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

A comprehensive examination of underrepresentation of ethnic minorities in Hollywood films

A logo with blue and yellow squares

Description automatically generated

**Name:** Bastiaan van Houten

**SNR:** 2082816

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

**Academic year:** 2023-2024

**Submission date:** January 5, 2024

**Name supervisor** Dr. Arjen van Lin

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

Management summary

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

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

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

Preface

Table of Contents

[Introduction 5](#_Toc150513286)

[Problem Indication 5](#_Toc150513287)

[Problem Statement 7](#_Toc150513288)

[Academic Contribution 7](#_Toc150513289)

[Managerial Contribution 8](#_Toc150513290)

[Structure of the Thesis 9](#_Toc150513291)

[Literature Review and Hypotheses 9](#_Toc150513292)

[Long-Term Audience Engagement (LTAE) 9](#_Toc150513293)

[Racial Representation in Films 10](#_Toc150513294)

[Authentic Inclusive Representation (AIR) in Racial Representation 13](#_Toc150513295)

[The Relationship of Racial Diversity and Long-Term Audience Engagement (LTAE) **Error! Bookmark not defined.**](#_Toc150513296)

[The Relationship Between Diversity and LTAE with AIR as Mediator 14](#_Toc150513297)

[Conceptual framework 16](#_Toc150513298)

[Research Methodology 16](#_Toc150513299)

[Data sources 16](#_Toc150513300)

[Sample 18](#_Toc150513301)

[Variable operationalization 19](#_Toc150513302)

[Models 27](#_Toc150513303)

[Results 28](#_Toc150513304)

[Descriptive Statistics 29](#_Toc150513305)

[Assumptions of log-log linear and logistic regression model 32](#_Toc150513306)

[Interpreting the results 33](#_Toc150513307)

[Bibliography 47](#_Toc150513308)

# 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; Bunche, 2018; Reporter, 2021). This suggests that racial diversity is both a moral and financial imperative for the film industry.

Efforts to increase diversity and representation in film have led to a significant increase in the proportion of films featuring minority actors. For example, 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). Most studies do not make a distinction between different minority groups and measure diversity based on the share of minorities in the cast. Additionally, some studies focus solely on one minority group. This narrow approach has caused two major issues.

Firstly, the approach of treating minority groups as a homogenous unit has resulted in uneven representation across the different minority groups. For instance, while African-Americans have been overrepresented in films for the past three consecutive years and the Asian community is ‘rightfully’ represented, the Latin community remains severely underrepresented, with 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 (AIR) 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 previous studies such as Kuppuswamy and Younkin (2016) and Weaver (2011) have 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, AIR is an essential factor that warrants consideration for racial diversity analysis.

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

## Problem Statement

Following the problem background this studies problem statement is formulated as follows: “*What is the relationship between racial Authentic Inclusive Representation (AIR) and Long term Audience Engagement (LTAE) 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. 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 (AIR) 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 (LTAE), this study uses a more comprehensive measure of success, as it takes into account factors such as word-of-mouth, social media engagement, and re-watching. This contributes to existing literature because it provides a more nuanced understanding of how audiences are engaging with films with diverse casts. Moreover, this causes that this study in contrast to other studies identifies films that are having a lasting impact on audiences, even if they may not have been box office blockbusters.

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

## 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 intentionally casting diverse roles and developing characters which authentically represent a variety of backgrounds can potentially attract a wider audience, which can boost a film's financial performance.

As the film industry transforms, the traditional reliance on box office revenue diminishes as streaming services surge. Consequently, the industry landscape likely evolves towards a reduced focus on short-term gains, signaling a more prominent role for long-term success. While this study does not delve deeply into whether Long-Term Audience Engagement (LTAE) is a superior metric for long-term financial success, it demonstrates its 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.

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.

## Structure of the Thesis

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

# Literature Review and Hypotheses

This literature review embarks on a comprehensive examination of the intricate relationship between racial diversity in Hollywood films and the concept of LTAE (long-term advertising effectiveness). In order to establish a clear foundation for this exploration, LTAE and racial representation are initially defined, followed by an in-depth discussion of the significance of AIR in ensuring effective racial representation. Subsequently, the existing body of research delving into the relationship between these concepts is thoroughly examined.

## Long-Term Audience Engagement (LTAE)

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

Before delving into the social attribute of LTAE it is important to clarify that LTAE has financial purposes. Similar to box office revenue LTAE is correlated with ancillary revenues such as engrossed viewing, longevity in distribution channels, and the purchase of film-related merchandise (Kumar, 2022). These factors demonstrate a persistent connection between the film and its audience, extending beyond the initial screening. This correlation between LTAE and revenue highlights its significance, surpassing box office revenue as a comprehensive measure of film performance. The social dimensions of LTAE will be explored in detail in the following paragraphs.

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

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

## Representation and racial identity

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

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

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

Our racial identity is not solely a product of our own perspective but is also shaped by how others perceive us (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. Therefore, the powerful impact of media representations on shaping perceptions cannot be underestimated (Kidd, 2015).

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

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

## 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). This suggests that when a film embraces racial diversity, it has the potential to attract a broader audience and foster greater overall LTAE (USC Annenberg, 2018).

Nevertheless, the demand for greater racial diversity in film is not universally embraced. A study by King (2020) identifies a group while not opposing racial representation in film, prioritizes the quality of the storyline. Arguing that when minorities are introduced without developed characters, through tokenism it influences the quality of the storyline.

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).

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

Nevertheless, it is essential to emphasize that the preference for whitewashing 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). The prevalence of whitewashing may be more a product of industry habit than an accurate reflection of audience preferences. Nonetheless, due to the underfunding of diverse films, accurately assessing their potential appeal to audiences becomes a challenging task.

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

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

Finally, it is argued that diverse casts can better reflect the diversity of the real world, which can help viewers to connect with the characters and the story (USC Annenberg, 2018). The general consensus is that increased racial diversity in films has a positive impact on LTAE. Nevertheless, as noted for racial representation to be successful industry forces and tokenism need to be taken into account the latter will be discussed in the next paragraph.

## Authentic Inclusive Representation (AIR) 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, overlooking authenticity in representation can result in films being mistakenly categorized as racially diverse while still perpetuating stereotypes and contributing to marginalization. As discussed previously these stereotypes in the film can perpetuate biases and misconceptions about minority groups (Umaña-Taylor AJ, 2014; Abraham and Appiah, 2006; Hurley et al., 2015) . Therefore, the concept of Authentic Inclusive Representation (AIR) emerges as a crucial element in racial representation.

The pursuit of authenticity in the representation of minorities in films is not a recent development. In 2014, Ralph Roughton stressed that genuine understanding and empathy, free from stereotypes, are key to changing attitudes. Effective representation requires viewers to truly understand and empathize with characters. (Roughton, 2014). AIR, as a concept, describes how accurately and respectfully a film portrays underrepresented groups in a nuanced manner.

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). Notably, Lazar et al. expanded the application of this test to assess ethnic representation in films in 2020. The test comprises three gradations:

(T1) A film must feature a minimum of two named ethnic minority characters.

(T2) Two ethnic minority characters must engage in a conversation.

(T3) This conversation must not revolve around or include a white character.

An example for such a conversation would be a conversation between the main character and his wife in the film ‘12 years a slave’ :

**ANNE**

Solomon...

**SOLOMON**

Come, Anne. Jump.

**ANNE**

I will not ruin my dress. Catch me!

**SOLOMON**

I will catch you, Anne. I will.

**ANNE**

You will.

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

## The Relationship Between Racial AIR and LTAE

There is an intuitive belief that racial diversity leads to authentic inclusive representation (AIR). The rationale behind this notion is straightforward. The more people from different ethnicities are included, the more likely it is that one of them will provide an authentic representation for that ethnicity. Nevertheless, including diverse characters in films is not without its challenges. One concern is tokenism, which occurs when underrepresented characters are introduced without fully developed storylines. This can lead to backlash and damage LTAE if viewers perceive the diversity as being insincere , which can ultimately affect the perceived quality of the film (Smith, 2016). For example, when minorities are introduced through tokenism, it can create the perception of a "racial agenda," which is often disliked by viewers (King, 2020).

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

Furthermore, these negative depictions can cause minority groups to disengage completely from the media. For 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.

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).

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

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 A black and white line with a plus symbol 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. 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

This study aims to investigate the relationship between racial Authentic and Inclusive Representation (AIR) and Long-Term Audience Engagement (LTAE) in Hollywood films. AIR is defined as the creation of characters and narratives that genuinely reflect the experiences and perspectives of marginalized groups. Given the subjective nature of AIR, the Bechdel Test was employed as a quantitative measure for this concept.

The following sections will provide a detailed overview of the data collection and sampling procedures, as well as a clear operationalization of the variables. Notably, AIR will be discussed across three distinct levels [[1]](#footnote-1), as defined by the Bechdel test, elaborated in the operationalization section.

## Data sources

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

*LTAE:* IMDB will be used to measure LTAE. While IMDb users and the general filmgoing population differ somewhat in demographics[[2]](#footnote-2), previous research suggests that IMDb's demographics are representative of the broader filmgoing audience (Ghiassi, 2017; Partha and Chakraborty, 2019; Apala, 2013). Therefore, given its large user base and previous applications in research, I assumed that IMDb data is reliable for producing generalizable and accurate findings for the dependent variable of this research.

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

*AIR determination*: The Bechdel test is an ideal tool for measuring AIR in films because of its adaptability to our specific focus (Lazar, 2020), its potential for automation, and its quantifiability (Argarwal, 2015). Previous research such as Argawal (2015) used screenplays to conduct the Bechdel test to determine the level of AIR in films. 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. I will discuss subtitle processing in the operationalization part.

## Sample

The films for the sample were chosen carefully and based on specific criteria. IMDB started in 1996 so the sample was limited to films released after this year. Nevertheless, the steep increase in ranking which can be seen in Figure 2 is not due to films becoming less popular over time, but rather to the fact that IMDb was not widely used until 2003. Therefore, there was an additional filter for films from those years.

Films produced outside the United States and animated films were also filtered out, because Hollywood is the focus of this study and voice actors are not represented on screen, 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 the criteria of release date (after 2000, elaborated upon in the next section), production country (United States), film type (live-action), and wide release (500 screens), the sample consisted of 2,319 unique films. Subsequently, subtitle files were gathered for this set. Concluding that the final sample consisted of 1,105 films in a processable format, which were then prepared for analysis.

A graph with a line going up

Description automatically generated

Figure 2 Average ranking across films from different years

## 

## Variable operationalization

**Long term audience engagement (DV)**

Long term audience engagement was operationalized using IMDb's MovieMeter, a on a weekly basis metric derived from popularity rankings. A score of one means that 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 LTAE, such as online discussions, reviews, and word-of-mouth conversations. By examining film scores over time, I aimed to measure sustained engagement, where lower scores indicate higher popularity.

Specifically, I measured LTAE as the average popularity 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 popularity. Figure 2 shows the average LTAE scores for films released in different years.

The declining popularity rankings of films over time are a logical consequence of the constant influx of new releases. Each year, the cinematic landscape welcomes new entries, while older films remain, intensifying competition for all films. The rise of online streaming services may have played a role for the rapid increase in ranking (decrease in popularity) between 2011 and 2016 seen in figure 2.[[3]](#footnote-3) Nevertheless, this is speculative, as further research is needed to determine the exact impact of online streaming and the actual reasons behind this steep drop. To control for the trend and the steep drop, year dummies are incorporated into the research.

**Ethnicity**

*Ethnicity determination* The Kairos API is chosen because of its efficiency, accuracy (99.63%), and ability to handle a large dataset (Kairos, 2023). In the films a total of 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 32,513 unique actors and actresses. [[4]](#footnote-4)

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%. Not having an image on IMDB goes co in hand with not having played a significant role in a film.[[5]](#footnote-5) Therefore, these people are less important for the analysis.

Using the Kairos API and Rethnicity, I extracted the ethnicity of each actor or actress in the dataset. The API and package had one challenge: it provides probabilities rather than an assignment that a person belongs to one of the ethnicities. Moreover, as shown in figure 3 the dataset had skewed distribution of probabilities towards White actors.

This is due to the abundance of White actors in the dataset, which creates a lot of low probability values for the other three ethnic groups. Nevertheless, the confidence associated with the different ethnicities did not differ (Asian: 0.9972, Black: 0.9988, Hispanic: 0.9990, White: 0.9985). Therefore, I established a process to assign ethnicity values to the actors and actresses.

The procedure involved assigning a percentage value to the two highest-probability ethnicities based on probability. This calculation involved dividing the probability of the highest-probability ethnicity by the sum of the probabilities of the two highest-probability ethnicities (Highestprob / (Highestprob + SecondHighestprob)). For illustration, if a person's two highest assigned probabilities were 0.53 and 0.19, respectively, the calculation would be (0.53 / (0.53 + 0.19)) = 0.736. If this percentage value exceeded 0.5, the person was classified as belonging to that ethnicity. This procedure successful identified the ethnicity for 55,583 characters.

For individuals who could not be definitively assigned to an ethnicity with the API, it is assumed to be challenging for audiences to associate them with any specific ethnic group as well. And without a clear indication of an individual's ethnicity, audiences may face ambiguity when attempting to make cultural or ethnic associations.

Ambiguous portrayals may lead to varied interpretations or misinterpretations, affecting the overall impact of the character in terms of representation. Due to this these characters were removed. The list of characters not assigned an ethnicity did not include any characters assigned based on their names with Rethnicity.

A graph of different sizes and shapes

Description automatically generated with medium confidence

Figure 3 Distribution probabilities ethnicities

**Authentic inclusive representation (AIR)**

The Bechdel test was modified to measure AIR for the three different ethnicities of this research. This approach enables to look at AIR on different levels. It will be dichotomous variables for all three different levels. I will refer to the specific Bechdel used test for this study further as the reformed Bechdel test. But I will refer to the specific gradient levels (T1, T2, T3) more often:

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

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

(T3) without a white character.

*(T1) Non named character filtering:* The first step in the reformed Bechdel test is to filter out non-named characters. To filter generic characters from the dataset, I identified frequently occurring tokens, where a token is a segment of a name divided by spaces. I then removed names which inly 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 revised Bechdel test using subtitle files. While detailed processing steps are available in Appendix B the following section outlines an overview of this procedure.

*(T2 ,T3) Subtitle file parching:* To assess who speaks to one another (T2) and whether they talk about something besides a white character (T3), I used subtitle files for the hearing impaired. These files can be extracted from a film in the form of .srt files, which are text files with strict formatting. Each subtitle in an .srt file has a unique identifier, precise start and end times, and one or two lines of text. 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 then 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. The variables considered alongside AIR 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, (SEQUALi) represents the number of predecessors for which the sequel is, which is equal to the number of predecessor films plus one. 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 that year. 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 popularity, 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.

Finally, it is crucial for this study to be cautious in every step of the process to possibly achieve causality. Therefore, every factor previously mentioned will also be analyzed on its influence on the racial diversity of the film before it is included as a control variable.[[6]](#footnote-6)

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 popularity score where a higher score means being less popular. | Average ranking over a one year period of time in the third year after release logged. | IMDB |  |
| Covariates |  |  |  |  |
| Log(SEQUALi) | Amount of sequel it is. | Amount of sequel it is logged. | IMDB | KRM, H , KY, CM |
| 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-transformed sum of the top 4 actors, the year of the film 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. In the year of release of the film. | Log-transformed sum of 100 divided through the director ranking + 0.01 (e.g., log(100/rank 2 +0.00)) for the director one year before the film release. | The Numbers | KRM, H, G, KY, CM |
| Log(CRITICi) | Log-transformed average rating the film received from professional film reviewers + 0.01. | Log-transformed average rating the film received from professional film reviewers | Metacritic | H, KRM, KY, G |
| Log(NOMINATIONSi) | Number of award nominations the film received. | Log transformed number of award nominations the film received + 0.01. | IMDb | KY,KRM, HTS |
| Log(WINSi) | Number of award wins the film received. | Log transformed number of award nominations the film received + 0.01. | 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 | Total amount of screens logged | Numbers | H , KY, KRM , HTS, CM |
| Log(BUDGETi) | Production budget of film | The average production budget of the data sources | IMDB, Numbers, OMDB | KY,KRM, HTS, CM |
| Log(RUNTIMEi) | Duration of a film in minutes | Log-transformed duration of a film in minutes | IMDb |  |
| 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 | Spring release = 1, Other = 0 [Mar, Apr, May] | IMDb | V |
| SUMMERi | Film released in the summer | Summer release = 1, Other = 0 [Jun, Jul, Aug] | IMDb | - |
| AUTUMi | Film released in the autumn | Autum release = 1, Other = 0 [Sep, Oct, Nov] | IMDb | - |
| WINTERi | Film released in the winter | Winter release = 1, Other = 0 [Dec, Jan, Feb] | IMDb | - |
| 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 |  |
| YEAR2002i | Year2002 | 2002 = 1, other = 0 | IMDb |  |
| YEAR2003i | Year2003 | 2003 = 1, other = 0 | IMDb |  |
| YEAR2004i | Year2004 | 2004 = 1, other = 0 | IMDb |  |
| YEAR2005i | Year2005 | 2005 = 1, other = 0 | IMDb |  |
| YEAR2006i | Year2006 | 2006 = 1, other = 0 | IMDb |  |
| YEAR2007i | Year2007 | 2007 = 1, other = 0 | IMDb |  |
| YEAR2008i | Year2008 | 2008 = 1, other = 0 | IMDb |  |
| YEAR2009i | Year2009 | 2009 = 1, other = 0 | IMDb |  |
| YEAR2010i | Year2010 | 2010 = 1, other = 0 | IMDb |  |
| YEAR2011i | Year2011 | 2011 = 1, other = 0 | IMDb |  |
| YEAR2012i | Year2012 | 2012 = 1, other = 0 | IMDb |  |
| YEAR2013i | Year2013 | 2013 = 1, other = 0 | IMDb |  |
| YEAR2014i | Year2014 | 2014 = 1, other = 0 | IMDb |  |
| YEAR2015i | Year2015 | 2015 = 1, other = 0 | IMDb |  |
| YEAR2016i | Year2016 | 2016 = 1, other = 0 | IMDb |  |
| YEAR2017i | Year2017 | 2017 = 1, other = 0 | IMDb |  |
| YEAR2018i | Year2018 | 2018 = 1, other = 0 | IMDb |  |
| YEAR2019i | Year2019 | 2019 = 1, other = 0 | IMDb |  |

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

Following the approach outlined by Clement, Wu, and Fischer (2014), this research employs log-log linear regressions. Because of the nature of log-log linear regressions 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 films characters 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

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. After the assumptions of the regression models, including multicollinearity and independence are assessed the multiple regression models and their results are described.

It is crucial to note that, as mentioned previously, popularity represents the average rank of a film in the third year post-release. Therefore, a higher value of the dependent variable signifies lower popularity. Consequently, negative estimates in the results denote positive effects, while positive estimates indicate negative effects on popularity.

## Descriptive Statistics

Table 2 presents descriptive statistics for the continuous 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 | 1105 | 3410 | 4387 | 75.92 | 103633 |
| *Control Variables* |  |  |  |  |  |
| Opening screens | 1105 | 2650.06 | 826.48 | 502 | 4662 |
| Runtime | 1105 | 108.69 | 17.19 | 75 | 219 |
| Budget | 1105 | 54,673,873.16 | 53,393,953.25 | 250,000 | 356,000,000 |
| Nominee | 1105 | 16.26 | 36.30 | 0 | 462 |
| Winner | 1105 | 5.99 | 23.02 | 0 | 490 |
| Director power | 1105 | 0.03 | 0.09 | 0 | 1.0 |
| Star Power | 1105 | 0.06 | 0.15 | 0 | 1.2 |
| Meta score | 1105 | 47.66 | 16.46 | 0 | 96 |
| Sequel | 1105 | 0.56 | 1.94 | 0 | 22 |
| Remake | 1105 | 0.09 | 0.32 | 0 | 3 |
| MPAA | 1105 | 3.24 | 0.69 | 1 | 4 |

The table presented below offers a comprehensive overview of the independent variable at levels T1, T2, and T3, drawing a comparison between AC (All Characters) and NC (Named Characters).

Table 2: Films passing T1, T2 and T3 condition, AC (All Characters) and NC (Named Characters)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ethnicity | AC (T1) | NC (T1) | AC (T2) | NC (T2) | AC (T3) | NC (T3) |
| Black | 901 | 503 (0.53) | 414 | 335 (0.81) | 215 | 206 (0.96) |
| Asian | 748 | 334 (0.39) | 180 | 149 (0.83) | 78 | 77 (0.99) |
| Hispanic | 751 | 326 (0.35) | 240 | 199 (0.83) | 104 | 97 (0.93) |

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

The percentage difference between conversations with all characters (AC) and those with named characters (NC) decreases at each stage from T1 to T3 for every ethnicity. This suggests that, at each step, named characters play a larger role in the observed films passing a consideration. In other words, when a conversation meets the criteria of (T2) and/or (T3), these conversations are typically done by named characters. Showcasing there might be some connection between these type of conversations and named characters.

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 |

## Assumptions of log-log regression model

The log-log model rests on four specific assumptions: no multicollinearity, homoscedasticity, independence, and normality.

*Normality:* Non-normality in a regression model refers to a violation of the assumption that the residuals (the differences between observed and predicted values) are normally distributed, potentially affecting the accuracy and reliability of statistical inferences. Non-normality of the residuals is considered unproblematic due to the substantial sample size.

*Independence:* Independence of observations is a key assumption of regression analysis. It means that the residuals should not be correlated with each other. In this study, cross-sectional data is used, where each film is represented by a single observation. This means that the observations are independent, satisfying the assumption.

*Multicollinearity:* To evaluate multicollinearity within the models, a correlation matrix was constructed, and variance inflation factors (VIFs) were computed for all models. The detailed correlation matrix can be found in Appendix C. The matrix uncovered instances of notable correlation among specific variables.[[7]](#footnote-7) These correlations suggest potential interdependencies between these sets of variables within the dataset. This can cause a few issues in the model.

Therefore, VIF analysis was conducted across models. None of the models used showed any VIF values exceeding the threshold of 10 displaying the assumption is met.

*Homoscedasticity:* The log-log regression model revealed heteroscedasticity. To address this issue, robust standard errors, adjusting for heteroskedasticity by using a different variance error estimator, were implemented. This adjustment does not alter point estimates but enhances the accuracy of standard errors and confidence intervals.

## Interpreting the results

With information available for different levels of AIR (T1, T2, and T3), regression analysis was conducted for all three levels. The following table presents the results for each model. The last rows show the researched variables and the values representing the model fit. To ensure consistency and clarity, it should be noted that the dependent variable, average ranking, is inversely related to film popularity. This means that a higher ranking corresponds to lower popularity. As a result, the estimated coefficients are interpreted in reverse: a negative coefficient indicates a positive effect of the variable on film popularity.

Table 7: Results regression analysis T3, T2, T1

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | T3 | T2 | T1 |
| Action | -0.162 \*\* | -0.066 | -0.063 |
| Adventure | -0.066 | -0.048 | -0.058 |
| Biography | 0.301 \*\*\* | 0.292 \*\*\* | 0.295 \*\*\* |
| Comedy | 0.373 \*\*\* | 0.15 \*\*\* | 0.146 \*\* |
| Crime | 0.037 | 0.032 | 0.055 |
| Documentary | 1.586 \*\*\* | 0.982 \*\*\* | 0.93 \*\* |
| Drama | 0.118 \* | 0.139 \*\*\* | 0.137 \*\*\* |
| Family | 0.357 \*\*\* | 0.247 \*\*\* | 0.247 \*\* |
| Fantasy | -0.027 | -0.029 | -0.039 |
| History | 0.015 | 0.152 | 0.137 |
| Horror | 0.26 \*\*\* | 0.004 | -0.006 |
| Music | 0.122 | 0.252 \*\* | 0.267 \*\* |
| Musical | 0.108 | 0.052 | 0.067 |
| Mystery | 0.004 | 0.027 | 0.006 |
| Romance | -0.204 \*\*\* | -0.143 \*\* | -0.152 \*\*\* |
| Sci.Fi | -0.014 | 0.006 | 0.013 |
| Sport | -0.163 | -0.008 | 0.027 |
| Thriller | 0.068 | 0.027 | 0.025 |
| War | -0.065 | 0.072 | 0.083 |
| Western | 0.203 | 0.476 \*\* | 0.495 \*\* |
| Fall | 0.042 | -0.014 | -0.013 |
| Spring | 0.021 | -0.015 | -0.02 |
| Summer | 0.005 | -0.042 | -0.044 |
| Based on book | -0.214 \* | -0.109 | -0.104 |
| Based on comic book | -0.302 \*\*\* | -0.181 \*\* | -0.198 \*\* |
| Based on novel | -0.238 \*\*\* | -0.124 \*\* | -0.149 \*\*\* |
| Based on play | 0.333 | 0.354 | 0.318 |
| Based on short story | -0.018 | 0.041 | -0.006 |
| Spinoff | 0.013 | -0.034 | -0.039 |
| Log(Sequel) | 0.004 | 0.022 \*\* | 0.02 \* |
| Log(Budget) | 0.035 | -0.033 | -0.032 |
| Log(Screens) | -0.04 | -0.459 \*\*\* | -0.469 \*\*\* |
| Log(Runtime) | 1.824 \*\*\* | -0.726 \*\*\* | -0.726 \*\*\* |
| Log(MPAA) | -0.344 \*\* | -0.512 \*\*\* | -0.507 \*\*\* |
| Log(Director Power) | -0.183 \*\*\* | -0.098 \*\*\* | -0.085 \*\*\* |
| Log(Metascore) | -0.187 \*\* | -0.347 \*\*\* | -0.334 \*\*\* |
| Log(Nominee) | -0.267 \*\*\* | -0.153 \*\*\* | -0.154 \*\*\* |
| Log(Remake) | -0.007 | 0.005 | 0.006 |
| Log(Star power) | -0.121 \*\*\* | -0.031 | -0.03 |
| Log(Winner) | -0.187 \*\*\* | -0.144 \*\*\* | -0.15 \*\*\* |
| Asian\_condition\_t3 | 0.032 ` |  |  |
| Hispanic\_condition\_t3 | 0.028 ` |  |  |
| Black\_condition\_t3 | 0.263 \*\*\* |  |  |
| Asian\_condition\_t2 |  | 0.015 |  |
| Hispanic\_condition\_t2 |  | 0.055 |  |
| Black\_condition\_t2 |  | 0.188 \*\*\* |  |
| Asian\_t1 |  |  | 0.001 |
| Hispanic\_t1 |  |  | -0.044 |
| Black\_t1 |  |  | 0.026 |
| R2 | 0.725 | 0.722 | 0.7141 |
| Adjusted R2 | 0.710 | 0.707 | 0.698 |
| Number of Observations | 1105 | 1105 | 1105 |
| F Statistic | 47.354\*\*\* | 46.664\*\*\* | 44.51\*\*\* |

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

Starting with an assessment of model fits, it is evident that the R-squared (R2) and adjusted R-squared values are notably high across all models. Specifically, the adjusted R2 values for T2 (0.707) and T3 (0.710) closely align with each other and exhibit slightly superior performance compared to model with T1 (0.698) as an explanatory variable. Nevertheless, compared to a R squared adjusted of model where none of the T variables were added which was 0.707 it can be seen that the variables of interest did not add additional explanatory power.

The outcomes of the models align with existing research, underscoring the significance of a substantial number of control variables. A total of 39 variables displayed significant effects across the different models, reinforcing their importance in explaining variations in a film’s success.

Delving into the findings of the dummy variables, 13 out of 19 genres exhibited statistical significance, demonstrating diverse positive and negative effects. Notably, with regards to genre Romance emerged with the most substantial positive effect across the models (β = -0.146, β = -0.138, and β = -0.147), while 'Documentary' portrayed the most significant negative impact (β = 0.942, β = 0.942, and β = 0.834).

A noteworthy observation is the insignificance of the seasonality variables. However, this can be rationalized by considering the nature of long-term audience engagement. The analysis spans over a year already controlling for seasonality in the measurement. Nevertheless, the result due 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 popularity at the time of release.

Films based on existing sources like books, novels, or comic books consistently exhibited higher popularity across most models, with estimates ranging from β = -0.124 to β = -0.302. It is worth noting that this trend was not observed for films based on short stories or plays.

An examination of the year dummies reveals a significant decline in popularity around 2009 compared to the baseline year of 2000. This trend aligns with the observations presented in Figure 2 from the research methodology section. To avoid redundancy and ensure clarity, the year dummy coefficients have been omitted from the table.

Concerning the interpretation of dummy variables, given that the dependent variable is logged, the estimates indicate that if a variable (x) changes by 1 unit, (y) is expected to change by 100 times β1 percent. For instance, films categorized under the genre 'Romance' were, on average, -0.146 (14.6 %) , -0.138 (13.8 %) , and -0.147 (14.7%) more popular across the models. To illustrate this effect, a 15% decrease in popularity would translate to an average popularity ranking of 2899 for a novel-based film, assuming the initial average ranking of 3410 (Table 2).

Among the continuous variables, sequels exhibited a negative effect on popularity in two out of the three models (β = 0.02, T1; β = 0.022, T2). This suggests that with each sequel, a film is less likely to maintain popularity over time. This could be due to audience fatigue or a decline in quality as the franchise progresses.

In contrast, the continuous variables related to critical acclaim, log(Nominee) (β = -0.267, β = -0.153, β = -0.154), log(Winner) (β = -0.187, β = -0.144, β = -0.15), and log(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 Metascore ratings, tend to enjoy greater popularity. Suggesting critical recognition can signal to audiences that a film is worth watching, leading to increased engagement.

Moreover, the continuous variables Log(Director Power) (β = -0.183, β = -0.098, β = -0.085) and Log(MPAA) (β = -0.344, β = -0.512, β = -0.507) demonstrated consistent positive effects across all models. This suggests that films with more powerful directors and higher MPAA ratings tend to enjoy greater popularity.

There were a few variables which did not show consistent effects. Log(Star Power) only exhibited a significant effect in the model of T3 (β = -0.163), while log(Screens) had a positive impact on popularity in both models T1 (β = -0.469) and T2 (β = -0.459). Log(Runtime) showed contradicting effects, with a positive coefficient in the model of T3 (β = 1.282) and negative coefficients in both models T1 (β = -0.726) and T2 (β = -0.726). Log(Budget) was not significant for any model, contradicting current research.

Due to the log-log nature of the model, interpreting results for continuous variables involves understanding them as elasticities. For instance, a one percent increase in Metascore, which ranges from 0 to 100, translates to an average ranking in the third year that is -0.187%, -0.347%, and -0.334% lower across the respective models. To contextualize, consider the average Metascore of 47.66. If we increase it by 10% to 52.43, using the T1 model and the average popularity ranking as a reference, the initial ranking of 3410 would decrease by 3.47%, resulting in a new ranking of 3292. While this effect is statistically significant, its magnitude is relatively small, prompting a consideration of its managerial relevance.

Before examining the specific variables of interest, it is noteworthy that the T3 model exhibits the most differences in variable significance Specifically, Star Power, Based on Book, Horror, and Action were all significant in T3 but not in the others. Conversely, Sequel, Western, and Music showed significance in T1 and T2 but not in T3. Interestingly, T1 and T2 did not differ in terms of variable significance. Without further research, it is unclear why this differentiation exists for model T3. Possibly, the small number of films (206, 77, and 97) passing the T3 condition for the ethnicities could be an explanation.

Finally, an examination of the ethnicity variables revealed an unexpected finding. The variables representing the Black ethnicity exhibited negative correlation with average ranking in two out of the three models, while Asian and Hispanic ethnicity showed no significant correlation. This contradicts the initial hypotheses, which anticipated a positive correlation for racial AIR. Due to this unexpected result, further analysis was conducted to gain a deeper understanding of the underlying factors.

# Additional analysis

This study emphasizes the importance of accounting for confounding variables, especially cultural influences, often overlooked in prior research such as Malik and others (2022) and Kuppuswamy and others (2016). In today's society, values like belonging, community, and personal growth, closely tied to representation, hold significant importance (Neufeld, 2020). Films and culture are inherently linked, with research showing that those effectively capturing the cultural zeitgeist tend to be more successful (Ettema, 2005). Therefore, films reflecting representation-related values are assumed to thrive today due to the cultural landscape. However, as societal norms evolve, this appreciation changes over time. This section aims to explore the impact of authentic inclusive representation (AIR) and its connection to the current cultural zeitgeist focused on representation.

## Culture as cofounding variable

Cultural resonance, the ability of a film to connect with an audience's cultural values and experiences, can be achieved by either reflecting or challenging those values in a thought-provoking manner (Ettema, 2005). Diverse cultural representation in films plays a pivotal role in broadening audiences and fostering meaningful connections (Bamford, 2018). For instance, the film "Black Panther" (2018) received widespread acclaim for its positive portrayals of African culture, aligning with a period of heightened cultural discourse surrounding representation. This demonstrates the film's role as both a product and a catalyst for cultural change.

Culture significantly influences audience engagement with films, particularly those that explore social and cultural issues through a racially diverse lens. These films can spark lasting discussions about race, identity, and social justice (Garrett, 2020). Current film critics prioritize representation and inclusion, leading films with authentic, identifiable, and relevant (AIR) representations to receive more favorable critical reception, contributing to Long-Term Audience Engagement (LTAE). The transformative impact of cinema extends beyond its on-screen influence, also shaping industry hiring practices and fostering increased racial diversity in film production (Khrebtan-Hörhager, 2011). In light of these considerations, the following hypothesis is proposed:

**H2**: The relationship between racial AIR and LTAE is cofounded by culture resonance.

*Variable operationalization*: Two approaches were considered to capture the moment when racial representation became a major topic:

Step Dummy: A step dummy variable would be designated for films released when racial representation became a significant cultural topic, serving as the interaction term.

Diff-in-Diff Analysis: A diff-in-diff analysis would compare the differences between a treatment group of racially diverse films and a control group of non-racially diverse films.

The final approach was a combination of both chosen due to the inherent challenges of applying a diff-in-diff analysis to films released at different points in time. Diff-in-diff analysis typically compares the performance of a treatment group and a control group before and after a specific event or intervention. However, with films, the release dates themselves serve as the "events," making it difficult to establish a clear pre- and post-treatment period for each film. Despite this, the operationalization incorporates a "natural shock" concept akin to a diff-in-diff analysis.

In this context, the "natural shock" refers to the #oscarssowhite phenomenon in January 2015[[8]](#footnote-8), which effectively quantified the onset of racial representation as a major cultural theme. A dummy variable, after\_jan\_2015, was created and assigned a value of one for films released in or after 2015. Most studies implementing a diff-in-diff analysis start by examining the differences between a treatment and control group. Because the following is solely for illustrative purposes and T1, T2 and T3 are expressed for multiple ethnicities the dataset was split into a 'treatment' and a 'control' group based on overall racial diversity of the cast. The methodology for this split is detailed in Appendix D.

It is solely important to know the treatment group represents racially diverse films, and as evident in Figure 4, while both groups exhibit similar trends between 2010 and 2011, the treatment group experiences a significant spike followed by a sharp drop. This indicates that the two groups differ in their popularity behavior over time. Therefore, it is worth investigating whether this effect is also reflected in the T1, T2, and T3 models.

A graph of a graph showing the number of years

Description automatically generated with medium confidence

Figure 4 Average ranking films treatment and control group

**Model**The model incorporates the interaction term after\_jan\_2015 with the three T1, T2, and T3 models to examine the potential confounding effect of culture resonance on the relationship between racial AIR and LTAE. The interaction term allows us to assess whether the effect of racial AIR on LTAE differs before and after the #oscarssowhite phenomenon in January 2015, which served as a proxy for the increased cultural significance of racial representation.

Here is the T3 model with the added interaction term:

*Model* : log(LTAEi) = β0 + β1 × AIR\_HispanicT3i + β2 × AIR\_BlackT3i + β3 × AIR\_AsianT3i + β4 × FILMSAFTER2015i × AIR\_HispanicT3i + β5 × FILMSAFTER2015i \* AIR\_BlackT3i + β6 × FILMSAFTER2015i \* AIR\_AsianT3i +

…. β65× YEAR2018i + β66 × YEAR2019i + εi

In this model, LTAEi similar to the previous models 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 us to capture whether the effect of racial AIR on LTAE differs before and after 2015.

## Results

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. The complete models can be referenced in Appendix D.

Table 8: Results step dummy regression analysis

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Model step T3 | Model step T2 | Model step T1 |
| after\_jan\_2015 | -0.12 | -0.094 | 0.066 |
| jan\_2015:Asian\_T1 |  |  | -0.181 \*\* |
| jan\_2015:Black\_T1 |  |  | -0.133 \* |
| jan\_2015:Hispanic\_1 |  |  | 0.091 |
| Asian\_T1 |  |  | 0.038 |
| Black\_T1 |  |  | 0.12 \*\*\* |
| Hispanic\_T1 |  |  | - 0.07 \* |
| jan\_2015:Asian\_T2 |  | -0.302 \*\* |  |
| jan\_2015:Black\_T2 |  | -0.065 |  |
| jan\_2015:Hispanic\_2 |  | 0.091 |  |
| Asian\_T2 |  | 0.095 |  |
| Black\_T2 |  | 0.194 \*\*\* |  |
| Hispanic\_T2 |  | -0.065 |  |
| jan\_2015:Asian\_T3 | -0.066 |  |  |
| jan\_2015:Black\_T3 | -0.045 |  |  |
| jan\_2015:Hispanic\_3 | 0.209 |  |  |
| Asian\_T3 | 0.035 |  |  |
| Black\_T3 | 0.246 \*\*\* |  |  |
| Hispanic\_T3 | 0.02 |  |  |
| R2 | 0.7363 | 0.7344 | 0.731 |
| Adjusted R2 | 0.7204 | 0.7183 | 0.7147 |
| Number of Observations | 1105 | 1105 | 1105 |
| F Statistic | 46.18\*\*\* | 45.73\*\*\* | 44.83\*\*\* |

All three models (T1, T2, and T3) demonstrated higher R-squared and R-squared adjusted values compared to their respective predecessor models. The R-squared adjusted values for T3, T2, and T1 were 0.72, 0.72, and 0.71, compared to the previous 0.71, 0.71 and 0.70 respectively. Indicating a small improvement in model fit.

Intriguingly, the interaction terms involving Asian T2, Asian T1, and Black T1 with post-2015 exhibited negative coefficients, suggesting a positive relationship between these factors and film popularity. However, further examination revealed that the negative coefficients for the Black ethnicity interaction terms were evaluated in the context of an existing negative relationship indicated by positive estimates. This implies that while the interaction terms suggest a positive influence of post-2015 release on film popularity for films passing the Black conditions, the overall the relationship cannot be determined to be significant.

Finally, the most interesting finding is the positive estimates of Asian in model T2 ( = -0.302 ) and T1 ( = -0.181). While showing no significance for the terms on their own for both the T2 and T1 model the effect is positive when a film is released after 2015. Because the terms on their own are not significant this is in contrary to the Black ethnicity suggesting an actual positive effect.

This study highlights the intricate relationship between cultural resonance and film popularity. The observed trends suggest that cultural factors, as reflected by the interaction between ethnicity and release year, can exert an influence on film perception and popularity.

Table 9 : Summary Results Proposed Hypotheses

|  |  |  |  |
| --- | --- | --- | --- |
| **Hypotheses** | **β** | **P-Value** | **Decision** |
| H1a: Racial Authentic Inclusive Representation (at T1 level) is positively correlated with Long Term Audience Engagement for films. | Black (0.026)  Asian (0.001 )  Hispanic(-0.044) | (P = 0.282)  (P = 0.982)  (P = 0.263) | Rejected |
| H1b: Racial Authentic Inclusive Representation (at T2 level) is positively correlated with Long Term Audience Engagement for films. | Black (0.188)  Asian (0.015)  Hispanic(0.055) | (P = < 0.001) \*\*\*  (P = 0.744)  (P = 0.183) | Partly contradicted |
| H1c: Racial Authentic Inclusive Representation (at T3 level) is positively correlated with Long Term Audience Engagement for films. | Black (0.263)  Asian (0.032)  Hispanic(0.028) | (P = 0.049) \*  (P = 0.074)  (P = 0.067) | Partly contradicted |
|  |  |  |  |
| H2a: The relationship between racial AIR (at T1 level) and LTAE is cofounded by culture resonance. | Black (-0.133) Asian (-0.181)  Hispanic(0.091) | (P = 0.695)  (P = 0.001) \*\*  (P = 0.278) | Accepted |
| H2b: The relationship between racial AIR (at T2 level) and LTAE is cofounded by culture resonance. | Black (-0.065 )  Asian (-0.302)  Hispanic(0.091) | (P = 0.985)  (P = 0.022) \*\*  (P = 0.552) | Accepted |
| H2c: The relationship between racial AIR (at T3 level ) and LTAE is cofounded by culture resonance. | Black (-0.045)  Asian (-0.066)  Hispanic(0.209) | (P = 0.531)  (P = 0.990)  (P = 0.184) | Accepted |

5. Dicsussion

1. Personal identity theory does not lead to increase in performance films not clear from diversity not significant. It is not a good measurement. It does not have a positive effect on films

What must be said is that this can also be caused because the sources which the films derive from also have new releases. If a new Batman comic book is released people might think of watching the film again. Also is important that when the film is based on a play this effect is negative.

Main effect: These findings suggest that higher diversity, as measured by the inverse Simpson diversity index, tends to be linked with lower popularity, signifying a potential inverse relationship between diversity and the ranking.

This somewhat contradictory with previous research (Kuppuswamy, 2016) who found a significant positive result between a film having multiple black actors and box office. They concluded that the amount of Black actors was positively correlated with Domestic box office and positively with World- Wide box office.

This negative correlation can have two causes.

Moreover, Hall 2022 found that people do generate more engagement with films that have characters with similar traits. There was research providing evidence for positive correlation with racial diversity and box office revenue. The finding of this model suggest that including people from different ethnicities might provide inclusion for the people who are now included but it might cause other demographics not being represented. Simply said if we would take race and age as two characteristics and you have two characters we keep everything else constant. Because the population in the states is still mostly white by removing certain characters because they are replaced by coloured actors etc.

Nevertheless, the reason can also as found by (Aumer, 2017) that there might be no preference for watching ourselves on the screen. 18 50, Hispanic , White

Current industry forces might have caused under budgeting of films which are racial diverse. Causing the negative effects seen by Black

**Entire dataset**

As mentioned before T2 and T3 could only be processed for around of all available observation due to lack of proper srt files. T1 on the other can be conducted for the entire dataset of 2319 observations. The entire regression can be seen in Appendix D. It can be seen however with the difference in explanatory power between T1 the entire dataset (R2 = 0.675) and the sample set (R2 = 0.698) that the model is better in predicting for the sample dataset with including T1.

There were no shocking differences between the two datasets. Most variables remained to have there significance and the same direction of effect. The only differences were that for the entire dataset average budget was significant (β = -0.04), which does coincide with existing research more. Starpower is significant in the entire dataset (β = -0.03), moreover Based on Book as well (β = -1.44), and the genre action was significant, (β = -1.06). history was also significant (β = 0.149),

The only variable significant in the model used for the sample which could be processed with subtitles was Music.

Aumer 2017 Look at the paper once again. Proofing no support for social identity theory.

Main points :   
AIR increases the explanatory power. Nevertheless, this is in a different group of sample. When Simpson diversity used on this sample the R2 squared (0.7158) and R2 adjusted (0.7005) increase. Where the Simpson diversity index has not a significant effect (β = 0.184 , p = 0.102) Therefore, the model is better in explaining this specific sample but there is something going on.

Moreover, this also shows that the inverted Simpson diversity index is not robust. While T1 remains significant in the entire dataset and solely the dataset used for mediation analysis. Nevertheless, this is a negative effect for the black ethnicity and almost a significant effect for Hispanic ethnicity.

But with all levels of T the black ethnicity remains to have negative effect.

Moreover AIR 2 and 3 are a bit arbitrary.

AIR 1 is better than racial diversity with the same sample.

C path

All results did not go along with the expected hypotheses. Therefore, an additional analysis was made.

Diversity solely not a proper measurement. AIR is indeed can be seen as a positive effect. And it compensates for the negative effect of diversity which positively influences AIR. So focus on AIR not diversity. When we added racial diversity with the T values there was not significance .

Nevertheless, this phenomenon has only appeared in the last few years and the positive effect might wonder off. Trend of more budget,

# Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix A: Stop word list

"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 thrershold of 0.85, aligning the character list with their respective ethnicities within the dataset.

Appendix C : multicollinearity matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LTAE = 1 Sequal = 2 Spinoff = 3 (B) Book = 4 (B) Play = 5 (B) C = 6 (B) CB = 7 (B) Novel = 8 (B) Short = 9 SI = 10 | Action = 11 Adventure = 12 Comedy = 13 Fantasy = 14 Crime = 15 Drama = 16 Mystery = 17 Thriller = 18 Romance = 19 Sci.Fi = 20 | Biography = 21 Sport = 22 War = 23 Family = 24 Musical = 25 History = 26 Horror = 27 Documentary = 28 Western = 29 2000 = 30 | 2001 = 31 2002 = 32 2003 = 33 2004 = 34 2005 = 35 2006 = 36 2007 = 37 2008 = 38 2009 = 39 2010 = 40 | 2011 = 41 2012 = 42 2013 = 43 2014 = 44 2015 = 45 2016 = 46 2017 = 47 2018 = 48 2019 = 49 (L)Nominee = 50 | Winner = 51 Remake = 52 D\_power = 53 (L)Metscore = 54 (L) Str power = 55 Spring = 56 Summer = 57 Fall = 58 Winter = 59 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| LTAE | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Sequal | -0.07 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Spinoff | -0.02 | 0.01 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (B) Book | -0.02 | -0.05 | -0.03 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (B) Play | -0.03 | -0.02 | -0.01 | -0.02 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (B) C | -0.10 | 0.04 | 0.09 | -0.04 | -0.02 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (B) CB | -0.13 | 0.02 | 0.06 | -0.06 | -0.02 | 0.71 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (B) Novel | -0.10 | -0.04 | -0.04 | 0.08 | 0.03 | -0.07 | -0.09 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |
| (B) Short | -0.03 | -0.02 | -0.01 | -0.02 | -0.01 | -0.02 | -0.02 | -0.04 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |
| SI | 0.12 | 0.06 | 0.06 | -0.04 | -0.06 | 0.06 | 0.06 | -0.11 | -0.02 | 1.00 |  |  |  |  |  |  |  |  |  |  |
| Action | -0.15 | 0.14 | 0.05 | -0.07 | -0.01 | 0.25 | 0.35 | -0.03 | -0.02 | 0.23 | 1.00 |  |  |  |  |  |  |  |  |  |
| Adventure | -0.14 | 0.26 | 0.10 | -0.02 | -0.01 | 0.23 | 0.27 | 0.06 | -0.02 | 0.06 | 0.41 | 1.00 |  |  |  |  |  |  |  |  |
| Comedy | 0.04 | -0.05 | -0.06 | -0.01 | 0.01 | -0.08 | -0.06 | -0.08 | -0.04 | -0.02 | -0.25 | -0.11 | 1.00 |  |  |  |  |  |  |  |
| Fantasy | -0.07 | 0.02 | -0.03 | -0.06 | -0.03 | 0.08 | 0.14 | 0.11 | 0.00 | -0.03 | 0.08 | 0.23 | 0.00 | 1.00 |  |  |  |  |  |  |
| Crime | -0.04 | -0.06 | -0.05 | -0.02 | -0.04 | -0.03 | -0.02 | -0.01 | 0.01 | 0.07 | 0.08 | -0.19 | -0.06 | -0.16 | 1.00 |  |  |  |  |  |
| Drama | 0.10 | -0.17 | -0.10 | 0.12 | 0.07 | -0.09 | -0.16 | 0.15 | 0.03 | -0.09 | -0.26 | -0.26 | -0.24 | -0.13 | 0.10 | 1.00 |  |  |  |  |
| Mystery | 0.01 | -0.03 | 0.03 | -0.01 | -0.03 | -0.07 | -0.05 | 0.07 | 0.05 | -0.08 | -0.12 | -0.10 | -0.24 | 0.00 | 0.07 | 0.01 | 1.00 |  |  |  |
| Thriller | -0.08 | 0.02 | -0.01 | -0.02 | -0.04 | -0.03 | -0.03 | 0.00 | 0.04 | 0.07 | 0.27 | -0.05 | -0.47 | -0.14 | 0.31 | 0.01 | 0.34 | 1.00 |  |  |
| Romance | -0.06 | -0.12 | -0.03 | 0.02 | 0.10 | -0.09 | -0.12 | 0.09 | 0.01 | -0.14 | -0.28 | -0.17 | 0.30 | 0.00 | -0.18 | 0.13 | -0.15 | -0.30 | 1.00 |  |
| Sci.Fi | -0.11 | 0.09 | 0.07 | -0.06 | -0.04 | 0.23 | 0.33 | -0.01 | 0.12 | 0.12 | 0.38 | 0.38 | -0.18 | 0.05 | -0.19 | -0.21 | -0.01 | 0.02 | -0.16 | 1.00 |
| Biography | 0.09 | -0.06 | -0.04 | 0.24 | -0.02 | -0.05 | -0.07 | -0.07 | -0.02 | -0.08 | -0.14 | -0.09 | -0.15 | -0.08 | 0.01 | 0.30 | -0.10 | -0.12 | -0.03 | -0.12 |
| Sport | 0.05 | -0.03 | -0.03 | 0.07 | 0.04 | -0.04 | -0.06 | -0.06 | -0.02 | -0.03 | -0.10 | -0.06 | -0.02 | -0.08 | -0.08 | 0.16 | -0.09 | -0.16 | -0.01 | -0.09 |
| War | 0.01 | -0.02 | -0.03 | 0.08 | 0.16 | 0.01 | -0.03 | 0.06 | -0.02 | -0.06 | 0.07 | 0.04 | -0.12 | -0.04 | -0.06 | 0.19 | -0.05 | -0.03 | -0.03 | -0.06 |
| Family | 0.09 | 0.03 | -0.03 | 0.03 | -0.03 | -0.06 | -0.09 | 0.13 | -0.03 | -0.06 | -0.14 | 0.24 | 0.20 | 0.36 | -0.14 | -0.11 | -0.06 | -0.24 | 0.00 | -0.03 |
| Musical | 0.03 | 0.10 | 0.01 | -0.03 | 0.17 | -0.02 | -0.03 | 0.02 | -0.01 | -0.06 | -0.09 | 0.02 | 0.05 | 0.07 | -0.03 | 0.01 | -0.04 | -0.08 | 0.08 | -0.06 |
| History | 0.04 | -0.04 | -0.03 | 0.10 | 0.16 | -0.04 | -0.05 | 0.03 | -0.02 | -0.13 | -0.03 | -0.05 | -0.11 | -0.08 | -0.05 | 0.21 | -0.09 | -0.09 | -0.04 | -0.09 |
| Horror | 0.03 | -0.02 | 0.03 | -0.02 | -0.03 | -0.03 | -0.02 | -0.09 | -0.03 | -0.04 | -0.13 | -0.15 | -0.24 | 0.02 | -0.12 | -0.17 | 0.41 | 0.27 | -0.18 | 0.05 |
| Documentary | 0.30 | -0.02 | 0.04 | -0.01 | -0.01 | -0.01 | -0.02 | -0.03 | -0.01 | 0.05 | -0.05 | -0.03 | 0.03 | -0.03 | -0.03 | -0.06 | -0.03 | -0.05 | -0.03 | -0.03 |
| Western | -0.02 | -0.02 | -0.01 | 0.03 | -0.01 | 0.03 | 0.05 | 0.04 | -0.01 | -0.02 | 0.10 | 0.07 | -0.03 | 0.05 | -0.04 | 0.00 | -0.04 | 0.03 | 0.00 | -0.02 |
| 2000 | -0.05 | -0.04 | -0.01 | 0.03 | -0.01 | -0.04 | -0.05 | -0.01 | -0.01 | -0.09 | -0.03 | 0.00 | 0.05 | -0.03 | 0.01 | 0.02 | -0.04 | -0.01 | 0.03 | 0.01 |
| 2001 | -0.08 | -0.02 | 0.00 | -0.01 | -0.02 | -0.04 | -0.04 | 0.01 | 0.08 | -0.02 | -0.05 | -0.06 | 0.02 | -0.04 | 0.05 | -0.02 | 0.01 | 0.02 | 0.05 | 0.00 |
| 2002 | -0.06 | 0.05 | -0.01 | 0.05 | -0.02 | -0.02 | -0.02 | 0.05 | 0.04 | -0.02 | 0.00 | 0.00 | 0.02 | -0.03 | 0.04 | 0.02 | 0.08 | 0.04 | -0.03 | -0.02 |
| 2003 | -0.07 | 0.00 | 0.00 | -0.01 | -0.02 | -0.02 | -0.02 | -0.03 | -0.02 | -0.03 | 0.00 | -0.04 | 0.08 | 0.01 | 0.00 | -0.05 | -0.02 | -0.03 | -0.02 | -0.04 |
| 2004 | -0.06 | -0.04 | -0.03 | -0.03 | 0.04 | 0.02 | 0.01 | 0.03 | -0.02 | -0.06 | -0.05 | -0.03 | 0.03 | -0.03 | 0.02 | -0.02 | 0.01 | -0.05 | 0.04 | 0.00 |
| 2005 | -0.05 | -0.02 | -0.01 | 0.00 | -0.01 | 0.01 | 0.01 | 0.01 | -0.02 | -0.04 | 0.01 | 0.06 | 0.01 | 0.03 | -0.02 | -0.02 | 0.00 | 0.01 | 0.02 | -0.03 |
| 2006 | -0.06 | 0.00 | 0.01 | -0.03 | 0.09 | -0.04 | -0.04 | -0.03 | -0.02 | -0.04 | -0.03 | -0.05 | 0.00 | -0.01 | 0.05 | -0.01 | -0.02 | 0.01 | 0.01 | -0.04 |
| 2007 | -0.05 | -0.02 | -0.03 | -0.02 | -0.02 | -0.02 | -0.02 | -0.01 | -0.02 | -0.04 | -0.07 | -0.06 | -0.02 | -0.01 | 0.08 | 0.03 | 0.01 | 0.09 | -0.03 | -0.07 |
| 2008 | -0.08 | 0.00 | -0.03 | -0.02 | -0.02 | 0.08 | 0.06 | 0.00 | -0.02 | -0.08 | 0.02 | -0.02 | 0.01 | 0.03 | -0.03 | 0.01 | -0.05 | -0.03 | 0.06 | -0.02 |
| 2009 | -0.07 | 0.00 | 0.02 | 0.05 | -0.02 | -0.01 | -0.02 | -0.01 | 0.03 | -0.08 | -0.05 | -0.01 | -0.01 | 0.05 | -0.03 | 0.01 | 0.02 | -0.03 | 0.04 | 0.01 |
| 2010 | -0.04 | 0.00 | 0.01 | -0.02 | -0.02 | 0.01 | 0.04 | 0.01 | 0.03 | -0.05 | 0.01 | -0.03 | 0.05 | 0.07 | 0.00 | -0.03 | -0.01 | -0.03 | 0.06 | -0.03 |
| 2011 | -0.05 | 0.02 | -0.03 | 0.03 | 0.06 | 0.00 | 0.02 | 0.03 | 0.02 | -0.09 | 0.00 | 0.00 | 0.01 | 0.02 | -0.07 | -0.03 | 0.01 | -0.02 | 0.04 | -0.02 |
| 2012 | -0.01 | 0.02 | 0.04 | -0.02 | 0.07 | -0.03 | -0.02 | 0.02 | 0.02 | -0.03 | 0.01 | 0.01 | 0.00 | -0.04 | -0.04 | 0.00 | -0.03 | -0.02 | 0.00 | 0.01 |
| 2013 | 0.02 | 0.01 | -0.02 | 0.00 | -0.02 | 0.01 | 0.04 | 0.00 | -0.02 | 0.02 | 0.03 | -0.01 | -0.02 | -0.04 | 0.06 | 0.00 | -0.04 | 0.04 | -0.05 | -0.01 |
| 2014 | 0.09 | 0.00 | 0.02 | -0.02 | -0.02 | -0.01 | 0.01 | 0.04 | 0.01 | 0.04 | 0.05 | 0.03 | -0.05 | 0.02 | 0.00 | 0.00 | 0.02 | -0.02 | -0.04 | 0.01 |
| 2015 | 0.05 | 0.01 | 0.01 | -0.04 | -0.02 | -0.01 | -0.02 | -0.05 | -0.02 | 0.03 | 0.04 | 0.03 | -0.01 | -0.05 | 0.00 | 0.04 | 0.01 | 0.00 | -0.02 | 0.04 |
| 2016 | 0.09 | -0.05 | -0.02 | -0.01 | -0.01 | -0.03 | -0.02 | -0.03 | -0.01 | 0.07 | 0.02 | -0.01 | -0.02 | 0.03 | 0.00 | 0.01 | -0.04 | 0.01 | -0.03 | 0.04 |
| 2017 | 0.08 | -0.02 | -0.03 | 0.01 | -0.01 | -0.01 | 0.03 | 0.01 | -0.01 | 0.11 | -0.01 | 0.03 | -0.05 | 0.05 | -0.04 | 0.03 | -0.01 | -0.02 | -0.02 | 0.06 |
| 2018 | 0.21 | 0.01 | 0.05 | 0.04 | -0.02 | 0.01 | 0.01 | -0.01 | -0.02 | 0.20 | 0.06 | 0.07 | -0.05 | -0.03 | 0.01 | -0.02 | 0.02 | 0.05 | -0.07 | 0.08 |
| 2019 | 0.13 | 0.05 | 0.02 | 0.02 | -0.02 | 0.09 | 0.04 | -0.05 | -0.02 | 0.16 | 0.01 | 0.04 | -0.05 | 0.00 | -0.07 | 0.04 | 0.05 | -0.02 | -0.05 | 0.02 |
| (L)Nominee | -0.38 | 0.18 | 0.07 | 0.06 | 0.06 | 0.17 | 0.23 | 0.05 | 0.03 | -0.03 | 0.11 | 0.26 | -0.18 | 0.10 | -0.03 | 0.00 | -0.01 | -0.03 | -0.07 | 0.20 |
| Winner | -0.29 | 0.14 | 0.02 | 0.05 | 0.07 | 0.15 | 0.17 | 0.09 | 0.01 | -0.03 | 0.06 | 0.22 | -0.16 | 0.09 | -0.05 | 0.04 | -0.02 | -0.04 | -0.05 | 0.14 |
| Remake | -0.03 | -0.08 | -0.02 | -0.05 | 0.02 | -0.02 | -0.05 | -0.07 | -0.02 | -0.03 | -0.03 | -0.04 | -0.04 | 0.02 | 0.03 | -0.04 | 0.01 | 0.04 | 0.00 | -0.01 |
| D\_power | -0.17 | 0.11 | 0.04 | -0.02 | 0.08 | 0.12 | 0.14 | 0.07 | 0.02 | 0.02 | 0.09 | 0.20 | -0.08 | 0.11 | 0.02 | -0.05 | -0.02 | -0.01 | -0.05 | 0.08 |
| (L)Metscore | -0.33 | 0.10 | 0.02 | 0.07 | 0.04 | 0.06 | 0.08 | 0.06 | 0.04 | 0.03 | 0.04 | 0.09 | -0.11 | -0.01 | 0.05 | 0.17 | -0.04 | 0.02 | -0.04 | 0.06 |
| (L) Str power | -0.17 | 0.04 | 0.08 | -0.05 | -0.02 | 0.15 | 0.16 | -0.05 | 0.00 | 0.07 | 0.16 | 0.16 | -0.03 | 0.04 | 0.00 | -0.08 | -0.05 | -0.01 | -0.03 | 0.12 |
| Spring | 0.01 | -0.01 | 0.03 | -0.01 | -0.02 | 0.05 | 0.05 | -0.02 | -0.03 | 0.02 | 0.01 | 0.02 | 0.01 | -0.03 | -0.01 | 0.01 | -0.01 | -0.05 | 0.05 | 0.07 |
| Summer | -0.01 | -0.01 | 0.01 | -0.02 | -0.02 | 0.07 | 0.09 | -0.06 | 0.04 | 0.05 | 0.10 | 0.07 | 0.08 | -0.01 | -0.06 | -0.15 | -0.04 | -0.03 | -0.04 | 0.10 |
| Fall | 0.03 | 0.04 | -0.05 | 0.02 | 0.03 | -0.09 | -0.10 | 0.04 | 0.00 | -0.01 | -0.08 | -0.09 | -0.06 | -0.02 | 0.05 | 0.10 | 0.05 | 0.04 | -0.06 | -0.12 |
| Winter | -0.03 | -0.02 | 0.01 | 0.00 | 0.01 | -0.03 | -0.04 | 0.04 | -0.02 | -0.06 | -0.03 | 0.00 | -0.04 | 0.07 | 0.02 | 0.04 | 0.00 | 0.05 | 0.05 | -0.05 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 22 | 23 | 24 | 2 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 |
| Biography | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Sport | 0.26 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| War | 0.06 | -0.02 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Family | -0.05 | 0.02 | -0.05 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Musical | 0.00 | -0.03 | 0.01 | 0.19 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| History | 0.34 | 0.05 | 0.42 | -0.07 | 0.01 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Horror | -0.10 | -0.09 | -0.03 | -0.10 | -0.03 | -0.08 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Music | 0.08 | -0.04 | -0.01 | 0.04 | 0.13 | 0.01 | -0.08 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Documentary | -0.02 | -0.01 | -0.01 | -0.02 | -0.01 | 0.06 | -0.03 | 0.13 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |
| Western | 0.02 | 0.03 | 0.14 | -0.03 | -0.01 | 0.03 | -0.01 | -0.02 | -0.01 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |
| 2000 | -0.03 | 0.04 | -0.01 | 0.01 | -0.02 | -0.01 | -0.03 | -0.01 | -0.01 | -0.02 | 1.00 |  |  |  |  |  |  |  |  |  |  |
| 2001 | -0.04 | -0.05 | 0.02 | -0.01 | -0.03 | -0.02 | -0.02 | 0.00 | -0.01 | -0.02 | -0.04 | 1.00 |  |  |  |  |  |  |  |  |  |
| 2002 | -0.04 | -0.02 | 0.01 | 0.02 | 0.01 | -0.02 | -0.04 | 0.03 | -0.01 | -0.02 | -0.04 | -0.05 | 1.00 |  |  |  |  |  |  |  |  |
| 2003 | -0.04 | 0.00 | 0.01 | 0.02 | 0.01 | 0.01 | 0.02 | -0.02 | -0.01 | 0.03 | -0.04 | -0.05 | -0.04 | 1.00 |  |  |  |  |  |  |  |
| 2004 | 0.01 | 0.04 | 0.00 | 0.05 | 0.01 | 0.04 | 0.02 | 0.05 | -0.01 | 0.07 | -0.04 | -0.05 | -0.05 | -0.04 | 1.00 |  |  |  |  |  |  |
| 2005 | 0.03 | 0.05 | -0.01 | 0.02 | 0.01 | 0.04 | -0.01 | 0.04 | 0.06 | 0.09 | -0.03 | -0.04 | -0.04 | -0.04 | -0.04 | 1.00 |  |  |  |  |  |
| 2006 | 0.03 | 0.04 | -0.02 | 0.00 | 0.01 | 0.06 | -0.01 | -0.02 | -0.01 | -0.02 | -0.04 | -0.05 | -0.05 | -0.04 | -0.05 | -0.04 | 1.00 |  |  |  |  |
| 2007 | 0.02 | 0.00 | 0.08 | 0.01 | 0.01 | 0.03 | 0.02 | -0.02 | -0.01 | -0.02 | -0.04 | -0.04 | -0.04 | -0.04 | -0.04 | -0.04 | -0.04 | 1.00 | 1.00 |  |  |
| 2008 | -0.01 | 0.01 | 0.01 | -0.01 | -0.03 | 0.01 | -0.01 | -0.03 | -0.02 | -0.02 | -0.04 | -0.05 | -0.05 | -0.05 | -0.05 | -0.04 | -0.05 | -0.05 | -0.06 | 1.00 |  |
| 2009 | 0.06 | 0.01 | -0.03 | -0.01 | -0.03 | 0.01 | 0.00 | 0.02 | -0.02 | -0.02 | -0.04 | -0.05 | -0.05 | -0.05 | -0.05 | -0.04 | -0.05 | -0.05 | -0.06 | -0.06 | 1.00 |
| 2010 | -0.06 | -0.05 | -0.01 | 0.03 | -0.03 | -0.03 | -0.01 | -0.01 | -0.02 | 0.02 | -0.04 | -0.05 | -0.05 | -0.05 | -0.05 | -0.05 | -0.05 | -0.05 | -0.07 | -0.07 | -0.07 |
| 2011 | -0.03 | 0.01 | -0.02 | 0.01 | -0.01 | 0.00 | 0.00 | -0.02 | 0.03 | 0.01 | -0.05 | -0.06 | -0.06 | -0.06 | -0.06 | -0.05 | -0.06 | -0.06 | -0.06 | -0.06 | -0.06 |
| 2012 | -0.02 | 0.02 | 0.03 | 0.00 | 0.03 | -0.01 | -0.02 | 0.01 | -0.02 | 0.02 | -0.05 | -0.06 | -0.05 | -0.05 | -0.05 | -0.05 | -0.06 | -0.05 | -0.06 | -0.06 | -0.06 |
| 2013 | 0.03 | -0.01 | -0.03 | -0.07 | 0.00 | -0.05 | -0.01 | -0.01 | 0.04 | 0.02 | -0.04 | -0.05 | -0.05 | -0.05 | -0.05 | -0.05 | -0.05 | -0.05 | -0.07 | -0.07 | -0.07 |
| 2014 | 0.09 | 0.00 | 0.02 | -0.02 | -0.02 | -0.01 | 0.01 | 0.04 | 0.01 | 0.04 | 0.05 | 0.03 | -0.05 | 0.02 | 0.00 | 0.00 | 0.02 | -0.02 | -0.04 | 0.01 | 0.02 |
| 2015 | 0.05 | 0.01 | 0.01 | -0.04 | -0.02 | -0.01 | -0.02 | -0.05 | -0.02 | 0.03 | 0.04 | 0.03 | -0.01 | -0.05 | 0.00 | 0.04 | 0.01 | 0.00 | -0.02 | 0.04 | 0.00 |
| 2016 | 0.09 | -0.05 | -0.02 | -0.01 | -0.01 | -0.03 | -0.02 | -0.03 | -0.01 | 0.07 | 0.02 | -0.01 | -0.02 | 0.03 | 0.00 | 0.01 | -0.04 | 0.01 | -0.03 | 0.04 | -0.02 |
| 2017 | 0.08 | -0.02 | -0.03 | 0.01 | -0.01 | -0.01 | 0.03 | 0.01 | -0.01 | 0.11 | -0.01 | 0.03 | -0.05 | 0.05 | -0.04 | 0.03 | -0.01 | -0.02 | -0.02 | 0.06 | 0.01 |
| 2018 | 0.21 | 0.01 | 0.05 | 0.04 | -0.02 | 0.01 | 0.01 | -0.01 | -0.02 | 0.20 | 0.06 | 0.07 | -0.05 | -0.03 | 0.01 | -0.02 | 0.02 | 0.05 | -0.07 | 0.08 | 0.04 |
| 2019 | 0.13 | 0.05 | 0.02 | 0.02 | -0.02 | 0.09 | 0.04 | -0.05 | -0.02 | 0.16 | 0.01 | 0.04 | -0.05 | 0.00 | -0.07 | 0.04 | 0.05 | -0.02 | -0.05 | 0.02 | 0.03 |
| Nominee | -0.38 | 0.18 | 0.07 | 0.06 | 0.06 | 0.17 | 0.23 | 0.05 | 0.03 | -0.03 | 0.11 | 0.26 | -0.18 | 0.10 | -0.03 | 0.00 | -0.01 | -0.03 | -0.07 | 0.20 | 0.08 |
| Winner | -0.29 | 0.14 | 0.02 | 0.05 | 0.07 | 0.15 | 0.17 | 0.09 | 0.01 | -0.03 | 0.06 | 0.22 | -0.16 | 0.09 | -0.05 | 0.04 | -0.02 | -0.04 | -0.05 | 0.14 | 0.09 |
| Remake | -0.03 | -0.08 | -0.02 | -0.05 | 0.02 | -0.02 | -0.05 | -0.07 | -0.02 | -0.03 | -0.03 | -0.04 | -0.04 | 0.02 | 0.03 | -0.04 | 0.01 | 0.04 | 0.00 | -0.01 | -0.06 |
| (L) D power | -0.17 | 0.11 | 0.04 | -0.02 | 0.08 | 0.12 | 0.14 | 0.07 | 0.02 | 0.02 | 0.09 | 0.20 | -0.08 | 0.11 | 0.02 | -0.05 | -0.02 | -0.01 | -0.05 | 0.08 | -0.03 |
| (L) Metscore | -0.33 | 0.10 | 0.02 | 0.07 | 0.04 | 0.06 | 0.08 | 0.06 | 0.04 | 0.03 | 0.04 | 0.09 | -0.11 | -0.01 | 0.05 | 0.17 | -0.04 | 0.02 | -0.04 | 0.06 | 0.11 |
| (L) Str power | -0.17 | 0.04 | 0.08 | -0.05 | -0.02 | 0.15 | 0.16 | -0.05 | 0.00 | 0.07 | 0.16 | 0.16 | -0.03 | 0.04 | 0.00 | -0.08 | -0.05 | -0.01 | -0.03 | 0.12 | -0.05 |
| Spring | 0.01 | -0.01 | 0.03 | -0.01 | -0.02 | 0.05 | 0.05 | -0.02 | -0.03 | 0.02 | 0.01 | 0.02 | 0.01 | -0.03 | -0.01 | 0.01 | -0.01 | -0.05 | 0.05 | 0.07 | -0.04 |
| Summer | -0.01 | -0.01 | 0.01 | -0.02 | -0.02 | 0.07 | 0.09 | -0.06 | 0.04 | 0.05 | 0.10 | 0.07 | 0.08 | -0.01 | -0.06 | -0.15 | -0.04 | -0.03 | -0.04 | 0.10 | -0.05 |
| Fall | 0.03 | 0.04 | -0.05 | 0.02 | 0.03 | -0.09 | -0.10 | 0.04 | 0.00 | -0.01 | -0.08 | -0.09 | -0.06 | -0.02 | 0.05 | 0.10 | 0.05 | 0.04 | -0.06 | -0.12 | 0.06 |
| Winter | -0.03 | -0.02 | 0.01 | 0.00 | 0.01 | -0.03 | -0.04 | 0.04 | -0.02 | -0.06 | -0.03 | 0.00 | -0.04 | 0.07 | 0.02 | 0.04 | 0.00 | 0.05 | 0.05 | -0.05 | 0.03 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 |
| 2010 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2011 | -0.07 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2012 | -0.06 | -0.07 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2013 | -0.06 | -0.07 | -0.06 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2014 | -0.07 | -0.08 | -0.07 | -0.07 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2015 | -0.06 | -0.07 | -0.06 | -0.06 | -0.07 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2016 | -0.04 | -0.05 | -0.04 | -0.04 | -0.05 | -0.04 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2017 | -0.04 | -0.05 | -0.05 | -0.04 | -0.05 | -0.04 | -0.03 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |
| 2018 | -0.07 | -0.08 | -0.07 | -0.07 | -0.08 | -0.07 | -0.05 | -0.05 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |
| 2019 | -0.06 | -0.07 | -0.06 | -0.06 | -0.07 | -0.06 | -0.04 | -0.04 | -0.07 | 1.00 |  |  |  |  |  |  |  |  |  |  |
| Nominee | -0.01 | 0.02 | -0.01 | 0.01 | -0.01 | 0.01 | 0.00 | 0.00 | 0.05 | 0.04 | 1.00 |  |  |  |  |  |  |  |  |  |
| Winner | -0.03 | -0.01 | -0.01 | -0.01 | 0.03 | 0.01 | 0.02 | -0.02 | 0.05 | 0.04 | 0.81 | 1.00 |  |  |  |  |  |  |  |  |
| Remake | -0.01 | -0.04 | -0.05 | 0.00 | -0.03 | -0.02 | -0.05 | -0.05 | 0.01 | 0.04 | 0.02 | 0.03 | 1.00 |  |  |  |  |  |  |  |
| (L) D power | 0.01 | 0.01 | -0.01 | -0.01 | 0.00 | -0.02 | -0.02 | 0.00 | 0.01 | 0.01 | 0.37 | 0.39 | -0.02 | 1.00 |  |  |  |  |  |  |
| (L) Metscore | -0.03 | -0.01 | 0.01 | 0.03 | -0.01 | 0.00 | 0.03 | 0.00 | 0.03 | 0.03 | 0.53 | 0.46 | -0.02 | 0.19 | 1.00 |  |  |  |  |  |
| (L) Str power | 0.02 | -0.04 | -0.03 | -0.01 | -0.02 | 0.00 | 0.02 | 0.04 | -0.01 | 0.02 | 0.23 | 0.19 | -0.01 | 0.23 | 0.14 | 1.00 |  |  |  |  |
| Spring | 0.03 | -0.01 | -0.05 | -0.01 | -0.05 | 0.03 | -0.07 | 0.10 | -0.01 | 0.01 | 0.01 | -0.02 | 0.04 | -0.06 | 0.03 | -0.01 | 1.00 |  |  |  |
| Summer | -0.01 | -0.05 | 0.02 | 0.00 | 0.00 | 0.06 | -0.06 | -0.07 | 0.04 | 0.00 | 0.05 | 0.07 | 0.00 | 0.06 | 0.03 | 0.05 | -0.36 | 1.00 |  |  |
| Fall | 0.00 | 0.06 | -0.01 | 0.04 | -0.02 | -0.06 | 0.09 | -0.05 | 0.01 | -0.01 | -0.03 | -0.03 | -0.04 | -0.03 | 0.02 | -0.06 | -0.35 | -0.36 | 1.00 |  |
| Winter | -0.03 | 0.00 | 0.04 | -0.03 | 0.07 | -0.03 | 0.05 | 0.02 | -0.05 | 0.00 | -0.03 | -0.02 | 0.00 | 0.02 | -0.09 | 0.01 | -0.30 | -0.31 | -0.30 | 1.00 |

Appendix D VIF values models:

|  |  |
| --- | --- |
| Variable | VIF value |
| simpson\_index | 1.23 |
| log(boxofficemojo.com\_openingtheaters) | 2.11 |
| log(imdb.com\_runtime) | 1.93 |
| log\_MPAA | 2.13 |
| log(average\_budget) | 2.63 |
| log\_sequel | 1.21 |
| imdb.com\_spinoff | 1.06 |
| log\_remake | 1.09 |
| imdb.com\_basedonbook | 1.09 |
| imdb.com\_basedonplay | 1.06 |
| imdb.com\_basedoncomicbook | 1.30 |
| imdb.com\_basedonnovel | 1.17 |
| imdb.com\_basedonshortstory | 1.05 |
| log\_Nominee | 1.30 |
| log\_Winner | 1.31 |
| log\_dir\_power | 1.62 |
| log\_metascore | 1.10 |
| log\_starpower | 1.37 |
| Action | 2.00 |
| Adventure | 1.69 |
| Comedy | 2.09 |
| Fantasy | 1.35 |
| Crime | 1.45 |
| Drama | 1.72 |
| Mystery | 1.37 |
| Thriller | 2.02 |
| Romance | 1.40 |
| Sci.Fi | 1.48 |
| Biography | 1.42 |
| Sport | 1.23 |
| War | 1.24 |
| Family | 2.05 |
| Musical | 1.08 |
| History | 1.36 |
| Horror | 1.78 |
| Music | 1.16 |
| Documentary | 1.19 |
| Western | 1.04 |
| Spring | 1.64 |
| Summer | 1.66 |
| Fall | 1.65 |
| `2001` | 1.97 |
| `2002` | 1.96 |
| `2003` | 1.94 |
| `2004` | 2.02 |
| `2005` | 2.09 |
| `2006` | 2.16 |
| `2007` | 2.10 |
| `2008` | 2.06 |
| `2009` | 2.02 |
| `2010` | 1.99 |
| `2011` | 2.02 |
| `2012` | 1.90 |
| `2013` | 1.93 |
| `2014` | 2.02 |
| `2015` | 1.94 |
| `2016` | 2.20 |
| `2017` | 1.95 |
| `2018` | 2.19 |
| `2019` | 2.00 |

ViF values three regression models

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable:** | **Black** | **Asian** | **Hispanic** |
| **SI** | 1.30 | 1.46 | 1.38 |
| **Opening screens** | 2.14 | 2.13 | 2.21 |
| **Log(Runtime)** | 2.33 | 2.58 | 2.49 |
| **Log(MPAA)** | 2.21 | 2.17 | 2.14 |
| **Log(Average\_Budget** | 3.73 | 4.15 | 3.93 |
| **log\_sequel** | 1.31 | 1.30 | 1.25 |
| **imdb.com\_spinoff** | 1.10 | 1.23 | 1.13 |
| **log\_remake** | 1.12 | 1.16 | 1.14 |
| **imdb.com\_basedonbook** | 1.17 | 1.28 | 1.15 |
| **imdb.com\_basedonplay** | 1.00 | 1.00 | 1.00 |
| **imdb.com\_basedoncomic** | 2.02 | 2.58 | 1.76 |
| **imdb.com\_basedoncomicbook** | 2.38 | 2.78 | 2.10 |
| **imdb.com\_basedonnovel** | 1.16 | 1.28 | 1.18 |
| **imdb.com\_basedonshortstory** | 1.00 | 1.24 | 1.00 |
| **log\_Nominee** | 4.14 | 4.27 | 4.16 |
| **log\_Winner** | 3.47 | 3.41 | 3.40 |
| **log\_dir\_power** | 1.39 | 1.44 | 1.32 |
| **log\_metascore** | 1.57 | 1.69 | 1.61 |
| **log\_starpower** | 1.27 | 1.48 | 1.34 |
| **Action** | 2.41 | 2.62 | 2.45 |
| **Adventure** | 2.03 | 2.30 | 2.34 |
| **Comedy** | 2.03 | 2.23 | 1.99 |
| **Fantasy** | 1.42 | 1.73 | 1.54 |
| **Crime** | 1.58 | 1.45 | 1.62 |
| **Drama** | 1.92 | 2.05 | 2.09 |
| **Mystery** | 1.39 | 1.50 | 1.44 |
| **Thriller** | 2.16 | 2.18 | 2.20 |
| **Romance** | 1.43 | 1.26 | 1.41 |
| **Sci.Fi** | 1.71 | 1.89 | 1.79 |
| **Biography** | 1.61 | 1.75 | 1.62 |
| **Sport** | 1.44 | 1.37 | 1.38 |
| **War** | 1.36 | 1.45 | 1.44 |
| **Family** | 2.22 | 2.57 | 2.23 |
| **Musical** | 1.23 | 1.25 | 1.12 |
| **History** | 1.47 | 1.76 | 1.92 |
| **Horror** | 1.86 | 1.99 | 1.87 |
| **Music** | 1.28 | 1.24 | 1.38 |
| **Documentary** | 1.00 | 1.00 | 1.00 |
| **Western** | 1.15 | 1.00 | 1.00 |
| **Spring** | 2.00 | 1.95 | 1.69 |
| **Summer** | 2.06 | 2.02 | 1.78 |
| **Fall** | 1.96 | 1.71 | 1.82 |
| **`2001`** | 2.39 | 1.00 | 4.53 |
| **`2002`** | 2.59 | 2.63 | 2.86 |
| **`2003`** | 2.32 | 2.01 | 1.00 |
| **`2004`** | 2.04 | 2.03 | 2.89 |
| **`2005`** | 1.91 | 1.45 | 3.65 |
| **`2006`** | 2.71 | 2.72 | 5.57 |
| **`2007`** | 2.42 | 2.06 | 2.90 |
| **`2008`** | 2.59 | 1.43 | 3.80 |
| **`2009`** | 2.39 | 2.08 | 5.89 |
| **`2010`** | 2.26 | 1.00 | 4.62 |
| **`2011`** | 2.36 | 2.14 | 6.27 |
| **`2012`** | 2.46 | 2.82 | 2.02 |
| **`2013`** | 2.75 | 2.42 | 5.33 |
| **`2014`** | 3.44 | 2.48 | 10.36 |
| **`2015`** | 2.82 | 2.77 | 7.35 |
| **`2016`** | 2.12 | 2.36 | 3.62 |
| **`2017`** | 2.50 | 2.45 | 3.69 |
| **`2018`** | 4.58 | 4.68 | 12.47 |
| **`2019`** | 4.10 | 3.45 | 7.56 |

Appendix D: Explanation how the graph was made

Appendix E : Protest 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>

Appendix F: Entire Regression analysis interaction terms

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Model step T3 | Model step T2 | Model step T1 |
| `2001` | -0.101 | -0.103 | -0.074 |
| `2002` | 0.202 | 0.199 | 0.09 |
| `2003` | 0.219 | 0.223 | 0.075 |
| `2004` | 0.051 | 0.059 | 0.005 |
| `2005` | 0.201 | 0.194 | 0.156 \* |
| `2006` | 0.155 | 0.161 | 0.107 |
| `2007` | 0.183 | 0.187 | 0.128 |
| `2008` | 0.043 | 0.054 | 0.05 |
| `2009` | 0.2 | 0.201 | 0.187 \* |
| `2010` | 0.38 \*\* | 0.38 \*\* | 0.215 \*\* |
| `2011` | 0.442 \*\*\* | 0.434 \*\*\* | 0.374 \*\*\* |
| `2012` | 0.58 \*\*\* | 0.571 \*\*\* | 0.521 \*\*\* |
| `2013` | 0.754 \*\*\* | 0.736 \*\*\* | 0.639 \*\*\* |
| `2014` | 0.878 \*\*\* | 0.879 \*\*\* | 0.832 \*\*\* |
| `2015` | 0.914 \*\*\* | 0.898 \*\*\* | 0.831 \*\*\* |
| `2016` | 1.06 \*\*\* | 1.075 \*\*\* | 0.921 \*\*\* |
| `2017` | 1.043 \*\*\* | 1.052 \*\*\* | 0.914 \*\*\* |
| `2018` | 1.234 \*\*\* | 1.275 \*\*\* | 1.107 \*\*\* |
| `2019` | 1.119 \*\*\* | 1.13 \*\*\* | 1.098 \*\*\* |
| Action | 0.003 | 0 | -0.032 |
| Adventure | -0.149 \*\* | -0.132 \* | -0.127 \*\* |
| after\_jan\_2015 | -0.12 | -0.094 | 0.066 |
| Comedy | 0.162 \*\* | 0.154 \*\* | 0.119 \*\*\* |
| Crime | 0.02 | 0.006 | 0.052 |
| Documentary | 0.982 \*\*\* | 0.972 \*\*\* | 1.155 \*\*\* |
| Drama | 0.139 \*\* | 0.134 \*\* | 0.132 \*\*\* |
| Fall | -0.067 | -0.056 | 0.036 |
| Family | 0.244 \*\* | 0.246 \*\* | 0.179 \*\* |
| Fantasy | -0.05 | -0.05 | -0.052 |
| Biography | 0.261 \*\* | 0.251 \*\* | 0.161 \*\* |
| History | 0.206 | 0.214 \* | 0.132 |
| Horror | 0.016 | 0.018 | 0.071 |
| Based on book | -0.13 | -0.129 | -0.138 \* |
| Based on comic book | -0.329 \*\* | -0.277 \*\* | -0.302 \*\*\* |
| Based on novel | -0.131 \*\* | -0.123 \* | -0.11 \*\* |
| Based on play | 0.208 | 0.191 | 0.351 \* |
| Based on short story | 0.037 | 0.056 | 0.017 |
| sequel | 0.012 | 0.011 | 0.019 \*\* |
| spinoff | -0.073 | -0.083 | -0.053 |
| log(Budget) | -0.045 | -0.053 | -0.044 \* |
| log(Screens) | -0.409 \*\*\* | -0.41 \*\*\* | -0.473 \*\*\* |
| log(Runtime) | -1.635 \*\*\* | -1.625 \*\*\* | -1.418 \*\*\* |
| log(MPAA) | -0.65 \*\*\* | -0.633 \*\*\* | -0.763 \*\*\* |
| Log(Dir Power) | -0.012 \*\*\* | -0.012 \*\*\* | -0.009 \*\*\* |
| Log(Meta score) | -0.067 \*\*\* | -0.065 \*\*\* | -0.039 \*\*\* |
| Log(Nominee) | -0.01 \*\*\* | -0.01 \*\*\* | -0.008 \*\*\* |
| Log(Remake) | 0 | 0 | 0.003 |
| Log(Star power) | -0.004 | -0.003 | -0.004 \* |
| Log(Winner) | -0.015 \*\*\* | -0.015 \*\*\* | -0.015 \*\*\* |
| Music | 0.275 \*\* | 0.297 \*\* | 0.062 |
| Musical | -0.176 | -0.145 | -0.321 \*\* |
| Mystery | 0.006 | -0.003 | -0.062 |
| Romance | -0.118 \* | -0.112 \* | -0.084 \* |
| Sci.Fi | -0.052 | -0.058 | -0.115 \*\* |
| Sport | 0.038 | 0.039 | 0.105 |
| Spring | -0.027 | -0.013 | 0.008 |
| Summer | -0.102 \* | -0.098 | -0.014 |
| Thriller | 0.023 | 0.027 | 0.037 |
| War | 0.166 | 0.175 | 0.173 \* |
| Western | 0.709 \*\*\* | 0.708 \*\*\* | 0.483 \*\*\* |
| Inverted Simpson index |  |  |  |
| simpson\_index:a\_jan\_2015 |  |  |  |
| jan\_2015:Asian\_T1 |  |  | -0.181 \*\* |
| jan\_2015:Black\_T1 |  |  | -0.133 \* |
| jan\_2015:Hispanic\_1 |  |  | 0.147 |
| Asian\_T1 |  |  | 0.038 |
| Black\_T1 |  |  | 0.12 \*\*\* |
| Hispanic\_T1 |  |  | - 0.07 \* |
| jan\_2015:Asian\_T2 |  | -0.302 \*\* |  |
| jan\_2015:Black\_T2 |  | -0.065 |  |
| jan\_2015:Hispanic\_2 |  | 0.091 |  |
| Asian\_T2 |  | 0.095 |  |
| Black\_T2 |  | 0.194 \*\*\* |  |
| Hispanic\_T2 |  | -0.065 |  |
| jan\_2015:Asian\_T3 | -0.066 |  |  |
| jan\_2015:Black\_T3 | -0.045 |  |  |
| jan\_2015:Hispanic\_3 | 0.209 |  |  |
| Asian\_T3 | 0.035 |  |  |
| Black\_T3 | 0.246 \*\*\* |  |  |
| Hispanic\_T3 | 0.02 |  |  |
| R2 | 0.7363 | 0.7344 | 0.6965 |
| Adjusted R2 | 0.7204 | 0.7183 | 0.6877 |
| Number of Observations | 1105 | 1105 | 1105 |
| F Statistic | 46.18\*\*\* | 45.73\*\*\* | 79.27\*\*\* |

Appendix G: Regression analysis T1 entire dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable |  | T1 entire dataset | |  |
| Action |  | -0.106 \*\*\* | |  |
| Adventure |  | -0.055 | |  |
| Biography |  | | 0.239 \*\*\* |  |
| Comedy |  | | 0.113 \*\*\* |  |
| Crime |  | | 0.057 |  |
| Documentary |  | | 1.176 \*\*\* |  |
| Drama |  | | 0.147 \*\*\* |  |
| Family |  | | 0.215 \*\*\* |  |
| Fantasy |  | | -0.043 |  |
| History |  | | 0.149 \* |  |
| Horror |  | | 0.066 |  |
| Music |  | | 0.081 |  |
| Musical |  | | -0.096 |  |
| Mystery |  | | -0.03 |  |
| Romance |  | | -0.123 \*\*\* |  |
| Sci.Fi |  | | -0.063 |  |
| Sport |  | | 0.062 |  |
| Thriller |  | | 0.017 |  |
| War |  | | 0.082 |  |
| Western |  | | 0.42 \*\*\* |  |
| Fall |  | | 0.057 |  |
| Spring |  | | 0.02 |  |
| Summer |  | | 0.023 |  |
| Based on book |  | | -0.144 \*\* |  |
| Based on comic book |  | | -0.181 \*\*\* |  |
| Based on novel |  | | -0.101 \*\* |  |
| Based on play |  | | 0.39 \*\* |  |
| Based on short story |  | | 0.073 |  |
| Spinoff |  | | -0.045 |  |
| Log(Sequel) |  | | 0.026 \*\*\* |  |
| Log(Budget) |  | | -0.04 \* |  |
| Log(Screens) |  | | -0.484 \*\*\* |  |
| Log(Runtime) |  | | -0.572 \*\*\* |  |
| Log(MPAA) |  | | -0.599 \*\*\* |  |
| Log(Director Power) |  | | -0.041 \* |  |
| Log(Metascore) |  | | -0.316 \*\*\* |  |
| Log(Nominee) |  | | -0.146 \*\*\* |  |
| Log(Remake) |  | | 0.014 |  |
| Log(Star power) |  | | -0.03 \* |  |
| Log(Winner) |  | | -0.177 \*\*\* |  |
| Asian\_t1 |  | | -0.007 |  |
| Hispanic\_t1 |  | | -0.046  (0.006) |  |
| Black\_t1 |  | | 0.092 \*\*\* |  |
| R2 |  | | 0.675 | 0.6965 |
| Adjusted R2 |  | | 0.666 | 0.6877 |
| Number of Observations |  | | 2319 | 2361 |
| F Statistic |  | | 77.669\*\*\* | 79.27\*\*\* |

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

   (T2) two named {ethnicity} characters speak to one another.

   (T3) about something besides a white character. [↑](#footnote-ref-1)
2. IMDb's user base is 62% male, which is comparable to the gender distribution of ticket sales (59% male). In terms of age distribution IMDb's demographics align with those of filmgoers, except that older age groups are underrepresented on IMDb (MPAA, 2022; Similaweb, 2023). Moreover, IMDb's user base is predominantly American, according to Similarweb's 2023 data. Nevertheless, given the study's focus on the Hollywood film industry and the cultural nuances of racial representation, it makes sense to focus on the United States for the analysis. The results can potentially be transferred to countries with a similar cultural landscape and other demographic similarities with the United States. [↑](#footnote-ref-2)
3. Online streaming services have made it easier for people to watch older films, which has helped to maintain their popularity. As of October 2023, films such as Inglourious Basterds (2009), The Da Vinci Code (2006), and L.A. Confidential (1997) are all listed on Netflix's "Popular on Netflix" tab. This increased competition can make it more difficult for newer films to gain traction, especially those that fall outside of the "blockbuster" category. In the past, less popular new films only had to compete with the few hundred films available in a film store; now, they have to compete with thousands of films offered on online streaming platforms. [↑](#footnote-ref-3)
4. In the complete dataset, 15% of individuals linked to characters did not have associated images. Among unique actors and actresses, this figure increased to 28%. Meaning that 72% of the people (the ones with images) played 85% of the characters. Meaning people without images were more likely to play fewer roles. [↑](#footnote-ref-4)
5. Previous data showed that imageless characters tended to play fewer roles. The filtering process showed that imageless characters were also more likely to play less prominent characters. After the filter for generic names, 77% of characters without images were removed, compared to 66% of characters with images.

   This suggests that named characters were mostly played by people with images. As a result, the percentage of characters that were assigned an ethnicity through their name instead of facial recognition decreased from 15% to 11% of the dataset after filtering for non-named characters. As a result, imageless people are also more likely to play characters that were not authentic and inclusive (AIR). This suggests that filtering for AIR characters might also include filtering for characters played by people without images. This is an area for further exploration. [↑](#footnote-ref-5)
6. Two significant factors have demonstrated influencing racial diversity in previous research. In 2019, action films exhibited a notably higher likelihood of including characters from underrepresented racial and ethnic groups. Additionally, films with larger budgets tend to have more diverse casts, as they possess greater resources to hire a wide range of actors (Smith, 2020). Note that this is different from what was previously claimed. Large budget films have the ability to higher more diverse casts. Nevertheless, diverse films have lower budgets compared to non-diverse films. [↑](#footnote-ref-6)
7. Notably, budget demonstrated positive correlations with Screens (0.598), Runtime (0.522), and (G) Adventure (0.530). This correlation pattern is explicable, as substantial budgets are often associated with large-scale blockbusters, while lower-budget films tend to have more restrained advertising budgets (Kumar, 2022). Furthermore, the correlation between advertising expenses and the number of screens is pronounced enough to be considered substitutes, suggesting that a considerable production budget is linked to an extensive advertising budget, and the advertising budget and number of screens are interchangeable (Hennig-Thurau, Housten, and Walsh, 2007).

   It is also understandable that longer films would require or receive a larger production budget. Finally, Adventure films, due to their frequent need for extensive location shooting and elaborate special effects. This is also shown with the genre ‘Action’ also exhibiting a relatively high correlation with average budget (0.46).

   Conversely, the genre ‘Family’ had a substantial negative correlation with (log)MPAA (-0.639). This relationship is straightforward to explain. Additionally, strong positive correlations were observed between (log) Director power and Screens (0.530). This relationship was also demonstrated in (Prag and Casavant, 1994), which showed that star power was positively related to production budget and marketing expenditures. Production companies such as Disney likely see it as a less financial risk to invest in successful directors, leading them to invest in screens for these films as well. [↑](#footnote-ref-7)
8. Picture of #oscarsowhite protest and article shown in Appendix E displaying the significance of the event. [↑](#footnote-ref-8)