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

A comprehensive examination of representation 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 Analytics

**Academic year:** 2023-2024

**Submission date:** January 5, 2024

**Name supervisor** Dr. Arjen van Lin

**Name co-reader** Prof.Dr. Marnik Dekimpe

Management summary

In recent years, the issue of racial minority representation in Hollywood films has become increasingly prominent. Numerous studies have been published on this topic, yet there remains a lack of common consensus regarding the relationship between racial representation and its effects.

This paper addresses the pressing issue of racial diversity and representation within the film industry, specifically focusing on Hollywood films. Films are powerful cultural influencers that shape societal beliefs. The absence of representation in films perpetuates negative stereotypes and marginalization, which can have a detrimental impact on individuals and communities.

Despite recent improvements in representation, this study underscores that existing measures often overlook the nuances of representation and the need for Authentic Inclusive Representation. Analyzing Authentic Inclusive Representation alongside racial representation is a comprehensive approach to representation that considers effective practices rather than mere tokenism. To better assess film success, this study shifts from the widely used Box Office to Long-Term Audience Engagement. This metric is considered more appropriate for evaluating a social concept such as racial representation.

While there are countless of different ethnicities each with their own values and experiences this study focusses on three umbrella ethnicities which are Asian, Black and Hispanic. Through the analysis, it was observed that the relationship between LTAE and racial Authentic Inclusive Representation is nuanced, exhibiting varying effects across different ethnicities. The dynamic nature of this relationship was further observed when an external factor, time, was incorporated into the analysis. This revealed that the relationship is also sensitive to societal shifts over time.

In addition to these findings, the most significant conclusions drawn are that simplistic measurements of racial representation may be inadequate. More comprehensive measures, such as AIR, need to be employed to accurately examine the impact of racial representation on a film's success. Moreover, Long-term Audience Engagement appears to be a sufficient concept to measure a film's success.

Preface

In the pursuit of knowledge and the exploration of the dynamic landscape of marketing analytics, I am delighted to present this master's thesis, which encapsulates my academic journey and reflections. I have also found an interest in films and therefore I was delighted to hear I was able to perform research in this inspiring an captivating industry.

2023 was an extraordinary year for me. Besides working on my master's degree at Tilburg University, fortunate to learn from some truly inspiring professors in the marketing department. I also experienced the joy of becoming an uncle for the first time when my sister gave birth to a beautiful baby girl. It was a whirlwind of emotions, and just three months later, my brother welcomed his first child, a son.

Firstly, I extend my deepest gratitude to my thesis advisor, Arjen van Lin for their invaluable insights, constructive feedback, and relentless encouragement throughout this research process. Their expertise and dedication have been a guiding force, steering this project towards excellence. I like Arjen as a person as well and I am grateful he also brought me under the attention of an opportunity in academics at Tilburg University.

To my family, friends and my amazing girlfriend, who have been a pillar of strength and a source of unwavering support, I extend my heartfelt thanks. Your encouragement and understanding have been a source of motivation during the challenging moments of this academic journey. I would like to thank my volleyball coach Kasper Langendoek , he gave me the opportunity to play on a reasonable high level providing the sometimes much needed break while also prioritizing my studies throughout the year.

Thank you all for your time and attention and I hope this thesis is fun to read and inspiring.

Table of Contents

[Introduction 6](#_Toc154331530)

[Problem Indication 6](#_Toc154331531)

[Problem Statement 8](#_Toc154331532)

[Academic Contribution 9](#_Toc154331533)

[Managerial Contribution 10](#_Toc154331534)

[Structure of the Thesis 10](#_Toc154331535)

[Literature Review and Hypotheses 10](#_Toc154331536)

[Long-term Audience Engagement 11](#_Toc154331537)

[The social consequences and opportunities of Racial Representation in Films 11](#_Toc154331538)

[Study 1 : The Relationship between racial Authentic Inclusive Representation and Long-term Audience Engagement 12](#_Toc154331539)

[Racial representation in films 12](#_Toc154331540)

[Authentic Inclusive Representation as essential part of Racial Representation 13](#_Toc154331541)

[Conceptual framework study 1 15](#_Toc154331542)

[Research Methodology study 1 16](#_Toc154331543)

[Results study 1 25](#_Toc154331548)

[Study 2 : Introducing an external factor to the relationship 31](#_Toc154331551)

[Research Methodology 2 32](#_Toc154331553)

[Results study 2 33](#_Toc154331554)

[Robustness checks study 1 and 2 36](#_Toc154331555)

[Discussion 38](#_Toc154331556)

[General conclusion and discussion 38](#_Toc154331557)

[Limitations and suggestions for further research 42](#_Toc154331558)

[Bibliography 45](#_Toc154331559)

# Introduction

## 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 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 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 & Company study found that films with casts which consisted of at least 30% minority 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; Reporter, 2021). This suggests that racial diversity is both a moral and financial imperative for the film industry.

Efforts to increase diversity and rep

\resentation in film have led to a significant increase in the proportion of films featuring minority actors. For example, UCLA data shows that between 2011 and 2021, the percentage of films with predominantly minority casts increased from 2% to 32.1%.

Nevertheless, the public, media, and most studies adopt a narrow approach measuring diversity (Malik, 2022; Kuppuswamy and Younkin 2016; Kim and others 2020). Most studies do not make a distinction between different ethnic groups and measure diversity based on the share of ethnic minorities in the cast. Additionally, some studies focus solely on one minority group instead researching the groups simultaneously. These narrow approaches 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 6.8 % of the actors and actresses being Hispanic compared to an actual population of 20% in the United States. (UCLA, 2021; MPAA, 2021). As a result, this group still experiences the social consequences of being underrepresented.

Secondly, while measuring diversity based solely on minority share is a useful first step, it fails to capture important information about the nature of the representation. (Malik, 2022; Lazar, 2020) To effectively address the social dimension of diversity, Authentic Inclusive Representation is essential. (Lazar, 2020; Roughton, 2014) This means that films must avoid tokenism and create meaningful storylines for characters from diverse backgrounds. (Lazar, 2020)

Building on this perspective the current literature has fallen short of conducting a comprehensive analysis of the effects of racial diversity. Therefore, while emphasizing diversity in the cast is a positive step towards showcasing racial diversity. Studies need to avoid using too simplistic methods and Authentic Inclusive Representation is an essential factor that warrants consideration for racial diversity analysis.

In addition to the limitations associated with the metrics used to gauge racial representation, there are also constraints related to the metrics employed to measure a film's success. As previously discussed, racial representation in film is a social construct. Despite this, numerous studies (Kim et al., 2020; Kuppuswamy & Younkin, 2016; Madongo, 2023) predominantly utilize box office revenue as the dependent variable to assess a film's success. While box office revenue is a valuable indicator of commercial success, it may not adequately capture the film's quality or its impact on audiences. This metric is significantly influenced by marketing, advertising, and hype (Eliashberg, 2014; Clement, 2014).

To comprehensively understand a film's influence in these aspects, Long-term Audience Engagement is chosen as the dependent variable. As elaborated later in this thesis, this concept encompasses both social and economic constructs, making it a more nuanced and fitting measure for a film's success within the framework of a social concept, such as racial representation.

Finally, through the analysis it was suggested that the main relationship is dynamic. For instance, the amount of films having AIR for an ethnicity was significantly different prior and after 2015 (Lazar, 2020). This suggested that the relationship might be sensitive to time. In this thesis the relationship between Authentic Inclusive Representation for ethnic minorities and Long-term Audience Engagement is the primary focus and this will be analyzed in study 1. After study 1 an additional study was conducted to capture the dynamic nature of the main relationship researched.

## Problem Statement

Following the problem background this studies problem statement is formulated as follows: “*What is the relationship between racial Authentic Inclusive Representation and Long-term Audience Engagement in Hollywood films?”*

## 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. The ethnic groups researched in this study are Black, Asian and Hispanic. This is important because different minority groups have different experiences and perspectives, and their representation in films should be considered separately. Moreover, analyzing the minority groups with this approach 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 builds on prior research on gender biases in film (Agarwal et al., 2015; Kagan et al., 2020), and seeks to standardize the concept of Authentic Inclusive Representation using the Bechdel-Wallace test (1985).

The Bechdel test, though not novel in its application to racial representation, was used in 2020 when the UCLA Center for Scholars & Storytellers introduced the REM test. However, UCLA did not create an automated process to conduct the REM test and the test fails to distinguish between different minority groups. By combining the work of previous researchers, this study aims to be as inclusive as possible while keeping the automated nature in its approach making it easily applicable to a large number of films.

Third, as mentioned before by using Long-term Audience Engagement, 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 racial representation and film success.

## 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. This research will show if casting diverse roles and developing characters which authentically represent a variety of backgrounds can help boost audience engagement which can in turn boost a film's financial performance. And while the financial capabilities of a film is not necessary very important to the societal purposes of racial representation this would be an incentive for the industry.

As streaming services surge, the traditional reliance on box office revenue diminishes. This study demonstrates that Long-term Audience Engagement potential as a viable measure. Thus, it provides film studios and creators with an additional metric to evaluate the success of their films, acknowledging the industry's shift towards a more comprehensive assessment of sustained impact.

## Structure of the Thesis

To explore the research questions posed earlier, this study adopts a quantitative research approach. Two distinct studies form the backbone of this investigation. The initial study centers on the primary relationship, while the second study delves into the potential influence of an additional factor on this primary relationship, a discussion that will be thoroughly explored.

The thesis unfolds as follows: it begins by presenting a theoretical framework to provide context for the research, followed by an exploration of the research methodology of study 1. The subsequent sections detail the data collection and processing procedures. The findings are then presented and discussed. Post the completion of the first study, the focus shifts to the second study, presented in the exact order as the first. The studies are followed with a few robustness checks after this a general synthesis and discussion conclude the thesis.

# Literature Review and Hypotheses

This literature review embarks on a comprehensive examination of LTAE and racial representation, followed by an in-depth discussion of the significance of AIR in ensuring effective racial representation.

## Long-term Audience Engagement

Audience engagement is a complex and multifaceted concept that encompasses the active involvement of an film’s audience. In this study, the focus is on Long-term Audience Engagement which goes beyond merely watching a film and encompasses factors such as word-of-mouth, cultural impact, and sustained interest over time (Broersma, 2019; Kumar, 2022). It is important to clarify that Long-term Audience Engagement has financial purposes for films as well. Long-term Audience Engagement is correlated with ancillary revenues such as engrossed viewing, longevity in distribution channels, and the purchase of film-related merchandise (Kumar, 2022). With regards to social relevance Long-term Audience Engagement is a key factor in a films 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; Bard , 2006). How this ability of changing people’s believes becomes very important is explained in the next section.

## The social consequences and opportunities of 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 and is not solely a product of our own perspective but is also shaped by how others perceive us (Kidd, 2015). 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. Mastro et al. in 2007 found that white people especially the ones who do not have much real-life contact with people of color which were exposed to negative racial stereotypes in the media were more likely to hold those stereotypes themselves.

Moreover, 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). When we're constantly exposed with negative portrayals of certain groups of people, it is hard not to start believing them. In contrast, 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) Therefore, the powerful impact of media representations on shaping perceptions cannot be underestimated.

Further in this study the ongoing debate regarding the extent to which films should mirror the racial 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.

# Study 1 : The Relationship between racial Authentic Inclusive Representation and Long-term Audience Engagement

In this section, we delve into the existing literature surrounding the focal relationship under examination. It kicks off with an general discussion of racial representation in the film industry, followed by an explanation of why Authentic Inclusive Representation is deemed a pivotal attribute.

## Racial representation in films

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).

Moreover, 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. This suggests that when a film embraces racial diversity, it has the potential to attract a broader audience and foster greater overall Long-term Audience Engagement (USC Annenberg, 2018).

Furthermore, minority groups tend to be more prolific creators of online content (Correa's 2011). When a film successfully resonates with these audiences through racial representation, it is more likely to stimulate online discussions, which leads to higher Long-term Audience Engagement (Kumar, 2022). 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).

Nevertheless, the demand for greater racial diversity in film is not universally embraced. 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 a film success (Weaver, 2011).But this preference does not hold true for all scenarios (Aumer, 2017).

Recent studies have raised doubts about the idea that white actors are necessary for financially success of films (Chow, 2016). Moreover, the assumption that white majority audiences prefer racially homogeneous casts have led studios to allocate smaller production budgets (Smith et al. 2020), a significant predictor box-office sales (Eliashberg, 2014; Michel Clement, 2014). Causing that 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.

Therefore, it seems that including a diverse number of ethnicities might broaden the audiences reach of a film. However, including diverse characters in films is not without its challenges. The most important concern, tokenism is discussed thoroughly in the subsequent section.

## Authentic Inclusive Representation as essential part of 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, underrepresented characters who are introduced withough fully developed storylines can have negative influences on a film’s success.

A study by King (2020) identifies a group that does not oppose racial representation in film, prioritizes the quality of the storyline. Arguing that when minorities are introduced without developed characters, through tokenism it influences the quality of the storyline. This also creates the perception of a "racial agenda," which is often disliked by viewers. 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 stereotypes in the film can perpetuate biases and misconceptions about minority groups (Umaña-Taylor AJ, 2014; Abraham and Appiah, 2006; Hurley et al., 2015). This can lead to backlash and damage Long-term Audience Engagement if viewers perceive the diversity as being insincere , which can ultimately affect the perceived quality of the film (Smith, 2016).Furthermore, these negative depictions can cause minority groups to disengage completely from the media. For instance, 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.

Besides the general audience these 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 Long-term Audience Engagement, misrepresentation is likely to lead to a decline in Long-term Audience Engagement (Hofmann, 2016; Ghiassi, 2017; Kuppuswamy, 2016; Kumar et al., 2022).

In contrast racial representation 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. 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).

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). This empathizing will in turn influence a films capability to hold Long-term Audience Engagement. Authentic Inclusive Representation, as a concept, describes how accurately and respectfully a film portrays underrepresented groups in a nuanced manner.

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.

All illustrate the necessity of authenticity in racial representation and its potential to generate positive outcomes. Therefore, the hypothesis will be the following:

**H1:** Racial Authentic Inclusive Representation is positively correlated with Long-term Audience Engagement for films.

## Conceptual framework study 1

A white rectangular object with a black line

Description automatically generated

*Figure 1 Conceptual model paper*

Figure 1 indicates the conceptual model consists of one relationship across different levels among different ethnic groups. The ethnic groups researched separately but simultaneously are Hispanic, Asian and Black[[1]](#footnote-1). All relationships should be positive. These relationships will be examined through regressions models. The model specification can be found in the next section.

## Research Methodology study 1

This study aims to investigate the relationship between racial Authentic Inclusive Representation and Long-Term Audience Engagement in Hollywood films. The following sections will provide a detailed overview of the data collection and sampling procedures, as well as a clear operationalization of the variables. Given the subjective nature of Authentic Inclusive Representation, the Bechdel Test was employed as a quantitative measure for this concept. Notably, it will be discussed across three distinct levels, as defined by the Bechdel test, elaborated in the operationalization section.

## Main data sources

The main data sources for this study are the Internet Movie Database (IMDb), The Numbers, and The Movie Database (TMDB). IMDB is the most important source as it will be used to measure Long-term Audience Engagement. While IMDb users and the general filmgoing population differ somewhat in demographicsprevious research has used IMDB data numerous times demonstrating the representability for the broader filmgoing audience (Ghiassi, 2017; Partha and Chakraborty, 2019; Apala, 2013).

## Sample

The films for the sample were chosen carefully and based on specific criteria. IMDB started in 1996 but was not widely used until 2003 so the sample was limited to films released after the year 2000 (for a further explanation, see the dependent variable operationalization).

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, undermining 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 eliminated. After applying these criteria, the sample consisted of 2,319 unique films. To measure Authentic Inclusive Representation, data on interactions between actors is required (for a further explanation, see below). This data was available for 1,178 films, such that the final sample consisted of 1,178 films.

## Variable operationalization

**Long-term Audience Engagement (DV)**

Long-term Audience Engagement was operationalized using IMDb's MovieMeter, a metric derived from popularity rankings. A score of one means that the film was the most popular with regards to clicks, page views and reviews on IMDB in that week compared to other films. Therefore, this film score includes direct indicators of Long-term Audience Engagement, such as online discussions, reviews, and word-of-mouth conversations. To measure sustained engagement, I measured Long-term Audience Engagement as the average ranking 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 the engagement of audiences with a film.

**Ethnicity**   
To determine actors’ ethnicity, I use the Kairos API, a deep learning algorithm that can detect ethnicity through facial recognition. I collected the profile pictures of the actors and actresses with a scraper from IMDb and processed these with the API. The Kairos API is chosen because of its efficiency, accuracy (99.63%), and ability to handle a large dataset (Kairos, 2023). In the films included in the sample of this study a total of 62834 characters were present. This count includes instances where an actor or actress appeared in multiple films. It also included uncredited people (7.251). The Kairos API analyzed 26,891 images found on IMDb of the total of 32,513 unique actors and actresses.

For the remaining imageless people the ethnicities were based on their first and last name using the R package Rethnicity, which has an accuracy of around 80%. People not having an image on IMDB is correlated with not having played a significant role in a film [[2]](#footnote-2) therefore they are assumed to be less important.

**Racial Authentic Inclusive Representation**

Quantifying the concept of representation in films has resulted in researchers using the Bechdel-Wallace test (1985), originally designed to measure the authentic representation of women in a film (Agarwal, 2015). The Bechdel test is chosen as tool for measuring Authentic Inclusive Representation in films because of its adaptability to this study’s specific focus (Lazar, 2020), its potential for automation, and its quantifiability (Argarwal, 2015).

The test comprises three gradations: (T1) The movie has to have at least two named women in it (T2) who talk to each other, (T3) about something other than a man (Agarwal, 2015). The Bechdel has been previously used by Lazar and others (2020). They focused on ethnic minorities rather than individual ethnicities. 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 ethnic minority characters, and no white people are present or mentioned in the conversation.

The Bechdel test was modified to measure Authentic Inclusive Representation for the three different minority ethnicities of this research enabling to look at the concept on three different levels per ethnicity (T1, T2, T3):

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

(T2) two named {ethnicity} characters appear in a scene together.

(T3) without a white character.

As can be seen the modified Bechdel test similar to the original requires the actors to be named. Therefore, the first step of conducting the modified Bechdel test was to create a process which removed ‘generic’ characters from the dataset.

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 names which only included 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 Appendix A.

I made an exception to the character filter: I did not remove a character if the stop word was the first token in their name "Colonel Rich Bron" , "e.g.," “Docter Johnson” were kept. However, there was an addition to this exception which was that if the character's name consisted of only one token. For instance, a character named "Colonel" it would be removed.

To delve deeper into the link between inclusive characters and racial diversity, I implemented steps 2 and 3 of the modified Bechdel test using subtitle files. Previous research such as Argawal (2015) used screenplays to conduct the Bechdel test. 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 the subtitles.org API I was able to find subtitles for half of the films of the entire available dataset which were in a format that could be standardized for testing. These subtitle 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. As illustration, 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.

After extracting subtitles, as detailed in Appendix B, the dataset comprised scene indices, speakers, individuals mentioned, and the ethnicities of the characters. The subsequent step involved applying the revised Bechdel Test to this dataset.

Films were classified based on whether they satisfied the conditions (T1) , (T2) and (T3) for different ethnicities. These classification served as the variables utilized in the regression models.

**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. By controlling for the impact of these variables, more accurate estimations can be derived for the variables under investigation. Moreover, by including these covariates the chances for an omitted variables biased is tried to be reduced

The variables considered alongside Authentic Inclusive Representation in this study include Sequel, the star power of both Actors and Directors, MPAA rating, Number of Opening Screens, Critical Acclaim, Awards, Budget, Genre, Source and Seasonality. Table 3 provides detailed information including which sources identified which variables , the following paragraphs briefly discuss how the variables are measured in this study.

In this study, (SEQUELi) is a dummy variable if the film is a sequel. The measure for star power score (STARPOWERi) is based on the measurement from Nelson and Glotfelty's (2012), it is the four highest-grossing actors' ranking on the website The Numbers for one year before the film. For (DIRECTORPOWERi) the highest-grossing directors' ranking on The Numbers are also used.

With regards to critical reclaim the (CRITICSi) value is the average rating on metacritic.com. Moreover, 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 mentioned on the website of IMDB which makes it very easy and accessible to account for a vast amount of awards internationally and national.

(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 theater screens the film had according to the Numbers. (BUDGETi) was the production budget available at one of the three data sources used (IMDB, The Numbers and TMDB), if multiple production budgets where available across the sources the average was taken.

Moreover, this study introduced 19 genres through dummy variables, 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 (2016), 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, variables representing the genres of the film these dummy variables are not mutually exclusive because a film could be based on multiple sources.

As previously noted, there is an observable trend indicating a continuous decline in Long-term Audience Engagement, particularly marked by a significant drop in the year 2011. In order to account for potential variations over time, dummy variables (YEARi) were introduced for each year within the sample, enhancing the model's ability to control for temporal effects.

Table 3: Measures of Variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Description | Measure | Data Source | Academic research |
| **Independent variable** |  |  |  |  |
| AIRi | Authentically Inclusive Representation. | Dichotomous variable with three separate levels | Subtitles.org | L, R |
| **Dependent variable** |  |  |  |  |
| Log(LTAEi) | IMDB ranking 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 |  |
| C**ovariates** |  |  |  |  |
| Log(STARPOWERi) | The top box office stars based on the cumulative worldwide box office of all the films a star has had a leading role in over their lifetime. | Log (Sum of the top 4 actors, one year before the release \* 100 + 1) | The Numbers | KRM, H, KY, G, CM |
| Log(DIRECTORPOWERi) | Derived from the list of the highest grossing Directors based on the worldwide box office of the films they worked on. | Log ( The director ranking \* 100 + 1) for the director, one year before the film release . | The Numbers | KRM, H, G, KY, CM |
| Log(CRITICi) | Average rating the film received from professional film reviewers. | Log (Average rating the film received from professional film reviewers + 1 ) | Metacritic | H, KRM, KY, G |
| Log(NOMINATIONSi) | Number of award nominations the film received. | Log(Number of award nominations the film received + 1) . | IMDb | KY,KRM, HTS |
| Log(WINSi) | Number of award wins the film received. | Log( Number of award nominations the film received + 1). | IMDB | KY,KRM, HTS |
| 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 | H, KY, L, BCR, KRM , SD, CM |
| Log(SCREENSi) | Number of screens at release | Log(Number of screens) | Numbers | H , KY, KRM , HTS, CM |
| Log(BUDGETi) | Production budget of film | Log(The average production budget of the data sources) | IMDB, Numbers, OMDB | KY,KRM, HTS, CM |
| Log(RUNTIMEi) | Duration of a film in minutes | Log( Duration of a film in minutes) | IMDb | KY |
| SEQUELi | Indiciation if the films is a sequel or not | Sequel = 1, Not a sequel = 0 | IMDB | KRM, H , KY, CM |
| ACTIONi | Genre Action | Genre Action = 1, Other = 0 | IMDb | KRM, H , KY, G,CM |
| 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 | [Mar, Apr, May]= 1, Other = 0 | IMDb | V |
| SUMMERi | Film released in the summer | [Jun, Jul, Aug]= 1, Other = 0 | IMDb | - |
| AUTUMi | Film released in the autumn | [Sep, Oct, Nov] = 1, Other = 0 | IMDb | - |
| WINTERi | Film released in the winter | [Dec, Jan, Feb]= 1, Other = 0 | IMDb | - |
| BOOKi | Film is based on a book | Film is based on a book = 1, other = 0 | IMDb | H |
| 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 |  |
| {…..} i | Years 2002 – 2017 |  | IMDb |  |
| YEAR2018i | Year2018 | 2018 = 1, other = 0 | IMDb |  |
| YEAR2019i | Year2019 | 2019 = 1, other = 0 | IMDb |  |

Notes: CM = Clement et al. (2014) , *KRM = Kumar et al., (2022) ;* H = Hofman et.al (2016) ; L = Lazar et al., (2020) ; KY = Kuppuswamy et al. (2016) ; HTS = Hunter et al. (2016; SD = Ramesh & Delen (2006); BCR = Basuroy et. al , (2003), L = Litman , (1983); G = Ghiassi et. al (2017), HL = Hall, (2022) , SM = Smith , (2020)

## Models study 1

Following the approach outlined by Clement, Wu, and Fischer (2014), this research uses a log-log regression model. As such, all variables that are not dummy variables were log-transformed. 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 '1’ was added to all variables which were continuous before taking the logarithm, ensuring meaningful results For the variables director power and star power which were numbers between 0 and 1.5 the value was first multiplied by one hundred before the number was added, assuring the small integer added did not have too much effect on the value. Because AIR provides different levels models were made for all three levels with the ethnicities as independent variables.

Model T1 :

log(LTAEi) = β0 + β1 × AIR\_HispanicT1i + β2 × AIR\_BlackT1i +

β3 × AIR\_AsianT1i + 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 + β42 x SPINOFFi + β43 x SERIESi +

β44 × log(STARPOWERi) + β45 ×YEAR2000i + β46 × YEAR2001i  + β47 × YEAR2002i +

β48 x YEAR2003i + β49 × YEAR2004i + β50 × YEAR2005i + β51 × YEAR2006i +

β52 × YEAR2007i + β53 × YEAR2008i + β54 × YEAR2009i + β55 × YEAR2010i +

β56 × YEAR2011i + β57 × YEAR2012i + β58 × YEAR2013i + β59 × YEAR2014i +

β60 × YEAR2015i + β61 × YEAR2016i + β62 × YEAR2017i + β63 × YEAR2018i + β64 × YEAR2019i + εi

Model T2:

log(LTAEi) = β0 + β1 × AIR\_HispanicT2i + β2 × AIR\_BlackT2i + β3 × AIR\_AsianT2i + β4 × log(DIRECTORPOWERi) + … β60 × YEAR2015i + β61 × YEAR2016i + β62 × YEAR2017i + β63 × YEAR2018i + β64 × YEAR2019i + εi

Model T3:

log(LTAEi) = β0 + β1 × AIR\_HispanicT3i + β2 × AIR\_BlackT3i + β3 × AIR\_AsianT3i + β4 × log(DIRECTORPOWERi) + … β60 × YEAR2015i + β61 × YEAR2016i + β62 × YEAR2017i + β63 × YEAR2018i + β64 × YEAR2019i + εi

In these models, LTAEi represents the Long-term Audience Engagement for film i, AIR\_HispanicTi, AIR\_BlackTi, and AIR\_AsianTi represent the authenticity, identification, of Hispanic, Black, and Asian representation in film i. The other variables are the control variables.

## Results study 1

In this section, the results of the study are discussed. First, to establish a comprehensive overview of the dataset, descriptive and frequency statistics for the variables are presented. It is crucial to note that, as mentioned previously, Long-term Audience Engagement represents the average rank of a film in the third year post-release. Therefore, a higher value of the dependent variable signifies lower Long-term Audience Engagement. Consequently, negative estimates in the results denote positive effects, while positive estimates indicate negative effects on Long-term Audience Engagement.

## Descriptive Statistics

Table 2 presents descriptive statistics for the dependent variable and the key control variables used in the models (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 2: Descriptive statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable Name** | **N** | **Mean** | **SD** | **Minimum** | **Maximum** |
| *Dependent variable* |  |  |  |  |  |
| Avg Rank Third Year | 1178 | 3410 | 4387 | 75.92 | 103633 |
| *Control Variables* |  |  |  |  |  |
| Opening screens | 1178 | 2650.06 | 826.48 | 502 | 4662 |
| Runtime | 1178 | 108.69 | 17.19 | 75 | 219 |
| Budget | 1178 | 54,673,873.16 | 53,393,953.25 | 250,000 | 356,000,000 |
| Nominee | 1178 | 16.26 | 36.30 | 0 | 462 |
| Winner | 1178 | 5.99 | 23.02 | 0 | 490 |
| Director power | 1178 | 0.03 | 0.09 | 0 | 1.0 |
| Star Power | 1178 | 0.06 | 0.15 | 0 | 1.2 |
| Metascore | 1178 | 47.66 | 16.46 | 0 | 96 |
| Remake | 1178 | 0.09 | 0.32 | 0 | 3 |
| MPAA | 1178 | 3.24 | 0.69 | 1 | 4 |

The table presented below offers a comprehensive overview of the independent variable at levels T1, T2, and T3. This table displays the amount of films passing the condition. In Appendix C tables representing the descriptives for the dummy control variables are presented.

Table 3: Number of films passing T1, T2 and T3 condition,

|  |  |  |  |
| --- | --- | --- | --- |
| Ethnicity | NC (T1) | NC (T2) | NC (T3) |
| Black | 503 | 335 | 206 |
| Asian | 334 | 149 | 77 |
| Hispanic | 326 | 199 | 97 |

## Log-log regression model estimates

With information available for different levels of AIR (T1, T2, and T3), regression analysis was conducted for all three levels. Before running the models, the VIFs of all variables in all models were checked to get an understanding of potential multicollinearity. None of the models showed any VIF values exceeding the threshold of 10. To deal with potential heteroscedasticity issues robust standard errors were implemented.

Table 4 presents the results for each model. The last rows show the focal variables and the model fits. It should be noted that the dependent variable, average ranking, is inversely related to film Long-term Audience Engagement. This means that a higher ranking corresponds to lower engagement. As a result, a negative coefficient indicates a positive effect of the variable on film Long-term Audience Engagement and vice versa.

Table 4: Results regression analysis T3, T2, T1

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | T1 | T2 | T3 |
| Action | -0.064 | -0.066 | -0.07 |
| Adventure | -0.054 | -0.041 | -0.054 |
| Biography | 0.293 \*\*\* | 0.289 \*\*\* | 0.291 \*\*\* |
| Comedy | 0.144 \*\* | 0.148 \*\*\* | 0.152 \*\*\* |
| Crime | 0.052 | 0.029 | 0.037 |
| Documentary | 0.936 \*\* | 0.986 \*\*\* | 0.989 \*\*\* |
| Drama | 0.137 \*\*\* | 0.138 \*\*\* | 0.137 \*\*\* |
| Family | -0.01 | -0.011 | -0.02 |
| Fantasy | 0.241 \*\* | 0.239 \*\* | 0.228 \*\* |
| History | -0.043 | -0.033 | -0.035 |
| Horror | 0.135 | 0.15 | 0.142 |
| Music | 0.263 \*\* | 0.248 \*\* | 0.233 \*\* |
| Musical | 0.102 | 0.091 | 0.073 |
| Mystery | 0.008 | 0.027 | 0.038 |
| Romance | -0.15 \*\*\* | -0.143 \*\* | -0.15 \*\*\* |
| Sci.Fi | 0.009 | 0.002 | 0.005 |
| Sport | 0.037 | 0.002 | -0.006 |
| Thriller | -0.016 | -0.012 | -0.022 |
| War | -0.045 | -0.044 | -0.052 |
| Western | 0.03 | 0.031 | 0.023 |
| Fall | -0.013 | -0.014 | 0.042 |
| Spring | -0.02 | -0.015 | 0.021 |
| Summer | -0.044 | -0.042 | 0.005 |
| Based on book | -0.101 | -0.108 | -0.102 |
| Based on comic book | -0.215 \*\* | -0.199 \*\* | -0.227 \*\*\* |
| Based on novel | -0.152 \*\*\* | -0.128 \*\* | -0.131 \*\* |
| Based on play | 0.311 | 0.343 | 0.352 |
| Based on short story | -0.002 | 0.042 | 0.028 |
| Spinoff | -0.036 | -0.033 | -0.031 |
| Sequel | 0.102 \* | 0.097 \* | 0.096 \* |
| Log(Budget) | -0.031 | -0.032 | -0.025 |
| Log(Screens) | -0.476 \*\*\* | -0.464 \*\*\* | -0.457 \*\*\* |
| Log(Runtime) | -0.737 \*\*\* | -0.733 \*\*\* | -0.733 \*\*\* |
| Log(MPAA) | -0.504 \*\*\* | -0.509 \*\*\* | -0.533 \*\*\* |
| Log(Director Power) | -0.089 \*\*\* | -0.102 \*\*\* | -0.108 \*\*\* |
| Log(Metascore) | -0.327 \*\*\* | -0.34 \*\*\* | -0.331 \*\*\* |
| Log(Nominee) | -0.154 \*\*\* | -0.152 \*\*\* | -0.155 \*\*\* |
| Log(Remake) | 0.007 | 0.006 | 0.007 |
| Log(Star power) | -0.031’ | -0.033’ | -0.031’ |
| Log(Winner) | -0.147 \*\*\* | -0.142 \*\*\* | -0.142 \*\*\* |
|  |  |  |  |
| Asian | 0.002 | 0.013 | 0.036 ` |
| Hispanic | -0.041 | 0.052 | 0.084` |
| Black | 0.034 | 0.183 \*\*\* | 0.247 \*\*\* |
| R2 | 0.7211 | 0.7294 | 0.7325 |
| Adjusted R2 | 0.7054 | 0.7142 | 0.7175 |
| Degrees of freedom | 1110 | 1110 | 1110 |
| F Statistic | 44.51\*\*\* | 46.664\*\*\* | 47.354\*\*\* |

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

Starting with an assessment of model fit, the R-squared (R2) values are notably high across all models. However, comparing the models adjusted R-squared values with the adjusted R-squared of a model without the T variables (0.707), it can be seen that the variables of focus did not substantially enhance explanatory power.

The control variables were based on studies which mostly used box office as their dependent variable. Most control variables show significant effects in the models from this study, emphasizing their importance not just for short-term success but also for understanding why a film keeps audiences engaged in the long term.

Focusing on the dummy control variables, out of 19 genres, 13 exhibited significant effects, demonstrating diverse positive and negative impacts. Moreover, films based on existing sources such as novels and comic books consistently exhibited higher Long-term Audience Engagement across most models, with estimates ranging from β = -0.131 to β = -0.227. Notably, this trend was not observed for films based on short stories, books or plays. Finally, Sequel the dummy variable which also indicates an existing audience base had a positive estimate in all three models.

Zooming in on the continuous variables, starting with the variables related to critical acclaim, Nominee (β = -0.267, β = -0.153, β = -0.154, in T1, T2 and T3), Winner (β = -0.187, β = -0.144, β = -0.15), and Metascore (β = -0.187, β = -0.347, β = -0.334) demonstrated a consistent positive effect across all three models. This indicates that films receiving critical recognition, whether through nominations, awards, or high critic ratings, tend to enjoy greater popularity. Suggesting that critical recognition can signal to audiences that a film is worth watching contributing to Long-term Audience Engagement or that a film of higher quality is positive correlated with Long-term Audience Engagement.

Moreover, the continuous variables Director Power (β = -0.183, β = -0.098, β = -0.085), Screens (β = -0.476, β = -0.464, β = -0.457), MPAA (β = -0.344, β = -0.512, β = -0.507), Star Power (β = -0.031, β = -0.031, β = -0.033), and Runtime(β = -0.737\*\*\*, β = -0.733\*\*\*, β = -0.733\*\*\*) demonstrated consistent positive effects across all models. While the estimate of Star Power was only marginally significant (0.05 < p-value < 0.10), it still points to the potential impact of star power on film success. These findings align with previous research on the impact of these variables on the film's box office success.

A noteworthy observation is the insignificance of the seasonality variables. However, this can be rationalized by considering the nature of Long-term Audience Engagement metric which spans over a period of a year controlling for seasonality. However, this result does suggest that the initial success of a film due to seasonality, as observed in previous research, may diminish over time and is likely more closely tied to Long-term Audience Engagement at the time of release.

Zooming in on the key variables, an examination of the ethnicity variables reveals several interesting findings. The variables Asian and Hispanic ethnicity showed no significant effect in all models. Moreover, the variables representing the Black ethnicity exhibited a positive coefficient with average ranking in two out of the three models indicating a negative effect. In the context of a logged dependent variable, dummy variables should be interpret as a exp(x) – 1 \* 100 is the percentage change. For films featuring two black characters engaging in a conversation in module the estimates suggest that these films experience an overall ranking increase of 20.6% in the T2 module and a 30% increase in the T3 module. Meaning that the films passing this condition seem to be less popular in the third year after release.

These results, particularly the negative correlation for the Black ethnicity, challenge the hypotheses, which anticipated a positive relationship between racial authentic representation (AIR) in a film and Long-term Audience Engagement. Therefore, this study has shed light on various factors influencing film success. However the unexpected outcomes, might indicate considering additional factors that may influence audience perceptions and preferences. As such, an additional study was conducted, incorporating an extra variable to assess the impact of cultural context. This study, which sheds light on the nuances of racial representation's influence on film success, is provided in the subsequent part of this thesis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hypotheses** | **Ethnicity** | **β (T1, T2 , T3)** | **P-Value (T1, T2 , T3)** | **Decision** |
| H1a: Racial Authentic Inclusive Representation is positively correlated with Long-term Audience Engagement for films. | Black  Asian  Hispanic | (0.026, 0.188, 0.263 )  (0.001, 0.015, 0.032 )  (-0.044, 0.055, 0.028) | (P = 0.282, < 0.001, < 0.001)  (P = 0.982, 0.744, 0.074)  (P = 0.263, 0.183, 0.067) | Partly contradicted |

# Study 2 : Introducing an external factor to the relationship

While the marketing field recognizes the growing influence of external factors on consumer behavior (Cruz-Cardenas et al., 2021), many film studies, particularly those investigating racial diversity, have confined their focus to inherent film characteristics, neglecting a broader context. While some studies attempt to address temporal variations by incorporating dummy variables, these methods primarily establish that film success fluctuates across time periods. The limited focus on inherent film characteristics, as demonstrated in Study 1 by Malik et al. (2022) and Kuppuswamy et al. (2016), may overlook significant external factors influencing the relationship between racial representation and Long-term Audience Engagement.

Study 1 revealed instances of a negative relationship between racial Audience Inclusivity Ratio (AIR) and long-term audience engagement (LTAE). To delve deeper into the dynamics of this relationship, I explored whether an external factor could moderate the impact of racial representation on long-term audience engagement.

By focusing on one moderator, the study maintained a focused scope while still addressing the central research question: the potential influence of external factors on the relationship between racial representation and Long-term Audience Engagement. The #oscarssowhite movement, which emerged in January 2015, marked a significant shift in public discourse on racial representation in the film industry. This movement served as a catalyst for increased awareness and dialogue about the importance of authentic racial representation in films. Therefore, the following hypothesis is proposed:

H2: The relationship between racial authenticity in representation (AIR) and Long-term Audience Engagement changed after #OscarsSoWhite.

A diagram of a function

Description automatically generated

The model replicates the original negative relationship observed in Study 1. This study investigates whether appreciation of racial representation is time-sensitive, testing how the effects change after the #oscarssowhitemovement started.

## Research Methodology 2

The research methodology of Study 2 follows that of Study 1. Below, I will discuss how it will be extended to measure the effect of a change in perception towards racial Authentic Inclusive Representation.

As mentioned before, a specific cultural moment for racial representation was selected to serve as a reference point for the study. This event was the #OscarsSoWhite movement. A step dummy variable, after\_jan\_2015, was created. This variable was assigned a value of one for films released in or after 2015, indicating the period when racial representation gained heightened cultural significance. It is assumed, given the continued emphasis on belonging and community in 2020 and Lazar’s findings, that this period of cultural significance lasted at least from 2015 to 2019. This assumption eliminated the need to explicitly model a moment when racial representation ceased to be a prominent cultural topic.

**Model**To investigate the potential moderating effect of cultural resonance on the relationship between racial authentic representation (AIR) and Long-term Audience Engagement the interaction term after\_jan\_2015 was incorporated into the three T1, T2, and T3 models. This interaction term enables us to determine whether the influence of racial AIR on Long-term Audience Engagement varies before and after the #oscarssowhite movement in January 2015, which served as a proxy for the heightened cultural significance of racial representation The revised T1 model with the added interaction term is presented below:

log(LTAEi) = β0 + β1 × AIR\_HispanicT1i + β2 × AIR\_BlackT1i + β3 × AIR\_AsianT1i + β4 ×FILMSAFTERJAN2015i\*AIR\_HispanicT1i + β5×FILMSAFTERJAN2015i\* AIR\_BlackT1i + β6 × FILMSAFTERJAN2015i \* AIR\_AsianT1i + …. β61 × SUMMERi + β62 × FALLi + εi

In this model, LTAEi similar to the previous study represents the Long-term Audience Engagement for film i, FILMSAFTER2015i is a dummy variable that takes a value of 1 if the film was released in or after 2015. The interaction terms FILMSAFTER2015i × AIR\_HispanicT3i, FILMSAFTER2015i \* AIR\_BlackT3i, and FILMSAFTER2015i \* AIR\_AsianT3i allow to capture whether the effect of racial AIR on Long-term Audience Engagement differs before and after January 2015.

What needs to be noted is that while the interaction terms are included in the model the main effect of the dummy variable of after January 2015 is not included. This is because it would have multicollinearity issues with the year dummies incorporated as control variables.

## Results study 2

The results of the models incorporating 2015 as an interaction term are presented in the following table, showcasing only the variables under investigation to avoid redundant display of the control variables for which significance and influence was thoroughly discussed in study 1. The complete models can be referenced in Appendix D.

Table 8: Results step dummy regression analysis

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Model step T1 | Model step T2 | Model step T3 |
| Asian | 0.073 ‘ | 0.099 ‘ | 0.038 |
| Black | 0.041 | 0.182 \*\*\* | 0.262 \*\*\* |
| Hispanic | - 0.067 | 0.041 | 0.031 |
|  |  |  |  |
| after\_jan\_2015:Asian | -0.256 \*\*\* | - 0.228\* | -0.014 |
| after\_jan\_2015:Black | -0.032 | 0.003 | -0.049 |
| after\_jan\_2015:Hispanic | 0.075 | 0.046 | 0.153 |
|  |  |  |  |
| Delta, Asian | -0.184\* | -0.129’ | 0.023 |
| Delta, Black | 0.009 | 0.184\* | 0.212\* |
| Delta, Hispanic | 0.008 | 0.087 | 0.184\* |
| R2 | 0.7240 | 0.7308 | 0.7330 |
| Adjusted R2 | 0.7074 | 0.7147 | 0.7171 |
| Degrees of Freedom | 1116 | 1116 | 1116 |
| F Statistic | 44.83\*\*\* | 45.73\*\* | 46.18\*\*\* |

The three models, representing different time periods, exhibited higher adjusted R-squared values compared to the respective models in Study 1 (T1:0.70 , T2:0.71, T3:0.71), suggesting that incorporating enabled the model to capture more of the variance of the dependent variable. Nevertheless, the R squared adjusted values are only slightly higher than from the first modules. Suggesting that incorporating the dummy variable did not provide a lot more explanatory power to the module.

Examining the parameter estimates for racial representation (T variables) before 2015, it is evident that all ethnicities exhibited positive coefficients in at least one of the models, indicating a negative relationship with Long-term Audience Engagement. This finding deviates from the insignificant relationships observed in Study 1.

For clarity purposes the estimates of the interactions terms are included. As can be seen, all relationships are either insignificant or have a negative estimate, meaning a positive relationship. Nevertheless, due to the interaction, the estimates of the term are compared to the existing coefficients of the main effects of the T variables. Therefore, to see if the effects indeed became insignificant or reversed, the delta rule was applied.

The most intriguing findings of this study are that the effect of the interaction term for the Asian ethnicity showed a reversed effect for model T1 applying the Delta rule. This means that the initial negative effect was in fact reversed. With T2, the effect was also reversed for the Asian ethnicity however these effects were significant with a p-value of 0.10

Through applying the Delta it can be observed that the estimates of the interaction terms related to Black T2 and T3 remain significant. While just looking at the estimates of the interaction term suggested as possible insignificance. While the estimate with the T3 module is reduced slightly overall it is observed prior and after 2015 having AIR for a film in general means a reduction in Long-term Audience Engagement.

Lastly, the Hispanic T3 interaction term actually became significant applying the Delta rule with an estimate of 0.184. This suggests that films with higher levels of Hispanic representation in 2015-2019 were more likely to experience less Long-term Audience Engagement. This finding is somewhat surprising, given the insignificant relationship observed for the Hispanic ethnicity in prior periods and the heightened focus on racial representation after January 2015.

One possible explanation for this change in effect for the Asian ethnicity is because, as mentioned by Lazar and others (2021), there is an increase of films with ethnicity AIR in 2015-2019 compared to prior to 2015. This may have led to more authentic and relatable portrayals of Asian characters, which in turn, resonated with audiences.

However, for Black ethnicity, the negative relationship between racial representation and long-term audience engagement remained consistent across all three time periods. And for the Hispanic ethnicity is actually occurred only after January 2015. This suggests that for Black and Hispanic actors, the on-screen representation may have a more challenging path to gaining acceptance and resonance with audiences.

Through the results this study primarily highlights the intricate relationship between a time factor, racial representation and a film Long-term Audience Engagement. The findings suggests that factors beyond film characteristics, shifts attitudes towards racial representation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hypotheses** | **Ethnicity** | **β (T1, T2 , T3)** | **P-Value (T1, T2 , T3)** | **Decision** |
| H2a: The relationship between racial AIR and Long-term Audience Engagement is moderated by culture. | Asian  Black  Hispanic | (-0.184, -0.129, 0.023)  (0.009,0.184,0.212) (0.008, -0.087, 0.184) | (P = 0.021, 0.008, 0.990)  (P = 0.695, 0.002, 0.004)  (P = 0.278, 0.552, 0.032) | Accepted |

# Robustness checks study 1 and 2

Due to both studies measuring AIR for different ethnicities and on different levels the studies already have a broader scope compared to previous film studies. To check the robustness and validity of the findings, I conducted two checks. These checks focus mostly on the two substantial differences between this study and previous studies. Therefore, one check changes the dependent variable into box office. The other check is seeing if simplistic methods would have given similar results to this study or previous studies.

*Short term success (box office as DV):* While previous studies have primarily examined box office success, this research emphasizes the broader social implications of racial representation, necessitating a broader metric than purely financial performance. Consequently, a robustness check was conducted by analyzing box office success as the dependent variable rather than Long-term Audience Engagement. In line with expectations, the ranking on IMDB during the release week of the film exhibited a significant correlation with box office success. The ranking of the release explained approximately 38% of the variation in box office performance (R² = 0.3824).

When applying the modules to box office success, the observed patterns for box office mirrored those from Long-term Audience Engagement . For Black ethnicity, all three modules (T1, T2, and T3) exhibited a negative relationship with box office performance. However, these effects vanished when incorporating the January 2015 dummy variable as interaction term.

For the Asian ethnicity similar to Long-term Audience Engagement the initial relationship seems to be insignificant. But after the introduction of the dummy variable a positive relationship emerged between T2 and box office success. For Hispanic ethnicity, none of the relationships were statistically significant. These findings suggest that the relationship between AIR and box office aligns with the relationship between AIR and Long-term Audience Engagement.

The changes in the metrics demonstrate the consistency of the findings, suggesting that the relationship between AIR and Long-term Audience Engagement is robust to methodological variations in measuring racial representation.

*Simplistic measure for racial representation:* The initial study's findings challenges previous research by Kuppuswamy and others (2016) and Kim and others (2020), who observed a positive correlation between the number of Black actors in a film and box office success. To address the contradictory findings between these studies and that of Kuppuswamy and Kim the focal variables in the model were replaced with variables representing the total number of Black, Asian, and Hispanic actors.

This analysis yielded similar positive effects the Asian (β = -0.016, p = 0.004), Black (β = -0.005, p = 0.012) and Hispanic (β = -0.018, p = 0.006) ethnicities all showed positive correlations with regard to Long-term Audience Engagement. This suggests that a higher number of actors from these ethnicities had a positive relationship with Long-term Audience Engagement, aligning with Kuppuswamy's and Kim’s findings.

However, examining only the total number of actors fails to consider the overall cast size, introducing a bias caused by the omission of this crucial factor. While Kim and others (2020) partially control for this issue, Kuppuswamy (2016) neglects the total cast size altogether. To overcome this limitation, I utilized the Simpson diversity index to quantify racial diversity throughout the entire cast. The complete method is outlined in Appendix J. Incorporating this metric uncovered a negative correlation between racial diversity and Long-term Audience Engagement (β = 0.268, p < 0.001).

Therefore, upon this examination, it became apparent that the discrepancy in the findings stemmed from the different metrics used for racial representation rather than a fundamental difference in the effects of racial representation itself.

# Discussion

## General conclusion and discussion

The study investigated the impact of racial representation on Long-term Audience Engagement in films, using the Authenticity in Representation (AIR) metric to measure racial representation and Long-term Audience Engagement as measured by ranking on IMDB. The study employed two research designs: the first examined the overall relationship between racial representation and Long-term Audience Engagement, while the second investigated the effect of an interaction of an external factor on this relationship.

The most intriguing finding of this study is that authentic racial representation can significantly impact a film's Long-term Audience Engagement, albeit in varying ways across different ethnicities. The Asian and Hispanic representation, for instance, demonstrates not having association with Long-term Audience Engagement in all contexts, whereas Black AIR generally exhibits a negative relationship.

Moreover, Lazar and other (2021) observed a positive correlation between box office and AIR on the condition with large budget films. And large budgets is on it self correlated with large casts. Therefore, the negative correlations observed for the AIR variables could be due to the possibility that prioritizing the inclusion of actors from a certain ethnicity through AIR might inadvertently limit the presence of actors from other ethnicities in the film. This raises the question of whether there might be a trade-off between authentic representation of one ethnicity and the authentic representation of another ethnicity.

The initial study, drawing on racial identity theory, had hypothesized a positive relationship between racial representation and Long-term Audience Engagement. This theory suggests that increased audience engagement stems from the ability of more individuals to connect with characters on the basis of shared demographic characteristics. However, if the authentic representation of one ethnicity is not accompanied by a broader expansion in the number of ethnicities represented authentically, it could lead to an underrepresentation of other groups. This would be particularly concerning if the underrepresented group were white characters, as they constitute the largest ethnicity in the United States.

The second study in this research suggests that the connection between racial representation and Long-term Audience Engagement likely shifts over time. Culture was considered a potential external influence. To quantify this concept, a specific point in time was identified, which was assumed to mark when racial representation gained prominence in the cultural landscape. Interestingly, while negative correlations between racial representation and Long-term Audience Engagement were prevalent before #oscarssowhite, these effects appeared to diminish or even reverse afterward. While the study does not directly link this change to #oscarssowhite, it emphasizes that the perception of racial representation is dynamic across time periods. Providing argumentation that solely looking at the characteristics of a film limits a study.

What is particularly promising is that the findings of the second study align with those of Malik et al. (2021), who used screen time as a proxy for racial representation. Their dataset, covering the period from 2007 to 2019, closely mirrors the observed correlations in this study's interaction term. When using International Rent (IntRent) as their dependent variable, they found that films featuring more East and South Asian characters were significantly more likely to generate higher IntRent compared to films with white characters. Conversely, films with Black and African American and Latin characters were significantly less likely to achieve higher IntRent compared to films with white characters. These findings mirror the relationships observed in the present research.

This study also is align with the findings of Lazar and others (2020) who identified 2015 -2019 as a heightened period of focus on racial AIR in films. This could have influenced the types of contexts in which these actors were portrayed. Alternatively, these films might have garnered more critical acclaim after #OscarsSoWhite, drawing in a broader audience seeking out diverse cinematic experiences. However, the shifting appreciation for racial representation over time raises questions about the underlying motivations driving audience engagement. If external factors can influence our preferences, it challenges the idea that our desire to see ourselves represented on screen is solely based on internal factors, as proposed by racial identity theory.

**Final note**

Interestingly, the comparison of studies reveals a positive correlation between Long-term Audience Engagement and the total number of actors, a metric that doesn't account for authenticity. However, studies employing a more inclusive metric exhibited negative correlations. While this suggests the need to scrutinize authenticity in the context of racial diversity, it also challenges the notion that authenticity is universally crucial for positive effects of racial representation.

**Theoretical implications.**

The study's critical evaluation of Kuppuswamy and others(2016)'s and Kim and others (2020) research prompts a re-examination of the suitability of simplistic metrics like actor count in capturing the intricacies of ethnic representation. The transition from mere actor counts to the more refined tool aligns with the theoretical underpinnings of authenticity in diversity and racial portrayals.

This study introduces a novel metric for assessing racial representation. It builds upon existing research by demonstrating how AIR can be quantified. While AIR was previously applied to studies of gender representation, this study demonstrates its applicability to racial representation as well. The process of introducing this metric and conducting an analysis unveiled a significant trade-off in the realm of racial representation metrics: the balance between scalability and comprehension. This trade-off is evident when comparing this study's metric to those employed by Lazar and others. (2021) and Malik and others (2021). While their metrics arguably provide a more comprehensive assessment of racial representation, they have only been applied to a small number of films (under 100). This suggests limitations in scalability. This scalability issue is further highlighted by the observation that nearly half of the films in the dataset used for the studies of this paper were unable to be processed due to the lack of a suitable SRT file.

Moreover, to gain a thorough understanding of racial representation and its impact on film Long-term Audience Engagement, it's crucial to consider external factors like cultural shifts and societal attitudes. The study's findings, revealing a declining negative correlation for Black representation and the emergence of a positive relationship for Asian representation, underscore the need for a broad approach to studying this connection. This goes past looking at film characteristics and also take into account the effect of potential external factors.

This study offers fresh theoretical insights into measuring film success. It demonstrates that Long-term Audience Engagement, as reflected by ranking on IMDB, can be used as a viable indicator of a film's success. While the study doesn't specifically compare its effectiveness to other measurements, it successfully captures a film's enduring Long-term Audience Engagement over time. Moreover, it shows that IMDB rankings can interact with established factors affecting a film's success and has some correlation with financial achievements.

In summary, this research underscores the necessity for a more nuanced and context-specific comprehension of ethnic representation in films. It challenges established theories by emphasizing the dynamic nature of the relationship, the significance of authenticity in measurement, and the potential influence of external factors on the connection between racial representation and audience engagement.

**Managerial implications.**

The findings of this study hold significant implications for decision-makers in the film industry, emphasizing the importance of a nuanced and forward-thinking approach to racial representation. Notably, this study introduces a novel metric for measuring film success, offering an alternative to the limitations of box office revenue. This metric, based on Long-term Audience Engagement on IMDB, captures the enduring impact of films. While there are many films created primarily for financial gain, which can be better measured through box office revenue, Long-term Audience Engagement can be a useful metric for films that aim to create a lasting impact on audiences and foster meaningful engagement.

The authentic inclusive representation (AIR) metric, introduced in this research, presents a valuable tool for measuring racial representation in a large body of films. While not as exhaustive as some studies, the AIR metric provides concrete measurements based on quantitative data rather than relying on subjective evaluations. Its simplicity and adaptability make it a highly practical tool for industry professionals, enabling them to assess the racial diversity of a vast array of films. However, the effects observed in this study regarding racial representation were relatively small, raising questions about its managerial implications for film success.

The temporal dimension of the study's findings underscores the fluidity of audience perceptions regarding racial representation. The study's longitudinal analysis revealed a change in the association between racial representation and audience engagement following the #oscarssowhite movement, suggesting a dynamic evolution of societal attitudes towards representation over time. This evolving dynamic underscores the importance of filmmakers adopting a long-term perspective, continuously monitoring trends and adapting casting strategies accordingly.

Finally, industry professionals must recognize that different metrics can lead to varying outcomes. Therefore, careful consideration should be given to the metrics employed to assess racial representation. Recognizing the limitations of exclusively relying on the total number of actors from a specific ethnicity is essential. A more expansive view of racial diversity should be embraced, encompassing the entire cast and utilizing metrics like AIR to ensure more inclusive representation, as they encompass a broader understanding of ethnic representation.

## Limitations and suggestions for further research

Despite its significant contributions, this study has certain limitations. Notably, the study did not account for the impact of a film's initial success on its Long-term Audience Engagement. Early success, often driven by factors like marketing strategies, can have a lingering influence on audience engagement. Incorporating a model that considers the cumulative effect of past marketing efforts could have improved the study's ability to isolate the effects of ethnic representation on long-term success.

The study acknowledged the negative impact of sequels on Long-term Audience Engagement, suggesting the possibility of audience fatigue or a decline in quality with successive installments. However, the analysis failed to consider the specific release time of sequels. Analyzing whether a sequel was released in the third year, which serves as the dependent variable, could significantly enhance a film's value. A more nuanced understanding of the effects of sequels could have been obtained by considering their temporal alignment with the third year measurement.

Moreover, as discussed previously, there exists a trade-off between the inclusivity of an AIR metric and its scalability. Lazar et al. (2021) proposed a fourth step to the Bechdel test, which involves evaluating whether conversations between two characters of the same ethnicity without a white character are free from stereotypes, either negative or positive. This step was excluded from this thesis due to time constraints, but it is feasible to automate using techniques similar to identifying 'stop words' based on specific ethnicities. Determining whether the conversation actually perpetuates stereotypes is a recommended step for future research.

Which was also explored by Lazar and other (2021) and not in this study was an interaction effect between a non-film characteristic and AIR, it's also plausible that interactions exist between AIR and film characteristics. Lazar et al. (2021) found that AIR positively influences the box-office of large-budget films. This suggests that other film characteristics might also impact the relationship between AIR and Long-term Audience Engagement, a possibility that wasn't explored in this study.

The study identified the challenges posed by data limitations, emphasizing the need for strategic decision-making in the absence of comprehensive data. The paucity of information could have resulted in exclusions or simplifications that might have affected the study's generalizability and depth, particularly in the second study's operationalization of culture. The constraints imposed by data availability reinforce the importance of exercising caution when drawing overarching conclusions and underscore the need for future research to address the gaps in our current knowledge base.

For instance, further research is needed to unravel the underlying reasons for the persistent negative association observed for Black representation. Promising avenues for exploration include examining the influence of 'overrepresentation' concerns and 'racial agenda' perceptions. Furthermore, examining the contexts in which Black ethnicities are represented compared to other ethnicities merits further study.

Finally, to further enhance our understanding of racial representation in films, I propose exploring the application of social network analysis as a sophisticated method to measure the centrality of characters. This approach leverages neural networks to capture the dynamic and intricate relationships between characters, providing a more nuanced and comprehensive assessment of racial representation. While the AIR metric has proven valuable, social network analysis offers a distinct perspective, enabling us to delve deeper into the impact of each character on the narrative and overall engagement.

Finally, if this study offers evidence indicating a preference for racially diverse films. It is highly probable heterogeneity in this preference exists among people. Establishing a general preference for racial diversity would pave the way for future research to explore individual differences in this preference. The findings from these studies could provide information with regards to the development of film recommendation systems.

Especially for the Black ethnicity films further research is warranted to delve into the reasons behind the consistent negative relationship**.** Some suggestions for this relationship might be indicated byBlack actors being the only minority group that showed overrepresentation in the amount of actors. Causing that this is sheer number of Black characters introduced in these films could be perceived as tokenism, making audiences feel like the characters are being forced into the film rather than being an organic part of the story, indicating ‘racial agenda’. Which negatively influences a film's potential success. Nevertheless, this is outside the scope of this study.

# Bibliography

Abraham, L. &. (2006). Framing news stories: The role of visual imagery in priming racial stereotypes. *Howard Journal of Communications, 17*(3), 183–203. doi:https://doi.org/10.1080/10646170600829584

Agarwal, A. &. (2015). Key Female Characters in Film Have More to Talk About Besides Men: Automating the Bechdel Test. *Conference: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. doi:10.3115/v1/N15-1084

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

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, 67*(4), 103–117. doi:10.1509/jmkg.67.4.103.18692

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

Broersma, M. (2019). Audience Engagement. *The International Encyclopedia of Journalism*, 1-6. 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

Castañeda, M. (2017). *Challenging Inequalities: Readings in Race, Ethnicity, and Immigration.*

Dalisay, F. &. (2009). Assimilation and Contrast Effects in the Priming of Asian American and African American Stereotypes through TV Exposure. *Journalism & Mass Communication Quarterly, 86*(1), 7–22. doi:https://doi.org/10.1177/107769900908600102

Dan Rubin, I. M. (2022). Beyond the box office: A conceptual framework for the drivers of audience engagement,. *Journal of Business Research, 151*(3), 473-488. doi:10.1016/j.jbusres.2022.07.021

Darnell Hunt, A.-C. R. (2021). *Hollywood Diversity Report 2021.* UCLA.

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

Eliashberg, J. H. (2014). Assessing Box Office Performance Using Movie. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 26*(11). doi:10.1109/TKDE.2014.2306681

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

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. doi:10.1037/aca0000363

Hunter, S. D. (2016)). Predicting Box Office from the Screenplay: A Text Analytical Approach. *Film eJournal - Forthcoming*. doi:10.1386/JOSC.7.2.135\_1

Hurley, R. J. (2015). Viewer Ethnicity Matters: Black Crime in TV News and Its Impact on Decisions Regarding Public Policy. *Journal of Social Issues, 71*(1), 155–170. doi:https://doi.org/10.1111/josi.12102

Jan-Benedict E. M. Steenkamp, I. G. (2013). Manufacturer and Retailer Strategies to Impact Store Brand Share: Global Integration, Local Adaptation, and Worldwide Learning. *33*(1), 6-26. doi:https://doi.org/10.1287/mksc.2013.0801

Jay 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:10.1007/BF01080227

Jorge Cruz-Cárdenas, E. Z.-L.-F.-G. (2021). COVID-19, consumer behavior, technology, and society: A literature review and bibliometric analysis,. *Technological Forecasting and Social Change, 173*. doi:https://doi.org/10.1016/j.techfore.2021.121179.

K. Fröber, R. T. (2019). In the dark cube: Movie theater context enhances the valuation and aesthetic experience of watching films. *Psychology of Aesthetics, Creativity, and the Arts*. doi:10.1037/aca0000295

Katherine Aumer, D. B. (2017). Assessing Racial Preferences in Movies: The Impact of Mere-Exposure and Social Identity Theory. *Psychology, 8*(9), 1314-132. doi:10.4236/psych.2017.89085

Katherine L. Neff, D. S. (2023). *Inequality Across 1,600 Popular Films: Examining Gender, Race/Ethnicity & Age of Leads/Co Leads From 2007 to 2022.* University of Southern California, USC Anneberg Inclusion Initiative. Retrieved from https://annenberg.usc.edu/sites/default/files/2019/02/11/MDSCI\_Inclusion\_in\_Netflix\_2.11.19.pdf

Kidd, M. A. (2015). Archetypes, stereotypes and media representation in a multi-cultural society. *Procedia, 236*, 25-28. doi:https://doi.org/10.1016/j.sbspro.2016.12.007

Kim, M. &.-B. (2020). Hollywood’s Global Expansion and Racialized Film Industry. *Humanity & Society, 44*(1), 37-66. doi:https://doi.org/10.1177/0160597619832045

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.

Krushikanth R. Apala, M. J.-C. (2013). Prediction of movies box office performance using social media. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* doi:10.1145/2492517.2500232

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

Kuppuswamy, V. &. (2016). Testing the theory of consumer discrimination as an explanation for the lack of minority hiring in Hollywood films. *Management Science, 66*(3). doi:10.1287/mnsc.2018.3241

Lazar, L. H. (2020). *Driving box office performance through authentically inclusive storytelling.* UCLA, The Center for Scholars & Storytellers. Retrieved from https://www.fullstoryinitiative.com/Full\_Story\_Research.pdf

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

Mastro, D. B.-M. (2007). The cultivation of social perceptions of Latinos: A mental models approach. *Media Psychology, 9*(2), 347-365. doi:https://doi.org/10.1080/15213260701286106

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

Michel Clement, S. W. (2014). Empirical generalizations of demand and supply dynamics for movies. *International Journal of Research in Marketing*, 207-223. doi:10.2139/ssrn.2359966

Mohammed El Hazzouri, L. K. (2019). Why Us?! How Members of Minority Groups React to Public Health Advertisements Featuring Their Own Group. *Journal of Public Policy & Marketing , 38*(3). doi:https://doi.org/10.1177/0743915619846555

MPA. (2022). *Theatrical and Home Entertainment Market Environment Report.*

Musa Malik, F. R. (2021, February). Representations of Racial Minorities in Popular Movies, 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

Partha Chakraborty, S. R. (2019). Movie Success Prediction using Historical and Current Data Mining. *International Journal of Computer Applications, 178*(47), 1-5. doi:10.5120/ijca2019919415

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, 18*(3), 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 Movies With Diverse Casts.

Roberts, J. (2021). Empathy Cultivation through (Pro)Social Media: A Counter to Compassion Fatigue. *Journal. Media, 2*(4), 819-829. doi:https://doi.org/10.3390/journalmedia2040047

Smelser, N. W. (2001). *America Becoming: The Growing Complexity of America's Racial Mosaic.* RAND Corporation.

Smith, S. L. (2020). *Inequality in 1,300 popular films: Examining portrayals of gender, race/ethnicity, LGBTQIA, and disability from 2007–2019.* University of Southern California , USC Annenberg Inclusion Initiative.

Sperling, N. (2021, March). Hollywood Loses $10 Billion a Year Due to Lack of Diversity, Study Finds.

Sung Moon Bae, S. C. (2014). Utilization of Demographic Analysis with IMDB User Ratings on the. *The Journal of Society for e-Business Studies, 19*(3), 125-141. doi:http://dx.doi.org/10.7838/jsebs.2014.19.3.125

Tan, E. S. (2018). A psychology of the film. *Humanities and Social Sciences Communications, 4*. doi:10.1057/s41599-018-0111-y

Teresa Correa, S. H. (2011). RACE AND ONLINE CONTENT CREATION. *Information Communication and Society*. doi:10.1080/1369118X.2010.514355

Thorsten Hennig-Thurau, M. B. (2007). Determinants of motion picture box office and profitability: an interrelationship approach. *Review of Managerial Science*, 65-92. doi:10.1007/s11846-007-0003-9

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. *Child Development, 85*, 21-39. doi:10.1111/cdev.12196.

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

"the", "and", "a", "an", "man", "woman", "girl", "boy", "person",

"doctor", "nurse", "teacher", "student", "police", "officer",

"captain", "detective", "driver", "waiter", "waitress",

"bartender", "chef", "pilot", "scientist", "engineer",

"professor", "lawyer", "firefighter", "writer", "artist",

"singer", "dancer", "actor", "actress", "model",

"president", "king", "queen", "prince", "princess",

"doctor", "sir", "madam", "lord", "lady",

"mr", "mrs", "ms", "miss", "(uncredited)", "self", "barman", "mom", "dad",

"teammate", "cop", "director", "character", "role", "extra", "child",

"player", "customer", "passenger", "soldier", "villager", "audience",

"guest", "customer", "victim", "stranger", "citizen", "neighbor",

"client", "assistant", "host", "companion", "artist", "musician", "performer",

"guest", "teacher's pet", "bodyguard", "consultant", "helper", "coworker",

"bystander", "partygoer", "attendee", "athlete", "official", "associate",

"patient", "driver", "veterinarian", "coach", "driver", "employee", "scientist",

"assistant", "stranger", "examiner", "salesperson", "waitstaff", "witness",

"shooter", "victim", "spy", "veteran", "trader", "guard", "hiker", "shopper",

"actor", "writer", "adventurer", "explorer", "admiral",

"professor", "conductor", "fireman", "sheriff", "reporter", "editor",

"thief", "student", "resident", "colleague", "reporter", "buyer", "driver",

"rider", "leader", "teacher", "hunter", "butler", "stewardess", "crew",

"manager", "waitstaff", "server", "waiter", "waitress", "administrator",

"scientist", "technician", "fighter", "runner", "visitor", "pedestrian",

"stranger", "witness", "operator", "instructor", "inspector", "spectator",

"visitor", "spectator", "celebrity", "policeman", "housekeeper", "housewife",

"househusband", "deliveryman", "scientist", "guide", "designer", "commander",

"commando", "cowboy", "scientist", "scholar", "gentleman", "nanny", "professor",

"missionary", "nun", "missionary", "chaplain", "coach", "airman", "admiral",

"navigator", "soldier", "fighter", "nobleman", "servant", "neighbor", "sailor",

"secretary", "assistant", "professor", "clerk", "nobleman", "parolee", "neighbor",

"kidnapper", "mentor", "nurse", "reporter", "maid", "student", "scientist",

"scholar", "trainee", "supervisor", "official", "deliverywoman", "reporter",

"neighbor", "lawyer", "assistant", "reader", "guest", "witness", "villager",

"visitor", "crew", "engineer", "guest", "helper", "professor", "villager",

"guest", "team", "scientist", "student", "assistant", "chef", "waiter",

"waitress", "restaurantgoer", "attendant", "interviewer", "officer", "professor",

"assistant", "administrator", "assistant", "mentor", "assistant", "escort",

"client", "professor", "manager", "celebrity", "photographer", "tourist",

"presenter", "listener", "expert", "scientist", "doctor", "professor",

"philosopher", "veterinarian", "wife", "homemaker", "househusband",

"housekeeper", "neighbor", "guest", "stranger", "assistant", "teacher",

"student", "servant", "companion", "assistant", "professor", "shopper",

"customer", "bystander", "visitor", "neighbor", "colleague", "assistant",

"seller", "customer", "pedestrian", "passerby", "tourist", "clerk", "singer",

"scientist", "actor", "detective", "professor", "chef", "photographer",

"model", "driver", "manager", "assistant", "attendant", "professor",

"guest", "client", "student", "visitor", "stranger", "neighbor", "witness",

"driver", "athlete", "reporter", "official", "scientist", "shopper",

"customer", "clerk", "manager", "employee", "guest", "assistant", "teacher",

"student", "scientist", "expert", "host", "participant", "judge", "lawyer",

"scientist", "professor", "scientist", "student", "assistant", "singer",

"dancer", "model", "reporter", "scientist", "professor", "attendant",

"professor", "guest", "audience", "technician", "athlete", "clerk", "professor",

"consultant", "buyer", "tourist", "shopper", "client", "customer", "expert",

"scientist", "performer", "actor", "actress", "student", "director", "waitstaff",

"bartender", "actor", "actress", "model", "singer", "dancer", "musician",

"athlete", "attendant", "manager", "agent", "waitstaff", "assistant", "clerk",

"seller", "guest", "customer", "professor", "client", "designer", "expert",

"expert", "participant", "scientist", "speaker", "professor", "manager",

"operator", "technician", "bartender", "bystander", "assistant", "observer",

"producer", "villian", "hero", "sidekick", "mentor", "architect", "on",

"#1", "#2","#3", '#4', '#5' , "commentator", "comedian", "in", "surgeon",

"host", "contestant", "reporter","executive", "(voice)", "guy", "receptionist", "tech", "anchor", "janitor", "ranger", "1", "2", "3", "4", "worker", "member", "banker", "stockbroker", "paperboy", "footman", "daughter", "principal", "school", "friend", "addict", "mechanic", "traveler", "officer", "chief", "owner", "uncle",

"deputy","aunt", "security" , "woman", "brother", "female", "priest", "walker", "thug", "master", "king", "girl", "vendor", "russian", "minister", "baker", "baby", "boss", "reporter", "husband", "mayor", "dr", "father", "mother",

"announcer", "sister", "girlfriend", "senator", "band", "party", "club", "teen", "teenage", "teenager" "gardner","archer", "warrior", "fbi", "teller", "councilman", "prostitute", "fan", "salesman", "attorney", "grandma", "grandpa", "footage)", "dealer", "cashier", "dj", "farmer", "german", "newscaster", "chinese", "dog", "cat", "cheerleader"

"trainer", "co-worker", "partner", "stewart", "restaurant", "translator", "queen", "with", "quarterback", "paramedic"

"quarterback", "paramedic", "veteran", "villager", "plumber", "electrician",

"mechanic", "butcher", "librarian", "historian", "accountant", "banker", "clerk",

"cashier", "waiter", "waitress", "hostess", "host", "presenter", "journalist",

"anchor", "receptionist", "janitor", "worker", "employee", "executive", "owner",

"manager", "partner", "assistant", "supervisor", "consultant", "scientist",

"researcher", "technician", "designer", "architect", "engineer", "doctor", "nurse",

"therapist", "patient", "pharmacist", "student", "teacher", "professor", "scholar",

"lawyer", "attorney", "judge", "jury", "witness", "reporter", "detective",

"officer", "sergeant", "captain", "lieutenant", "commander", "director", "producer",

"actor", "actress", "musician", "singer", "dancer", "artist", "model", "athlete",

"coach", "player", "character", "hero", "villain", "sidekick", "comedian", "commentator",

"host", "contestant", "participant", "audience", "listener", "viewer", "reader",

"spectator", "guest", "fan", "customer", "client", "shopper", "buyer", "seller",

"vendor", "consumer", "passerby", "pedestrian", "tourist", "traveler", "driver",

"rider", "pilot", "navigator", "adventurer", "explorer", "admiral", "sheriff",

"thief", "fireman", "paramedic", "surgeon", "doctor", "nun", "missionary",

"chaplain", "missionary", "gentleman", "lady", "sir", "madam", "lord", "woman",

"man", "girl", "boy", "child", "baby", "infant", "toddler", "teen", "teenager",

"elder", "senior", "junior", "boss", "manager", "owner", "chief", "deputy",

"sarge", "capt", "sergeant", "colleague", "associate", "partner", "expert",

"specialist", "professional", "novice", "amateur", "enthusiast", "expertise",

"authority", "expert", "guru", "entertainer", "performer", "singer", "dancer",

"musician", "artist", "comedian", "actor", "actress", "model", "character",

"role", "persona", "identity", "individual", "personality", "citizen", "resident",

"inhabitant", "neighbor", "neighborhood", "community", "colony", "society",

"culture", "civilization", "world", "planet", "earth", "universe", "cosmos",

"galaxy", "star", "celestial", "space", "astronomy", "galactic", "extraterrestrial",

"alien", "creature", "being", "entity", "thing", "object", "item", "element",

"substance", "material", "matter", "stuff", "commodity", "product", "good",

"merchandise", "service", "assistance", "help", "aid", "support", "benefit",

"advantage", "profit", "gain", "value", "worth", "quality", "feature", "attribute",

"trait", "characteristic", "property", "possession", "ownership", "possession",

"belonging", "object", "thing", "item", "artifact", "creation", "invention",

"innovation", "discovery", "finding", "uncovering", "revelation", "disclosure",

"knowledge", "wisdom", "insight", "information", "data", "fact", "reality",

"truth", "certainty", "certitude", "surety", "confidence", "assurance", "guarantee",

"promise", "commitment", "responsibility", "obligation", "duty", "requirement",

"necessity", "essential", "must", "need", "want", "desire", "wish", "hope", "dream",

"ambition", "goal", "objective", "purpose", "intention", "plan", "strategy",

"tactic", "method", "approach", "technique", "way", "means", "mode", "manner",

"fashion", "style", "form", "expression", "communication", "message", "idea",

"thought", "notion", "concept", "perception", "understanding", "cognition",

"knowledge", "wisdom", "insight", "intelligence", "genius", "brilliance", "creativity",

"imagination", "innovation", "invention", "discovery", "inspiration", "influence",

"impact", "effect", "outcome", "result", "consequence", "significance", "importance",

"relevance", "meaning", "value", "worth", "quality", "feature", "characteristic",

"trait", "aspect", "dimension", "factor", "element", "component", "ingredient",

"constituent", "part", "portion", "segment", "section", "piece", "fragment",

"bit", "particle", "molecule", "atom", "microscopic", "subatomic", "infinitesimal",

"imperceptible", "unobservable", "indistinguishable", "invisible", "hidden",

"secret", "mysterious", "unknown", "obscure", "ambiguous", "unclear", "uncertain",

"(segment", "dentist", "pharmacist", "veterinarian", "plumber", "electrician",

"carpenter", "mechanic", "pilot", "engineer", "architect", "designer",

"artist", "musician", "athlete", "coach", "sir", "madam", "lord", "lady",

"king", "queen", "prince", "princess", "duke", "duchess", "mother", "father",

"daughter", "son", "brother", "sister", "aunt", "uncle", "cousin", "niece",

"nephew", "grandmother", "grandfather", "robot", "alien", "monster", "creature",

"spirit", "ghost", "vampire", "witch", "wizard", "sorcerer", "fairy", "elf", "double",

"dentist", "pharmacist", "veterinarian", "plumber", "electrician",

"carpenter", "mechanic", "pilot", "engineer", "architect", "designer",

"artist", "musician", "athlete", "coach", "sir", "madam", "lord", "lady",

"king", "queen", "prince", "princess", "duke", "duchess", "mother", "father",

"daughter", "son", "brother", "sister", "aunt", "uncle", "cousin", "niece",

"nephew", "grandmother", "grandfather", "robot", "alien", "monster", "creature",

"spirit", "ghost", "vampire", "witch", "wizard", "sorcerer", "fairy", "elf",

"god", "goddess", "deity", "myth", "legend", "hero", "heroine", "warrior",

"saint", "martyr", "historical", "mythical", "legendary", "ancient", "wise",

"sage", "prophet", "messenger", "angel", "demon", "creature", "beast", "minion",

"avatar", "monarch", "sovereign", "emperor", "empress", "dictator", "president",

"chancellor", "prime", "minister", "mayor", "governor", "senator", "congressman",

"congresswoman", "representative", "ambassador", "diplomat", "envoy", "delegate",

"official", "officer", "commander", "admiral", "captain", "lieutenant", "sergeant",

"private", "colonel", "general", "marshal", "sheriff", "deputy", "officer",

"agent", "inspector", "detective", "investigator", "trooper", "sergeant", "captain",

"officer", "agent", "inspector", "detective", "investigator", "trooper", "sergeant",

"captain", "soldier", "fighter", "warrior", "combatant", "brawler", "champion",

"winner", "loser", "villain", "sidekick", "henchman", "assistant", "partner",

"companion", "friend", "foe", "enemy", "rival", "opponent", "stranger", "neighbor",

"citizen", "resident", "vagrant", "wanderer", "traveler", "tourist", "explorer",

"adventurer", "pioneer", "settler", "migrant", "nomad", "drifter", "visitor",

"spectator", "audience", "viewer", "listener", "watcher", "reader", "auditor",

"reviewer", "critic", "commentator", "comedian", "host", "presenter", "interviewer",

"moderator", "panelist", "expert", "analyst", "commentator", "reporter", "journalist",

"correspondent", "photographer", "cameraman", "director", "producer", "editor",

"executive", "manager", "administrator", "coordinator", "supervisor", "foreman",

"boss", "leader", "chief", "owner", "proprietor", "landlord", "landlady",

"tenant", "resident", "guest", "visitor", "client", "customer", "shopper",

"buyer", "seller", "merchant", "salesman", "clerk", "cashier", "teller",

"baker", "butcher", "grocer", "chef", "waiter", "waitress", "bartender", "barista",

"hostess", "steward", "stewardess", "attendant", "server", "employee", "worker",

"laborer", "technician", "engineer", "scientist", "researcher", "scholar",

"professor", "teacher", "instructor", "lecturer", "educator", "tutor", "trainer",

"coach", "mentor", "advisor", "counselor", "therapist", "psychologist", "psychiatrist",

"doctor", "physician", "nurse", "pharmacist", "therapist", "patient", "client",

"resident", "inmate", "prisoner", "detainee", "hostage", "victim", "survivor",

"sufferer", "patient", "witness", "bystander", "spectator", "onlooker", "observer",

"participant", "contestant", "competitor", "candidate", "nominee", "winner",

"loser", "champion", "challenger", "challenger", "contender", "opponent", "rival",

"opponent", "adversary", "foe", "enemy", "opponent", "combatant", "athlete",

"runner", "jogger", "swimmer", "cyclist", "athlete", "participant", "competitor",

"winner", "loser", "champion", "medalist", "record", "holder", "challenger",

"contender", "favorite", "underdog", "outsider", "front-runner", "candidate",

"nominee", "competitor", "contender", "champion", "medalist", "record", "holder",

"challenger", "contender", "favorite", "underdog", "outsider", "front-runner",

"candidate", "nominee", "player", "contestant", "competitor", "challenger",

"champion", "finalist", "sem", "dude"

Appendix B: Subtitle files parching

I cleaned the subtitle files and applied a logic for scene identification. In these subtitle files, when it is 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 threshold of 0.85, aligning the character list with their respective ethnicities within the dataset.

Appendix C : Dummy variables descriptives

Tables 4, 5, and 6 present the frequency statistics for the control dummy variables for year, genre, seasonality, and the 'based-on' variables. These tables provide the number of observations per variable (N) and the percentage this dummy variable represents compared to the total number of 1105 observations.

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 | 489 | 42.30 |  | Fantasy | 160 | 13.84 |
| Thriller | 430 | 37.20 |  | Family | 115 | 9.95 |
| Comedy | 428 | 37.02 |  | Biography | 74 | 6.40 |
| Action | 408 | 35.29 |  | Sport | 50 | 4.33 |
| Adventure | 253 | 21.89 |  | History | 44 | 3.81 |
| Romance | 236 | 20.42 |  | Music | 43 | 3.72 |
| Crime | 220 | 19.03 |  | War | 42 | 3.63 |
| Sci.Fi | 202 | 17.47 |  | Musical | 18 | 1.56 |
| Mystery | 181 | 15.66 |  | Western | 10 | 0.87 |
| Horror | 162 | 14.01 |  | Documentary | 5 | 0.43 |

Table 6 : Remaining dummies

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Season** | **N** | **Percentage** |  | **Variable** | **N** | **Percentage** |
| Fall | 311 | 26.90 |  | Based on novel | 190 | 16.44 |
| Summer | 304 | 26.30 |  | Based on comic book | 75 | 6.49 |
| Spring | 301 | 26.03 |  | Based on book | 56 | 4.84 |
| Winter | 240 | 20.76 |  | Spinoff | 41 | 3.55 |
|  |  |  |  | Based on book series | 23 | 1.99 |
|  |  |  |  | Based on short story | 8 | 0.69 |
|  |  |  |  | Based on play | 7 | 0.61 |

Appendix D : Protests Oscars so white

A person holding a microphone and holding a poster

Description automatically generated

New York times article:

A black and white cover with text

Description automatically generated

Can be found at <https://www.nytimes.com/2020/02/06/movies/oscarssowhite-history.html>

Article from the university of Berkeley

A screenshot of a movie

Description automatically generated

Can be found at <https://greatergood.berkeley.edu/article/item/diverse_films_make_more_money_at_the_box_office>

Appendix E: Racial diversity through Simpson diversity index

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 ni(ni-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. These were the results of the regression for clarity I only added the significant coefficients. No violations of the assumptions of a regression analysis were detected.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Variable* | *Coefficient* | *Sign* |  | Variable | *Coefficient* | *Sign* |
| *Inverse Simpson index* | 0.268 | < 0.001 \*\*\* |  |  |  |  |
|  |  |  |  |  |  |  |
| *Log(Screens)* | -0.461 | < 0.001 \*\*\* |  | *Biography* | 0.163 | 0.006 \*\* |
| *Log(Runtime)* | -1.351 | < 0.001 \*\*\* |  | *Sport* | 0.136 | 0.025 \* |
| *Log(MPAA)* | -0.711 | < 0.001 \*\*\* |  | *War* | 0.143 | 0.043 \* |
| *Log(Average budget)* | -0.047 | 0.008 \*\* |  | *Family* | 0.208 | < 0.001 \*\*\* |
| *Log(Sequal)* | 0.006 | < 0.001 \*\*\* |  | *Musical* | -0.292 | 0.007 \*\* |
| *Based on book* | -0.132 | 0.034 \* |  | *History* | 0.159 | 0.028 \* |
| *Based on play* | 0.375 | 0.012 \* |  | *Documentary* | 0.749 | < 0.001 \*\*\* |
| *Based on comic book* | -0.314 | < 0.001 \*\*\* |  | *Western* | 0.492 | < 0.001 \*\*\* |
| *Based on novel* | -0.108 | 0.002 \*\* |  | *2009* | 0.172 | 0.023 \* |
| *Log(Nominee)* | -0.009 | < 0.001 \*\*\* |  | *2010* | 0.201 | 0.009 \*\* |
| *Log(Winner)* | -0.014 | < 0.001 \*\*\* |  | *2011* | 0.351 | < 0.001 \*\*\* |
| *Log(Director Power)* | -0.010 | < 0.001 \*\*\* |  | *2012* | 0.502 | < 0.001 \*\*\* |
| *Log(Metascore)* | -0.036 | < 0.001 \*\*\* |  | *2013* | 0.617 | < 0.001 \*\*\* |
| *Log(Star power)* | -0.004 | 0.032 \*\* |  | *2014* | 0.791 | < 0.001 \*\*\* |
| *Adventure* | -0.135 | < 0.001 \*\*\* |  | *2015* | 0.791 | < 0.001 \*\*\* |
| *Comedy* | 0.130 | < 0.001 \*\*\* |  | *2016* | 0.879 | < 0.001 \*\*\* |
| *Drama* | 0.149 | < 0.001 \*\*\* |  | *2017* | 0.823 | < 0.001 \*\*\* |
| *Romance* | -0.085 | 0.013 \* |  | *2018* | 1.021 | < 0.001 \*\*\* |
| *Sci.Fi* | -0.109 | 0.006 \*\* |  | *2019* | 0.968 | < 0.001 \*\*\* |
| *R2* | 0.626 |  |  | *Adj R2* | 0.617 |  |
| *F Statistic* | 65.980\*\*\* | (df = 60; 2364) |  |  |  |  |

1. While it is recognized that this study focuses on three ethnicities (Black, Hispanic, and Asian) due to the limitations of the software programs used (Rethnicity and Kairos), it is important to acknowledge that there are many nuances in ethnicity. [↑](#footnote-ref-1)
2. The filtering process showed that imageless characters were also more likely to play non named characters. After the filter for generic names, 83% 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, imageless people are also more likely to play characters that were not authentic and inclusive (AIR) with regards to the criteria used in this study. This suggests that the decision to use Rethenicity for the imageless people is less troublesome due to these actors probably not being important for the analysis regardless. [↑](#footnote-ref-2)