite recent improvements in representation, the study underscores that existing measures often overlook the nuances of representation and the need for authentic inclusive representation (AIR). AIR is a comprehensive approach to representation that accounts for effective representation rather than mere tokenism. The study argues that AIR is essential for creating films that resonate with diverse audiences and positively impact society.

Preface

Table of contents

Chapter 1 . Introduction

* 1. Problem Indication

The film industry is a major part of our global economy. Films generate nearly $100 billion annually through theaters and home channels. Moreover, the economic impact is evident in successful products. For instance, Disney's "The Force Awakens" earned a net profit of $780 million (MPA, 2022), displaying that films are a serious, high-return industry despite their light-hearted nature. Furthermore films do not only hold economic significance, they are related to cultural aspects of society.

Films are not only a form of entertainment, they also reflect and shape the cultural context in which they are created. They mirror societal values, attitudes, and perspectives and can effectively convey ideas, emotions, and perspectives through their visual and narrative nature. As a result, they have the potential to shape how people perceive the world and the cultures around them (Belton, 1995).

In today's increasingly diverse and socially conscious world, this responsibility currently translates that the Hollywood film industry is under increasing pressure to be socially conscious and to address issues of representation, particularly in relation to gender and people of color (Sperling, 2021). The common consensus is that representation in films is important because it counteracts narrow viewpoints regarding groups depicted in the media, which can reinforce negative stereotypes and biases that result in discrimination and marginalization (The Annenberg Foundation, 2018; Castañeda, 2015; Ross, 2019; Kubrak, 2020; Buchanan, 2005).

Moreover, representation in films is not only important for social factors, studies show that films with diverse casts appeal to broader audiences and perform better at the box office (Annenberg Foundation, 2018). For example, a 2018 study by the University of Southern California found that films with diverse casts were 1.4 times more likely to be seen by broader audiences. Additionally, a 2021 McKinsey and Company study found that films with casts that had at least 30% minority actors were 1.3 times more likely to be profitable. By prioritizing diversity and representation, filmmakers can create more inclusive and authentic stories, attract larger audiences, and boost their bottom line (Whitten, 2019; Bunche, 2018; Reporter, 2021).This suggests that racial diversity is both a moral and financial imperative for the film industry.

Efforts to increase diversity and representation in film have led to a significant increase in the proportion of films featuring minority actors. For example, a study by UCLA found that in the years from 2011 to 2021, films went from 51.2% having less than 11% of colored actors to 28.8% of films featuring a cast that has a majority of colored actors.

Nevertheless, the public, media, and most studies adopt a narrow approach measuring diversity (Malik, 2022). Most studies do not make a distinction between different minority groups and measure diversity based on the share of minorities in the cast. This approach has caused two major issues.

Firstly, the approach of treating minority groups as a homogenous unit has resulted in uneven representation across the different minority groups. For instance, while African-Americans have been overrepresented in films for the past three consecutive years and the Asian community is ‘rightfully’ represented, the Latin community remains severely underrepresented, with a gap of 13.2% points (compared to an actual population of 20%) respectively. (UCLA, 2021; MPAA, 2021). As a result, this group still experiences the social consequences of being underrepresented.

Secondly, measuring diversity based solely on minority share overlooks crucial information about the nature of the representation (Malik, 2022; Lazar, 2020). To effectively address the social dimension associated with diversity, “authentic inclusive representation” (AIR) is necessary (Lazar, 2020; Roughton , 2014). That is, to accurately and respectfully depict a character from a diverse background, films need to go past tokenism and need to create meaningful storylines for the character (Lazar, 2020).

With AIR being arguably crucial for effective representation this study argues that by neglecting to consider AIR alongside diversity, previous studies such as (Kuppuswamy, 2016; Weaver, 2011) have fallen short to conduct a comprehensive analysis of the effects of racial diversity. This causes that the true effects on both the social dimensions and financial implications have not being properly studied previously.

Lastly, as discussed racial representation in film is a social construct. Nevertheless, most studies use box office revenue as the dependent variable to analyze a film’s success (Malik, 2022; Kuppuswamy, 2016;Madongo, 2023). While box office revenue is a useful measure of commercial success, it does not necessarily reflect the quality of a film or its impact on audiences, as it is highly influenced by marketing , advertising and hype (Eliashberg, 2014; Clement, 2014). Therefore, this study will choose long-term audience engagement (LTAE) as dependent variable. As discussed later this concept entails both social and economical constructs which makes it a better measurement for a film’s success in the context of a social concept such as racial representation.

1.2. Problem Statement

Following the problem background this studies problem statement is formulated as follows: “*What is the relationship between racial diversity in Hollywood films and LTAE, and to what extent does the presence of AIR mediate this relationship?”*

1.3 Academic Contribution

Previous studies on racial diversity in films have taken a narrow approach, either by focusing on only one or two ethnic groups (Patel, 2015; Hall, 2020; Dixon, 2000 ; Kuppuswamy , 2016) or by grouping all minority groups together (Aumer, 2017). This study addresses these limitations by including multiple minority groups and distinguishing between them. This is important because different minority groups have different experiences and perspectives, and their representation in films should be considered separately. Moreover, analyzing all minority groups separately improves the generalization of the results.

Second, as noted by Malik (2021, p. 1), 'there are no clearly defined, standardized, and scalable metrics for taking stock of racial minorities’ cinematographic representation'. Previous studies such as Weitzman et al. (1972) and Smith et al. (2013) have used manual annotation, which produces high-quality insights, but is time-consuming and expensive. This study uses an automated process to measure the concept of AIR and racial diversity. By doing so, this study not only aims to be as inclusive as possible, going beyond mere inclusion of actors but also makes the process applicable for automation. Therefore, it can be easily applied to a large number of films.

Building on prior research on gender biases in film (Agarwal et al., 2015; Kagan et al., 2020), this study seeks to standardize the concept of authentic inclusive representation (AIR) using the Bechdel-Wallace test (1985). This is not the first time the Bechdel test has been applied to racial representation. In 2020, the UCLA Center for Scholars & Storytellers developed the REM test, but they have not yet automated it. Moreover, the REM test does not distinguish between different minority groups. By combining the work of previous researchers, this study makes a significant contribution to the academic literature on racial representation in films.

Third, as mentioned before by using Long-term audience engagement (LTAE), this study uses a more comprehensive measure of success, as it takes into account factors such as word-of-mouth, social media engagement, and re-watching. This contributes to existing literature because it provides a more nuanced understanding of how audiences are engaging with films with diverse casts. Moreover, this causes that this study in contrast to other studies identifies films that are having a lasting impact on audiences, even if they may not have been box office blockbusters.

Using this approach, this study builds upon previous research and provides a more comprehensive and in-depth analysis of the relationship between representation and long term audience engagement in films.

1.4 Managerial Contribution

This study offers valuable insights for filmmakers, studios, and stakeholders, helping them make informed decisions that can significantly impact a film's success in both financial and cultural terms. By intentionally casting diverse roles and developing characters that authentically represent a variety of backgrounds, ethnicities, and cultures, filmmakers can attract a wider audience, which can boost a film's financial performance.

Furthermore, measuring the long-term success of films with diverse casts and themes allows studios to assess whether embracing diversity can expand their audience reach over time. This would require a shift in focus from short-term box office results to long-term audience engagement. Ancillary revenues, such as streaming and DVD sales, may be just as important as box office success. Measuring the long-term financial performance of films with diverse casts and themes can help studios make better investment decisions.

Finally, if this study offers evidence indicating a potential preference for racially diverse films. It is highly probable that there is heterogeneity in this preference among people. By establishing that racially diverse films receive greater favor, this paves the way for further research to investigate the extent of variation among different individuals. The findings from these studies could potentially be leveraged to develop movie recommendation systems tailored to individuals with a preference for racial diversity or individuals who have diversity or inclusivity as a key attribute in their profile.

1.5 Structure of the Thesis

To address the research questions stated earlier, a quantitative research approach was employed in this study. This paper will first present a theoretical framework to contextualize the research, followed by a detailed description of the research methodology. Afterwards it will be discussed how the data collection and processing went. Finally, the results will be discussed.

Chapter 2. Literature Review and Hypotheses

This literature review delves into the current existing literature on the relationship between racial diversity in Hollywood films and LTAE. To provide clarity before delving into the relationships we first define LTAE and racial representation, and discuss the importance of AIR in the latter.

2.1 Long-Term Audience Engagement (LTAE)

Audience engagement is a complex and multifaceted concept that encompasses the active involvement of an audience. In this study, the focus is on long-term audience engagement (LTAE), which goes beyond passive consumption and encompasses factors such as word-of-mouth, cultural impact, and sustained interest over time (Broersma, 2019; Kumar, 2022).

Before delving into the social attribute of LTAE it is important to clarify that LTAE has financial purposes. Similar to box office revenue LTAE is correlated with ancillary revenues such as engrossed viewing, longevity in distribution channels, and the purchase of film-related merchandise (Kumar, 2022). These factors signify a lasting connection with the film beyond its initial screening and shows that LTAE is correlated with revenue. Due to the financial dimensions related with LTAE it makes box office revenue to become obsolete as component. These financial constructs exist through the social components of LTAE.

LTAE is a key factor in the media's capacity to shape attitudes, beliefs, and behaviors. It allows for a deeper and more sustained connection between the audience and the film, which can have a lasting impact on their lives (Tan, 2018). The effect of LTAE and changing behavior is demonstrated by Bard research in 2006, which discovered that individuals highly interested in violent video games for a longer period of time were more prone to displaying heightened levels of aggressive behavior, aggressive cognitions, feelings of anger, and physiological arousal.

Furthermore, the effects of long term engagement with media and changing our beliefs was displayed by Mastro et al. in 2007 which found that white people who were constantly exposed to negative racial stereotypes in the media were more likely to hold those stereotypes themselves. This was especially true for white people who don't have much real-life contact with people of color. When we're constantly exposed with negative portrayals of certain groups of people, it is hard not to start believing them. Therefore, the importance of racial representation in media will be discussed in the following paragraph.

2.2 Racial Representation in Films

Representation of minorities in films pertains to the presence and portrayal of characters from minority groups, including people of color, people with disabilities, LGBTQ+ individuals, and other marginalized groups (Buckingham, 2008). Representation is essential because it allows people from different groups to see themselves on screen, which can lead to a greater sense of inclusion, empowerment, and validation (Annenberg Foundation., 2018; Dixon, 2000).

In contrast, the absence of representation has negative impacts on self-worth, as individuals from underrepresented groups do not see themselves or their experiences reflected in mainstream media (Castañeda, 2015; Ross, 2019; Kubrak, 2020; Buchanan, 2005). The spread of positive and accurate portrayals in the media is therefore essential for people to explore their identities with regards to race.

Racial identity is a complex concept that is constantly evolving. This study uses the defintion of Umaña-Taylor AJ, (2014 p. 3) which defines racial identity as “a multidimensional psychological construct that reflects the beliefs and attitudes that individuals have about their ethnic group memberships”. The four ethnicities discussed in this study are Black, Hispanic, White and Asian.

Our racial identity is not solely a product of our own perspective but is also shaped by how others perceive us. In multicultural societies, building positive relationships among diverse groups presents a significant challenge. In densely populated areas characterized by ethnic segregation, individuals often encounter other cultures and ethnicities solely through media portrayals (Kidd, 2015).Therefore, the powerful impact of media representations on shaping perceptions cannot be underestimated.

Positive depictions of communities of color can diminish feelings of threat and social distance among white audiences (Dalisay and Tan, 2009; Ortiz and Harwood, 2007), whereas negative portrayals, such as associations with criminality, can exacerbate negative stereotypes and widen divisions among ethnic groups (Abraham and Appiah, 2006; Hurley et al., 2015).

Further in this study the ongoing debate regarding the extent to which films should mirror the diversity of the real world will be discussed. Ultimately, filmmakers must understand the potential impact their creative choices can exert on viewers. Moreover, as highlighted in the next few paragraphs for filmmakers to understand how to represent ethnic minorities in a successful manner is as equally important.

2.2.1 Authentic Inclusive Representation (AIR) in Racial Representation:

The representation of minority groups in films takes on various forms. Some films and studies focus on the experiences of minority characters, while others simply include them as part of a larger cast (Malik, 2021). Nevertheless, overlooking authenticity in representation can result in films being mistakenly categorized as racially diverse while still perpetuating stereotypes and contributing to marginalization. As discussed previously these stereotypes in the film can perpetuate biases and misconceptions about minority groups (Umaña-Taylor AJ, 2014; Abraham and Appiah, 2006; Hurley et al., 2015) . Therefore, the concept of Authentic Inclusive Representation (AIR) emerges as a crucial element in racial representation.

The pursuit of authenticity in the representation of minorities in films is not a recent development. In 2014, Ralph Roughton stressed that genuine understanding and empathy, free from stereotypes, are key to changing attitudes. Effective representation requires viewers to truly understand and empathize with characters. (Roughton, 2014). AIR, as a concept, describes how accurately and respectfully a film portrays underrepresented groups in a nuanced manner. To make this concept more tangible, researchers often use the Bechdel-Wallace test (1985), which was originally designed to measure the authentic representation of women in a film.

The Bechdel-Wallace test can be adapted to measure the representation of ethnic underrepresented groups. The test can be summarized as follows: Does the film contain a conversation between two named characters of ethnicity X that is not about a white character? An example for such a conversation would be a conversation between the main character and his wife in the film ‘12 years a slave’ :

**ANNE**

Solomon...

**SOLOMON**

Come, Anne. Jump.

**ANNE**

I will not ruin my dress. Catch me!

**SOLOMON**

I will catch you, Anne. I will.

**ANNE**

You will.

Even though the conversation is not very thorough or meaningful, it is between two named African American characters, and no white people are present or mentioned in the conversation. Therefore, this film would have AIR for black people. With an understanding of the concepts, the following paragraphs will explain the relationships between one another.

2.3 The Relationship of Racial Diversity and Long-Term Audience Engagement (LTAE),:

The power of storytelling lies in its ability to create connections between viewers and characters who possess relatable qualities and admirable traits (Murray, 1999; Appiah, 2001; Hall, 2020). This connection is strengthened when there are similarities in demographic factors such as ethnicity, age, and gender, creating a sense of affinity between viewers and the on-screen portrayals (Hall, 2020).

As a result, when individuals see themselves or their own experiences represented in a story, they are more likely to form a strong emotional bond and become engaged with the film and its characters (Murray, 1999; Appiah, 2001; Hall, 2020). This suggests that when a film embraces racial diversity, it has the potential to attract a broader audience and foster greater overall LTAE (USC Annenberg, 2018).

Nevertheless, the demand for greater racial diversity in film is not universally embraced. A study by King (2020) identifies a group while not opposing racial representation in film, prioritizes the quality of the storyline. Arguing that when minorities are introduced without developed characters it influences the quality of the storyline. An effect discussed more in detail in the next section of this paper.

Patel (2015) argues that efforts to increase racial diversity have faced criticism, particularly from those who fear change and the increased visibility of people of color due to a persistent culture of colonialism or systematic racism. This idea of preference has led to the industry's practice of whitewashing, which is based on the assumption that white majority audiences prefer racially homogeneous casts and that diversity would not positively impact LTAE (Weaver, 2011).

The assumption that white majority audiences prefer racially homogeneous casts is believed to have led to studios allocating smaller budgets to projects with higher racial diversity. This assumption is supported by research by Smith et al. (2020), which found that films featuring racial minorities in lead roles often receive significantly less production budget. It is important to note that lower production budgets are a significant predictor of lower box-office sales (Eliashberg, 2014; Michel Clement, 2014), which can make it difficult to assess the true potential appeal of diverse films.

Nevertheless, it is essential to emphasize that the preference for whitewashing doesn't hold true for all scenarios (Aumer, 2017). Recent studies have also raised doubts about the idea that white actors are necessary for financially success of films (Chow, 2016). The prevalence of whitewashing may be more a product of industry habit than an accurate reflection of audience preferences. Nonetheless, due to the underfunding of diverse films, accurately assessing their potential appeal to audiences becomes a challenging task.

Furthermore, Teresa Correa's study in 2011 delved into the connection between racial diversity and social media engagement, revealing that minority groups tend to be more prolific creators of online content. This heightened interaction between minority communities and media is further evidenced by the active involvement of African-American and Latin communities in both traditional and digital entertainment, surpassing their proportional representation in the U.S. population (Gonzalez, 2014; MPAA, 2014).

This trend is likely to become even more impactful in the future as minority groups continue to grow as a percentage of the total U.S. population (Desilver, 2015). When a film successfully resonates with these audiences through representation, it is more likely to stimulate online discussions, which can lead to higher LTAE (Kumar, 2022).

Finally, it is argued that diverse casts can better reflect the diversity of the real world, which can help viewers to connect with the characters and the story (USC Annenberg, 2018). Despite the challenges posed by current industry forces and tokenism (which will be discussed in the next paragraph), the general consensus is that increased racial diversity in films has a positive impact on LTAE. Based on these considerations, the following hypothesis is proposed  
**H1:** Increased racial diversity has a positive influence on LTAE.

2.4 The Relationship Between Diversity, LTAE, and AIR as a Mediator:

Including diverse characters in films is not without its challenges. One concern is tokenism, which occurs when underrepresented characters are introduced without fully developed storylines. This can lead to backlash and damage LTAE if viewers perceive the diversity as being insincere , which can ultimately affect the perceived quality of the film (Smith, 2016). For example, when minorities are introduced through tokenism, it can create the perception of a "racial agenda," which is often disliked by viewers (King, 2020).

Moreover, negative stereotypes are seen as a gross misrepresentation by balanced critics who advocate for avoiding films with stereotyping regardless of representation. Since critic reception plays a crucial role in shaping LTAE, misrepresentation is likely to lead to a decline in LTAE (Hofmann, 2016; Ghiassi, 2017; Kuppuswamy, 2016; Kumar et al., 2022).

Furthermore, these negative depictions can cause minority groups to disengage completely from the media. For example, a study by El Hazzouri (2019) found that ethnic minorities who saw public health ads featuring people from their own ethnic group were less likely to follow the advice in the ads than those who saw ads featuring white people. The authors explained this by saying that minorities felt like they were being negatively stereotyped by the advertisers.

In contrast AIR can have a positive effect if done correctly. According to research, large-budget films in 2021 performed better when they had more racial authentic representation (Lazar, 2020). The study emphasizes that racial representation must be authentic in order to foster empathy, understanding, and connection among viewers. There is an intuitive belief that racial diversity leads to authentic inclusive representation (AIR). The rationale behind this notion is straightforward. The more people from different ethnicity are included, the more likely it is that one of them will provide an authentic representation for that ethnicity.

Moreover, others argue that diversity encompasses a variety of perspectives and experiences, which can enhance authenticity of characters. This is because diverse casts can help to challenge stereotypes and assumptions about different cultures. As a result, they can help to create more nuanced and complex representations of people from different backgrounds (Smith, 2020).

Furthermore, there is a growing body of research that suggests that authentic representation of diverse individuals on screen, free from stereotypes, has a profound educational and socially engaging impact on audiences (Bamford, 2018). This transfers to all types of media, as shown by a study by Roberts, (2021) which found that news stories that portrayed diverse cultures and identities in an authentic way were more likely to foster empathy and understanding among readers. This is likely because seeing oneself and one's own experiences reflected in the media can lead to a deeper emotional connection with the characters and the story, making the media more engaging. Therefore, the hypothesis will be the following:

**H2:** The relationship between racial diversity and LTAE is mediated through AIR.

2.5 Conceptual framework

A white rectangular sign with black text

Description automatically generated

Figure 1 Conceptual model paper

Figure 1 indicates the conceptual model consists of three relationships. AIR being the mediator of the relationship between diversity and LTAE. As shown in the figure all relationships should be positive. These relationships will be examined through regressions models. Thorough model specification can be found in the research methodology.

3. Research Methodology

This study aims to investigate the relationship between racial diversity and Long-Term Audience Engagement (LTAE) in films, mediated by Authentic and Inclusive Representation (AIR). AIR is defined as the creation of characters and narratives that genuinely reflect the experiences and perspectives of marginalized groups.

The following sections will provide a detailed overview of the data collection and sampling procedures, as well as a clear operationalization of the variables, including AIR. Notably, AIR will be discussed across three distinct levels, as defined by the Bechdel test, which was discussed in the literature.

3.1. Data sources

*Film characteristics*: I collected data on film characteristics from three trusted sources: IMDb, The Numbers, and The Movie Database (TMDB). These platforms have millions of registered users, ensuring credibility and diverse perspectives (Ghiassi, 2017).

*LTAE:* IMDB will be used to measure LTAE. The dependent variable of this study is rooted in IMDb user behavior which would be clarified in the operationalization. To ensure the generalizability of the findings, I compared demographic data on IMDb users to filmgoers.

While IMDb users and the general moviegoing population differ somewhat[[1]](#footnote-1), previous research suggests that IMDb's demographics are representative of the broader filmgoing audience (Ghiassi, 2017; Partha and Chakraborty, 2019; Apala, 2013). Therefore, given its large user base and diverse perspectives, I assumed that IMDb data is reliable for producing generalizable and accurate findings.

*Ethnicity determination:* I used the Kairos API, a deep learning algorithm that can detect ethnicity through facial recognition, to determine the ethnicity of actors and actresses in the films that were studied. I collected the profile pictures of the actors and actresses from IMDb.

*AIR determination*: This study builds on previous work on AIR and implements the Bechdel test. The Bechdel test is an ideal tool for measuring AIR in films because of its adaptability to our specific focus (Lazar, 2020), its potential for automation, and its inherent quantifiability (Argarwal, 2015). Furthermore, the test comprises three gradations seen below. Offering a nuanced assessment of representation by examining AIR from various angles:

(T1) The film must have at least two named female characters.

(T2) The two female characters must talk to each other.

(T3) The conversation between the two female characters must not be about a man.

Traditionally, screenplays have been used to conduct the Bechdel test to determine the level of AIR in films (Argawal, 2015). However, film scripts can be changed during production, so they may not match the final film. Therefore, I decided to take a different approach and use subtitles for the hearing impaired. Subtitles represent the final version of the film's dialogue, capturing it exactly as it appears in the film. By using subtitles.org I was able to find subtitles for half of the films which were in a format that could be standardized for testing. I will discuss subtitle processing in the operationalization part.

3.2 Sample

The films for the sample were chosen carefully and based on specific criteria. IMDB started in 1996 so the sample was limited to films released after 1996. Nevertheless, The steep increase in ranking which can be seen in Figure 4 is not due to films becoming less popular over time, but rather to the fact that IMDb was not widely used between 1998 and 2000. Therefore, there was am additional filter for films from those years.

Films produced outside the United States and animated films were also filtered out, because Hollywood is the focus of this study and voice actors are not represented on screen and this undermines representation. Following Joshi and Mao (2012), the analysis only included films that received a wide release, requiring a minimum of 500 screens at their launch. Observations with missing data were also eliminated.

After applying the criteria of release date (after 2000), production country (United States), film type (live-action), and wide release (500 screens), the final sample consisted of 2,315 unique films.

A graph with a line going up

Description automatically generated

Figure 4 Average popularity per year

3.3: Variable operationalization

*3.3.1 LTAE (DV):* Long term audience engagement was operationalized using IMDb's MovieMeter, a metric derived from popularity rankings. The score means that film was the most popular with regards to clicks, page views and reviews on IMDB. Therefore, this film score includes direct indicators of LTAE, such as online discussions, reviews, and word-of-mouth conversations. By examining film scores over time, I aimed to measure sustained engagement, where lower scores indicate higher popularity.

To measure popularity, I used the average popularity rankings of a film over a one-year period, starting in the third year after its release. I chose this metric because it shows how engaged audiences are with a film after a few years, and the average reduces the influence of any spikes or certain drops in popularity. Figure 4 shows the average LTAE scores for films released in different years.

That the average popularity after three years decreases overtime is rather logical. Every year when films are released they have to compete with more films that are previously released. The only explanation to the fast increase in ranking (meaning a decrease in popularity) between 2011-2016 I believe is due to the entrance of online streaming services. Due to this surpassing the focus of this paper I have provided the argumentation in the footnote. [[2]](#footnote-2)

*3.3.2 Racial diversity (IV)*:

*Ethnicity determination* The Kairos API is chosen because of its efficiency, accuracy (99.63%), and ability to handle a large dataset (Kairos, 2023). In the films a total of 121,898 characters were present. This count includes instances where an actor or actress appeared in multiple films. It also included uncredited people (13.491). The Kairos API analyzed 28,564 images found on IMDb of the total 39,400 unique actors and actresses. [[3]](#footnote-3)

In the few cases where all actors were used for an analysis, such as for robustness checks, the ethnicities of the imageless people were based on their first and last name using the R package Rethnicity, which has an accuracy of around 80%.

Using the Kairos API, I extracted the ethnicity of each actor or actress in the dataset. The API had one challenge: it provides probabilities rather than an assigent that a person belongs to one of the ethnicities. Moreover, the dataset was skewed towards White actors. Figure 5 shows the distribution of probabilities for the different ethnicities, which is unequal. This is due to the abundance of White actors in the dataset, which creates a lot of low probability values for the other three ethnic groups. Nevertheless, the confidence associated with the different ethnicities did not differ (Asian: 0.9972, Black: 0.9988, Hispanic: 0.9990, White: 0.9985). Therefore, I established a process to assign ethnicity values to the actors and actresses

* A probability exceeding 0.6 led to an ethnicity assignment value of 1.

Using this threshold, I was able to successfully identify the ethnicity of 98,905 characters. For the remaining actors and actresses I took an additional step for ethnicity assignment.[[4]](#footnote-4)

A graph of different sizes and shapes

Description automatically generated with medium confidence

Figure 5 Distribution probabilities ethnicities

*Inverse Simpson diversity index:* After determining the ethnicities of individuals, following previous research I assessed the racial diversity of the entire cast. To measure racial diversity, I used the inverse Simpson index (1949). The Simpson diversity index is a measure of diversity that takes into account the number of ethnicities present, as well as the relative abundance of each ethnicity. It is calculated using the following formula:

D = 1 - / (Σ *n*i(*n*i-1)/N(N-1))

Where: *n* is the number of actresses and actors that belong to an ethnicity i and N is the total number of actresses and actors.

*3.3.3 Mediator AIR*

The Bechdel test was modified to measure AIR for three different ethnicities. This approach enables to look at AIR for the three different ethnicities. It will be dichotomous variables for all three different levels. I will refer to the specific Bechdel test for this study further as the reformed Bechdel test:

(T1) two named characters of {ethnicity} ethnicity appear in film X.

(T2) who speak to one another.

(T3) about something besides a white character.

*(T1) Non named character filtering:* The first step in the reformed Bechdel test is to filter out non-named characters. To filter generic characters from the dataset, I identified frequently occurring tokens, where a token is a segment of a name divided by spaces. I then removed these tokens using a stop word list. The stop word list eventually contained 915 words, such as "doctor," "agent," and "the." The list can be seen in the Appendix.

I made an exception to the stop word filter: I did not remove a character if the stop word was the first token in their name. For example, "Colonel Rich Bron" was kept. However, there was an addition to this exception which was that if the character's name consisted of only one word. For example, a character named "Colonel" would still be removed.[[5]](#footnote-5)

Table 2 shows that when characters from a particular ethnicity are required to be named (T1), the number of films with two or more characters is noticeably lower. Films that exist in the first column (featuring two or more characters from the same ethnicity) but not in the second column (where at least two of these characters are named) can be characterized as engaging in tokenism. These films are including characters of color without giving them proper names, which could be an indicator of tokenism, where superficial diversity is prioritized over meaningful and authentic representation.

|  |  |  |
| --- | --- | --- |
| Ethnicity | Two characters all cast | Two named characters |
| Black | 1857 | 982 |
| Asian | 1471 | 568 |
| Hispanic | 1508 | 531 |

Table 1 (T1) all characters versus named characters

To delve deeper into the link between inclusive characters and racial diversity, I implemented steps 2 and 3 of the revised Bechdel test using subtitle files. While detailed processing steps are available in Appendix A, the following section outlines an overview of this procedure.

*(T2 ,T3) Subtitle file parching:* To assess who speaks to one another (T2) and whether they talk about something besides a white character (T3), I used subtitle files for the hearing impaired. These files can be extracted from a film in the form of .srt files, which are text files with strict formatting. Each subtitle in an .srt file has a unique identifier, precise start and end times, and one or two lines of text. For example, here is the opening of the film "300: Rise of an Empire": 1

00:00:38,363 --> 00:00:40,698

(HORSE NICKERS)

2

00:01:02,654 --> 00:01:07,024

QUEEN GORGO: The oracle's

words stand as a warning.

3

00:01:07,026 --> 00:01:08,225

A prophecy.

4

00:01:08,227 --> 00:01:11,796

"Sparta will fall.

Consequently, my dataset now consisted of scene indices, speakers, individuals mentioned, and the ethnicities of the characters. The final stage involved applying the revised Bechdel Test to this dataset. Films were then categorized based on whether they met the conditions (T2) and (T3) for different ethnicities. This categorization serves as the foundation for regression models.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ethnicity | AC (T1) | NC (T1) | AC (T2) | NC (T2) | AC (T3) | NC (T3) |
| Black | 901 | 503 (0.53) | 804 | 260 (0.32) | 445 | 128 (0.29) |
| Asian | 748 | 334 (0.39) | 561 | 108 (0.20) | 205 | 142 (0.69) |
| Hispanic | 751 | 326 (0.35) | 602 | 156 (0.26) | 254 | 191 (0.75) |

Table 3 shows the different step for the reformed Bechdel test. AC being All characters and NC meaning named characters. The interesting to see is that the percentage of films passing T3 with named characters related to the film passing the test if all characters are considered is proportionally higher than for T1 expect for the Black ethnicity. This means that when people from an ethnicity are added they are mostly talking to each other. This is confirmed with comparing T3 with T2. What can be seen is that for Hispanic and Asian T3 is higher in obsolete values than T2. This means it occurs more often that people from the same ethnicity talk to each other without a white person than with a white person. This shows some type of segregation in the films.

Moreover, the difference between T1 all , T1 named and T3 all and T3 named shows shows that in the cases that if a conversation happened for all the characters this conversation also trigger it for named characters. Meaning that this conversation was more likely to be held by named characters. This means that when proceeding in the step from characters talk to each other there are also more likely to be named. Providing support that named characters more often proceed in conversations with each other. Meaning also some support for using the Reversed BEchdel Test as measurement.

In more than 60% of the films that have a conversation meeting the criteria it is held by two people who are named. This is quite substantial looking at that named characters only exist for parts of the cast.

3.4 Covariates:

It is important to account for additional factors that have been identified as influencing a film’s success, IN doing so this study draws upon previous research. The variables considered in this study include Series the star power of Actors and Directors, MPAA rating, Number of screens, Critical Acclaim, Awards, Budget, Genre, Source and Seasonality. Table 3 provides detailed information which sources identified which variables the following paragraphs briefly discuss how the variables are measured in this study.

In this study, (SERIESi) represents the number of sequels associated with a film. The star power score (STARPOWERi) is derived from Nelson and Glotfelty's (2012) measure, based on the four highest-grossing actors' ranking on The Numbers. (DIRECTORPOWERi) is similarly measured using the highest-grossing directors' ranking on The Numbers. The (CRITICSi) value is the average rating on metacritic.com

This research will use the actual number of award (NOMINATIONSi) as a proxy for award nomination. (WINSi) will also be added to the model to represent awards wins. Because there is an abundance of available film awards this study uses the awards used in the research of Gemser, Leenders and Wijnberg (2008). [[6]](#footnote-6)

(MPAAi) rating is given by the Motion Picture Association of America and is used to rate a film's suitability for certain audiences based on its content. These ratings are encoded as an interval variable as [0 = unrated; 1 = G; 2 = PG; 3 = PG-13; 4 = R;5 = C-17].

(SCREENSi) is the amount of opening theaters the film had according to the Numbers. (BUDGETi) was the production budget available at one of the three data sources used, if multiple production multiple budgets where available across the sources the average was taken.

Moreover, this study introduces 19 genres , Action, Adventure, Comedy, Crime, Drama, Family, Fantasy, Horror, Romance, Musical, Sci-Fi, Mystery, Thriller, Western, Biography, Documentary, History, Music, Sport and War. Because a film could have multiple genres, these dummy variables are not mutually exclusive.

(SPRINGi, SUMMERi, FALLi, WINTERi) Within this study the four seasons are encoded as the following. Spring[March, April, May] Summer [June, July, August] Fall [September, October, November], Winter [December, January, February]. (RUNTIMEi) is included as the actual numerical value in minutes, following (Holbrook, 1999).

Furthermore, following Hofmann, Clement, Völckner, and Hennig-Thurau (2017), multiple dummy variables were added to control for whether the film was ({BASED ON}i) a book, comic, novel, short story, or TV series. Moreover, whether the film is a (REMAKEi) or (SPINOFFi) Similar, to the genre variables these dummy variables are not mutually exclusive because a film could be based on multiple sources. To control for possible differences over time, (YEARi) dummies for every year in the sample were added to the model.

Because log-log linear regressions are used in this research, all control variables that are not dummy variables were log-transformed. Most of these variables had zero values in this situation a small number (0.01) was added.

Finally, it is crucial for this study to be cautious in every step of the process to possibly achieve causality. Therefore, every factor previously mentioned will also be analyzed on its influence on the racial diversity of the film before it is included as a control variable.[[7]](#footnote-7)

Table 3: Measures of Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Description | Measure | Data Source |
| Simpson diversity indexi | Measure of diversity which takes into account the number of species present, as well as the relative abundance of each species. | Continuous value between 0-1. 1 being complete diversity. | IMDB, TMDB , Kairos |
| AIRi | Authentically Inclusive Representation. | Dichotomous variable with three separate levels | Subtitles.org |
| Log(LTAEi) | IMDB popularity score where a higher score means being less popular. | Average ranking over a one year period of time in the third year after release logged. | IMDB |
| Covariates |  |  |  |
| SERIESi | Number of films associated with the film |  | IMDB |
| Log(STARPOWERi) | Log-transformed score of the top three actors in the film + 0.01. | Log-transformed sum of the top 4 actors, the year of the film release. | The Numbers |
| Log(DIRECTORPOWERi) | Log-transformed score of the number one director in the film + 0.01. | Log-transformed sum of 100 divided through the director ranking + 0.01 (e.g., log(100/rank 2 +0.00)) for the director one year before the film release. | The Numbers |
| Log(CRITICi) | Log-transformed average rating the film received from professional film reviewers + 0.01. | Log-transformed average rating the film received from professional film reviewers | Metacritic |
| Log(NOMINATIONSi) | Number of award nominations the film received. | Log transformed number of award nominations the film received + 0.01. | IMDb |
| Log(WINSi) | Number of award wins the film received. | Log transformed number of award nominations the film received + 0.01. | IMDB |
| Log(MPAAi) | MPAA rating, the film received 1 = Not Rated, 2 = G, 3 = PG, 4 = PG-13, 5 = R, 6 = NC-17 | Log transferred of the interval ranking | IMDb |
| Log(SCREENSi) | Number of screens at release | Total amount of screens logged | Numbers |
| Log(BUDGETi) | Production budget of film | The average production budget of the data sources | IMDB, Numbers, OMDB |
| ACTIONi | Genre Action | Genre Action = 1, Other = 0 | IMDb |
| ADVENTUREi | Genre Adventure | Genre Adventure = 1, Other = 0 | IMDb |
| ANIMATIONi | Genre Animation | Genre Animation = 1, Other = 0 | IMDb |
| COMEDYi | Genre Comedy | Genre Comedy = 1, Other = 0 | IMDb |
| CRIMEi | Genre Crime | Genre Crime = 1, Other = 0 | IMDb |
| DRAMAi | Genre Drama | Genre Drama = 1, Other = 0 | IMDb |
| FAMILYi | Genre Family | Genre Family = 1, Other = 0 | IMDb |
| FANTASYi | Genre Fantasy | Genre Fantasy = 1, Other = 0 | IMDb |
| HORRORi | Genre Horror | Genre Horror = 1, Other = 0 | IMDb |
| MUSICALi | Genre Musical | Genre Musical = 1, Other = 0 | IMDb |
| MYSTERYi | Genre Mystery | Genre Mystery = 1, Other = 0 | IMDb |
| ROMANCEi | Genre Romance | Genre Romance = 1, Other = 0 | IMDb |
| SCI-FIi | Genre Sci-Fi | Genre Sci-Fi = 1, Other = 0 | IMDb |
| THRILLERi | Genre Thriller | Genre Thriller = 1, Other = 0 | IMDb |
| WESTERNi | Genre Western | Genre Western = 1, Other = 0 | IMDb |
| BIOGRAPHYi | Genre Biography | Genre Biography = 1, Other = 0 | IMDb |
| DOCUMENTARYi | Genre Documentary | Genre Documentary = 1, Other = 0 | IMDb |
| MUSICi | Genre Music | Genre Music = 1, Other = 0 | IMDb |
| HISTORYi | Genre History | Genre History = 1, Other = 0 | IMDb |
| SPORTi | Genre Sport | Genre Sport = 1, Other = 0 | IMDb |
| SPRINGi | Film released in the spring | Spring release = 1, Other = 0 [Mar, Apr, May] | IMDb |
| SUMMERi | Film released in the summer | Summer release = 1, Other = 0 [Jun, Jul, Aug] | IMDb |
| AUTUMi | Film released in the autumn | Autum release = 1, Other = 0 [Sep, Oct, Nov] | IMDb |
| WINTERi | Film released in the winter | Winter release = 1, Other = 0 [Dec, Jan, Feb] | IMDb |
| Log(RUNTIMEi) | Duration of a film in minutes | Log-transformed duration of a film in minutes | IMDb |
| BOOKi | Film is based on a book | Film is based on a book = 1, other = 0 | IMDb |
| COMICi | Film is based on a comic | Film is based on a comic = 1, other = 0 | IMDb |
| NOVELi | Film is based on a novel | Film is based on a novel = 1, other = 0 | IMDb |
| SHORTSTORi | Film is based on a short story | Film is based on a short story = 1, other = 0 | IMDb |
| TVSERIESi | Film is based on TV seriess | Film is based on a TV series = 1, other = 0 | IMDb |
| REMAKEi | Film is a remake | Film is a remake = 1, other = 0 | IMDb |
| SPINOFFi | Film is a spinoff | Film is a spinoff = 1, other = 0 | IMDb |
| YEAR2000i | Year2000 | 2000 = 1, other = 0 | IMDb |
| YEAR2001i | Year2001 | 2001 = 1, other = 0 | IMDb |
| YEAR2002i | Year2002 | 2002 = 1, other = 0 | IMDb |
| YEAR2003i | Year2003 | 2003 = 1, other = 0 | IMDb |
| YEAR2004i | Year2004 | 2004 = 1, other = 0 | IMDb |
| YEAR2005i | Year2005 | 2005 = 1, other = 0 | IMDb |
| YEAR2006i | Year2006 | 2006 = 1, other = 0 | IMDb |
| YEAR2007i | Year2007 | 2007 = 1, other = 0 | IMDb |
| YEAR2008i | Year2008 | 2008 = 1, other = 0 | IMDb |
| YEAR2009i | Year2009 | 2009 = 1, other = 0 | IMDb |
| YEAR2010i | Year2010 | 2010 = 1, other = 0 | IMDb |
| YEAR2011i | Year2011 | 2011 = 1, other = 0 | IMDb |
| YEAR2012i | Year2012 | 2012 = 1, other = 0 | IMDb |
| YEAR2013i | Year2013 | 2013 = 1, other = 0 | IMDb |
| YEAR2014i | Year2014 | 2014 = 1, other = 0 | IMDb |
| YEAR2015i | Year2015 | 2015 = 1, other = 0 | IMDb |
| YEAR2016i | Year2016 | 2016 = 1, other = 0 | IMDb |
| YEAR2017i | Year2017 | 2017 = 1, other = 0 | IMDb |
| YEAR2018i | Year2018 | 2018 = 1, other = 0 | IMDb |
| YEAR2019i | Year2019 | 2019 = 1, other = 0 | IMDb |

3.5 Models:

Following the approach outlined by Clement, Wu, and Fischer (2014), this research employs log-log linear regressions. However, it is worth noting that numerous continuous variables had zero values, and taking the logarithm of '0' would result in an error. To address this, a small constant value of '+0.001' was added to all films characters which were continuous before taking the logarithm, ensuring meaningful results.

The mediation analysis consists of two models. First the path between AIR and LTAE involves a linear regression. Since AIR is a dichotomous variable, the model relating racial diversity and AIR follows a logistic regression framework.

The Simpson index, being a number between 0 and 1, will it simplifies the interpretation of the results if it will not be logged. The detailed model specifications are provided below.

Main relationship:

log(LTAEi) = β0 + β1 × SIMPSONDIVERSITYINDEXi + β2 × SERIESi +

β3 × log(STARPOWERi) + β4 × log(DIRECTORPOWERi) +

β5 × log(CRITICi) + β6 × log(NOMINATIONSi) + β7 × log(WINSi) +

β8 × log(MPAAi) + β9 × log(SCREENSi) + β10 × log(BUDGETi) +

β11 × ACTIONi + β12 × ADVENTUREi + β13 × ANIMATIONi +

β14 × COMEDYi + β15 × CRIMEi + β16 × DRAMAi + β17 × FAMILYi +

β18 × FANTASYi + β19 × HORRORi + β20 × MUSICALi +

β21 × MYSTERYi + β22 × ROMANCEi + β23 × SCI-FIi +

β24 × THRILLERi + β25 × WESTERNi + β26 × BIOGRAPHYi +

β27 × DOCUMENTARYi + β28 × MUSICi + β29 × HISTORYi +

β30 × SPORTi + β31 × SPRINGi + β32 × SUMMERi +

β33 × AUTUMNi + β34 × WINTERi + β35 × log(RUNTIMEi) +

β36 × BOOKi + β37 × COMICi + β38 × NOVELi + β39 × SHORTSTORi

β40 x TVSERIESi  + β41 x REMAKEi + SPINOFFi + β42 ×YEAR2000i +

β43 × YEAR2001i  + β44 × YEAR2002i + β45 x YEAR2003i +

β46 × YEAR2004i + β47 × YEAR2005i + β48 × YEAR2006i +

β49 × YEAR2007i + β50 × YEAR2008i + β51 × YEAR2009i +

β52 × YEAR2010i + β53 × YEAR2011i + β54 × YEAR2012i +

β55 × YEAR2013i + β56 × YEAR2014i + β57 × YEAR2015i +

β58 × YEAR2016i + β59 × YEAR2017i + β60 × YEAR2018i +

β61 × YEAR2019i + εi

Mediation models:

Path A:  
AIRi = β0 + β1 × SIMPSONDIVERSITYINDEXi + β2 × SERIESi + β3 × log(STARPOWERi) + β4 × log(DIRECTORPOWERi) + … β61× YEAR2018i + β62 × YEAR2019i + εi

Path B:

log(LTAEi) = β0 + β1 × AIRi + β2 × SERIESi + β3 × log(STARPOWERi) + β4 × log(DIRECTORPOWERi) +… β60× YEAR2018i + β61 × YEAR2019i + εi

4. Results

In this section, we begin by presenting the research findings. However, prior to delving into the results, we first provide descriptive and frequency statistics for the variables utilized in the regressions. Subsequently, we assess the assumptions for log-log linear regressions, followed by the estimation of the regression models and a check for multicollinearity and correlations. Finally, we interpret the results of the log-log linear regression models and perform robustness checks.

Descriptive Statistics  
Table 2 presents descriptive statistics for the continuous variables before being logged. This table displays the number of observations per variable (N), as well as the mean, standard deviation (SD), minimum, and maximum values.

Table 3: Descriptive statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable\_Name** | **N** | **Mean** | **SD** | **Minimum** | **Maximum** |
| *Dependent variable* |  |  |  |  |  |
| Avg Rank Third Year | 2360 | 3410 | 4387 | 75.92 | 103633 |
| *Independent variable* |  |  |  |  |  |
| Inverse Simpson index | 2360 | 0.41 | 0.15 | 0 | 1 |
| *Control Variables* |  |  |  |  |  |
| Opening screen | 2360 | 2650.06 | 826.48 | 502 | 4662 |
| Runtime | 2360 | 108.69 | 17.19 | 74 | 219 |
| Budget | 2360 | 49,442,970 | 48,828,780 | 12 | 356,000,000 |
| Nominee | 2360 | 14.10 | 32.37 | 0 | 462 |
| Winner | 2360 | 5.39 | 20.62 | 0 | 490 |
| Director power | 2360 | 0.03 | 0.09 | 0 | 1.0 |
| Total Star Power | 2360 | 0.06 | 0.15 | 0 | 1.5 |
| Metascore | 2360 | 47.66 | 16.46 | 1 | 96 |
| Sequal |  |  |  |  |  |

Following this, Table 4, Table 5 and Table 6 provide the frequency statistics for the dummy variables for year, genre, seasonality and the based on varibales. The table displays the number of observations per variable (N), and for the dummy variables, it showcases the count of occurrences where the dummy variable yielded a '1' as its outcome.

Table 4: Year dummies

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Year** | **N** | **Percentage** |  | **Year** | **N** | **Percentage** |
| 2000 | 117 | 4.82 |  | 2010 | 119 | 4.91 |
| 2001 | 122 | 5.03 |  | 2011 | 121 | 4.99 |
| 2002 | 120 | 4.95 |  | 2012 | 106 | 4.37 |
| 2003 | 116 | 4.78 |  | 2013 | 109 | 4.49 |
| 2004 | 127 | 5.24 |  | 2014 | 118 | 4.87 |
| 2005 | 134 | 5.53 |  | 2015 | 110 | 4.54 |
| 2006 | 143 | 5.90 |  | 2016 | 134 | 5.53 |
| 2007 | 133 | 5.48 |  | 2017 | 104 | 4.29 |
| 2008 | 130 | 5.36 |  | 2018 | 132 | 5.44 |
| 2009 | 125 | 5.15 |  | 2019 | 105 | 4.33 |

Table 5: Genre dummies

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Genre** | **N** | **Percentage** |  | **Genre** | **N** | **Percentage** |
| Drama | 1064 | 43.88 |  | Fantasy | 325 | 13.40 |
| Comedy | 949 | 39.13 |  | Family | 246 | 10.14 |
| Thriller | 863 | 35.59 |  | Biography | 146 | 6.02 |
| Action | 801 | 33.03 |  | Sport | 119 | 4.91 |
| Romance | 518 | 21.36 |  | Music | 102 | 4.21 |
| Adventure | 495 | 20.41 |  | History | 91 | 3.75 |
| Crime | 465 | 19.18 |  | War | 87 | 3.59 |
| Sci.Fi | 374 | 15.42 |  | Musical | 32 | 1.32 |
| Mystery | 349 | 14.39 |  | Documentary | 32 | 1.32 |
| Horror | 341 | 14.06 |  |  |  |  |

Table 6 : Remaining dummies

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Season** | **N** | **Percentage** |  | **Variable** | **N** | **Percentage** |
| Fall | 649 | 26.76 |  | Based on novel | 387 | 15.96 |
| Summer | 616 | 25.40 |  | Based on comic book | 124 | 5.11 |
| Spring | 613 | 25.28 |  | Based on book | 99 | 4.08 |
| Winter | 547 | 22.56 |  | Based on comic | 63 | 2.60 |
|  |  |  |  | Spinoff | 52 | 2.14 |
|  |  |  |  | Based on short story | 23 | 0.95 |
|  |  |  |  | Based on play | 16 | 0.66 |

Assumptions model   
With a linear regression model there are some assumptions that need to be tested.

*Independence:* Independence of observations is a key assumption of regression analysis. It means that the residuals (the difference between the actual and predicted values of the dependent variable) should not be correlated with each other. In this study, we use cross-sectional data on movies, where each movie is represented by a single observation. This means that the observations are likely to be independent, and the assumption of independence is satisfied.

*Collinearity:* With regards to collinearity I made a correlation matrix and a and model to check the VIF’s values. With regards to the correlation matrix there was not a variable which was highly correlated to each other >0.5. Moreover, with regards to the VIF values there was not a value >10 indicating there is no multicollinearity in the model.

*Homoscedastic:* When running the log-log regression the homoscedastic assumption is not met. When doing a white test the result is < 2.2e-16 indicating there is heteroskedasticity. When heteroskedasticity is present, ordinary least squares (OLS) standard errors are not reliable, and can lead to incorrect inferences about the statistical significance of the regression coefficients.

Robust standard errors adjust the OLS standard errors to make them more robust to heteroskedasticity. This is done by using a different estimator of the variance of the errors. Robust standard errors are particularly useful when the assumption of homoskedasticity is violated, as is often the case in real-world data.

This method doesn't change the point estimates but provides more accurate standard errors and confidence intervals.

*Normality:* Non-normality of the residuals is considered unproblematic due to the substantial sample size and taking the log-log transformation model.

Interpreting the results

Table 3 shows the significant variables in the main analysis the relationship between racial diversity of the cast and LTAE the following are the results of the model. Only included are all significant variables. With regards to interpreting the results I will discuss first the result of the main analysis afterwards the mediation effects is discussed.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Coefficient** | **Robust.SE** | **Sign** |
| Inverse Simpson index | 0.268249 | 0.0854869 | < 0.001 \*\*\* |
|  |  |  |  |
| Log(Screens) | -0.4607796 | 0.0508223 | < 0.001 \*\*\* |
| Log(Runtime) | -1.3511415 | 0.1143878 | < 0.001 \*\*\* |
| Log(MPAA) | -0.7109994 | 0.0729009 | < 0.001 \*\*\* |
| Log(Average budget) | -0.0470878 | 0.0171892 | 0.008 \*\* |
| Log(Sequal) | 0.0058112 | 0.0014671 | < 0.001 \*\*\* |
| Based on book | -0.1316908 | 0.0615430 | 0.034 \* |
| Based on play | 0.3750325 | 0.1386266 | 0.012 \* |
| Based on comic book | -0.3142000 | 0.0785232 | < 0.001 \*\*\* |
| Based on novel | -0.1080797 | 0.0352643 | 0.002 \*\* |
| Log(Nominee) | -0.0091412 | 0.0016273 | < 0.001 \*\*\* |
| Log(Winner) | -0.0143761 | 0.0011203 | < 0.001 \*\*\* |
| Log(Director Power) | -0.0096619 | 0.0016879 | < 0.001 \*\*\* |
| Log(Metascore) | -0.0355427 | 0.0082175 | < 0.001 \*\*\* |
| Log(Starpower) | -0.0038060 | 0.0018459 | 0.032 \*\* |
| Adventure | -0.1351720 | 0.0358245 | < 0.001 \*\*\* |
| Comedy | 0.1301336 | 0.0383721 | < 0.001 \*\*\* |
| Drama | 0.1486707 | 0.0320064 | < 0.001 \*\*\* |
| Romance | -0.0847929 | 0.0367200 | 0.013 \* |
| Sci.Fi | -0.1089913 | 0.0372327 | 0.006 \*\* |
| Biography | 0.1629182 | 0.0625266 | 0.006 \*\* |
| Sport | 0.1355867 | 0.0636215 | 0.025 \* |
| War | 0.1428677 | 0.0705197 | 0.043 \* |
| Family | 0.2083408 | 0.0534861 | < 0.001 \*\*\* |
| Musical | -0.2922735 | 0.1308821 | 0.007 \*\* |
| History | 0.1588541 | 0.0690280 | 0.028 \* |
| Documentary | 0.7488378 | 0.1260843 | < 0.001 \*\*\* |
| Western | 0.4924241 | 0.1199375 | < 0.001 \*\*\* |
| 2009 | 0.1722148 | 0.0772681 | 0.023 \* |
| 2010 | 0.2008708 | 0.0787402 | 0.009 \*\* |
| 2011 | 0.3513528 | 0.0799067 | < 0.001 \*\*\* |
| 2012 | 0.5024574 | 0.0793373 | < 0.001 \*\*\* |
| 2013 | 0.6166931 | 0.0823798 | < 0.001 \*\*\* |
| 2014 | 0.7911409 | 0.0847740 | < 0.001 \*\*\* |
| 2015 | 0.7911298 | 0.0807096 | < 0.001 \*\*\* |
| 2016 | 0.8791387 | 0.0790009 | < 0.001 \*\*\* |
| 2017 | 0.8228709 | 0.0901873 | < 0.001 \*\*\* |
| 2018 | 1.0214914 | 0.0841829 | < 0.001 \*\*\* |
| 2019 | 0.9679212 | 0.0984391 | < 0.001 \*\*\* |

*Significance levels: \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001*

The results of the study are consistent with previous research, highlighting the importance of certain variables. While not all genres show significance, a few are noteworthy. One intriguing finding is that seasonality is insignificant, which is understandable given the nature of long-term engagement. The analysis spans a year in the third year after release, which mitigates the initial impact of seasonal influences, which tend to wane over time.

However, the most unexpected finding is the subtle but significant negative impact of racial diversity on popularity. In a linear regression model, a one-unit increase in the Simpson diversity index is associated with a 0.268249-unit decrease in the predicted average popularity ranking in the third year after release, holding other variables constant.

Logistic regression A path

Additional analysis

The last remark this study would like to make is that as explained previously, films and culture are deeply intertwined. In today’s society the following values are seen as highly important belonging, community, and personal growth, which are all closely related to representation. (Neufeld, 2020) It is proven that films that properly represent the cultural zeitgeist in which the film is made are more successful. (Ettema, 2005)

Therefore, it is reasonable to assume that films that reflect the values related to representation will be more successful in today's cultural landscape. However, as societal norms and values evolve, this appreciation might wonder off. This study aims to explore how much of a film's success is due to the effects caused by representation, and how much is due to the degree the films is displaying the current cultural zeitgeist which is focused on representation.

Finally, prior studies (Malik, 2022; Kuppuswamy, 2016) have overlooked cofounding variables and macro environmental factors, and have largely focused solely on the relationship between racial diversity and film success. This study introduces cultural resonance to the analysis, arguing that it is a crucial factor in the success of films with diverse casts, as it may influence how one's cultural values shape their perception of these films.

2.5 Cultural resonance

Cultural resonance, as discussed in this study, refers to how well a film connects with the cultural values, experiences, and zeitgeist of its audience. It can be achieved by either reflecting or challenging the audience's own culture, values, and experiences in a thought-provoking way. (Ettema, 2005) Representation is key, as films that represent different cultures and experiences can broaden audiences and create a sense of connection. (Bamford, 2018)

For the sake of clarity and illustration, the film Black Panther was praised for its positive portrayal of African culture and its representation of African people. It was released at a time when representation was a major topic of discussion, and its popularity amplified the conversation about race and representation. Additionally, its positive portrayal of African culture helped to promote African fashion. Thus, Black Panther is both a product of culture, parts of its success is caused by culture and it is a force for change in culture.

2.5.1 Culture as cofound/moderator

Culture has a big impact on how long audiences stay engaged with a film. Racially diverse films often explore important social and cultural issues, which can lead to lasting discussions about race, identity, and social justice. As Garrett (2020) points out, "Films about race offer us the chance to grapple with past and present constructs of racism, power, and oppression." This shows that cultural shifts are driving conversations about racial diversity, which are essential for keeping audiences engaged in the long term. This underscores the idea that when films activate these conversations they will remain more under the attention for longer periods of time.

*Brokeback Mountain, https://www.tandfonline.com/doi/full/10.1080/19359705.2013.792128*

Moreover, contemporary film critics place a high value on representation and inclusion in media. (Akser, 2021) As a result, films featuring racially diverse casts tend to receive more favorable critical reception, which in itself contributes to LTAE. Finally, cinema's transformative capacity on culture extends to its influence on the hiring practices within the industry, resulting in increased racial diversity in films. (Khrebtan-Hörhager, 2011) With regards to the above considerations, the following hypothesis is proposed:

**H4**: The relationship between racial diversity and LTAE is cofounded by culture resonance.

.2.3 Study 3: Cultural resonance as cofounding variable

Oscars so white   
https://www.tandfonline.com/doi/full/10.1080/15205436.2017.1409356?src=recsys

*Variable operationalization* : for the relationship studied a diff in diff analysis was seen as the most fitting. Quantifying whenever racial representation started to become a major topic in today’s culture. Eventually the event #oscarssowhite was chosen as natural schock. n the context of a difference-in-differences (DiD) analysis, a "natural shock" refers to an external event or occurrence that affects one group or entity being studied but not another group, and it is not caused by any deliberate intervention or treatment by the researchers.

*Model* :

# Bibliography

Agarwal, A. &. (2015). Key Female Characters in Film Have More to Talk About Besides Men: Automating the Bechdel Test.

Akser, M. (2021). Diversity and Inclusion in Film, Television and Media Sector: Policy Alternatives for an Inclusive Film Industry and Training. *CINEJ Cinema Journal, 9*(1), 1-13. doi:10.5195/cinej.2021.414

Alexander Dhoest, N. S. (2015). Still ‘Watching’ TV? The Consumption of TV Fiction by. *Media and Communication*, 176-184. doi:10.17645/mac.v4i3.427

Association., M. P. (2021). Theatrical Market Statistics. Retrieved from https://www.motionpictures.org/wp-content/uploads/2022/03/MPA-2021-THEME-Report-FINAL.pdf

Bamford, N. (2018). The Strangers’ Case: harnessing the power of screen entertainment to. *In: Cross-Cultural Communication Conference*. Retrieved from https://eprints.bournemouth.ac.uk/30284/

Basuroy, S. C. (2003). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. *Journal of Marketing*, 103–117. Retrieved from http://www.jstor.org/stable/30040552

Belton, J. (1995). *Films and Mass Culture.* The Athlone Press. Retrieved from https://www.goodreads.com/book/show/1639322.Films\_and\_Mass\_Culture

Benshoff, H. M. (2016). America on Film: Representing Race, Class, Gender, and Sexuality at the Films.

Boboltz, S. &. (2017). Why On-Screen Representation Actually Matters.

Brad J. Bushman, P., & L. Rowell Huesmann, P. (2006). Short-term and Long-term Effects of Violent Media. *Personality and Social Psychology*.

Broersma, M. (2019). *Audience Engagement. .* doi:10.1002/9781118841570.iejs0060.

Buckingham, D. (2008). *Youth , identity and digital media.* MIT Press, Cambridge, Mass. Retrieved from https://library.oapen.org/bitstream/handle/20.500.12657/26085/1/1004001.pdf

Bunche, R. J. (2018). "The Business Case for Diversity in the Entertainment Industry". Retrieved from https://bunchecenterold.pre.ss.ucla.edu/2018/02/28/new-hollywood-diversity-report-2018/

Castañeda, M. (2015). The power of (mis)representation: Why racial and ethnic stereotypes in the media matter.

Dan Rubin, I. M. (2022). *Beyond the box office: A conceptual framework for the drivers of audience engagement,* (Vol. 151). Journal of Business Research.

Dittmar, L. &. (2016). Racial and gender biases in magazine advertising persisted, and in some cases increased, from 1983 to 2014. *Marketing Science*(35), 911-926.

Dixon, T. L. (2000). Overrepresentation and underrepresentation of African Americans and Latinos as lawbreakers on television news. *Journal of Communication,, 50*, 131-154. doi:https://doi.org/10.1111/j.1460-2466.2000.tb02866.x

Einav, L. (2007). Seasonality in the U.S. Motion Picture Industry. *The RAND Journal of Economics, 38*, 127–145. doi:http://www.jstor.org/stable/25046296

Elberse, A. &. (2003). Demand and Supply Dynamics for Sequentially Released Products in International Markets: The Case of Motion Pictures. *Marketing Science., 22*, 329-354. doi:10.1287/mksc.22.3.329.17740.

Eliashberg, J. H. (2014). Assessing Box Office Performance Using Film.

Ettema, J. S. (2005). Crafting cultural resonance: Imaginative power in everyday journalism. *Journalism , 6*(2). doi:https://doi.org/10.1177/1464884905051005

Foundation, T. A. (2018). Inclusion in the Director’s Chair? Gender, Race & Age of Directors across 1,000 films from 2007-2017. Retrieved from https://www.bing.com/ck/a?!&&p=6955b1041ed8beadJmltdHM9MTY4MzY3NjgwMCZpZ3VpZD0wZTJmYjE3ZS04MWM3LTZiOWMtMWU2OC1hMzE5ODA0NDZhOGMmaW5zaWQ9NTE5Ng&ptn=3&hsh=3&fclid=0e2fb17e-81c7-6b9c-1e68-a31980446a8c&psq=Inclusion+in+the+Director%e2%80%99s+Chair%3f+Gender%2c

Garrett, G. (2020). *A Long, Long Way.* Oxford University Press. Retrieved from https://global.oup.com/academic/product/a-long-long-way-9780190906252?cc=nl&lang=en&

Gemser, G. V. (2007). The impact of film reviews on the box office performance of art house versus mainstream motion pictures. *Journal of Cultural Economics*, 43-63.

Ghiassi, M. S. (2017). Predicting Film Success Based on IMDB Data. *Journal of Big Data*, 1-14. doi:https://doi.org/10.1186/s40537-017-0072-4

Govindaraju, W. W. (2010). Explaining Global Box-Office Tastes in Hollywood Films: Homogenization of National Audiences’ Film Selections. *Communication Research, 37*, 215 - 238.

Hall, A. E. (2022). Audience Responses to Diverse Superheroes: The Roles of Gender and Race in Forging Connections With Media Characters in Superhero Franchise Films. *Psychology of Aesthetics, Creativity, and the Arts, 16*, 414–425. Retrieved from Psychology of Aesthetics, Creativity, and the Arts

Hunter, S. D. (2016), June 1). Predicting Box Office from the Screenplay: A Text Analytical Approach .

Initiative., A. I. (2018). Inclusion in Netflix scripted series and films. Retrieved from https://annenberg.usc.edu/sites/default/files/2019/02/11/MDSCI\_Inclusion\_in\_Netflix\_2.11.19.pdf

Jackson, L. A. (1995). Stereotypes, Emotions, Behavior, and.

Katherine Aumer, D. B. (2017). Assessing Racial Preferences in Films: The Impact of Mere-Exposure and Social Identity Theory. Retrieved from https://www.scirp.org/pdf/PSYCH\_2017071814195150.pdf

Khrebtan-Hörhager, J. (2011). The Art of Creating a Cross-Cultural Dialogue on Screen: "The Way Home "& "The Colour of Fear". *The International Journal of the Arts in the Society*. Retrieved from https://www.academia.edu/29706441/The\_Art\_of\_Creating\_a\_Cross\_cultural\_Dialogue\_on\_Screen\_The\_Way\_Home\_and\_The\_Colour\_of\_Fear\_

Kidd, M. A. (2015). Archetypes, stereotypes and media representation in a multi-cultural.

King, J. &. (2020). Representing race: the race spectrum subjectivity of diversity in film. *Ethnic and Racial Studies, 44*, 1-18. doi:10.1080/01419870.2020.1740290.

Kohl, B. R. (1989). Predicting financial success of motion pictures: The '80s experience. *Journal of Media Economics}, 2*(2), 35-50. doi:10.1080/08997768909358184

Kubrak, T. (2020, May 2). Impact of Films: Changes in Young People's Attitudes after Watching a Film. *Behav Sci (Basel)*. doi:10.3390/bs10050086.

Kumar, D. I. (2022). Beyond the box office: A conceptual framework for the drivers of audience engagement. *Journal of Business Research*.

Kuppuswamy, V. &. (2016). Testing the theory of consumer discrimination as an explanation for the lack of minority hiring in Hollywood films. Retrieved from https://ssrn.com/abstract=2738728

Kwon, O. L. (2020). Film Success Prediction using Historical and Current Data Mining. *Advances in Intelligent Systems and Computing*, 491-498. doi:https://doi.org/10.1007/978-981-15-5566-4\_59

Lazar, L. H. (2020). Driving box office performance through authentically inclusive storytelling. Retrieved from https://www.fullstoryinitiative.com/Full\_Story\_Research.pdf

Lemmens, J. V. (2011). Psychosocial causes and consequences of pathological gaming. Computers in Human Behavior. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0747563210002116

Leonard, D. J. (2006). *Screens Fade to Black:.* Praeger Publishers. Retrieved from https://cultureinjection.files.wordpress.com/2020/07/david-j.-leonard-screens-fade-to-black-contemporary-african-american-cinema.pdf

Litman, B. (1983). Predicting Success of Theatrical Films: An Empirical Study. *The Journal of Popular Culture}, 16*, 159-175.

McKinsey. (2021). The bottom line: films with more diverse casts perform better at the box office.

Mestyán, M. Y. (2013). Early Prediction of Film Box Office Success Based on Wikipedia Activity Big Data. doi:https://doi.org/10.1371/journal.pone.0071226

Michel Clement, S. W. (2014). Empirical generalizations of demand and supply dynamics for films. *International Journal of Research in Marketing*, 207-223.

MPA. (2022). *Theatrical and Home Entertainment Market Environment Report.* Retrieved from https://www.motionpictures.org/wp-content/uploads/2022/03/MPA-2021-THEME-Report-FINAL.pdf

Musa Malik, F. R. (2022, February). Representations of Racial Minorities in Popular Films, A Content-Analytic Synergy of Computer Vision and Network Science. *Computational Communication Research, 4*(1). doi:https://doi.org/10.5117/CCR2022.1.006.MALI

Neufeld, D. (2020). LEADERSHIPThe World’s Most Influential Values, In One Graphic. *Visual Capital*. Retrieved from https://www.visualcapitalist.com/most-influential-values/

NMAAHC. (n.d.). Popular and Pervasive Stereotypes of African Americans. Retrieved from https://nmaahc.si.edu/explore/stories/popular-and-pervasive-stereotypes-african-americans

Patel, L. (2015, November). Desiring Diversity and Backlash: White Property Rights in Higher Education. *Regional & Urban Planning, 47*. doi:https://doi.org/10.1007/s11256-015-0328-7

Prag, J. C. (1994). An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry. . *Journal of Cultural Economics*, 217–235. doi:https://doi.org/10.1007/BF01080227

Ramesh Sharda, D. D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications, 30*(2), 243-254. doi:https://doi.org/10.1016/j.eswa.2005.07.018.

Reporter, T. H. (2021). Audiences Prefer Hollywood Films With Diverse Casts. Retrieved from https://newsroom.ucla.edu/releases/2021-hollywood-diversity-report

Roberts, J. (2021). Empathy Cultivation through (Pro)Social Media: A Counter to Compassion Fatigue. *Journal. Media*, 819-829. doi:https://doi.org/10.3390/journalmedia2040047

Smith, S. L. (2020). Inequality in 1,300 popular films: Examining portrayals of gender, race/ethnicity, LGBTQIA, and disability from 2007–2019.

Sperling, N. (2021, March). Hollywood Loses $10 Billion a Year Due to Lack of Diversity, Study Finds. *New York times*. Retrieved from https://www.nytimes.com/2021/03/11/films/hollywood-black-representation.html#:~:text='Ahsoka'-,Hollywood%20Loses%20%2410%20Billion%20a%20Year%20Due%20to%20Lack%20of,representation%20can%20lead%20to%20change.

Tan, E. S. (2018). A psychology of the film. . *Humanities and Social Sciences Communications*. Retrieved from https://doi.org/10.1057/s41599-018-0111-y

Teresa Correa, S. H. (2011). RACE AND ONLINE CONTENT CREATION. *Information Communication and Society*. Retrieved from https://www.researchgate.net/publication/232924586\_RACE\_AND\_ONLINE\_CONTENT\_CREATION

UCLA. (2021). Hollywood Diversity Report 2021. Retrieved from https://socialsciences.ucla.edu/hollywood-diversity-report-2021/

Umaña-Taylor AJ, Q. S.-D. (2014, February). Ethnic and Racial Identity in the 21st Century Study Group. Ethnic and racial identity during adolescence and into young adulthood: an integrated conceptualization. 21-39. doi:10.1111/cdev.12196.

Weaver, A. J. (2011). The Role of Actors’ Race in White Audiences’ Selective Exposure to. *Journal of Communication, 61*, 369-385.

Whitten, S. (2019). More diversity, more money: Theater owners want more inclusive film. Retrieved from https://www.cnbc.com/2019/04/04/more-diversity-more-money-theater-owners-want-more-inclusive-films.html

Appendix A: Stop word list

Appendix B: Subtitle files parching

I cleaned the subtitle files and applied a logic for scene identification. In these subtitle files, when it's unclear who is speaking, the person is identified and labeled (e.g., "QUEEN GORGO"), referred to in the datasets as the "speaker." To ascertain the individuals in the scenes, I employed the en\_core\_web\_md model from the spaCy Natural Language package, which can recognize entities in the text.

To identify scenes, I wrote code which went through the subtitle file and looked for pauses in dialogue. If a pause exceeded five seconds, the code identified it as a new scene. I chose five seconds because it allows the audience to mentally adjust to the new scene without disrupting the narrative flow. Because this is a threshold, even if a scene switch took longer, I would still be counted it as one scene switch.

Next, I conducted a fuzzy merge with a thrershold of 0.85, aligning the character list with their respective ethnicities within the dataset.

Appendix C: Entire main regression analysis

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | **Robust.SE** | Sign |
| (Intercept) | 18.7634364 | 0.6172377 | < 0.001 \*\*\* |
| Inverse simpson index | 0.2788352 | 0.0854869 | < 0.001 \*\*\* |
|  |  |  |  |
| Log\_screens | -0.4607796 | 0.0508223 | < 0.001 \*\*\* |
| Log runtime | -1.3511415 | 0.1143878 | < 0.001 \*\*\* |
| log\_MPAA | -0.7109994 | 0.0729009 | < 0.001 \*\*\* |
| Log\_average\_budget | -0.0470878 | 0.0171892 | 0.008 |
| log\_sequel | 0.0058112 | 0.0014671 | < 0.001 \*\*\* |
| spinoff | -0.0564512 | 0.0420592 | 0.203 |
| log\_remake | 0.0035104 | 0.0017161 | 0.063 |
| basedonbook | -0.1316908 | 0.0615430 | 0.034 |
| basedonplay | 0.3750325 | 0.1386266 | 0.012 |
| basedoncomic | -0.0560893 | 0.1070822 | 0.590 |
| basedoncomicbook | -0.3142000 | 0.0785232 | < 0.001 \*\*\* |
| basedonnovel | -0.1080797 | 0.0352643 | 0.002 |
| basedonshortstory | -0.0100308 | 0.1037285 | 0.936 |
| log\_Nominee | -0.0091412 | 0.0016273 | < 0.001 \*\*\* |
| log\_Winner | -0.0143761 | 0.0011203 | < 0.001 \*\*\* |
| log\_dir\_power | -0.0096619 | 0.0016879 | < 0.001 \*\*\* |
| log\_metascore | -0.0355427 | 0.0082175 | < 0.001 \*\*\* |
| log\_starpower | -0.0038060 | 0.0018459 | 0.032 |
| Action | -0.0410699 | 0.0345181 | 0.247 |
| Adventure | -0.1351720 | 0.0358245 | < 0.001 \*\*\* |
| Comedy | 0.1301336 | 0.0383721 | < 0.001 \*\*\* |
| Fantasy | -0.0493828 | 0.0369585 | 0.220 |
| Crime | 0.0458143 | 0.0361069 | 0.205 |
| Drama | 0.1486707 | 0.0320064 | < 0.001 \*\*\* |
| Mystery | -0.0519966 | 0.0383778 | 0.187 |
| Thriller | 0.0376747 | 0.0337026 | 0.282 |
| Romance | -0.0847929 | 0.0367200 | 0.013 |
| Sci.Fi | -0.1089913 | 0.0372327 | 0.006 |
| Biography | 0.1629182 | 0.0625266 | 0.006 |
| Sport | 0.1355867 | 0.0636215 | 0.025 |
| War | 0.1428677 | 0.0705197 | 0.043 |
| Family | 0.2083408 | 0.0534861 | < 0.001 \*\*\* |
| Musical | -0.2922735 | 0.1308821 | 0.007 |
| History | 0.1588541 | 0.0690280 | 0.028 |
| Horror | 0.0693812 | 0.0452779 | 0.126 |
| Music | 0.0622507 | 0.0674210 | 0.325 |
| Documentary | 0.7488378 | 0.1260843 | < 0.001 \*\*\* |
| Western | 0.4924241 | 0.1199375 | < 0.001 \*\*\* |
| Spring | 0.0167648 | 0.0346022 | 0.629 |
| Summer | -0.0165910 | 0.0351252 | 0.635 |
| Fall | 0.0425621 | 0.0344176 | 0.214 |
| 2001 | -0.0835118 | 0.0841020 | 0.271 |
| 2002 | 0.0797040 | 0.0809338 | 0.296 |
| 2003 | 0.0671009 | 0.0822956 | 0.384 |
| 2004 | -0.0175361 | 0.0869597 | 0.816 |
| 2005 | 0.1337888 | 0.0785376 | 0.074 |
| 2006 | 0.0907173 | 0.0785327 | 0.219 |
| 2007 | 0.1149531 | 0.0726303 | 0.126 |
| 2008 | 0.0317699 | 0.0826234 | 0.673 |
| 2009 | 0.1722148 | 0.0772681 | 0.023 |
| 2010 | 0.2008708 | 0.0787402 | 0.009 |
| 2011 | 0.3513528 | 0.0799067 | < 0.001 \*\*\* |
| 2012 | 0.5024574 | 0.0793373 | < 0.001 \*\*\* |
| 2013 | 0.6166931 | 0.0823798 | < 0.001 \*\*\* |
| 2014 | 0.7911409 | 0.0847740 | < 0.001 \*\*\* |
| 2015 | 0.7911298 | 0.0807096 | < 0.001 \*\*\* |
| 2016 | 0.8791387 | 0.0790009 | < 0.001 \*\*\* |
| 2017 | 0.8228709 | 0.0901873 | < 0.001 \*\*\* |
| 2018 | 1.0214914 | 0.0841829 | < 0.001 \*\*\* |
| 2019 | 0.9679212 | 0.0984391 | < 0.001 \*\*\* |

1. IMDb's user base is 62% male, which is comparable to the gender distribution of ticket sales (59% male). In terms of age distribution, which can be seen in Figure 2 and 3, IMDb's demographics align with those of filmgoers, except that older age groups are underrepresented on IMDb (MPAA, 2022; Similaweb, 2023).

   Moreover, IMDb's user base is predominantly American, according to Similarweb's 2023 data. Nevertheless, given the study's focus on the Hollywood film industry and the cultural nuances of racial representation, it makes sense to focus on the United States for the analysis. The results can potentially be transferred to countries with a similar cultural landscape and other demographic similarities with the United States. [↑](#footnote-ref-1)
2. Online streaming services causes that older films are maintaining their popularity. In October 2023 Inglorious Basterds(2009), the Davinci Code(2006), L.A. Confidential(1997) and numerous other films that are years old are under the “Popular on Netflix” tab. Newer films in the end of the long tail might lose competition against the older films in the ‘blockbuster region’. While back in the days less popular new films had to compete with just the few hundred films in a movie store they now have to compete with the thousands of films offered on the online streaming platforms. [↑](#footnote-ref-2)
3. In the complete dataset, 15% of individuals linked to characters did not have associated images. Among unique actors and actresses, this figure increased to 28%. Meaning that people without images were more likely to play fewer roles. [↑](#footnote-ref-3)
4. If there was no ethnicity assigned through the threshold. The two highest scoring ethnicities on probability where assigned the value of percentage relating to each other. Highestprob / Highestprob + SecondHighestprob. As illustration if a person was assigned for the two highest probabilities 0.53 for white and 0.19 for black. It would have (0.53/0.53+0.19) = 0.736 for white. [↑](#footnote-ref-4)
5. Previous data showed that imageless characters tended to play fewer roles. The filtering process showed that imageless characters were also more likely to play less prominent characters. After the filter for generic names, 77% of characters without images were removed, compared to 66% of characters with images.

   This suggests that named characters were mostly played by people with images. As a result, the percentage of characters that were assigned an ethnicity through their name instead of facial recognition decreased from 15% to 11% of the dataset after filtering for non-named characters. As a result, imageless people are also more likely to play characters that were not authentic and inclusive (AIR). This suggests that filtering for AIR characters might also include filtering for characters played by people without images. This is an area for further exploration. [↑](#footnote-ref-5)
6. The awards used in this research are the following: Academy Awards, Critics’ Choice Awards, Directors Guild Awards, Golden Globe Awards, Golden Laurel Awards, Independent Spirit Awards, Los Angeles Film Association Awards, MTV Awards, National Board of Review Awards, National Society of Film Critics Awards, New York Film Critics Circle Awards, People’s Choice Awards and the Screen Actors Guild Awards. [↑](#footnote-ref-6)
7. Two significant factors have demonstrated a correlation with racial diversity in previous research. In 2019, action films exhibited a notably higher likelihood of including characters from underrepresented racial and ethnic groups. Additionally, films with larger budgets tend to have more diverse casts, as they possess greater resources to hire a wide range of actors (Smith, 2020). [↑](#footnote-ref-7)