AN EXTRA TREES REGRESSOR TO PREDICT CONTENT POPULARITY ON THE NETFLIX PLATFORM IN KENYA

By

ESTHER G KARUKU

MASTER OF SCIENCE IN DATA ANALYTICS

KCA UNIVERSITY

2024

AN EXTRA TREES REGRESSOR TO PREDICT CONTENT POPULARITY ON THE NETFLIX PLATFORM IN KENYA

By

ESTHER G KARUKU

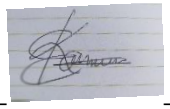
A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF MASTER OF SCIENCE IN DATA ANALYTICS TO THE SCHOOL OF TECHNOLOGY OF KCA UNIVERSITY

SEPTEMBER 2024

DECLARATION

I declare that this paper is my original work and has not been previously published or submitted elsewhere for an award of master's degree. I further declare that this dissertation contains no material written or published by other people except where reference is made, and the author is duly recognized.

Student Name: Esther Gathoni Karuku Reg No: 21/00662

Sign: _____  ___ Date: 20th, September, 2024

This dissertation has been submitted for examination with my approval as the appointed university supervisor.

Digitally signed by Dr. Lucy Waruguru Mburu
Date: 2024.10.22 19:23:35 +03'00'

Date: 20/ 09/2024

Dr. Lucy Waruguru Mburu

ABSTRACT

This study aimed to develop an Extra Trees Regressor model to predict content popularity on the Netflix platform in Kenya. The data used in this study was collected from June 28, 2021 to March 24, 2024.

The experimentation yielded compelling results, with the Extra Trees Regressor demonstrating superior performance compared to both Linear Regression and Ridge Regression. Extra Trees Regressor showed consistently lower error rates across all metrics (MAE, MSE, RMSE, and MAPE) except RMSLE suggesting a high degree of accuracy in predicting content popularity for Kenyan audiences.

A high R^2 value (0.9140) indicates the Extra Trees Regressor model effectively captured the relationship between content attributes and content popularity.

The study revealed the two most important predictors of content popularity are the show title and the director contributing to the ongoing investigation of the content popularity problem globally.

Keywords: SVOD demand, local content, Netflix Kenya top 10, popular shows in Kenya

ACKNOWLEDGMENT

I want to express my gratitude to my friends, current and future family for their support through the years. You are the sum of all the good parts of me.

Thank you to Dr. Lucy W. Mburu for being my supervisor and taking time to help at each stage. Thank you to my colleagues for their moral support and active participation in group work projects.

I would like to thank the School of Computing for enabling access to this course to all through remote learning.

ACRONYMS AND ABBREVIATIONS

AI - Artificial Intelligence

CDN - Content Delivery Network

ETR - Extra Trees Regressor

ML - Machine Learning

OTT – Over The Top

SVOD – Subscription Video on Demand

GLOSSARY

Bias identification - Uncovering unfair or prejudiced tendencies within a model's predictions.

Generalizability testing - Evaluating how well a model's performance translates to unseen data.

Model comparison - Assessing the relative strengths and weaknesses of different models for a specific task.

Mise-en-scène - The arrangement of elements within a scene or frame, influencing visual storytelling.

Pre-processing - A pre-processor is a program that interprets its input data to generate a result that is used as output to another program such as a compiler.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

The television landscape in Kenya has significantly transformed over the past 5–10 years. Traditional players have adapted their content strategies and revenue models in response to evolving audience behaviours. While free-to-air and direct-to-home broadcasting once dominated, the rise of pay-TV and Over The Top (OTT) streaming services has diversified audience choices, challenging the industry's established norms(Kacungira & Owuor, 2023). The transitioning from the conventional to digital formats has fundamentally revolutionised and popularise the streaming platforms such as the Netflix to a great extent in the landscape of movie culture, the patterns of media consumption and the strategies, change in acceptance mode and acceptance psychology of traditional movie and television, and hence breaking the confinement of time and space(Susilo & Harliantara, 2023). Netflix is a pioneering and successful streaming company and its growth is replacing television, radio and cinema (Alsuhaim, 2024); with the world's leading entertainment services, with millions of subscribers all over the world. Almost 15% of Internet downlink traffic is attributed to Netflix and is poised to control 26.58% of the global market for video streaming services (Liu, 2022). This platform has the largest subscription video on-demand platform in more than 190 countries and over 30 languages subscribed with 195 million subscribers environments (Kamarudin et al., 2022; Wayne, 2022). The popularity of Netflix is growing tremendously (Laban et al., 2020) and the number of VOD subscribers of the biggest platforms is growing year by year, with the so-called online time-shift (Pluta & Siuda, 2022). The collaboration of United States video platform

Netflix with South Korea on the web series 'The Glory' in the early years of 2023, has increased Korean subscriptions (Wei, 2024). With the rise of OTT (Over the top) platforms' popularity, has established itself as an advocate of multiculturalism unlike any other contemporary in its time. One of the contributing factors for its increase in subscription is the diversity of content, high availability, quality of service provision and also the ability to avoid advertising forcing young people to become Netflix subscribers in today's world's two movie watching (2023; Akinci & Başer, 2020). The emergence of Netflix during the global outbreak of the Covid-19 epidemic in 2020 (Wang et al., 2022) has led to over 237.5 million paid members by the end of 2022 across 190 countries hence transforming customers' experience and expectations as well. According to Netflix, revenue was $32 billion in fiscal year-on-year and $7.85 billion in the fourth quarter. Netflix forecast revenue of $8.17 billion in the new quarter (Jiang, 2024). OOT players have enabled the consumers to choose their diverse choices of content, watching time, and place independently hence increasing their popularity across locations and demographics (Dastidar, 2021; Q. Li & Yi, 2022). Better device compatibility with a wider and volatile catalogue, along with better user experience and audio has increased the usage of Netflix on online video streaming space hence making it competitive among several over-the-top (OTT) platforms to capture the viewer's attention (Lad et al., 2020). The ability of Netflix to differentiate itself from its competitors in terms of creating various content for different markets, blending cultural differences, and eliminating geographical boundaries has gained prominence from producers in different branding contents such as series, movies, and documentaries (Yilmaz & Erdem, 2022).

The emergence of on-demand content consumption from major platforms like Netflix, Amazon, and HBO has led to greater audience fragmentation due to the vast selection of

available content. The transition to subscription-based models, which remove advertising, along with the rise of mobile and out-of-home viewing, has established a new ecosystem. In this context, traditional audience measurement systems are inadequate for evaluating success(Neira et al., 2021).

The rise of streaming platforms (video-on-demand [VOD] services), such as Netflix, HBO Now (or HBO Go in Poland), Amazon Prime, Hulu or Disney Plus has revolutionized the entertainment education industry with slightly over 20 percent claim that they have taken some pro-health actions(Pluta & Siuda, 2022). According to research published by Media Partners Asia (MPA), Netflix has led a premium video share of 40%, broadening the appeal of its international catalogue. The massive and escalate use of Netflix in Indonesia has been provided the Ministry of Communications and Informatics as an addition from the Indonesian regulator/government (Djamzuri & Mulyana, 2022). The entertainment landscape in India has transformed significantly due to the rapid growth of Over-The-Top (OTT) platforms such as Netflix, Amazon Prime Video, Disney+ Hotstar, SonyLIV, and Zee5. These new platforms offer a mix of international and local content which is appealing to the diverse Indian audiences providing a competitive market featuring both global and local providers, increased internet access and affordable data plans, leading to a demand for on-demand, personalised content (Panda et al., 2023). Subscription video-on-demand (SVOD) players such as Netflix are not only changing the way we define television, but also the weight we ascribe their circulation power, as they put pressure on audio-visual financing and increase competition for distribution and content acquisition and commissioning.

Moreover, these services operate as platforms, creating closed circuits that facilitate their control not only over their distribution and network infrastructure, but also over financing,

consumer relationships and programming rights (Iordache et al., 2023). Majority of the audiences prefer Netflix platforms because they lead in digital space, focusing on teaching foreign languages, supporting foreign language, and translation applications working on these programs (Türkmen, 2020). Spearman's correlation analysis was performed for each indicator of RACOI and deduced that Netflix popularity score was correlated with the digital video views indicator of RACOI. As a result of analysing the platforms, it was found that the popularity of the content provided by Netflix and Tving was relatively high (Hong et al., 2021).

Additionally, a survey was conducted to examine the Netflix behaviour and its digital wave in Portugal, Spain, Belgium, Italy, Turkey, Georgia and Malaysia. Of the people who answered the survey, 90.1% were stream consumers, but only 59.1% had premium TV channels. From those 90.1%, 58.3% also said that they watched streams between two and four times per week, but the majority of premium TV channel subscribers (63.8%) replied that they watch TV less than twice in a week. This affirms that consumer habits are changing, and people are getting used to the digitalization era (Au-Yong-Oliveira et al., 2020).

Nevertheless, Netflix has an important role in the business world, more than just being the popular streaming company and the provider of great impact in the new trends of content delivery. However, the company fights a tough rivalry because there is cutthroat and never-ending competition from a number of new platforms. These competitive dynamics suggest that Netflix has to constantly reinvent its strategies in order to sustain competitive advantage.

For instance, business analysis has indicated a continual saturation in the existing revenue generation in post COVID time and that tremendously increasing the significant activity of competitors in the industry of streaming platforms for the distribution of video content (Lozić et al., 2024). Studies on the dubbing of English children's cartoons into Modern

Standard Arabic highlighted the use of negative face speech acts, including compliments and rejections. The dubbing process was shaped by linguistic and cultural factors, particularly social norms and power dynamics in Arab culture, such as respect for elders. Translators employed politeness strategies, including off-record and bald on-record approaches, to ensure cultural appropriateness and preserve the integrity of familial relationships depicted in the content (Alsuhaim, 2024).

Similarly, Korean media industry has been facing aggressive international content strategies. By acquiring all IP rights to its Korean originals and global streaming rights to many dramas, production companies and Korean television stations struggle to profit despite the global popularity of these shows. This strategy may lead to a consolidation of platform imperialism, undermining local media entities (J. H. Park, 2022). Intense competition among subscription video on-demand (SVOD) providers compels Netflix to focus on subscriber growth, particularly in high-potential markets like Indonesia and Korea. One of its key strategies is producing Netflix originals, which are films and series created in collaboration with local producers and exclusively distributed on the platform (Adiprabawa, 2024).

The findings of Netflix's rapid expansion in European markets and its growing investments in original content points to an increasing diversity in catalogue composition, as well as growing investments in European works. It was also noted that Netflix contributes to existing power imbalances between markets with the US content dominates the four European catalogues, while investments in European original content considerably favour strong media markets over weaker ones (Catalina Iordache, 2022, 2024). Netflix executives have framed the service's content commissioning process as one offering a greater equality of opportunity than other global buyers. This is achieved through the repeated invocation of the notion that Netflix

commissioned series ‘come from anywhere’ and are then ‘loved everywhere’ (Penmatcha, 2022). Netflix has not yet complied with the existing regulations in Indonesia. The films shown on Netflix have a wide variety of genres and content (Phillo & Ruchimat, 2022). There are instances where several film makers have failed to regulate what they broadcast because the content is not within their mandate. Such film’s content contains elements of violence and pornography which infringes and violates the concept of broadcasting based on Broadcasting Law (Putri & Kleden, 2022).

The exploration below delves into a few real-life cases, spanning diverse genres, to illuminate the multifaceted challenges of predicting content success (Gandasari et al., 2023). “John Carter”: Despite its Martian setting and rich source material, the film's marketing and execution failed to capture the imagination of a broad audience, resulting in a $200 million loss. “Mars Needs Moms”: This animated family film aimed to charm both children and parents, but its motion-capture animation and marketing missteps led to a $100 million loss (Kamarudin et al., 2022).

“The Lone Ranger”: This Western-inspired action film, burdened by casting controversies and a bloated budget, failed to attract a large enough audience, resulting in a $150-190 million loss. “Cutthroat Island”: Despite the enduring popularity of pirate films, this action-adventure offering succumbed to poor critical reception and marketing issues, culminating in a $125-147 million loss. “Gigli”: This romantic comedy-crime film, despite its high-profile stars, suffered from a convoluted plot and negative reviews, leading to a $75-80 million loss (Kamarudin et al., 2022).

SVOD platforms have had their share of failings as well. "Girlboss" (Netflix): A comedy-drama series, inspired by the life of Nasty Gal founder Sophia Amoruso, was canceled

after one season due to mixed reviews and reportedly low viewership. "Haters Back Off" (Netflix): Based on the YouTube character Miranda Sings, this comedy series failed to attract a broad audience and was cancelled after two seasons (Gandasari et al., 2023).

"Take Me Home" (Showmax): This Kenyan dating show explored the complexities of love by offering six couples in committed relationships the chance to date their exes. Filled with emotional revelations and challenging decisions, it sparked controversy but ultimately didn't secure a second season (Abaya, 2022).

"Mom vs Wife" (Showmax) is a Kenyan reality cooking show hosted by media personality Betty Kyallo. The show garnered interest but also faced criticism for its superficial and stereotypical portrayal of family dynamics and gender roles. Whether it gets renewed for a second season remains to be seen (Akpan, 2023).

These unfortunate incidents showcase the diverse ways forecasting errors can manifest across movie studio blockbusters and SVOD services' original releases. From misreading audience preferences to marketing misfires, the road to financial failure is paved with complex challenges.

This study primarily focused on the application of machine learning techniques to predict the popularity of content on the Netflix platform. Netflix is a streaming service offering movies, TV shows, documentaries, and more, all accessible on demand for a monthly subscription. With content in over 190 countries, it reaches a massive audience, constantly adding new titles and even producing its own award-winning original content.

This research venture, while inspired by the potential for broader applications in diverse contexts, placed its immediate focus on the realm of the Netflix audience in Kenya, specifically analysing the platform's weekly top ten lists.

It is important to note that although this study concentrated on the Netflix audience, the applications for this study extended far beyond the domain of Netflix. The learnings could be applied by individual content creators on digital platforms like YouTube, stakeholders and investors in movie production studios, producers in traditional linear television, creative directors crafting commercial advertisements, and social media content makers.

Generally, prior research looking at algorithmic culture as a user-centred initiative has utilised Netflix's recommender algorithm. In this respect, this study has been able to recommend available content to subscribers (Varela & Kaun, 2024). Extra Trees regressor algorithm has been used to predict accurately from the BRMS stock price data. From the outcome derived, it can be postulated that Extra Trees achieves a rather reasonable accuracy for up to the 6th day after training set with a MAPE of <0. 1%. Further, tree-based methods including regression tree and classification tree have been used extensively in numerous analyses. The level of accuracy of the algorithms, and small error is one of the reasons why this method is popular (Park & Lee, 2022).

In conclusion, Netflix is an important player on the business market, being the leader of the market in streaming services and an eminent contributor to the changes that occur in the field of content consumption. However, it has some threats such as the emergence of new and relevant platforms that put pressure on this site to change its strategies and methods regularly. Therefore, by using algorithms such as the Extra Trees algorithm the predictive analysis can be further improved for example to predict content popularity and the extent to which users will engage with it. Thus, it is possible to avail such technologies for enhancing recommendations and, in turn, enhancing the overall impression to sustain competitiveness in the rapidly evolving

market. If the firm is to continue to be successful, the firm needs to adapt to the trends seen in this industry.

1.2 Statement of Problem

Despite the notable growth of Netflix in Kenya, there is limited knowledge about the factors that influence content popularity on the platform in this market (Yao, 2023). A notoriously difficult task, typically for the glamorous film and TV industry, is the ability to forecast content populism before the content's inauguration (Gutiérrez et al., 2020). The model will examine various factors to provide a comprehensive understanding of what drives content popularity on Netflix within the Kenyan context. In the quest of understanding the content that drives the Netflix popularity, a study was conducted to rate the performance of the Netflix. The results depicted that Naïve Bayes algorithm had an overall accuracy of 72% compared to the Decision Tree with an average of 70% and KNN with an average of 61%. This suggests that out of the above-discussed algorithms, namely K-Nearest Neighbour, Naïve Bayes, and Decision Tree algorithms, Naïve Bayes performs best in the rating classification.

A similar study revealed that Extra Trees outperformed the Naïve Bayes because it utilises the ensemble of decision trees which are less likely to over fit and can accurately handle non-linear relationships between the key factors unlike the Naïve Bayes which does not perform well once the independence assumption between the predictors holds. Additionally, Naive Bayes is more applicable when used for simple models as it does not have the capacity to handle interactions while Extra Trees is more acceptable for the prediction of content popularity on Netflix. In the case of streaming data in Netflix (Martiello Mastelini et al., 2023), proposed that

Online Extra Trees algorithm takes less time and space than Adaptive Random Forest algorithm and the method is applicable in real-world applications. While both the Extra Trees algorithm and recommender systems can aid in understanding popularity of content on Netflix.

Author	**Study**	**Results**	**Gap**
Zulkarnain et al.	Performance Comparison of K-Nearest Neighbour, Naive Bayes, and Decision Tree Algorithms for Netflix Rating Classification	Naïve Bayes algorithm had an accuracy of 72%, Decision Tree with 70% and KNN with 61%	72% accuracy is low
Dissanayake et al.	Early Prediction of Movie Success Using Machine Learning	Multiple Linear Regression, Polynomial Regression, SVR, Decision Tree Regression, Random Forest Regression	Mean Squared Error and Root Mean Squared Error had high values due to overfitting and outliers
Yuan Ni et al.	Movie Box Office Prediction Based on Multi-Model Ensembles	XGBoost, LightGBM, CatBoost, GBDT, Support Vector Regression, Random Forest	Average R2 was low at 0.8476 MAPE, MAE, MSE, RMSE were abnormally high due to presence of outliers
Mastelini et al.	Online Extra Trees Regressor	Extra Trees Regressor takes less time and space than Adaptive Random Forest algorithm in real-world applications	Study comprised 22 datasets and none included any data on content popularity

Table 1 Summary of previous studies

This study seeks to fill this gap by creating a predictive model to forecast the popularity of Netflix content in Kenya which completely distinguishes it from the recommender system which is used to purely suggest content to users based on their preferences and behaviour.

The results of this study will assist Netflix and other streaming services in optimising their content acquisition, marketing, and distribution strategies to better align with the preferences of Kenyan audiences, ultimately enhancing subscriber growth and retention in the country. Furthermore, the predictive model developed in this research can serve as a framework for other researchers and practitioners looking to explore content popularity in different markets or across other streaming platforms.

1.3 Main Objective

The main objective of this study is to develop an Extra Trees Regressor model to predict the popularity of Netflix content in Kenya.

1.4 Specific Objectives

The research objectives that facilitated the achievement of the aim above included:

1. To identify the key factors that influence the popularity of Netflix content in the Kenyan market.
2. To address the challenges in forecasting content popularity in the film and TV industry.
3. To evaluate the performance results of the machine learning algorithms for predicting Netflix popularity in Kenya.

1.5 Research Questions

The research aimed to answer the below questions:

1. What are the key factors that influence the content popularity of Netflix Kenyan market?
2. What are the challenges in forecasting content popularity in the film and TV industry?

3. Which machine learning algorithms were evaluated for predicting Netflix popularity in Kenya and what were their performance results?

1.6 Significance of Study

The SVOD landscape thrives on captivating content that resonates with viewers. Yet, predicting what sparks audience engagement remains an elusive puzzle, with countless shows succumbing to the silent abyss of low viewership. In this complex ecosystem, a machine learning model to predict content popularity on the Netflix platform in Kenya is quite useful. This exploration delves into the benefits such a model could offer various stakeholders, both within and beyond the entertainment industry.

For content creators and producers, navigating the murky waters of audience preferences can be daunting. A ML model, trained on historical data could assist in optimising resource allocation. By predicting content popularity with greater accuracy, studios can prioritise projects with higher projected viewership, maximising their return on investment (ROI) (Chen et al., 2023). This data-driven approach could help avert costly flops and allocate resources more efficiently towards content with a stronger chance of success.

This model could also assist to mitigate financial risk. Green lighting a show involves significant financial risk. A prediction tool could provide crucial information about potential audience reception, enabling studios to make more informed decisions about project selection and budgeting (Brundage et al., 2020). By anticipating potential viewership levels, studios could mitigate financial risks associated with producing content that may not resonate with audiences.

This tool could help tailor content for specific audiences. The model could identify specific audience segments with high predicted demand for certain genres, themes, or character

types. Creators could then strategically leverage this information to tailor their narratives and characters, catering to specific demographics and interests, potentially increasing the likelihood of connecting with viewers (Hu et al., 2020).

For SVOD owners, this model could streamline the content acquisition process. Platforms could leverage predictions to prioritise acquiring content with high projected viewership, potentially outbidding competitors and securing exclusive rights to shows with broad audience appeal (Karp, 2020). This could give them a competitive edge in attracting and retaining subscribers.

In marketing and promotion, precise ML tools could inform targeted marketing campaigns. Resources could be focused on promoting shows with higher predicted viewership, maximising the effectiveness of marketing efforts, and ensuring resources are not wasted on content with limited audience potential (Lee et al., 2022). This could also assist in subscription pricing. SVOD services could utilise ML forecasts to adjust pricing strategies based on predicted subscriber behaviour and content popularity. This data-driven approach could optimise revenue generation and potentially personalise subscription plans based on individual viewing preferences.

Investing in the SVOD industry carries inherent risks. For investors and analysts, accurate ML forecasts could provide valuable insights to reduce investment risk. By identifying platforms and studios with strong content pipelines and predicted audience growth, investors could make more informed investment decisions, potentially reducing the risk of backing ventures with limited success prospects.

This model could also optimise capital allocation. ML forecasts could inform the allocation of investments across different platforms and content creators. Investors could

prioritise companies with more promising content strategies and higher predicted return on investments, contributing to the overall growth and stability of the SVOD landscape.

While the primary beneficiaries reside within the SVOD industry, the ripple effects of a ML forecasting model could extend outward to the public. More efficient content production, informed by precise ML predictions, could lead to a wider variety of high-quality shows available on streaming platforms, catering to diverse interests and preferences. This could enhance the overall viewing experience for the public, offering a broader range of engaging content to choose from (Liu et al., 2022).

Advertisers are also potential beneficiaries. Identifying shows with specific audience demographics through ML forecasts could enable advertisers to place their products more effectively. By targeting ads based on predicted viewership patterns, they could reach their target demographics with greater precision, potentially increasing the effectiveness of their advertising campaigns. Lastly, are researchers and policy makers. Accurate data could inform research on audience behaviour, content trends, and the impact of SVOD on the broader media landscape. This data could be invaluable for researchers studying media consumption patterns, content production trends, and the evolving dynamics of the entertainment industry.

1.7 Motivation of Study

The burgeoning realm of Subscription Video on Demand (SVOD) has fundamentally reshaped entertainment consumption worldwide. Yet, while research diligently maps viewer preferences and content trends, a significant population gap persists, understanding SVOD demand within the Kenyan context remains largely unexplored. This study, fuelled by a potent

confluence of personal and academic motivations, aimed to illuminate this uncharted territory, unveiling the unique dynamics shaping SVOD demand in Kenya.

Firstly, a personal connection to Kenya ignited the researcher's passion for this study. Having witnessed first-hand the growing popularity of streaming platforms like Netflix in Kenyan households, the researcher was deeply curious about the specific content that resonates with local audiences. Witnessing diverse viewing habits within social circles, each influenced by cultural nuances and individual preferences, sparked a keen desire to delve deeper into this multifaceted phenomenon (Akpan, 2023).

However, this personal interest goes together with a broader academic imperative. Research on SVOD demand exhibits a concerning population gap: Kenyans and, by extension, many other African populations, were often absent from studies focusing on audience preferences and content consumption patterns (Hegde et al., 2020). This lack of representation created a skewed understanding of the global SVOD landscape, failing to capture the diverse voices and viewing habits present across different cultural contexts. This study aspired to bridge this gap, offering valuable insights into the Kenyan SVOD market's unique characteristics and contributing to a more comprehensive understanding of global viewing trends.

Beyond simply filling a research void, Kenya presented a unique and dynamic study brimming with cultural richness and evolving consumption patterns. The nation boasts a rapidly growing youth population, increasing internet penetration, and a burgeoning appetite for digital entertainment (Asu, 2021). These factors, coupled with Kenya's diverse cultural tapestry, contribute to a unique viewing landscape ripe for exploration. Understanding the interplay between these cultural nuances and SVOD demand necessitated research tailored to the Kenyan context.

For instance, local language preferences, storytelling traditions, and socio-economic realities influenced content choices and viewing habits in ways not captured by studies focused on Western audiences (Nguyen, 2022). Kenyan filmmakers and creators are increasingly finding their voices on global platforms, further necessitating research that acknowledged and valued their contributions (Abaya, 2022). By examining Kenyan SVOD demand, we gained valuable insights into the evolving preferences of African audiences, contributing to a richer and more inclusive understanding of the global media landscape.

Furthermore, this study transcended mere academic inquiry. Its findings held the potential to benefit diverse stakeholders. SVOD services have benefited from understanding Kenyan viewer preferences to inform content acquisition strategies, marketing campaigns, and platform development efforts, and catered to the specific needs and desires of the local audience (Cheng et al., 2021).

Knowing what resonates with Kenyan viewers guided content creators in tailoring their narratives and productions to better connect with the local market (Akpan, 2023). Policymakers and researchers also found this study helpful. Insights into content consumption patterns informed cultural policy and media research initiatives within Kenya and beyond (Hegde et al., 2020).

The researcher's motivation for this study extended beyond personal curiosity and academic rigour. It was driven by a desire to bridge the population gap, celebrate the diversity of the SVOD landscape, and empower stakeholders with knowledge that can enhance the viewing experience for Kenyan audiences and contribute to a more inclusive global understanding of media consumption trends. This research illuminated the Kenyan SVOD scene and contributed

to a richer and more representative dialogue on the future of entertainment in a culturally diverse world.

1.8 Scope of Study

This research investigated the potential of a regression model in predicting the popularity of content on the Netflix platform. While acknowledging its unique SVOD landscape, the research posited that the challenges and opportunities it presented resonated with broader global contexts. Insights gleaned from this focused investigation hold relevance and applicability across diverse SVOD settings and regions.

By delving into the Kenyan population, the research strived to evaluate the effectiveness of the model in predicting popularity of content. The study also aimed to identify the practical challenges and considerations associated with integrating regression techniques into such models within the SVOD context. Third, to uncover the potential benefits of the prediction model for enhancing success of content.

Ultimately, this research aimed to contribute to the development of more effective and contextually relevant movies and TV shows. By doing so, it envisioned advancing the entertainment experiences and outcomes of content creation in Kenya.

1.9 Structure of Research

This research was structured into distinct chapters, each serving a specific purpose in achieving the overall research objective. This section describes the content of each chapter that is in the study. The forthcoming chapter, literature review, presents a comprehensive analysis of existing research relevant to the study's problem statement. The review aimed to provide a

thorough understanding of the current knowledge base on the topic, serving as the foundation for the study.

Next is methodology. This chapter delved into the research design and methods employed in the study. It detailed the chosen research approach, data collection strategies, and the overall framework guiding the empirical investigation (Newhart & Pattern, 2023). The chapter explicitly discussed the rationale behind the chosen methodologies, allowing readers to grasp the research process and its alignment with the study's objectives.

The subsequent chapter formed the core of the research, presenting the key findings and their interpretation. It detailed the development and refinement of the proposed model, followed by a rigorous analysis of the collected data to uncover significant insights. Extensive discussions contextualised the findings, providing a deeper understanding of their implications.

The last chapter synthesised the findings of the entire study, offering a concise and conclusive report. It presented conclusions drawn from the results, encapsulating the core contributions, experiences, and challenges encountered during the research process. Notably, the chapter extended beyond mere summarization by offering valuable recommendations stemming from the research outcomes. These recommendations aimed to inform future actions and potential advancements within the study's domain. Appended to the core chapters are supplementary materials deemed essential for understanding the research. This section provided access to critical information such as the project budget, anticipated timelines for various phases, and selected raw data, thereby enriching the overall comprehensiveness of the research undertaking.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Netflix is regarded as a leader among the global OTT service providers with over 167 million streaming subscribers. Forecasters have estimated that the number of worldwide Netflix streaming service subscribers will reach nearly 237 million by 2025 (Lee et al., 2021). Similarly, academic studies have considered OTT as an innovative service based on cutting-edge technology, and scholars have employed the diffusion of innovation theory or technology acceptance framework to understand consumers' OTT adoption behaviour (Shim et al., 2022).

The widespread use of social media as a marketing tool during the last decade has been responsible for attracting a significant volume of academic research, which, however, can be described as highly fragmented to yield clear directions and insights (F. Li et al., 2023). Additionally, the great connectivity offered by the Internet has given the entrepreneurs powers of utmost importance for the good achievement of concrete actions according to their personality types and for relevant success in their entrepreneurial projects (González-Padilla et al., 2024).

This section will review an existing theory and model that explain the popularity of Netflix at the theoretical framework. It then reviews existing techniques used in content

popularity prediction providing a detailed analysis of their strength and weakness as well. The literature review will also address the challenges facing the content popularity in Netflix by exploring the dynamic nature viewer preferences and market trends as well as the bias and limitation that accrues from the consumer behaviour.

2.2 Theoretical Background

This section will review an existing theory and model that is related to the content popularity in Netflix. Among the many key theories in film theory, this section will investigate the cultural proximity theory and Audience Engagement model.

2.2.1 Cultural proximity Theory.

According to cultural proximity theory the consumers have a higher tendency of identifying with such contents due to having similar cultural experience as the contents portrayed. This theory is very relevant if considered from the perspective of NetFlix where content can be classified into local and global. The theory concerning the content trends in Netflix has been grouped into local content, global content, audience preference, cultural content and implication for media strategy amongst others.

The content of media has local and global dimensions that are important for the communication with a wide range of audiences. The distinction between local and global programming can be used where local programming engages viewers with familiarity while global programming does this with novelty and exoticism. This way, the elements involved would be recognized and well balanced by a media creator thus a perfect mix would be made for a bigger and more diversified audience experience. This particular approach enhances the view

of programmes but goes further in the promotion and acknowledgment of various cultures and groups.

Local content assumes a reflection of the cultural, historical and social setting of the given society. By watching this picture, it is possible to recognize the life, problems, and achievements of the audience members, and thus establish an identity. For example, a television drama that focuses on the daily life of families in a certain city may incorporate local and regional language, traditions, or social problems. It seems to capture the true local experiences and emotions hence the need of local people to identify with them. Besides, it is quite noteworthy that concerns like family or community relations, problems and issues of the locals are often the focus of the local content. Such audiences will be more likely to participate and commit on the account of their own experiences in the administrations.

For instance, a series that focuses on the social change as a result of economic transformation Prime focus of residents of a certain town may make a particular audience develop certain feelings such as nostalgia or empathy to the events demonstrated. Local content may also be associated with the development of pride in culture and historical traditions of a region. If people recognize their cultural practices, holidays and culturally sanctioned behavior as portrayed in the media, their ethnic identity and group cohesion is boosted. It may also create larger audiences since people are willing to support their culture hence be a call for production of more programs that showcase their culture.

Global content involves issues that concern everyone regardless of their cultural background and they include: love, friendship, enmity and other related issues, adventurous issues for instance. These are general stories, often relevant for people irrespective of their geographic location as regressed by C. Lee & Ji (2024). For example, a romantic chemical

comedy in which important events of a couple's life are shown will be interesting to viewers of different ages, as the idea of love is familiar to everyone. Thus, focused global content can feed local specificities that increase the chances of identification into successful productions. This can be related to culture, comedy involving a certain country, state or city, or issues affecting a certain society. For instance, action films of the world can depict the scene in some famous landmark of that area or use the native cast with the regional accent. These details can give a reader a perception that the contents are more appropriate for his/her understanding. Global contents increase awareness among the people of different cultures, customs, and attitudes towards life. What this exposure can do is help in creating cultural sensitivity and embrace of other cultures by the learners. For example, educational programs like the cooking show depicting different nations' cuisines not only make viewers laugh, but also enhance their knowledge about different cultures, providing people with an opportunity to have something in common regardless of the country they are from.

This preference plays a role in viewing patterns and thus creates a market for programming which targets certain groups in the society. Media consumers' preferences are crucial to know by media producers and social networks intending to share content. By understanding cultural, emotional, social and contextual influences in movie choice among the viewers, they can offer products that will appeal to the clients. This strategic alignment is not only beneficial to the viewers in terms of satisfaction with the program but is also key in building popularity and growth of content especially in this society with a lot of emerging media. Audience preference plays a crucial role in shaping media content and can be analysed through various dimensions: Viewers or listeners always are more inclined to look at or listen to something that they feel is familiar to them in terms of culture, beliefs and experiences. This

identification builds up the feelings of identification, which makes people feel as if they belong to the society or network. This way, when people watch their 'reflective selves' on screen, they stand higher chances of being interested in the content. For instance, series that have colourful cast or depict stories that are personal to those groups of people can be moving to those sets of people.

A factor that affects the audience preference will delve on content that focuses on aspects of life that the audience can relate to such as family influence, friendship, or challenges in life fosters the development of emotions. This makes viewers feel personally involved in the story and characters thereby making them more interested in the show. Further, audiences have embraced storytelling that comes with the aspect of identified emotions where people can feel the emotions of the characters. In return, this move can foster emotional involvement on a series or film and possibly boost its viewers' loyalty. It is also important to take into consideration that the audience can prefer different genres of movies (drama, comedy, thriller, and fantasy). Such preferences as stated above help the platform providers align themselves with the viewers' requirements. The preferences for the cultural influence genres can also be in some way dictated by cultural elements. For instance, some cultures may favour specific topics including melodrama or horror films that may impact content development. Subtle things, like recommendations, interfaces and data-driven watching modes could also shape audiences' choices on streaming platforms. Users may go towards machines that provide the content in simple and convenient ways.

Cultural context deals with the manner in which culture and other factors such as historical background influences the audience's perception of media. It is crucial for the creators and marketers to consider the contextualising cultural factors because they impact the

audience's perception, choice, and reaction. Cultural antecedents refer to the history and culture that colours the environment in which a society finds itself in. Examples of media that aim at portraying historical events or cultural practices can create sentiment that can range from nostalgia or pride among the people. In the same manner, certain matters regarding socio-political reality of a society like political regimes, economical setting or popularity of certain kinds of discursive social movements can influence themes and storytelling that would make connections with the audiences.

For instance, civil rights content can generate a lot of viewers during times of social unrest. It is also equally important to understand that every culture possesses a different code of ethics and moral standards which determines what is socially acceptable or not acceptable hence determining the way they package and present their stories to their audiences. Also, cultural context defines the relations within the family in one or another way, and media that reflects the roles and expectations of a family according to a cultural setting will be more preferred. It is thus evident that language performs a very strategic place in culture. The injection of regional loan words, slangs, proverbs, and jokes into a story can go a long way in making the piece render local comprehension by consumers of content.

Because of these differences such as body language, gestures and expressions also play a big role in the manner in which messages are delivered and side by side how they are received. These differences are important in the need to create content good for global use. Clothing, food, and rituals are a part of culture, which is great to use when creating a narrative because it has a certain meaning associated with it. For example, when it comes to the shooting of a scene such as a wedding scene, two different cultural settings would create different perceptions of the scene.

Culture thus plays a huge role and has far reaching effects in media strategy. Its awareness will go a long way in assisting creators, producers and marketers to avoid falling into the traps created by these implications and instead create works that connect with the audience at a deeper level. Media practitioners must try to find out the cultural framing of the targeted audiences and tell stories that can easily resonate with them. This entails a process of carrying out an extensive study with the intention of getting acquainted with the culture of the people, their beliefs, and the prevailing societal concerns. Inclusion of diverse characters, backgrounds, and issues into a story could make the stories more realistic and familiar. It should be realistic and free from stereotyping in order to permit deserved and accurate depiction of diverse cultures.

Translation is however not enough; localization is a function of redesigning content for a certain population demographic or geographic location. This may involve changing a part of the plot, a character's background, and even comedy to fit the customs of the region. Media strategies also have to ensure that it does not offend any culture or portray issues in a wrong manner. While designing the content, it is best to use local consultants or those with specialised knowledge in the culture to avoid offending anyone or giving out wrong information. Incorporation of cultural factors should be taken into consideration while adopting marketing strategies, that is, the language to be used together with other appealing factors such as images or concepts within the target culture region.

Engaging with the influencers who are familiar with the regional culture would help in improving the marketing appeals. Such promotions will be more relatable whenever the influencers act as intermediaries between the content creators and the target audiences. Some cultural parameters may be embraced more by the people of a certain culture to take media

content; therefore, strategies may have to factor this to understand the best way for arriving at the preferred way of taking media content. This is because relying on global streaming services which are generally available for everyone may not be effective, the use of regional platforms such as specific cultural television networks or specific streaming services can enhance the users' access and interest. Implementing ways for the audience response can clarify the perception of the content with regards to its culture. Such feedback can be beneficial for the future work and usage of audiences' preferences regarding the further projects of creators. This feedback can inform future projects and help creators adapt to evolving audience preferences. This kind of feedback is useful in the development of other projects, and adjusting the creative to the demands of the society.

Cultural trends as well as changes in the society thus enable the media firm to continue being relevant. Evaluating the data of viewers and the discussions at social platforms more often enables them to reveal new themes and interests. Thus, there is a need for the media strategies to be responsive during social unrest or cultural sensitivity. The management should always consider cultural changes and overall public opinion since it helps to retain customers' trust. Where the backlash is concerned, getting too confident and forgetting what culture the content belongs to can certainly be detrimental, and the only way to avoid this is by keeping the audience adequately informed. Concerns are listened to and audiences are responded to and it is highly culturally relevant to do so. establishing a brand that is interpreted as culturally appropriate can go a long way to guaranteeing audience's loyalty in the long term. Culturally relevant interaction may help a media company to be seen as a pioneer in multicultural narratives.

Engagement can be reached through certain events, sponsorships or partnerships in local communities that might enhance the association between the brand and its public. This type of approach can broaden the visibility of the business which in return reinforce brand images and associations.

In conclusion, it can be stated that cultural proximity theory stresses the need to relate media contents with the experience and cultures of audiences. In the case of platforms similar to Netflix, the ability to harness this theory will improve viewers' engagement, satisfaction, and brand loyalty, which in turn will improve content popularity and platform success in a rather saturated market. When the media companies are able to acknowledge and address different cultural origins of the consumers, better experiences are facilitated as well as the production of more relevant content.

2.2.2 Audience Engagement Model

In the context of Netflix and content popularity, Audience Engagement Model means the techniques and indicators that determine and facilitate increasing the level of viewers' engagement with the content posted on the site. Among them it singles out the process of audience interaction with shows or movies which may greatly affect their appreciation and appeal. These are the variables such as the viewership and the analysis of users' behaviours and their activity on the site or social networks.

Optimising the ratings and the viewership is important in order to determine the overall effectiveness of the content which is posted on the platforms such as Netflix. Through this lens, such metrics can be used in decision substantiations concerning production of content, marketing, and improvement of user experience by Netflix. This approach is useful to discover

the trending titles to guide the future course and investment in manner that the platform is relevant. These metrics tells about the audience, their choices and their interaction with the shows and movies being aired. Some of the basic key performance indicators that Netflix used include; total viewership of the title. This metric measures a show or movie's popularity and points at the numbers of subscribers that it has. Further, relative to the duration of the viewed title, percentages of viewers who watch a title—from beginning to end—are reflected in the completion rate.

A high completion rate implies that the users have found the content interesting and are likely to see it to its end whereas a low completion rate implies that the content they are presenting is not interesting. The number of minutes that viewers spent watching a title also shows how engaging a title is because "longer watch time" by the viewers rating means that the nature of the show satisfies viewers very much. The popularity indicators also allow for determining subscriber churn – whether or not specific new shows or new seasons lead to churn after the period of time. The 'binge watching', or the number of viewers who continue to watch multiple episodes of the particular series or continue watching the particular title after they have started is also essential since retention rates reflect on whether audiences continue to desire the series over time.

Users' interaction analysis process is very crucial for Netflix since it will help recommend how viewers engage with content on the platform. It will help Netflix to understand the web users' preferences and habits, and then offer customised and entertaining videos. This way, Netflix has an opportunity to carefully study the users' preferences in terms of content viewing patterns, interaction, feedback etc., and adjust the offers and approaches accordingly. User behaviour analysis it's a key aspect of Netflix's strategy, to know how its users or

audiences consume the content offered in the platform. This analysis is going to be useful for Netflix to be able to present the viewer with a sensual and intriguing procedure. Using such information from viewing patterns, content interaction, users' preferences and feedback, Netflix can further improve its operation and methods.

This approach works also for the benefit of the viewer satisfaction and increases the popularity of the content, and thus – the retention of the subscribers, which is important for sustaining the platform's competitiveness. Thus, by studying users' behaviour, including what the viewers watch, when they watch it, and how often they come back to a particular piece of content, Netflix can optimise recommendations and user satisfaction. Even its user viewing patterns can tell when and how they have been using the Service to watch, at what time or during the day, and whether they have been binge-watching.

This will help find out when the greatest number of people use the internet and hence help Netflix in planning on which dates to release some of the most awaited movies or shows. For example, based on ratings, it may be seen that viewership is higher over the weekends, hence Netflix might release its products over the weekend, for instance, on Fridays. User behaviour analysis can also relate to activity of users like, how they navigate, search or select content on the platform. Thus, information busted from the experiment concerning which genres or titles are more likely to be clicked can be used for planning of further selection of materials to be purchased and created. For instance, if the data shows that particular genre gains popularity, Netflix will feed users more content from this genre to profit from the trend.

Netflix has the ability to filter and analyse the user base as more users rate shows, interact and watch, Netflix can guarantee that users are served contents that would appeal to them. This personalization also improves user satisfaction as well as the loyalty relating to the

use of social networks. Interaction parameters like likes, shares, and the number of posts and comments as to how the users interact with content in addition to the view. High levels of engagement usually imply that the content is well received by the target audiences hence the formation of word-of-mouth publicity.

The factor of the engagement of the content with the social environment has a significant importance for the popularity and success of content delivery services like Netflix. It includes the talk and conversation that happens around shows and movies on social media platforms and in the fans' groups. Interaction is another determinant of content trending on Netflix as it enhances viewership and people's relations. Through social networks and SMM, commercials, user-generated content, satisfying partnerships, and community development, Netflix can improve the viewers' interest and frequency. Beyond helping in content marketing, and in social interaction analysis that is crucial in strategy formulation, they assist Netflix in designing an integrated viewing environment to engage the subscribers.

The social aspects and relevance of social engagement are as follows; the social mentions, Twitter, Facebook and Instagram, Netflix content shares, likes, comments and TikTok banning. Such social media promotion can greatly impact a title's popularity, as the shows that people talk the most often are the ones that receive the most viewership since others will be interested in finding out more about the content discussed by others. For instance, if there is a popular hashtag or a meme that is hot at the moment but is in some way related to a show then more people would tune into the show. Posts or fan-generated material that includes Netflix shows and movies, creates the feeling that the shows belong to users and they ought to watch them. Such a word of mouth promotion can prove to go down well than an actual advert, such as the friendships we have seen earlier. Watching this content is also beneficial for Netflix to

understand the audience's sentiments about their products and modify the marketing plan. Endorsements by social media personalities and content creators who share their programs with fans using the social media channels are also useful since the fans trust the social media personalities. In targeting the right audience, Netflix can work with the influencers who are in sync with some of the titles, and this will increase exposure on any new releases. Also, by creating online communities and forums in which fans can discuss, analyse and share their ideas and opinions regarding Netflix content, such a content can develop a strong fan base which can not only enjoy the content but also recommend it to others. Reddit, and other forums and fan site continue the conversation by presenting ideas, theories and fan-led projects after the viewer has moved on.

All in all, it is crucial to consider viewership figures; examine users' behaviours, and-address social impact in the course of furthering an understanding of content popularity in such sites as Netflix. Rating data offer qualitative information about the audiences giving more details concerning the reactions of viewers towards the programs in terms of the number of viewers, the time spent on the programs, and the percentage level of viewership. That is why these metrics make it possible for Netflix to assess the potential success of its offerings and make relevant choices when it comes to the creation and promotion of new content. User behavior analysis comes in handy with regards to these metrics since it goes further into details concerning the users' tastes and exercise of their liberties. Through patterns of viewing, options of the users and their reviews, Netflix is able to make suitable recommendations and thus keep the interest of the viewers high and thus, increasing the number of viewers. Audience participation enlarges the power of content by engaging the public to discuss shows and movies in their social networks. When combined with others sharing their opinion in their social

networks and live events with influencers and other parties it should increase the viewers even further and alter their views on it. In combination, these components provide the networked environment, which helps to further Netflix's audience research and assist the platform to evolve and succeed in a constantly competitive market. Addressing viewership metrics, users' behaviour analysis, and social activity, Netflix will be able to maintain the interest of the audience and develop new and qualitatively different content.

2.3 Existing Techniques used in Content Popularity Prediction

Machine learning, a vibrant branch of artificial intelligence (AI), has revolutionised numerous fields, from healthcare to finance, by empowering computers to learn without explicit programming. This ability to learn and adapt from data has propelled ML to the forefront of technological advancements, prompting the need for a comprehensive understanding of its origins, key players, and fundamental paradigms. This introduction delved into the history of ML, exploring its seminal moments and the visionary minds that shaped its course, before shedding light on its two principal learning paradigms: supervised and unsupervised learning.

The seeds of ML were sown in the fertile ground of mathematics and statistics in the mid-20th century. Alan Turing, a legendary figure in computer science, laid the groundwork with his seminal paper "Computing Machinery and Intelligence" (Nguyen, 2023), proposing the Turing test as a benchmark for machine intelligence. Concurrently, Arthur Samuel, an American pioneer in AI, coined the term "machine learning" in 1959, marking the official christening of this nascent field.

Early advances in ML were driven by ground-breaking algorithms. Frank Rosenblatt's Perceptron, a single-layer neural network, paved the way for learning algorithms inspired by the

human brain. Marvin Minsky and Seymour Papert, however, demonstrated the limitations of the Perceptron, leading to a period of relative stagnation in the field.

The resurgence of ML began in the 1980s with the development of powerful learning algorithms and the availability of computational resources. Geoffrey Hinton, David Rumelhart, and Ronald Williams revived neural networks with their influential work on backpropagation, a learning algorithm that enabled deeper networks to learn complex patterns. Other significant contributions came from Vapnik's work on Support Vector Machines and Breiman's development of Random Forests, all of which expanded the ML toolbox and paved the way for the field's current explosion (Montgomery, 2023).

Machine learning operates under two main paradigms: supervised and unsupervised learning. In supervised learning, the algorithm learns from labelled data, where each data point comes with a predefined label or output. For instance, an image classification algorithm might be trained on images labelled as "cat" or "dog," enabling it to learn to classify new images accurately. The key figures in supervised learning include Vladimir Vapnik (aforementioned), Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, who laid the foundations for deep learning, a powerful supervised learning technique.

In contrast, unsupervised learning deals with unlabelled data, where the algorithm must discover patterns and structure without guidance. Clustering, a common unsupervised learning task, groups similar data points together, uncovering hidden relationships within data. Notable contributors in this area include David Arthur and Andrew Harter for their work on k-means clustering and Judea Pearl for his contributions to Bayesian networks (Powers, 2020).

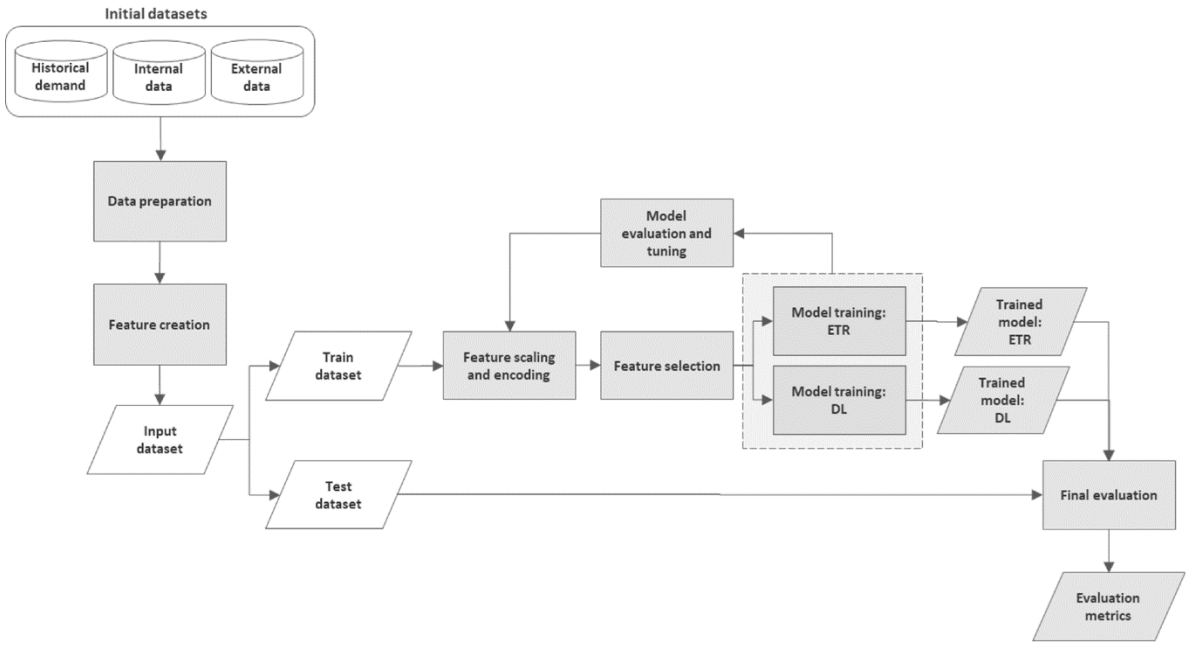


Figure 2. Machine Learning Model

2.3.1 Supervised Learning Techniques

Within the supervised learning paradigm, where labelled data guides the model's learning process, various techniques flourish. Regression, at its core, focuses on regressively predicting continuous values, be it estimating housing prices or forecasting sales trends. Linear regression, the foundational technique, employs a linear relationship to map inputs to outputs. More complex models like decision trees and ensemble methods like random forests introduce non-linearity and feature interaction, enhancing their predictive power (Usama et al., 2019).

Classification, on the other hand, tackles categorising data points into predefined classes, forming the essence of tasks like image recognition or email spam filtering. Logistic regression, a generalised linear model, calculates the probability of an instance belonging to a specific class.

2.3.1.1 Regression Models

Predicting continuous outcomes based on input features forms the core of regression analysis, a critical tool in supervised machine learning. While numerous techniques exist, three prominent models stood out regarding this study: linear regression, ridge regression, and extra trees regression. This exploration delved into their core principles, advantages, and disadvantages.

2.3.1.1.1. Linear Regression

Linear regression, the cornerstone of regression analysis, reigns supreme in its simplicity and interpretability. This technique excels at modelling linear relationships between independent features and a continuous dependent variable, making it a trusted tool in content popularity prediction (Montgomery & Peck, 2021). This section embarks on a comprehensive exploration, delving into its mathematical foundations, applications, and nuances.

Linear regression assumes a linear relationship between features (represented by the vector X) and the target variable (y). Mathematically, this relationship is encapsulated in the following equation:

$$y = f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

where:

β_0 is the intercept, representing the predicted value of y when all features are zero.

β_1 to β_n are the regression coefficients, quantifying the impact of each feature X_i on the target variable.

ε is the error term, accounting for the difference between the predicted and actual values, often assumed to be normally distributed with a mean of zero (Maulud & Abdulazeez, 2020).

Estimating the regression coefficients is the crux of linear regression. The most common approach minimises the sum of squared errors (SSE), defined as:

$SSE = \Sigma(y_i - f(X_i))^2$

where the summation covers all data points (i = 1 to n).

This minimization problem leads to a system of linear equations, which can be solved using least squares techniques, resulting in the estimated coefficients (denoted by $\hat{\beta}$). The resulting equation becomes:

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}\square X\square$

This equation represents the fitted regression line, capturing the linear relationship between the features and the target variable.

The regression coefficients hold immense value in understanding the model's behaviour. Each coefficient ($\hat{\beta}_i$) estimates the average change in the predicted target variable associated with a one-unit increase in the corresponding feature (X_i), holding all other features constant (Bartlett, 2020). Positive coefficients indicate a positive relationship, while negative coefficients denote an inverse relationship.

It's crucial to remember that linear regression thrives on certain assumptions such as linearity defined as the relationship between features and the target variable must be truly linear. Homoscedasticity is the variance of the error term and should be constant across all data points. Independence means the errors should be independent of each other and the features. Normality implies the error term should be normally distributed.

Violations of these assumptions can lead to biased and unreliable predictions. Visualising the data through scatter plots and normality checks are essential steps in model validation (Filzmoser & Nordhausen, 2021).

Linear regression provides a solid foundation, but its realm extends further. Bayesian Linear Regression incorporates prior knowledge about the parameters to provide richer insights and uncertainty quantification. Elastic Net Regularization combines L1 and L2 penalties to handle multicollinearity while promoting sparsity in the coefficients, potentially leading to improved interpretability (Powers, 2020).

Generalised Linear Models (GLMs) extend the linear framework to handle non-normal target variables, such as binary outcomes (Logistic Regression) or count data (Poisson Regression). Multivariate Linear Regression models relationships between multiple target variables and a set of features (Maulud & Abdulazeez, 2020). Polynomial regression presents polynomial terms to capture non-linear relationships. Regularisation includes techniques like L1 (LASSO) and L2 (ridge) regression are employed to handle multicollinearity and improve model generalizability. Bayesian linear regression incorporates prior knowledge about the parameters for richer insights.

Despite its strengths, linear regression encounters limitations that require careful consideration. These include multicollinearity which occurs when features are highly correlated, estimating coefficients accurately becomes difficult. Regularisation techniques like LASSO or ridge regression can mitigate this issue. Non-linearity occurs when the relationship between features and the target variable deviates from linear, predictions become inaccurate. Polynomial regression or non-linear models like decision trees offer alternatives.

Another limitation is outliers. Extreme data points can significantly impact the model and lead to biased predictions. Techniques like outlier detection and robust regression can help address this challenge. Overfitting occurs when the model memorises the training data too well,

it loses its ability to generalise to unseen data. Regularisation, cross-validation, and early stopping are crucial strategies to prevent overfitting (Filzmoser & Nordhausen, 2021).

Several studies have employed linear regression for content popularity prediction. Huberman et al. investigated user access patterns on YouTube and Digg, demonstrating that early viewership data could be used with linear regression to forecast content's long-term popularity on these platforms. Their findings suggested that content with high initial engagement was more likely to sustain popularity, with the optimal timeframe for initial data collection varying depending on the content type (e.g., shorter for news stories compared to videos).

Similarly, Bao et al. utilised linear regression to predict movie box office success in China. They incorporated features like director reputation, genre, cast, and budget, achieving moderate accuracy in their predictions. These studies highlight the potential of linear regression to capture basic relationships between content characteristics and popularity metrics.

Despite its initial appeal, linear regression has limitations that hinder its effectiveness in complex content popularity prediction scenarios. Oversimplification: Linear regression assumes linear relationships between features and the target variable (popularity). In reality, these relationships can be non-linear and more intricate. This oversimplification can lead to inaccurate predictions, particularly for content popularity, which is influenced by a multitude of factors that interact in non-linear ways.

Limited Feature Handling: Linear regression struggles with a high number of features, a common scenario in content popularity prediction. Including too many features can lead to overfitting, where the model performs well on the training data but fails to generalise to unseen

data. Feature selection or dimensionality reduction techniques are often required to mitigate this issue.

Inability to Capture Complex Interactions: Linear regression cannot capture complex interactions between features. For instance, the combined effect of a renowned director and a popular genre might have a greater impact on popularity than the sum of their individual effects. Linear models have missed these crucial interaction terms.

2.3.1.1.2 Ridge Regression

Linear regression, while powerful, can succumb to the perils of overfitting, particularly when dealing with high-dimensional data or multicollinearity. This exploration delves into the intricate world of ridge regression, uncovering its origins, mathematical foundations, benefits, and limitations in content popularity prediction.

Ridge regression, also known as Tikhonov regularisation, has its roots in the works of Andrey Tikhonov, a Soviet mathematician who published his seminal paper in 1963 (Hoerl, 2020). However, its journey into the mainstream of statistics can be attributed to Hoerl and Kennard, who, in 1970, laid the groundwork for its practical application, establishing its now-famous formula. Since then, numerous researchers have contributed to its refinement and understanding, solidifying its place as a cornerstone of regularised regression techniques.

At its core, ridge regression modifies the traditional least squares cost function used in linear regression by introducing a penalty term:

Minimize $(y - \beta Tx)^2 + \lambda \Sigma \beta j^2$ (Tsigler, 2021)

where:

y is the dependent variable vector

β is the vector of regression coefficients

X is the design matrix

λ is the regularization parameter (alpha in some literature)

Σ βj^2 is the L2-norm penalty term

This L2-norm penalty, essentially the sum of squared coefficients, acts as a ridge, shrinking the coefficients towards zero as λ increases. This shrinkage not only reduces the complexity of the model but also mitigates the impact of multicollinearity, leading to more stable and generalizable predictions.

One of the benefits presented by ridge regression is reduced overfitting. By shrinking coefficients, ridge regression effectively reduces model complexity, preventing it from overfitting to the training data and improving its ability to generalise to unseen data. Another benefit is improved multicollinearity handling. When features are highly correlated, estimating individual coefficients in linear regression becomes unreliable. Ridge regression's shrinkage helps alleviate this issue, leading to more stable and interpretable coefficient estimates (James et al., 2021). The third advantage is enhanced numerical stability: In cases of ill-conditioned design matrices, ridge regression can improve numerical stability during coefficient estimation, preventing issues like matrix inversion problems.

However, ridge regression has several limitations. Tuning the Regularization Parameter is the first. Selecting the optimal value for λ, the regularisation parameter, is crucial. Choosing too low a value negates the benefits of regularisation, while too high a value can lead to underfitting and biased predictions (Tsigler, 2021). Techniques like cross-validation or AIC/BIC can guide this selection process.

Another limitation is variable selection. Unlike LASSO regression, which can set coefficients to zero, ridge regression shrinks but doesn't eliminate them. This can limit its ability

to perform automatic variable selection (Xie & Deng, 2020). Interpretation of coefficients may also become a limitation. While still interpretable, the shrunken coefficients in ridge regression might not directly represent the true feature effects due to the shrinkage phenomenon (Qasim, et al., 2021).

Some studies have explored ridge regression for content popularity prediction. For instance, Sariyildiz et al. investigated movie revenue prediction using ridge regression. They incorporated features like cast, director, genre, and budget. While achieving moderate accuracy, the study highlighted the challenge of interpreting the model's results due to the regularisation term. Unlike linear regression, where coefficients directly represent feature importance, understanding the impact of individual features in ridge regression becomes less straightforward.

Despite its ability to mitigate overfitting, ridge regression presents limitations that hinder its effectiveness in content popularity prediction. Reduced Interpretability: The regularisation term in ridge regression shrinks the coefficients towards zero, making it difficult to directly interpret the relative importance of each feature on the predicted popularity. This lack of interpretability can be a significant drawback, as understanding which features drive popularity is often valuable for content creators and distributors.

Limited Explanatory Power: Ridge regression assumes primarily linear relationships between features and popularity. While it can handle some non-linearity through regularisation, complex interactions and non-linear patterns often characterise content popularity. Ridge regression might not capture these nuances effectively, leading to suboptimal predictions compared to more flexible models.

2.3.1.1.3 Extra Trees Regressor Model

Random forests, a cornerstone of ensemble learning for both classification and regression tasks, have revolutionised the field of machine learning. The origin story can be traced back to the pioneering work of Leo Breiman and the concept of bagging (Fernández-Delgado et al., 2019).

The groundwork for random forests was laid by Leo Breiman in his seminal paper "Bagging predictors." Bagging, short for bootstrap aggregating, is a powerful ensemble technique. It involves creating multiple versions (iterations) of a learning model by drawing random samples with replacement from the original dataset. These "bootstrap samples" are then used to train individual models. Finally, the predictions from all the models in the ensemble are aggregated (e.g., averaged for regression, majority vote for classification) to make a final prediction (Eslami et al., 2020).

Bagging addressed the issue of overfitting, a common problem where a model performs well on the training data but fails to generalise to unseen data. By introducing diversity through random sampling and training multiple models, bagging ensembles could achieve better generalisation performance.

Building upon the success of bagging, Leo Breiman introduced the concept of random forests in his paper "Random forests." This work extended the idea of bagging by incorporating an additional layer of randomness during the tree building process. In a standard decision tree, at each node, the best split among all features is chosen to partition the data. Random forests introduce an element of chance by randomly selecting a subset of features (mtry < total features) at each node. The best split is then chosen only from this random subset. This additional

randomness injects diversity into the ensemble, further enhancing its ability to avoid overfitting and improve generalisation (Alsariera et al., 2020).

Random forests quickly gained popularity due to their effectiveness, ease of use, and robustness to various data types. They consistently outperformed single decision trees and other learning models across diverse tasks. The research community actively explored random forests, leading to advancements in the following areas.

Feature Importance Measures: Techniques were developed to assess the relative importance of features within a random forest, providing valuable insights into the factors influencing the model's predictions.

Parameter Tuning: Strategies were devised to optimise the hyperparameters of random forests, such as the number of trees in the ensemble (n_estimators) and the number of features considered at each split (mtry) (Eslami et al., 2020).

Variations: Researchers explored variations of the random forest algorithm, such as Extremely Randomised Trees (Extra Trees Regressors), where an additional random split point is selected within the chosen feature at each node, further enhancing diversity.

The Extra Trees Regressor (ETR) leverages the power of ensemble learning by combining multiple decision trees for regression tasks. While the core concept is relatively straightforward, the underlying mathematical formulation involves several intricate components. Below is the mathematical details of the ETR algorithm:

Bootstrap Aggregation: Let D be the original dataset containing N data points (x_1 , y_1), ..., (x_N, y_N), where x_i is a d-dimensional feature vector and y_i is the corresponding target value. The ETR employs bootstrap aggregation (bagging) to create B ensemble trees. Each ensemble tree is built from a bootstrap sample D_b (b = 1, ..., B) drawn with replacement from

the original data D. The size of each bootstrap sample (|D_b|) is equal to the size of the original data (N) (Arya et al., 2022).

Decision Tree Construction: For each bootstrap sample D_b, a decision tree T_b is constructed iteratively. Here's the breakdown of the tree building process: Initialization: Start with the entire bootstrap sample D_b as the root node. Recursive Splitting: At each non-terminal node t: Randomly select a subset of features $F_t \subseteq \{1, ..., d\}$ containing mtry features (mtry < d) without replacement (Devi et al., 2019).

Among the features in F_t, find the best split that partitions the data at node t into two child nodes (left and right). The "best split" can be determined using a splitting criterion like mean squared error reduction for regression tasks.

The splitting criterion function (Ψ) takes the current node data D_t and a candidate split S as arguments and evaluates the improvement in prediction performance achieved by the split. The split with the highest improvement (ΔΨ) is chosen as the best split: $\Delta\Psi(D_t, S) = \Psi(D_t) - (|D_{tl}| * \Psi(D_{tl}) + |D_{tr}| * \Psi(D_{tr})) / |D_t|$ where D_tl and D_tr represent the data points in the left and right child nodes, respectively, resulting from the candidate split S (Fernández-Delgado et al., 2019).

The recursive splitting process continues until a stopping criterion is met. Common stopping criteria include reaching a maximum tree depth or having a minimum number of data points in a leaf node.

For a new data point x, the ETR makes a prediction by passing it through each of the B decision trees in the ensemble. Each tree T_b predicts a target value y_b based on the terminal leaf where x falls within the tree structure. The final prediction of the ETR ($\hat{y}$) is the average of the individual tree predictions: $\hat{y} = (1/B) * \Sigma(b=1 \text{ to } B)\ y_b(x)$ (Gupta et al., 2019).

Model Complexity Parameter (mtry): mtry, the number of features randomly selected at each node for splitting consideration, is a crucial parameter in the ETR algorithm. It controls the diversity of the trees in the ensemble. Lower values of mtry lead to more diverse trees, which can improve generalisation but potentially reduce prediction accuracy. Conversely, higher values of mtry can result in more correlated trees, potentially impacting model performance. Tuning mtry is often done through techniques like grid search or cross-validation (Fernández-Delgado et al., 2019).

In content popularity prediction, where a multitude of factors influence audience engagement (genre, director, etc.), the ETR's ability to handle complex relationships can be advantageous.

Several studies have explored the ETR for content popularity prediction. For instance, Wang et al. utilised an ETR model to predict movie viewership on a Chinese online video platform. They incorporated features like genre, director reputation, cast, and release date. The ETR achieved superior performance compared to other regression models like Support Vector Regression, demonstrating its effectiveness in this context.

Similarly, Luo et al. employed an ETR model to predict the click-through rate (CTR) for news articles on a social media platform. Their model included features such as news category, sentiment, and user demographics. The ETR outperformed other models like Logistic Regression and Gradient Boosting Machines, showcasing its potential for predicting user engagement with various content types.

Despite its advantages, the ETR has limitations that require consideration. These include the Black Box Nature. Ensemble models like ETRs can be challenging to interpret. While feature importance scores can be extracted, understanding the complex interactions between

features within the ensemble remains difficult. This lack of interpretability can limit insights into the factors driving content popularity (Devi et al., 2019).

Data Dependence: ETR performance is highly dependent on the quality and quantity of training data. For content popularity prediction, access to diverse content data with accurate popularity metrics is crucial. In scenarios with limited or imbalanced data, the ETR might underperform (Gupta et al., 2019).

Parameter Tuning: While the ETR typically requires less tuning compared to other models, the parameter mtry (number of features considered at each split) still needs optimization. Choosing the optimal mtry value can significantly impact prediction accuracy (Fernández-Delgado et al., 2019).

The Extra Trees Regressor offers a powerful approach to content popularity prediction due to its ensemble learning nature and ability to handle complex relationships between features. Studies have shown its effectiveness in predicting movie viewership and news article popularity. However, limitations like black box interpretability, data dependence, and parameter tuning require careful consideration.

2.4. Challenges facing the content popularity in Netflix

The business landscape is in a constant state of evolution, shaped by technological advancements, globalisation, and shifting consumer preferences (Heydarova, 2024). It is possible to argue that throughout this time, some aspects, both positive and negative, can be observed (De la Garza Montemayor et al., 2023) in how people interact with entertainment, news, and information which has changed dramatically as a result of these platforms, which distribute audio, video, and another media content online (Yadav & Jain, 2024). Netflix as a

global streaming behemoth referred to as over-the-top video, video-on-demand or online television services that are considered the major competitors of traditional terrestrial television services with a valuation exceeding $200 billion and an enormous subscriber base has revolutionised the way we consume and produce content (X. Li, 2023). As consumers demand high-quality, diverse content, Netflix has been forced to invest heavily in both licensing existing properties and developing new ones(X. Li, 2023). Netflix has been presented with a myriad of challenges which include long standing refusal to comment on viewers' geographic location and its unwillingness to publicise data for most of its content library reflects the reality that anti-transparency policies are still very much the norm (Wayne, 2022b).

With the coming up with other streaming other video streaming Disney, other players like Amazon Prime can take advantage of brand recognition and scale economies have increased stiff competition and equally damaged Netflix reach and potential customers. Whereas it is true that mismatched, overacted performances and unnatural scripts certainly compromise the quality of the dubbed version and should be kept to a minimum, the lack of habituation to dubbing might also affect the way audiences perceive the final product. Viewers' feedback suggests that those least exposed to dubbed programmes might find it difficult to enjoy the cinematic experience and to forget the uncanny inherent to the dubbing practice (Sánchez-Mompeán, 2021).

Further, other challenges in the facet of content in Netflix include the other regulatory dynamics which Netflix is facing across the globe. This has been due to the differences in censorship laws, data protection regulations and even tax laws on digital services across the globe. In addition, globalisation complicates the matter of audience tastes with programming in

part due to the need for a much more targeted and focused approach to global content curation and customer outreach.

2.5. Conceptual Framework

This study delved into the factors influencing a show's inclusion and duration within the coveted top ten rankings on the Netflix platform. It employed a diverse range of independent variables to paint a comprehensive picture of potential determinants.

Temporal Factors such as date and release year are frequently included features in previous studies on content popularity. Research suggests recency can significantly impact popularity. Content characteristics such as genre, theme, and category consistently emerge as important predictors, indicating audience preferences for specific content types (Liu et al., 2019).

Production details such as director reputation and country of origin are sometimes influential. The impact might vary depending on the platform and content category. Established show titles with dedicated fan bases likely influence popularity through the community building that occurs with popular releases (Chen et al., 2019).

This study included temporal factors like date and release year, content characteristics like genre, theme, and category, production details like country of origin, director, and lead cast, and includes the show's own title. Examining how these independent variables interact and contribute to a show's cumulative weeks in the top ten as the dependent variable offered valuable insights into audience preferences, content strategies, and the dynamic landscape of entertainment consumption within the SVOD market.

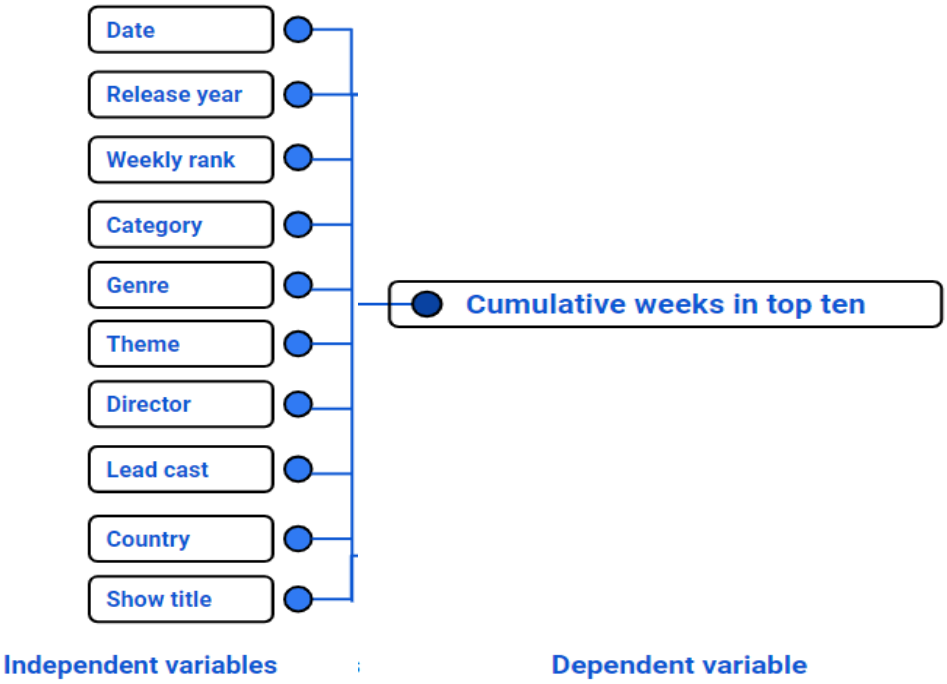


Figure 3. Conceptual Framework

2.6 Summary

This section explored the theoretical background of this study and then outlined one theory and a model to highlight the content popularity in Netflix. It then described the best practices in machine learning and delved into four current techniques and challenges facing the content popularity in Netflix. The section then concluded with an outline of the independent and dependent variables that will be investigated in the study.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This section describes the research design and demonstrates the conceptual framework in detail, highlighting the various variables that were of importance to this study. It expounded on the kind of data selected, target population, data analysis process, ethical considerations and intended reporting for this study.

3.2 Research Design

Machine learning models thrive on data, but not just any data – the right data, carefully collected and analysed, was crucial for building robust and effective models. While diverse research methodologies contribute to this endeavour, experimental research played a pivotal role in evaluating and refining these ML models, offering a controlled environment to assess their performance and generalizability. Experimental research was the research design this study was based on.

The defining characteristic of experimental research is its manipulative nature. Unlike observational methods that simply observe existing relationships, experiments actively manipulate independent variables (potential causes) to observe their impact on dependent variables (outcomes) of interest (Ledyard, 2020). This allowed the researcher to establish causal relationships between variables, providing compelling evidence for whether a specific factor truly influences the model's performance.

Within the realm of ML, experimental research serves several key purposes, the first of which is model comparison. Different ML algorithms or variations of the same model can be

compared in controlled settings to identify the best performing option for a specific task (Ledyard, 2020). This allowed the researcher to systematically evaluate the strengths and weaknesses of different approaches and select the most suitable model for the intended application.

Second is generalizability testing where experiments can evaluate how well a model trained on one dataset performs on another, unseen dataset. This generalizability assessment was crucial for ensuring the model's effectiveness in real-world scenarios beyond the training data (Gönenç et al., 2020). Third is bias identification where carefully designed experiments can help identify and mitigate potential biases within the model. This is crucial for ensuring fairness and ethical considerations in ML development, particularly when dealing with sensitive data or applications (Bolukbasi et al., 2019).

3.3 Data

This study delved into the dynamic landscape of content popularity on Netflix, specifically focusing on the coveted top ten rankings across various countries. The data used in this study was collected from June 28, 2021, to March 24, 2024. To achieve this, a rich dataset encompassing both readily available and meticulously collected information formed the foundation of the analysis. The initial dataset comprised weekly top ten rankings released by Netflix, categorised by country, specifically in Kenya. This served as the core foundation, providing essential variables like date, category, weekly rank, and show title. However, to gain a deeper understanding of the factors influencing content popularity, additional independent variables were deemed necessary.

Therefore, the data was meticulously enriched by incorporating the following variables from credible online sources: release year, genre, theme, country of origin, director, and lead cast. This additional information allowed for more nuanced analysis, considering not only temporal factors but also intrinsic content characteristics, production details, and even the potential influence of creative talent. Recognizing the inherent complexities of dealing with potentially duplicated titles, meticulous cross-checking was conducted against the official Netflix platform. This crucial step ensured the accuracy and consistency of the enriched dataset, particularly where multiple shows or movies shared the same title.

This comprehensive data preparation process transformed the initial rankings data into a rich and informative resource. By combining readily available information with supplementary data from trusted sources and ensuring accuracy through cross-checking, the enriched dataset lays the groundwork for insightful exploration of the multifaceted interplay between various factors and their impact on content popularity within the Netflix top ten data in Kenya.

3.4 Selection of Data

The choice of variables for content popularity prediction in Kenya was guided by a combination of theoretical considerations, domain knowledge, and data availability. Temporal factors such as date and release year were crucial due to the recency effect. Newer content often attracts more attention. Studies have consistently shown that recent releases tend to have higher initial popularity (Huberman et al., 2021).

Content variables such as genre, theme, and category are essential. Understanding audience preferences for specific genres can guide content creation strategies (Chen et al.,

2019). Identifying popular themes and categories can help tailor content to resonate with specific audience interests (Liu et al., 2019).

Production variables such as director reputation and lead cast can influence viewer expectations and attract audiences (Deloitte, 2023). Country of origin can impact popularity due to cultural preferences and language barriers.

Performance metrics such as weekly rank and cumulative weeks on top ten provided insights into audience engagement levels. These variables offered a comprehensive understanding of the factors influencing content popularity in Kenya, enabling more accurate predictions and informed content creation decisions.

The current Netflix system has several data limitations being that it relies on historical data to create the weekly top 10 rankings instead of using a predictive model like in the recommender engine it employs when interacting with its subscribers.

Also, the Netflix platform doesn't allow for direct user feedback on the kind of experience that they receive through feedback forms and user surveys.

3.5 Extra Trees Regressor Model

This study embarked on a mission to develop a robust ML model capable of predicting content popularity. To achieve this objective, the research focused on regression techniques, specifically the Extra Trees Regressor.

The Extra Trees Regressor does not use a simple formula but instead builds an ensemble of trees that collectively make predictions based on the patterns they learn from the data.

3.6 Target population

This research aspired to offer valuable insights into the Kenyan SVOD market's unique characteristics. Kenya presents a unique and dynamic population brimming with cultural richness and evolving consumption patterns. The nation boasts a rapidly growing youth population, increasing internet penetration, and a burgeoning appetite for digital entertainment (Asu, 2021). These factors, coupled with Kenya's diverse cultural tapestry, contributed to a unique viewing landscape ripe for exploration. Understanding the interplay between these cultural nuances and SVOD demand necessitated research tailored to the Kenyan context.

3.7 Data Analysis

The model development followed a structured process involving several key steps (Géron, 2019). At the outset, data cleaning played a crucial role, that aimed to address inconsistencies, missing values, and outliers that could negatively impact model performance. This involved techniques like imputation, normalisation, and transformation (Dua & Graff, 2019).

Once the data was prepped, feature engineering involved the creation of new features based on existing ones, to enhance the model's ability to learn complex relationships Model selection followed, where the researcher chose an appropriate technique based on factors like task type, data characteristics, and computational resources (Géron, 2019).

Before model development, the researcher undertook a correlation analysis to identify highly correlated features and remove redundant ones to avoid multicollinearity. Low p-values (less than 0.05) indicate a strong likelihood that the observed correlation is not due to random

chance. High p-values (greater than 0.05) indicate that the observed correlation could potentially be due to random chance.

Show title, weekly rank, release year, category, and country displayed statistically significant correlations with cumulative weeks in top 10, suggesting these factors might influence how long content stays on the top 10 list.

Month, year, genre, and theme either show no statistically significant correlation or a very weak effect, implying these attributes likely have a minimal influence on a title's performance on the top 10 list.

Hyperparameter Tuning is the process of systematically searching for the optimal combination of hyperparameters to achieve the best possible performance for a given machine learning model. Bayesian Optimization is a powerful technique for hyperparameter tuning in Extra Trees Regressor (ETR). Unlike traditional grid search or random search methods, Bayesian Optimization leverages a probabilistic model to intelligently explore the hyperparameter space, focusing on regions with higher potential for improved performance.

By constructing a surrogate model that approximates the relationship between hyperparameters and model performance, Bayesian Optimization efficiently identified hyperparameter combinations. In the context of content popularity prediction, Bayesian Optimization helped fine-tune the hyperparameters of the ETR model to achieve optimal accuracy and generalisation.

Model training then involved feeding the cleaned and preprocessed data to the Extra Trees Regressor, allowing it to learn the underlying patterns and relationships. The data was divided into a 70:30 train and test split where 70% of the data was used to train the model and 30% was used to test the performance of the model.

3.8 Validation & Usability Test

Evaluation assessed the model's effectiveness in unseen data using the following metrics. Mean Absolute Error (MAE): This metric calculates the average of the absolute differences between predicted and actual popularity scores. Lower MAE values indicate better model performance.

Mean Squared Error (MSE): MSE squares the individual differences between predicted and actual values, then calculates the average. While informative, MSE can be sensitive to outliers.

Root Mean Squared Error (RMSE): The square root of MSE, offering an error measure in the same units as the original data (popularity scores in this case).

R-squared (R^2): This coefficient of determination reflects the proportion of the variance in the dependent variable (popularity) that the model explains. R^2 values closer to 1 indicate a stronger model fit.

Root Mean Squared Logarithmic Error (RMSLE): This metric is particularly useful for data with skewed distributions, potentially encountered with viewership counts. It transforms the errors into a logarithmic scale before squaring and taking the root mean.

Mean Absolute Percentage Error (MAPE): This metric expresses the error as a percentage of the actual popularity score, facilitating comparisons across content with varying popularity levels.

By combining these metrics, we can gain a more comprehensive understanding of the model's performance. MAE provided a direct measure of average error. MSE and RMSE highlight the impact of outliers. R^2 assessed the overall model fit. RMSLE was useful for

skewed data distributions. This combination of metrics addressed the limitations of individual metrics and provided a more robust evaluation of the model's predictive capabilities.

If performance fell short, iterative adjustments to data cleaning, feature engineering, model selection, hyperparameter tuning, and training were made until satisfactory results were achieved. Overall, model development required careful attention to each step, from data cleaning to evaluation, to ensure a robust and effective model (Géron, 2019).

3.9 Data integration, interpretation, and reporting

A detailed research paper was published at the culmination of this study. It included extensive discussion on the different techniques applied and the results of each. A detailed recommendation was developed and made available to stakeholders in the SVOD space and entertainment sector in general.

3.10 Ethical considerations

Within the realm of research, web scraping presents a powerful tool for data acquisition and analysis. However, wielding this tool ethically necessitates careful consideration of potential pitfalls. This study adhered to a rigorous ethical framework by exclusively scraping data from Wikipedia, a website renowned for its transparency and explicit allowance of data scraping within its terms of service (Wikimedia Foundation, 2023).

3.11 Conclusion

The section above described the research design chosen for this study. It demonstrated the machine learning techniques that were used in detail, highlighting the various variables that

were of importance to this study. It explained the kind of data selected, target population, data analysis process, ethical considerations and intended reporting for this study.

CHAPTER FOUR

DATA ANALYSIS, FINDINGS AND DISCUSSIONS

4.1 Introduction

This section delves into the core elements of the research. Here, a detailed breakdown of the study variables employed in the investigation is presented. Subsequently, the analysis explores the relationships between content attributes, utilising correlation coefficients and p-values to assess the statistical significance of these connections.

Next, the section provides a comprehensive overview of the descriptive results obtained from the experiment. This includes statistics that summarise the key content consumption patterns among the Kenyan audience. Following this, the model development process is outlined, detailing the steps involved in constructing the machine learning model.

The section culminates in a discussion of the model's performance, evaluating its effectiveness in predicting the target variable. This discussion critically examines the metrics used to assess the model's accuracy and generalizability. Finally, the section addresses the extent to which the research objectives, established at the outset of the investigation, have been successfully achieved.

Finally, this section will also highlight some of the problems Netflix has faced in the quest of popularising it’s content as a global stream provider (X. Li, 2023).

4.2 Study variables

Figure 6 below reveals interesting connections between various content attributes in the dataset. There are weak negative correlations between date and month (-0.03), date and year (-0.01), suggesting dates tend to be spread across months and years without a strong seasonal or yearly pattern. Date shows weak positive correlations with genre (0.06) and country (0.04), which might indicate a slight tendency for certain genres or countries to release movies more consistently throughout the year.

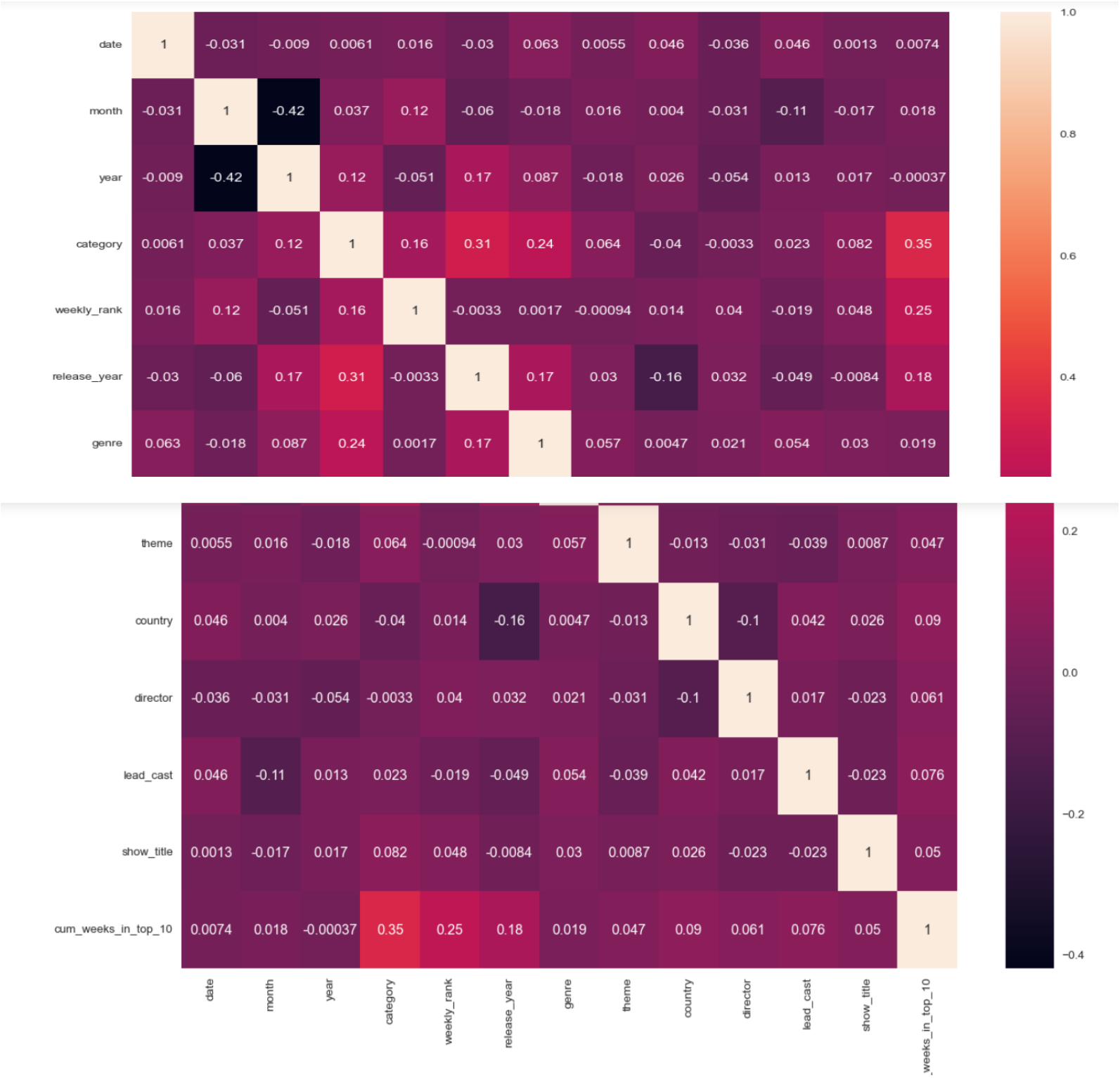


Figure 6. Correlation Matrix

Release year has a strong positive correlation with category (0.31) and a moderate positive correlation with genre (0.17), suggesting newer movies might fall into specific categories or genres more frequently. There's a weak negative correlation between release year and country (-0.16), possibly indicating a decline in movies from some countries over time.

A moderate positive correlation exists between genre and category (0.24) and theme (0.06), implying some overlap in how movies are categorized based on genre, theme, and broader classification.

Weekly rank has a moderate positive correlation with cum weeks in top 10 (0.25), signifying that movies ranking higher on weekly charts tend to stay in the top 10 list for longer and are therefore more popular. There are some negative correlations between month and lead cast (-0.11) and release year and lead cast (-0.05). Show title exhibits weak correlations with most other variables, suggesting titles might not be strongly indicative of other content attributes in this dataset.

Many correlations between variables fall below 0.1 (weak), indicating limited linear relationships. This suggests that most content attributes may not have a strong influence on each other.

Complementing the correlation coefficients, the p-values in the correlation and p-value chart (Appendix 3) provide crucial information about the statistical significance of the observed relationships between movie attributes. By considering both correlation coefficients and p-values, we can distinguish between potentially meaningful relationships and those potentially arising from random noise in the data. This strengthens the interpretation of the correlation matrix and allows us to focus on statistically significant associations for further investigation.

```
correlation_matrix = {}
```

```
for i in range(len(encoded_df.columns)):
  for j in range(i+1, len(encoded_df.columns)):
    var1 = encoded_df.columns[i]
    var2 = encoded_df.columns[j]
    correlation, p_value = calculate_correlation(var1, var2)
    correlation_matrix[(var1, var2)] = (correlation, p_value)
```

```
# Create a DataFrame from the dictionary
correlation_df = pd.DataFrame.from_dict(correlation_matrix, orient='index', columns=['Correlation', 'p-value'])

# Create a heatmap with annotations
fig, ax = plt.subplots(figsize=(15, 40))
sns.heatmap(correlation_df, annot=True, cmap="coolwarm")
plt.show()
```

Figure 7. Creating Correlation and P-Value Chart

Low p-values (less than 0.05): These indicate a strong likelihood that the observed correlation is not due to random chance. For instance, the positive correlation (0.31) between release year and category has a very low p-value, suggesting a statistically significant association. This implies a non-random pattern where newer content might be more likely to fall into specific categories.

Similarly, the moderate positive correlation (0.25) between weekly rank and cum weeks in top 10 has a low p-value, signifying a statistically significant connection. Content with higher weekly rankings are more likely to stay in the top 10 list for longer durations, based on a statistically robust observation.

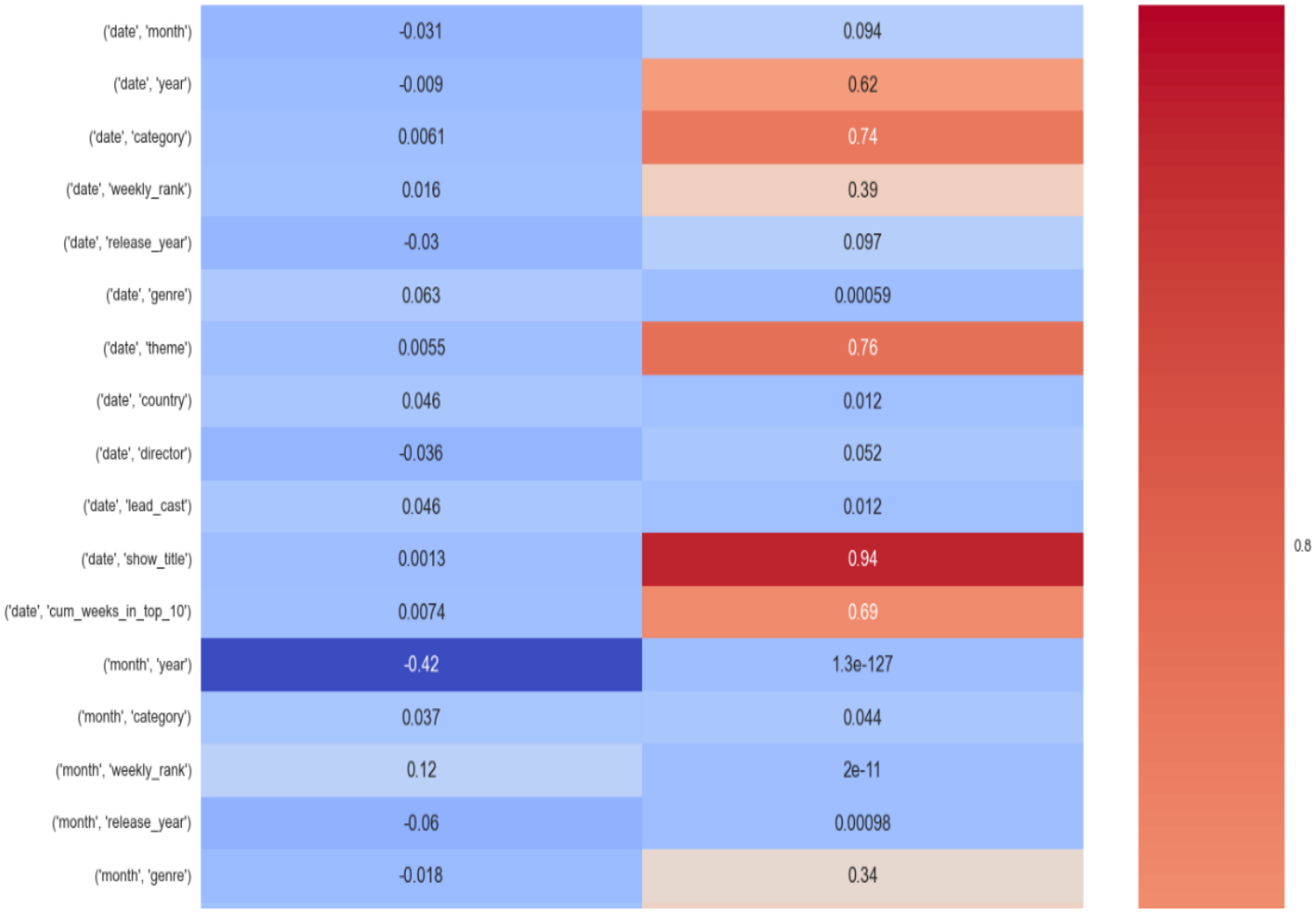


Figure 8. Correlation and P-value Chart

High p-values (greater than 0.05): These indicate that the observed correlation could potentially be due to random chance. For example, the weak negative correlation (-0.03) between date and month has a high p-value, suggesting this observed pattern might not be statistically significant. Dates might be spread across months without a strong seasonal trend.

Weekly Rank (0.2499, p-value < 0.0001): This is the strongest positive correlation, indicating a statistically significant relationship. TV shows and movies with higher weekly ranks tend to stay in the top 10 list for longer durations.

Release Year (0.1794, p-value < 0.0001): This moderate positive correlation suggests a statistically significant association. Newer content might have a slightly higher chance of staying on the top 10 list for a longer time.

Category (0.3486, p-value < 0.0001): This is the strongest correlation overall, highlighting a statistically significant link. TV shows and movies belonging to specific categories might be more likely to remain on the top 10 list compared to others.

Country (0.0901, p-value = 0.0000): This weak positive correlation shows a statistically significant association. TV shows and movies from certain countries might have a slightly higher chance of staying on the top 10 list.

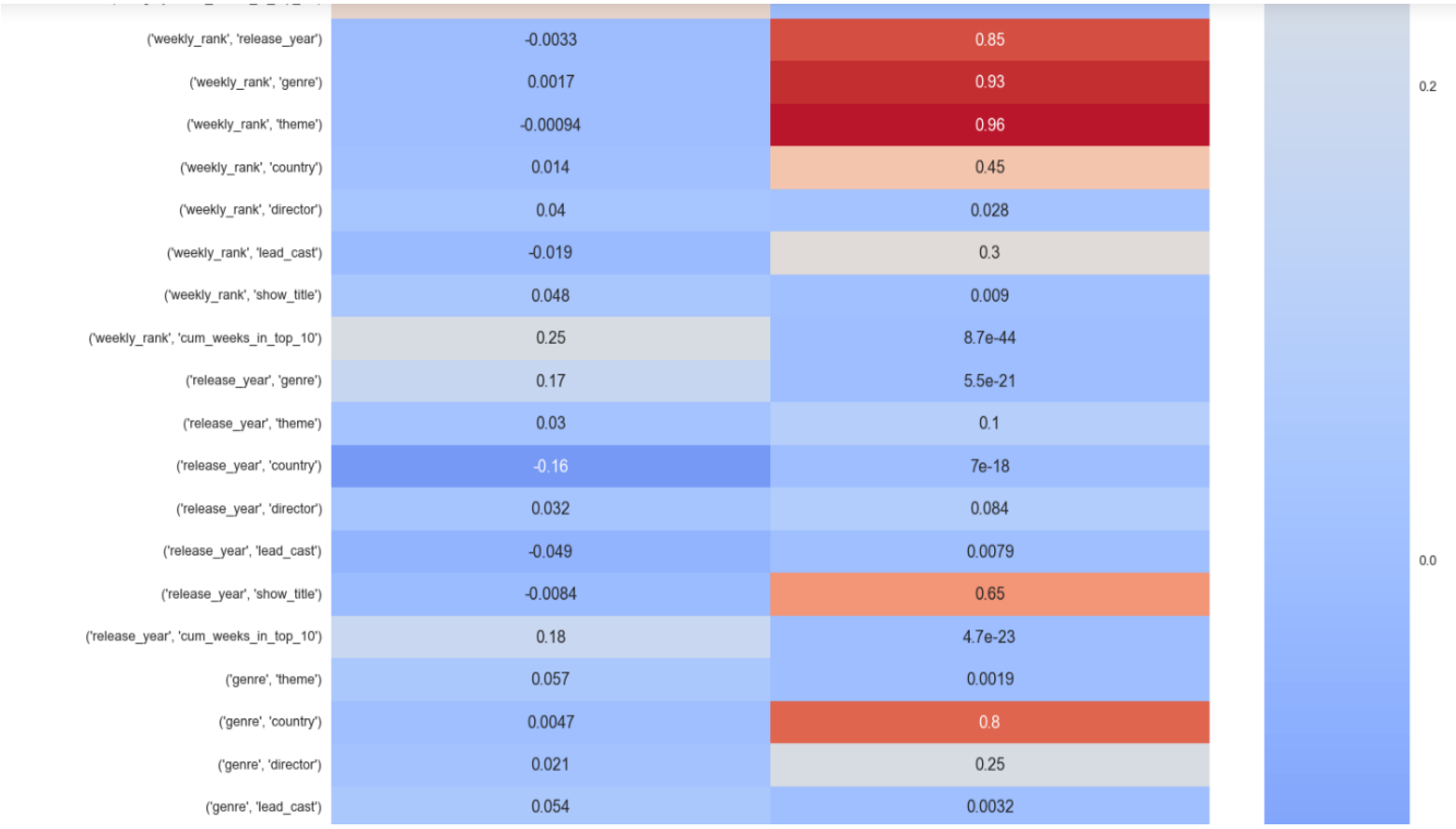


Figure 9. Correlation and P-Value Chart

Lead Cast (0.0759, p-value = 0.0000): This weak positive correlation is statistically significant. Movies featuring specific actors in the lead cast might have a minor influence on their performance in the top 10 list.

Month (0.0179, p-value = 0.3271): The weak positive correlation here is not statistically significant. There's no clear link between the month of release and a title's duration on the top 10 list.

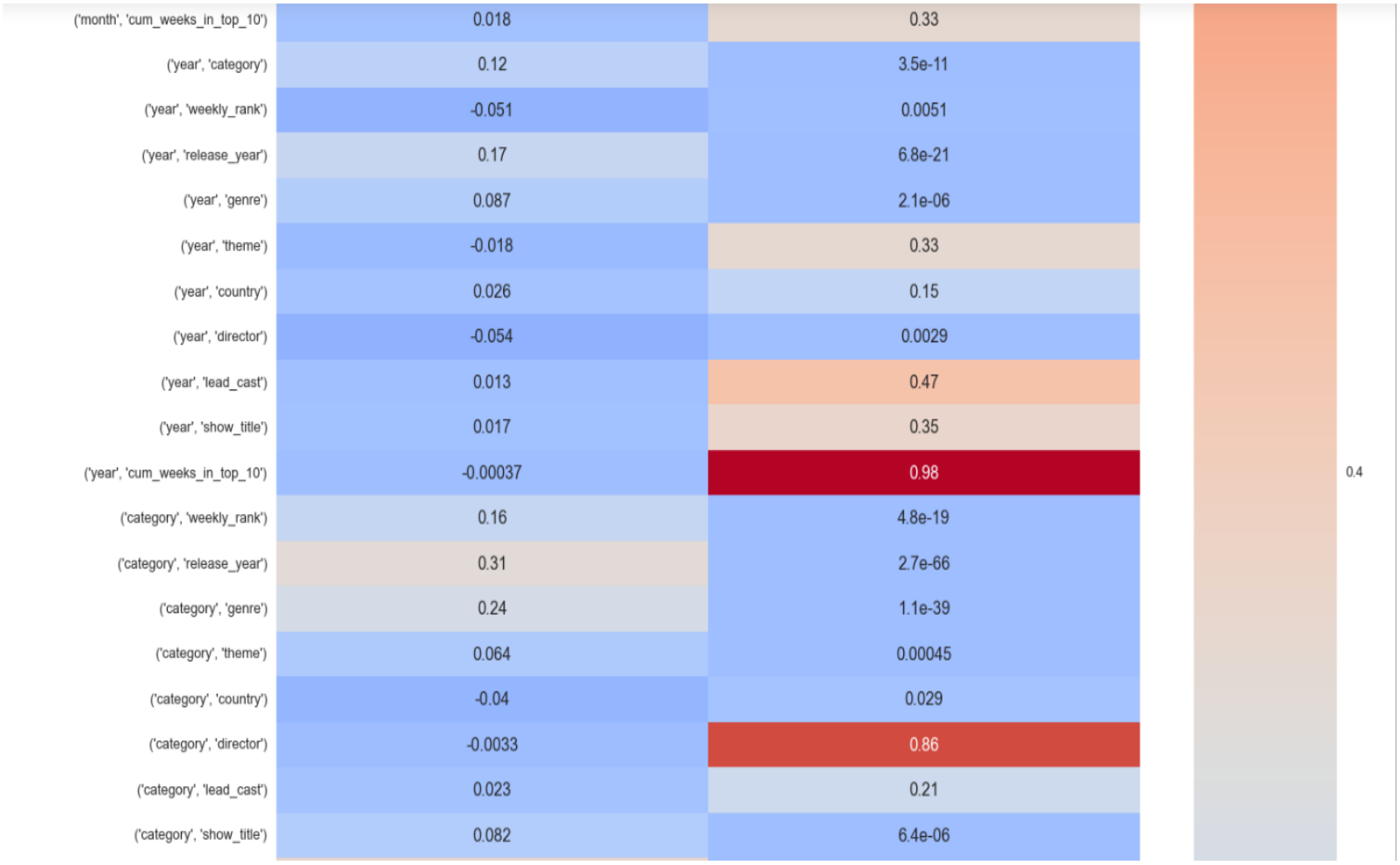


Figure 10. Correlation and P-Value Chart

Year (0.0004, p-value = 0.9840): This near-zero correlation suggests no statistically significant relationship between a title's year of release and its performance on the top 10 list.

Genre (0.0187, p-value = 0.3056): The weak positive correlation is not statistically significant. Content genre likely doesn't have a major influence on a title's presence on the top 10 list.

Theme (0.0470, p-value = 0.0102): This weak positive correlation might be statistically significant, but the effect size is very small. Theme likely has a minimal influence at best on a title's duration on the top 10 list.

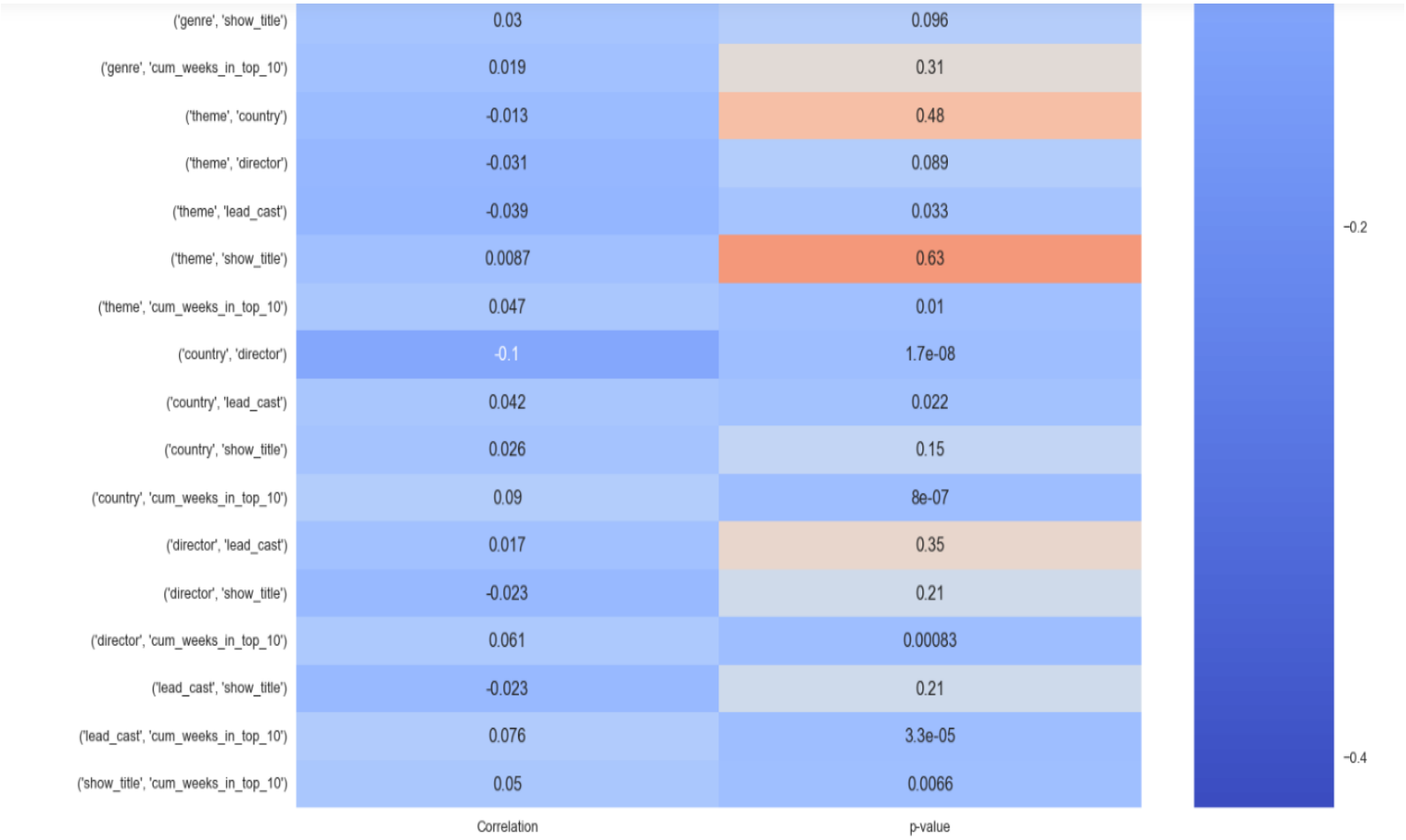


Figure 11. Correlation and P-Value Chart

Show title (0.0497, p-value = 0.0066): While the correlation coefficient is weak, the statistically significant p-value suggests that there might be a systematic influence of movie

titles on a film's performance in the top 10 list. While a title might hold some influence, it's likely one factor among many that contribute to a movie's performance on the top 10 list

Show title, weekly rank, release year, category, and country display statistically significant correlations with cumulative weeks in top 10, suggesting these factors might influence how long content stays on the top 10 list.

Month, year, genre, and theme either show no statistically significant correlation or a very weak effect, implying these attributes likely have a minimal influence on a title's performance on the top 10 list.

```
import numpy as np
from statsmodels.stats.power import TTestIndPower
import matplotlib.pyplot as plt

# variables for power analysis
effect_size = -0.5
alpha = 0.05
power = 0.8
p_analysis = TTestIndPower()
sample_size = p_analysis.solve_power(effect_size=effect_size, alpha=alpha, power=power)
print("Required Sample Size: " + str(sample_size))
```

```
Required Sample Size: 63.76561177540986
```

```
#power vs. number of observations
fig = plt.figure()
fig = TTestIndPower().plot_power(dep_var='nobs',
                                 nobs= np.arange(50, 2990),
                                 effect_size=np.array([-0.2, -0.5, -0.8]),
                                 alpha=0.01,
                                 title='Power of t-Test at variable effect sizes\n' + r'$\alpha = 0.01$')
plt.show()
```

```
<Figure size 800x550 with 0 Axes>
```

Figure 12. Plotting Effect Size

The graph below depicts the relationship between power and sample size in a t-test, with varying effect sizes as a differentiating factor. This visualisation serves as a crucial tool for the researcher to determine the appropriate sample size for this experiment, ensuring sufficient statistical power to detect a true effect as stated by Cohen.

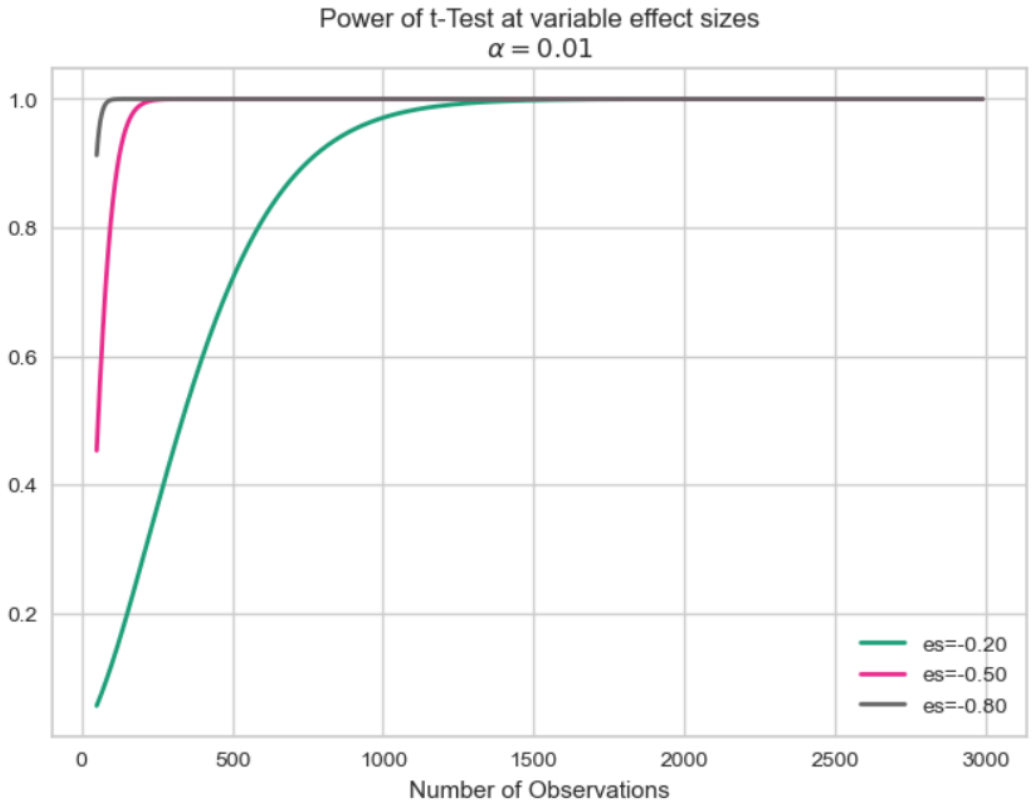


Figure 13. Power of t-Test Graph

The X-axis represents the number of observations (n), which reflects the sample size of the experiment. Larger sample sizes are depicted towards the right. The Y-axis represents the power (1 - β) of the t-test. Power signifies the probability of correctly rejecting the null hypothesis (H_0) when it is truly false. Higher power values on the y-axis indicate a greater likelihood of detecting an existing effect.

The graph typically includes multiple lines, each representing a different effect size (denoted as "es"). Effect size quantifies the magnitude of the difference between the means of the two groups being compared in the t-test.

Larger effect sizes (represented by higher es values) correspond to higher power for a given sample size. In simpler terms, it's easier to detect a substantial difference between groups even with a smaller sample.

Another critical observation from the graph is the positive association between sample size and power. As the number of observations increases (moving to the right on the x-axis), the power of the t-test increases for all effect sizes. This implies that with a larger sample size, there's a greater chance of detecting a real effect, even if it's relatively small.

The graph reveals that at a desired power level of 0.80 (indicating a high probability of detecting an effect), an effect size of 0.80 (a large effect) requires a near-zero sample size. This suggests even a very small sample could effectively detect such a substantial effect. As the effect size decreases (moving to 0.50 and 0.20, representing moderate and small effects, respectively), the required sample size increases (100 and 600, respectively).

This experiment uses a sample size of 2990. Since this value is considerably larger than what would be required to detect even a small effect size (given the high-power level of 0.80), we can conclude that the sample size is more than adequate.

4.3 Descriptive results of the experiment

The researcher began with measures of frequency in this analysis as it was important to find out what kind of content is consumed most by Kenyans on the Netflix platform. The visualisation for this was carried out using Tableau.

4.3.1 Category

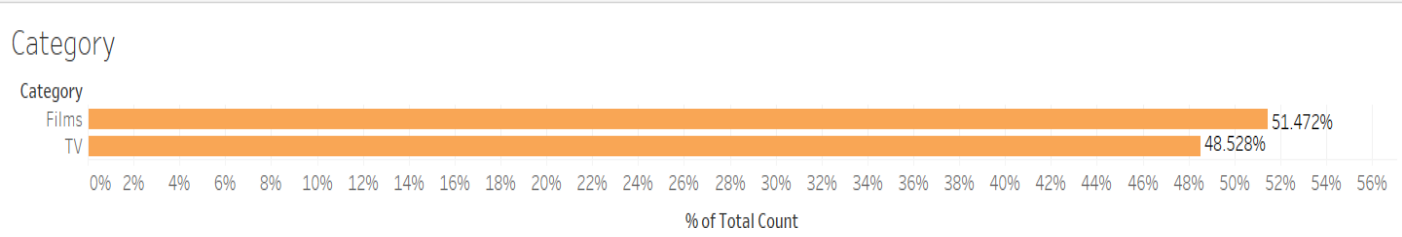


Figure 14. Content Category

Figure 14 depicts the percentage viewership of films and TV shows on Netflix among Kenyan users. The data provides insights into the content preferences of this audience segment.

The chart illustrates that films, (51.47%) account for a larger share of total viewership compared to TV shows (48.53%) on Netflix in Kenya. This suggests a stronger preference for movies among Kenyan viewers on the platform.

Several factors might contribute to the observed preference for films. Cultural Factors: Cultural norms or preferences in Kenya might favour cinematic storytelling over episodic narratives. Mobile Viewing: If mobile devices are the primary viewing platform for many Kenyan users, the format might be more suitable for shorter films than TV show episodes. Content Availability: The selection of films on Netflix Kenya might be more extensive or cater more effectively to local tastes compared to TV shows.

While films hold the majority viewership, TV shows still represent a significant portion (almost half) of viewing time on Netflix in Kenya. This indicates a substantial audience base that enjoys episodic content on the platform.

4.3.2 Country of Origin

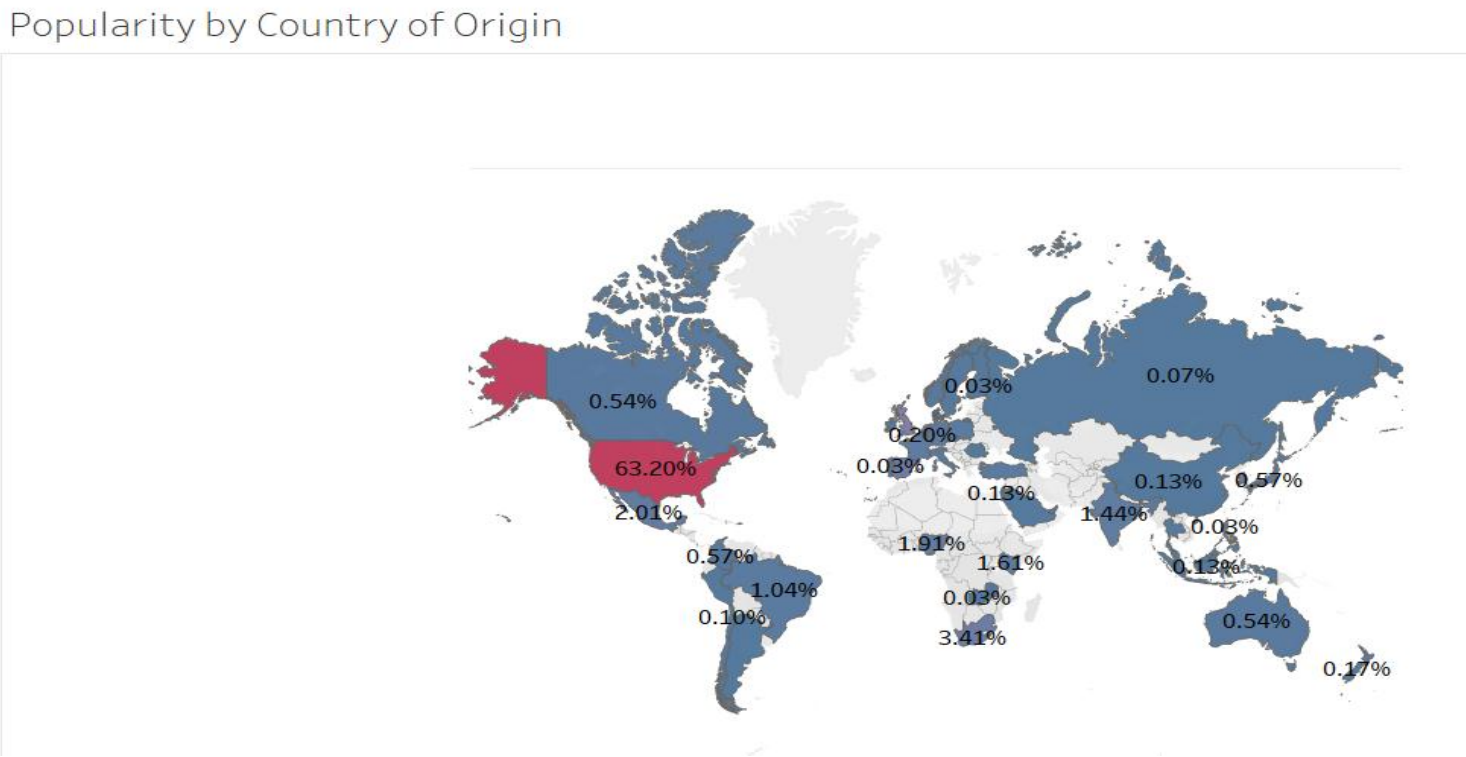


Figure 15. Popularity by Country of Origin

Figure 15 visualises the popularity of Netflix content by country of origin among Kenyan viewers. The data offers valuable insights into content preferences and cultural influences shaping viewership patterns on the platform.

The map reveals a clear dominance of American content, capturing a substantial 63.20% of viewership in Kenya. This significant skew suggests a strong preference or greater accessibility to American media productions.

Several potential factors might contribute to this phenomenon. Global reach of Hollywood: The longstanding influence of Hollywood as a global cultural powerhouse likely plays a significant role (Shafer, 2019). Production value: American productions are often

associated with high production values, which can be a draw for audiences seeking visually appealing content (Gan et al., 2019).

Marketing reach: Extensive marketing campaigns by major American studios might contribute to increased awareness and viewership of U.S. content on Netflix in Kenya. Content library size: The possibility of a more extensive library of American titles on Netflix in Kenya compared to other regions could influence viewership choices.

Following the U.S., Nigerian content emerges as a notable presence, capturing 3.41% of viewership. This reflects the growing popularity of Nollywood, the Nigerian film industry, which has established a strong cultural footprint across Africa (Arogundade, 2019). This regional preference suggests a cultural resonance with Kenyan audiences, potentially due to shared linguistic and cultural elements within the continent.

The presence of content from other countries like India (1.44%) and the United Kingdom (2.01%) adds to the diversity of content consumed by Kenyan viewers. Indian media consumption likely indicates an appreciation for Bollywood productions, known for their unique storytelling and vibrant aesthetics, appealing to diverse audience segments (Mitra & Sarkar, 2021). Similarly, British content might attract viewers due to historical ties, linguistic familiarity, and the distinct style of British productions.

Smaller percentages from countries like Brazil (1.91%) and South Korea (1.61%) reveal a broader international interest among Kenyan viewers. The presence of South Korean content, in particular, points to the rising global influence of Korean dramas (K-dramas) and K-pop culture, which have seen a surge in popularity worldwide (Nguyen, 2019).

The map highlights a complex and multifaceted viewership landscape on Netflix in Kenya. While American media maintains a dominant position, the data also illustrates a growing

openness to diverse content from various regions. This trend reflects the ongoing globalization of media consumption and the dynamic nature of cultural exchange facilitated by digital streaming platforms like Netflix.

4.3.3 Genre

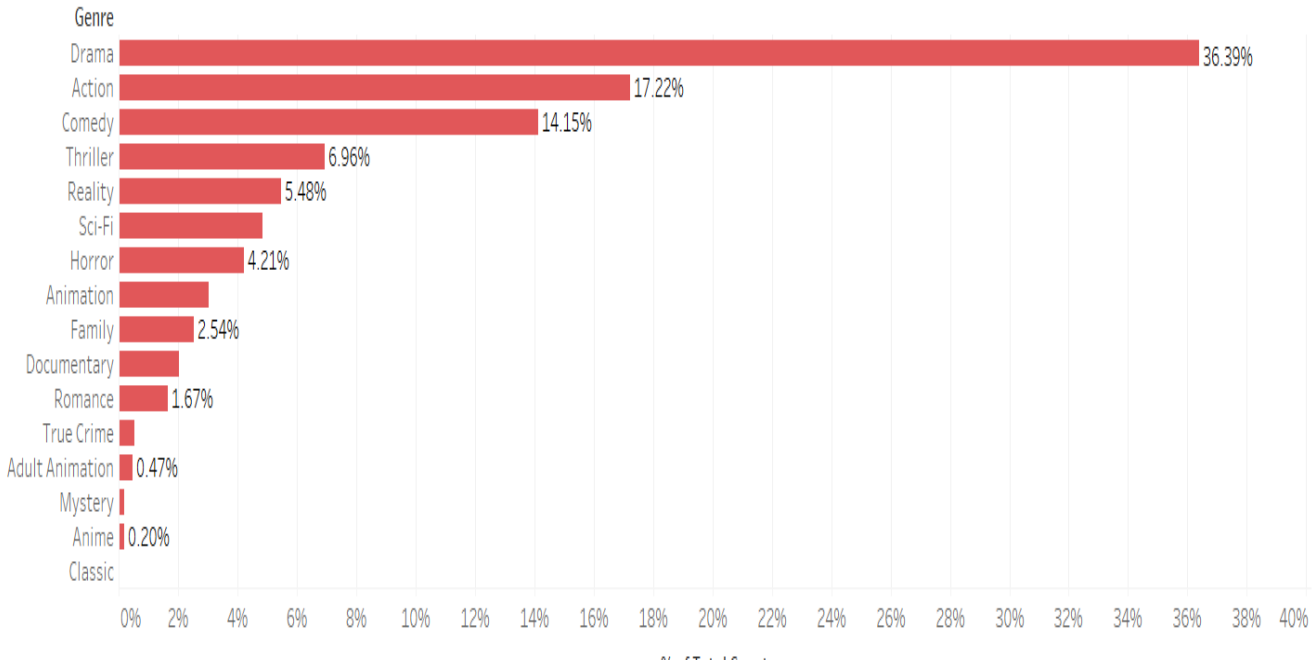


Figure 16. Popularity By Genre

Figure 16 delves into the genre preferences of Kenyan viewers on Netflix. The data offers valuable insights into the content choices and entertainment styles that resonate with this audience segment.

Drama emerges as the leading genre, capturing a substantial 36.39% of total viewership in Kenya. This strong preference suggests a Kenyan audience drawn to narrative complexity and character-driven storytelling, a hallmark of dramatic content. The emotional depth and

exploration of human experiences within dramas appear to resonate significantly with Kenyan viewers (Green & Brockmeier, 2020).

Following Drama, Action and Comedy genres hold significant shares of viewership, accounting for 17.22% and 14.15% respectively. The popularity of Action films can be attributed to their fast-paced nature and visually engaging elements, offering escapism and excitement for viewers (Cohen, 2019). The appeal of Comedy highlights a strong audience preference for humor and light-hearted entertainment, potentially serving as a counterpoint to the intensity of dramatic content.

Genres like Thriller (6.96%) and Reality TV (5.48%) also exhibit notable viewership. The interest in Thrillers suggests a Kenyan audience that enjoys suspense and intricate plotlines, elements that keep viewers engaged and intellectually stimulated (Tan, 2021). Meanwhile, the popularity of Reality TV reflects a fascination with unscripted, real-life scenarios and personalities, offering relatability and sometimes a voyeuristic appeal.

Sci-Fi and Horror genres, with 5.16% and 4.21% respectively, indicate a moderate interest in speculative fiction and supernatural themes. These genres cater to niche audiences that enjoy imaginative and often dystopian narratives, alongside the adrenaline rush and fear induced by horror content (Jost, 2021). Genres such as Animation (2.54%), Family (2.28%), Documentary (2.11%), and Romance (1.67%) capture smaller segments of the viewership. These figures suggest more specific audience interests, with Animation and Family content appealing to younger demographics or those seeking family-friendly options.

Documentaries reflect an interest in informative and real-world content, while Romance caters to viewers drawn to romantic and emotional narratives (Vorderer & Dietrich, 2006). Less popular genres include True Crime, Adult Animation, Mystery, Anime, and Classic films, each

holding less than 1% of viewership. These cater to specialised tastes and likely have limited but dedicated audience bases.

In conclusion, the dataset reveals a diverse range of genre preferences among Kenyan Netflix viewers, with a clear dominance of Drama, followed by Action and Comedy. This distribution highlights a balanced appetite for both intense and light-hearted content, along with an appreciation for various narrative forms and themes. The data suggests a complex and multifaceted viewership landscape on Netflix in Kenya, reflecting the evolving entertainment preferences of a global audience.

4.3.4 Theme

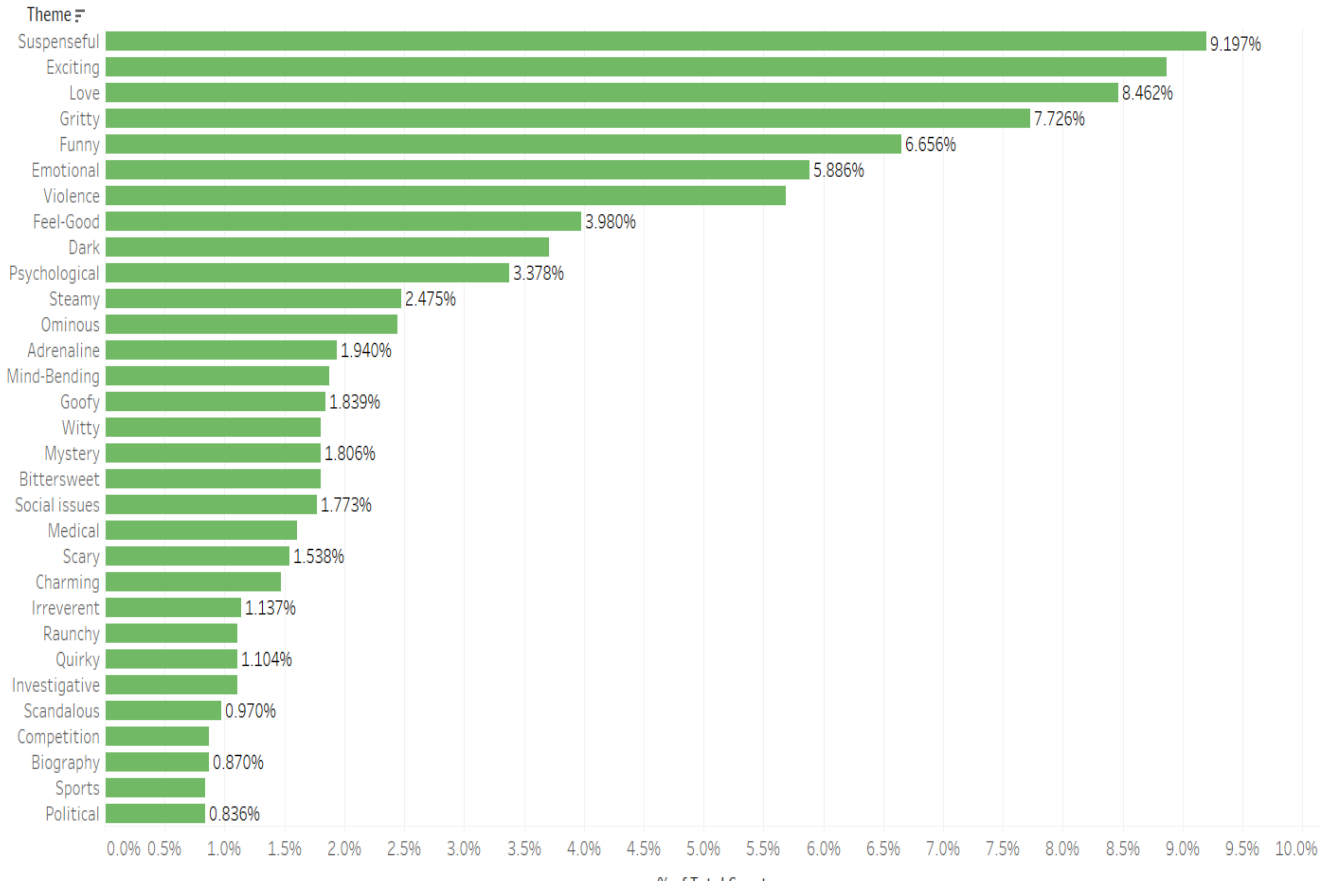


Figure 17. Popularity By Theme

Figure 17 depicts the thematic preferences of Kenyan viewers on Netflix. The data offers insights into the types of narratives and emotional experiences Kenyan audiences seek from their streaming content. The chart reveals a clear preference for themes that create anticipation and emotional engagement.

"Suspenseful" content emerges as the leader, capturing 9.197% of viewership. This suggests a strong Kenyan audience interest in narratives that maintain tension and uncertainty, keeping viewers engaged through a sense of the unknown (Tan, 2021). Following closely is the "Exciting" theme (9.104%), highlighting a preference for high-energy and fast-paced content that provides thrills and adrenaline. The significant presence of "Love" themes (8.462%) underscores a substantial affinity for romantic narratives that explore relationships and emotional connections. This indicates a Kenyan audience that values stories that resonate on a personal and emotional level.

Themes like "Gritty" (7.726%) and "Funny" (6.656%) reveal a duality in viewer preferences. The presence of "Gritty" content suggests an interest in realistic portrayals of life, even when intense. However, "Funny" themes highlight a simultaneous desire for content that provides humour and light-hearted entertainment (Johnson, 2024). These findings suggest Kenyan viewers appreciate diverse narratives, seeking both serious and comedic elements.

"Emotional" content (5.886%) further underscores the preference for narratives that explore the complexities of human emotions. This theme's significance suggests Kenyan viewers seek stories that resonate with them on a deeper emotional level. "Violence" (4.986%) indicates a noteworthy interest in more graphic and intense content, potentially appealing to viewers drawn to action-packed and confrontational scenarios.

Themes like "Feel-Good" (3.980%) and "Dark" (3.378%) capture the balance in viewer interests between positive, uplifting narratives and darker, more somber storylines. The presence of "Psychological" themes (2.475%) suggests an appreciation for content that explores the human mind and behavior, providing intellectual stimulation.

Themes like "Steamy" (1.940%), "Ominous" (2.093%), "Adrenaline" (2.246%), and "Mind-Bending" (2.475%) hold smaller percentages of viewership, reflecting niche interests in sensual, foreboding, action-driven, and intellectually challenging content. Lower percentages for "Political," "Sports," "Biography," and "Competition" (all below 1%) suggest these cater to more specific, albeit smaller, audience segments.

In conclusion, the thematic preferences of Kenyan Netflix viewers exhibit a broad spectrum. There's a pronounced inclination towards suspenseful, exciting, and emotionally charged content, alongside a significant appreciation for romantic and humorous themes. This diversity in thematic interests underscores the varied and multifaceted nature of the Kenyan viewership landscape on Netflix, reflecting the evolving content preferences of a global audience.

4.3.5 Director

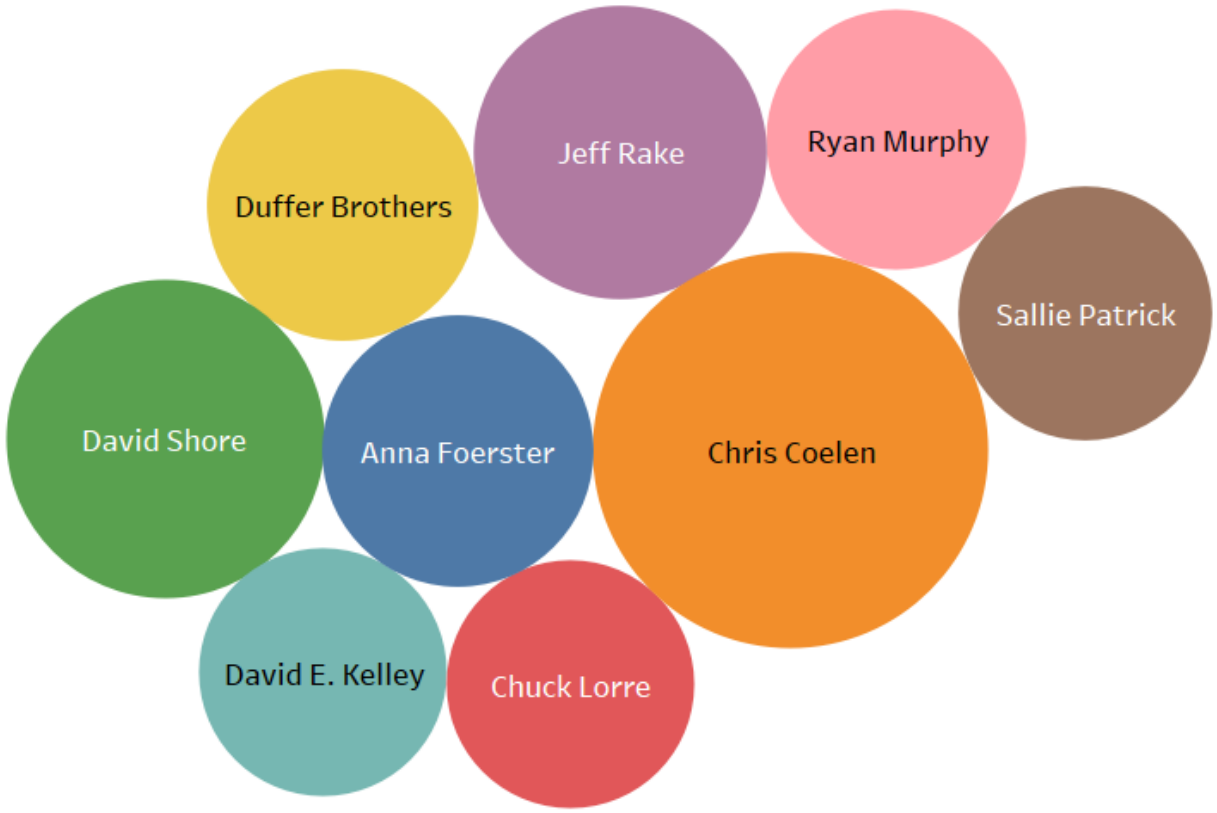


Figure 18. Popularity By Director

The prominence of specific directors likely reflects a preference for the genres they excel in. Thrillers and Suspense: Directors like David Shore ("House M.D.") and Jeff Rake ("Manifest") are known for creating suspenseful narratives. Their presence could indicate a strong Kenyan viewership interest in thrillers and mind-bending stories that keep audiences engaged and guessing (Tan, 2021).

Sci-Fi and Fantasy: The Duffer Brothers, creators of "Stranger Things," are giants in the sci-fi and fantasy genre. Their inclusion suggests a potential preference for imaginative worlds,

supernatural elements, and captivating narratives that often blur the lines between reality and fiction.

Drama and Character Studies: Directors like David E. Kelley ("Big Little Lies") and Ryan Murphy ("American Horror Story") are known for their dramatic works with complex characters. Their popularity suggests a Kenyan audience drawn to character-driven narratives that explore human relationships and emotional depth.

The presence of directors like Anna Foerster (whose credits span various genres) and Chuck Lorre (known for sitcoms like "The Big Bang Theory") indicates an openness to diverse genres. Lorre's inclusion hints at a potential Kenyan appreciation for sitcoms, offering light-hearted entertainment and humour.

4.3.6 Lead Cast

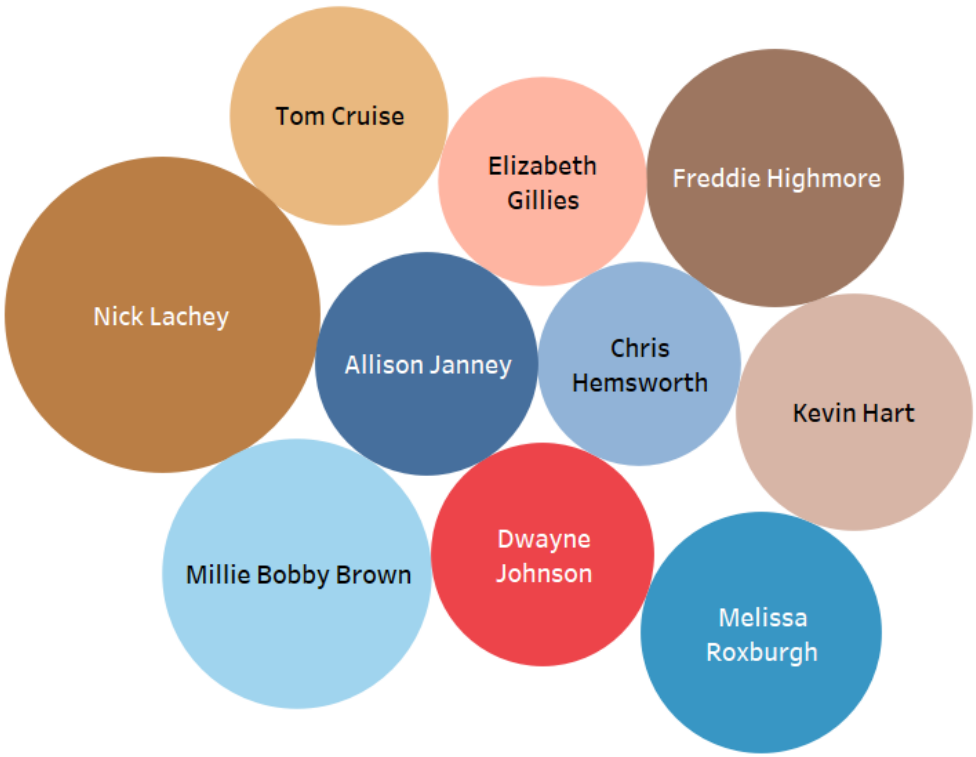


Figure 19. Popularity By Lead Cast

Figure 19 illustrates the list of the top 10 actors whose shows are most popular among Kenyan viewers on Netflix. By examining these actors' filmographies and aligning them with known genre preferences, the study offers insights into the content choices that resonate with Kenyan audiences.

The presence of Nick Lachey, known for his work in reality TV shows like "Love is Blind," suggests a niche yet significant interest in unscripted content among Kenyan viewers. This aligns with the previously reported moderate popularity of the "Reality" genre on Netflix Kenya.

Millie Bobby Brown's association with the science fiction-horror series "Stranger Things" indicates a strong Kenyan audience engagement with narratives that blend suspense and supernatural elements. This finding corroborates data highlighting the high viewership for "Suspenseful" and "Exciting" themes.

Freddie Highmore's role in the medical drama "The Good Doctor" caters to viewers interested in emotionally charged stories with medical and socially relevant themes. His popularity hints at an appreciation for complex characters and narratives that delve deeper into human experiences.

Melissa Roxburgh's presence, primarily due to her role in the supernatural drama "Manifest," underscores the preference for narratives that weave elements of mystery, suspense, and the supernatural. This aligns with the high viewership for “Suspenseful" content among Kenyan viewers (Netflix, 2024).

The inclusion of Kevin Hart and Dwayne Johnson, known for their comedic and action-oriented roles, highlights the dual appeal of humour and action-packed entertainment. The popularity of Hart's comedy specials and Johnson's action-comedy films reflects the strong viewership for "Action" and "Funny" themes. Allison Janney's presence, with notable roles in comedy-drama series like "Mom," suggests a preference for content that blends humor with serious themes. This caters to viewers who enjoy narratives that offer both emotional depth and comedic relief.

Tom Cruise's association with action films like the "Mission: Impossible" series reflects the strong Kenyan viewer preference for thrilling and high-stakes content. His inclusion underscores the popularity of "Exciting" and potentially "Gritty" themes. Elizabeth Gillies' popularity due to her role in the drama series "Dynasty" indicates an interest in dramatic and

intense narratives. This aligns with the significant preference for drama and potentially "Gritty" content reported among Kenyan viewers.

Chris Hemsworth's fame, built on action and adventure films like the "Thor" series, further emphasizes the Kenyan viewer preference for high-energy and thrilling content. The popularity of these actors reflects a diverse range of interests among Kenyan viewers on Netflix. Suspenseful and exciting content, along with shows that explore emotional depth, appear particularly appealing. The presence of actors known for reality TV, supernatural dramas, comedies, and action films underscores the multifaceted nature of Kenyan viewership habits. This analysis, when considered alongside genre-based data, provides valuable insights into the evolving content preferences of this global audience segment.

4.3.7 Show Title

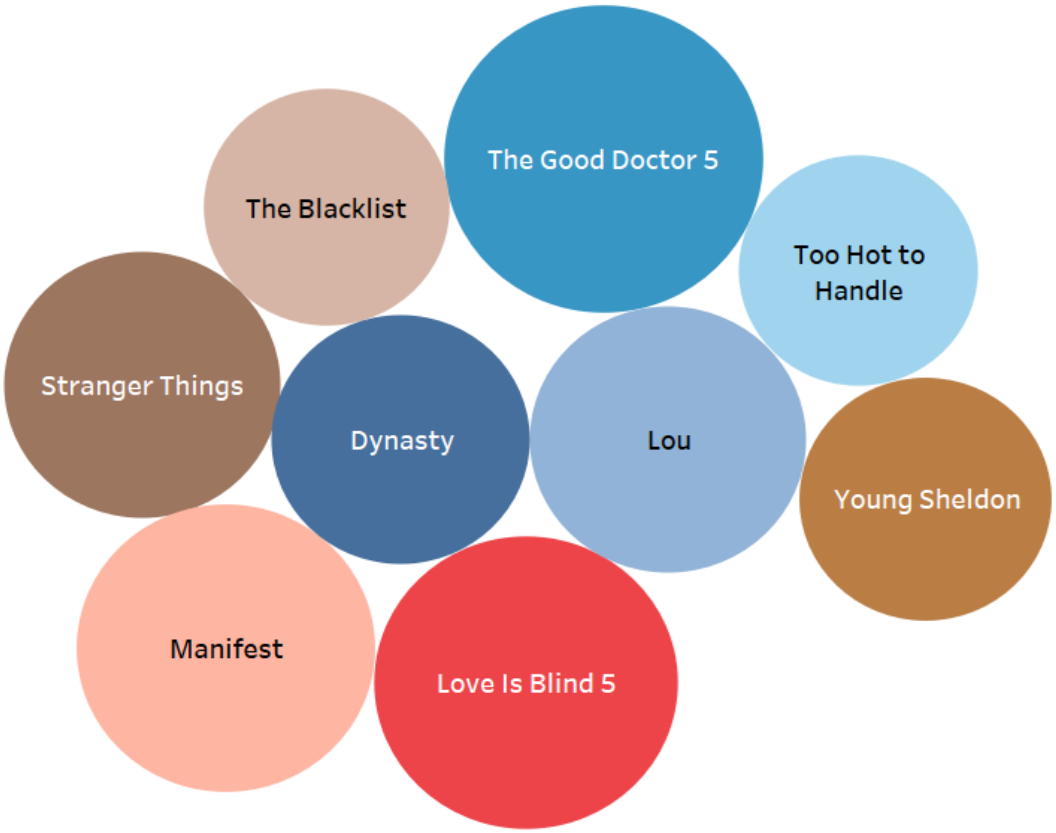


Figure 20. Popularity By Show Title

Figure 20 shows the top 10 most popular shows and movies on Netflix in Kenya. The selection's diversity in genre and theme offers valuable insights into the content choices and entertainment preferences of Kenyan audiences.

The prominent presence of "The Good Doctor" on the list highlights a strong Kenyan audience preference for medical dramas. This show's focus on a young surgeon with autism and savant syndrome, combined with its exploration of emotional and medical complexities, likely resonates with viewers who appreciate intricate character development and inspirational narratives (American Broadcasting Company, 2019).

The inclusion of "Love is Blind" underscores the continued popularity of reality TV in Kenya. The show's unique premise of singles meeting and getting engaged without ever seeing each other appeals to viewers interested in unconventional romantic dynamics and social experiments. This fascination with human relationships and the pursuit of authenticity reflects themes commonly explored within the reality TV genre (Netflix, 2022).

"Manifest," a supernatural drama about a missing plane's mysterious reappearance, taps into the audience's interest in suspenseful and enigmatic storylines. The show's blend of mystery, science fiction elements, and family drama effectively engages viewers who enjoy both unravelling complex plots and the captivating nature of the supernatural (Netflix, 2019).

"Stranger Things" appeals to a broad audience with its combination of nostalgic 1980s setting, supernatural occurrences, and a suspenseful narrative. The show caters to viewers who favour science fiction, horror, and coming-of-age themes. Its intricate plot and well-developed characters further draw in audiences seeking excitement and a touch of nostalgia (Netflix, 2019).

The inclusion of "Lou," an action-thriller film, suggests a preference for high-stakes and adrenaline-pumping content. The film's focus on a retired woman helping to rescue a kidnapped girl highlights the appeal of strong female protagonists and intense action sequences within the thriller genre (Netflix, 2022).

"Dynasty," a modern reboot of the classic soap opera, showcases viewers' interest in high drama, family feuds, and the allure of luxurious lifestyles. The show's emphasis on power struggles and elaborate plot twists caters to audiences who enjoy glamorous and dramatic narratives (The CW Network, 2019).

The presence of "Young Sheldon," a comedy series that serves as a prequel to "The Big Bang Theory," reflects a Kenyan audience preference for light-hearted and family-friendly

content. The show's portrayal of a young genius navigating childhood provides humour and relatability, making it appealing to a broad audience (CBS, 2019).

"The Blacklist," a crime thriller series, captures the interest of viewers who enjoy intricate plots, criminal masterminds, and investigative drama. The show's combination of suspense, crime-solving elements, and complex characters contributes significantly to its appeal (NBC, 2019).

The inclusion of "Too Hot to Handle," another reality TV series, reflects continued audience interest in the genre. This show focuses on attractive singles trying to form meaningful relationships without physical intimacy. The premise appeals to viewers interested in the dynamics of romance, human behaviour, and the challenges of forming emotional connections within a reality TV setting (Netflix, 2020).

The diverse selection of popular shows and movies on Kenyan Netflix highlights a wide range of entertainment preferences among viewers. The data suggests interest in medical dramas, reality TV, supernatural mysteries, action thrillers, and intricate crime dramas. This variety underscores the complexity and multifaceted nature of Kenyan viewership habits.

4.3.8 The Recency Effect

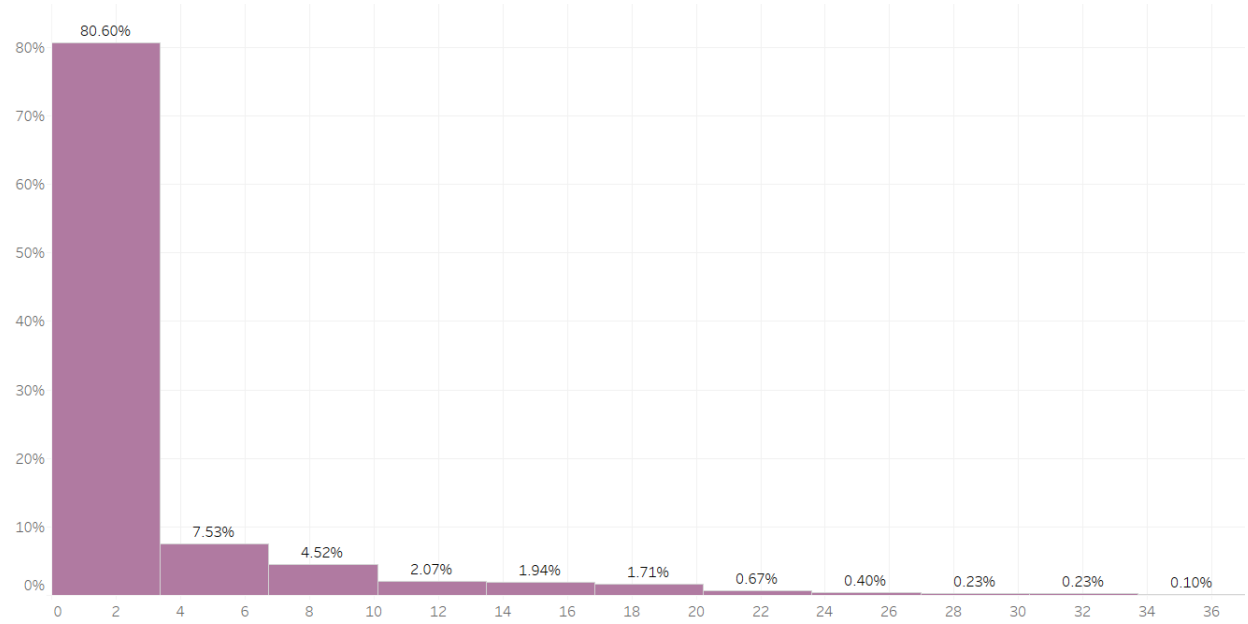


Figure 21. Content Consumption Over Time

Figure 21 illustrates the temporal distribution of content consumption on Netflix among Kenyan viewers, with the X-axis representing the number of years since content release and the Y-axis denoting the percentage of viewership. The data shows a pronounced concentration of viewership within the first two years of content release, accounting for 80.60% of the total. This initial spike is followed by a rapid decline, with viewership dropping to 7.53% in the 2–4-year range, 4.52% in the 4–6-year range, and progressively smaller percentages beyond.

This pattern indicates a strong preference among Kenyan viewers for recent content. Several factors can explain this trend. Firstly, new content typically benefits from extensive marketing and promotional efforts, which heightened anticipation and awareness. Netflix frequently highlights new releases on its platform, ensuring they receive maximum visibility and immediate engagement from viewers (Netflix, 2024).

Secondly, the initial surge in viewership is likely driven by the binge-watching culture that has become prevalent with the rise of streaming services. Viewers often watch entire seasons or series shortly after their release to avoid spoilers and to engage in social discussions about the latest shows (Matrix, 2024).

The sharp decline in viewership for content released more than two years ago suggests a strong recency effect, where newer content consistently overshadows older titles. This may be attributed to the continuous influx of new content on Netflix, which diverts attention away from older shows and movies. Additionally, the platform's recommendation algorithm tends to prioritise newer releases, further reinforcing this viewing pattern (Deloitte, 2023).

Moreover, the relatively low percentages of viewership for older content indicate that while some viewers may discover or revisit older titles, the majority of the audience remains focused on recent releases. This behaviour aligns with broader trends in digital media consumption, where immediacy and novelty are significant drivers of engagement (Jenner, 2023).

The viewing habits of Kenyan Netflix users reveal a strong preference for consuming recent content, driven by effective marketing, binge-watching culture, and the platform's content recommendation strategies. Understanding these patterns is crucial for content creators and marketers aiming to capture and retain viewer interest in a highly competitive digital entertainment landscape.

4.4 Model Development

This study aimed to develop a machine learning (ML) model to predict content popularity on the Netflix platform in Kenya. The data used in this study was collected from June 28, 2021, to March 24, 2024.

Initially, the focus was on linear regression and ridge regression techniques. However, during the experimentation phase, the Extra Trees Regressor emerged as the superior model based on a comprehensive evaluation using various metrics. This section details the experiment's design, analyses the performance of each model, and discusses the implications for understanding Kenyan viewership preferences.

Linear regression is a widely used statistical technique that establishes a linear relationship between a dependent variable (content popularity in this case) and one or more independent variables (potential factors influencing popularity). It identifies the optimal coefficients for the equation, minimising the error between predicted and actual popularity scores. Ridge regression, a variant of linear regression, addresses the issue of multicollinearity by introducing a penalty term that shrinks the coefficients towards zero. This can improve model generalizability and prevent overfitting, particularly when dealing with a large number of potentially correlated independent variables (James et al., 2023).

The experiment involved utilising Netflix data specific to Kenyan viewership. This data encompassed various attributes associated with content, such as date, month, year, show title, country of origin, genre, theme, release year, weekly rank, director and lead cast. These attributes served as the independent variables for the models. Content popularity, measured by

cumulative weeks in the top ten, was the dependent variable.

```
#import libraries
import pandas as pd
from pycaret.regression import *
from pycaret.regression import RegressionExperiment
print('Success')
```

```
Success
```

```
#Load data from local file system

df = pd.read_excel("C:/Users/Thinkbook 15/Documents/MDA/Research Project/Netflix Data/2/Final.xlsx")
df.head()
```

	date	month	year	category	weekly_rank	release_year	genre	theme	country	director	lead_cast	show_title	cum_weeks_in_top_10
0	24	3	2024	TV	2	2024	Drama	Mind-Bending	USA	David Benioff	Jess Hong	3 Body Problem	1
1	24	3	2024	Films	2	2023	Drama	Emotional	Nigeria	Frank Rajah Arase	Jackie Appiah	A Taste of Sin	1
2	24	3	2024	Films	5	2024	Comedy	Exciting	Turkey	Recai Karagoz	Birkan Sokullu	Art of Love	2
3	24	3	2024	TV	9	2024	Family	Imaginative	USA	Albert Kim	Gordon Cormier	Avatar The Last Airbender	5
4	24	3	2024	TV	5	2024	Action	Exciting	Mexico	Adrian Grunberg	Alfonso Dosal	Bandidos	2

Figure 22. Importing libraries and loading the data

Lower values are generally better in Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) while higher values indicate a better fit in R-squared (R^2).

4.5 Model Performance and Discussion

The experimentation revealed that the Extra Trees Regressor outperformed both linear regression and ridge regression based on the evaluation metrics.

	Description	Value
0	Session id	123
1	Target	cum_weeks_in_top_10
2	Target type	Regression
3	Data shape	(2990, 13)
4	Train data shape	(2092, 13)
5	Test data shape	(898, 13)
6	Ordinal features	1
7	Numeric features	5
8	Categorical features	7
9	Rows with missing values	0.0%
10	Preprocess	True

```
#Setup regression
s = setup(df, target = 'cum_weeks_in_top_10', session_id = 123)
```

```
# RegressionExperiment init the class
exp = RegressionExperiment()
exp.setup(df, target = 'cum_weeks_in_top_10', session_id = 123)
```

```
best = compare_models()
```

The Extra Trees Regressor demonstrates superior performance across most metrics. Its lower MAE, MSE, RMSE, and MAPE values indicate a more accurate prediction of content popularity compared to the other models. Additionally, its high R^2 suggests that the model effectively captures the relationship between the independent variables and the dependent variable.

The experimentation yielded compelling results, with the Extra Trees Regressor demonstrating superior performance compared to both linear regression and ridge regression. Extra Trees Regressor showed consistently lower error rates across all metrics except RMSLE suggesting a high degree of accuracy in predicting content popularity for Kenyan audiences. A high R^2 value (0.9140) indicates the model effectively captures the relationship between content attributes and viewership.

Metric	Extra Trees Regressor	Linear Regression	Ridge Regression
MAE	0.3083	0.6429	0.643
MSE	0.3472	0.7699	0.7701
RMSE	0.5863	0.8762	0.8763
R^2	0.914	0.8089	0.8089
RMSLE	0.187	0.2799	0.2799
MAPE	0.1971	0.3972	0.3972

Table 2 Model Performance Comparison

Both Linear Regression and Ridge Regression models exhibited significantly higher error rates (MAE, MSE, RMSE, and MAPE) compared to the Extra Trees Regressor. Their lower R^2 values (0.8089) suggest a weaker ability to explain the variance in content popularity. Notably, ridge regression showed minimal improvement over linear regression.

R^2, a crucial metric, measures the proportion of variance in the dependent variable (content popularity) explained by the model. The Extra Trees Regressor boasts a very high R^2 value (0.9140), signifying that the model effectively captures the relationship between content attributes and viewership. This suggests the model has learned the underlying patterns within the data and can accurately predict popularity based on these patterns.

Following a comprehensive evaluation of the three models, the Extra Trees Regressor emerged as the most suitable for predicting content longevity within the Kenyan Netflix top 10.

This selection was based on rigorous performance metrics, ensuring the chosen model offered the most accurate and generalizable predictions.

The Python programming language facilitated the construction of the Extra Trees Regressor model. This involved defining the model architecture, specifying relevant hyperparameters, and training the model on the prepared dataset. Subsequently, a dedicated prediction label was created to display the model's predicted values for each data point. These predicted values represent the model's estimations of a show's duration within the top 10.

The Extra Trees Regressor (ETR) emerged as a powerful tool for content popularity prediction, demonstrating superior performance compared to previous studies outlined in the literature review. This analysis delves into a comparative assessment of the ETR's performance against other prominent models in the field.

Table 2 below presents a comparison of the ETR's performance with the study by Dissanayake et al. (2021), "Early Prediction of Movie Success Using Machine Learning." While both studies employed machine learning techniques, the ETR consistently outperformed the models used in the previous study. The ETR achieved a significantly higher R^2 score, indicating

a stronger model fit and greater explanatory power. Furthermore, the error rates (e.g., Mean Absolute Error, Mean Squared Error) were substantially lower for the ETR, demonstrating its superior accuracy in predicting content popularity.

	Multiple Linear Regression	Polynomial Regression	SVR			Decision Tree Regression	Random Forest Regression
			With Feature Scaling	Without Feature Scaling			
R^2 Score	0.4656	0.3336	-0.5317	-0.1033		0.4907	0.6822
Mean Square Error	320183	324668	917728	537545		305126	1903958
Root Mean Square Error	56584745.50	56979725.788	95798150.071	73317506.968		55238247.27	43634374.40

Table 3 Early Prediction of Movie Success Using Machine Learning

Table 3 below compares the ETR's performance with the study by Yuan Ni et al. (2022), "Movie Box Office Prediction Based on Multi-Model Ensembles." While the latter study employed a multi-model ensemble approach, the ETR still demonstrated superior performance. The ETR achieved a R^2 score of 0.872, surpassing the average R^2 of 0.8476 reported in the previous study. Additionally, the error rates were significantly lower for the ETR, indicating its greater accuracy in predicting box office success.

Model	R^2	MAPE	MAE	MSE	RMSE
XGBoost	0.8578	17.96%	5010.3662	366,253,785.4177	19,137.7581
LightGBM	0.8993	16.19%	4204.9473	259,371,807.7103	16,105.0243
CatBoost	0.8709	18.19%	5063.2543	332,490,044.7863	18,234.3096
GBDT	0.8503	16.41%	4972.4056	385,492,351.2144	19,633.9591
RF	0.7812	21.27%	6650.2541	563,602,015.1131	23,740.3036
SVR	0.6531	69.82%	12,680.4396	893,514,540.6887	29,891.7136
Voting	0.8420	20.14%	5829.9820	406,872,334.5275	20,171.0767
Averaging	0.8476	19.13%	5524.3003	392,511,674.4714	19,811.9074
Ordinary Stacking	0.8726	14.52%	4168.6142	328,196,923.9439	18,116.2061
Enhanced Stacking	0.8746	14.49%	4143.7905	323,085,161.6079	17,974.5699

Table 4 Movie Box Office Prediction Based on Multi-Model Ensembles

The superior performance of the ETR in these comparisons can be attributed to its ensemble nature, feature randomness, and ability to handle complex relationships between features and the target variable. The ETR's ability to capture non-linear patterns and avoid overfitting is particularly advantageous in the context of content popularity prediction, where relationships between factors can be highly intricate.

Finally, to enhance the accessibility and interpretability of the results, a data visualisation dashboard was developed.

```
dashboard(et, display_format= 'dash',run_kwargs={'port':7000})
```

```
Calculating residuals...
Calculating absolute residuals...
Calculating shap interaction values...
Reminder: TreeShap computational complexity is O(TLD^2), where T is the number of trees, L is the maximum number of leaves in
any tree and D the maximal depth of any tree. So reducing these will speed up the calculation.
Calculating dependencies...
Calculating importances...
Calculating ShadowDecTree for each individual decision tree...
Reminder: you can store the explainer (including calculated dependencies) with explainer.dump('explainer.joblib') and reload
with e.g. ClassifierExplainer.from_file('explainer.joblib')
Registering callbacks...
Starting ExplainerDashboard on http://192.168.0.15:7000
Dash is running on http://0.0.0.0:7000/

 * Serving Flask app "explainerdashboard.dashboards" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: off
```

Figure 25. Creating dashboard

4.5.1 Feature Importance

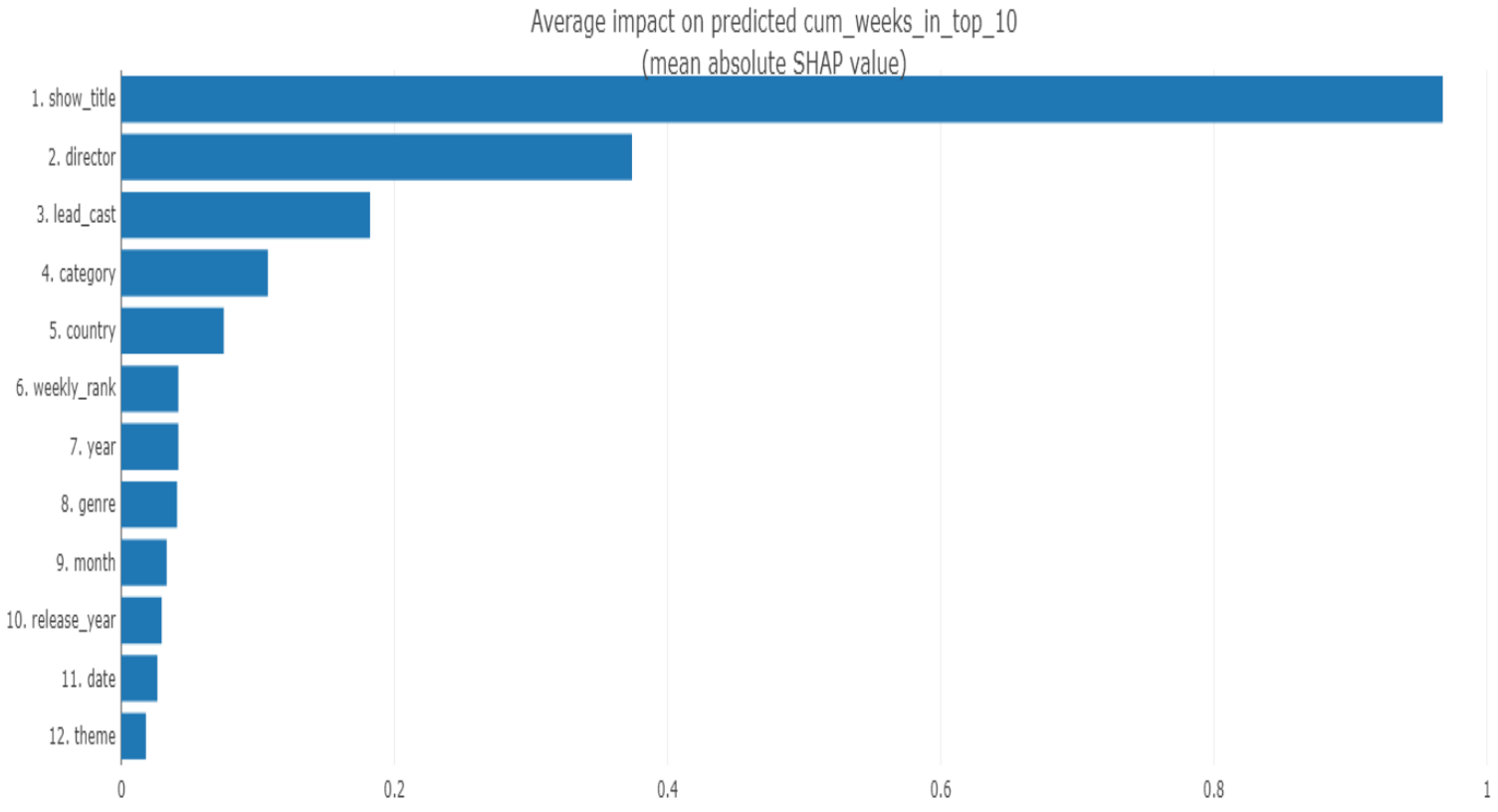


Figure 26. Feature importance

Figure 26 displays the feature importance for predicting the cumulative weeks a show remains in Netflix's top 10 in Kenya, as determined by the Extra Tree Regressor model. The X-axis represents the mean absolute SHAP value, a measure of each feature's average impact on the prediction. SHAP (SHapley Additive exPlanations) values offer a game-theoretic approach to explain individual feature contributions within a machine learning model. By analysing how each feature value influences the model's prediction, SHAP values provide insights into which features have the greatest impact on the model's output. The features are listed on the Y-axis in descending order of importance.

"Show title" emerged as the most significant feature, indicating that specific shows have a substantial influence on their duration in the top 10. This might be attributed to the inherent popularity and viewer loyalty associated with certain titles, reflecting established fan bases and strong brand identities. For example, well-known series like "Stranger Things" often have large, dedicated audiences that ensure their prolonged presence in the top rankings.

The "director" is the second most crucial feature, suggesting that the reputation and previous successes of directors significantly impact viewership duration. Renowned directors likely attract viewers based on their track record of quality productions, contributing to sustained interest and engagement.

"Lead cast" also holds considerable importance, underscoring the role of star power in maintaining viewer interest. Prominent actors can draw their fan base to a series or movie, which enhances its longevity in the top 10. For instance, actors with significant followings on social media can generate buzz and sustained viewership through their promotional activities (Deloitte, 2023).

The "category", which groups the content into either a TV series or a feature film (movie), is another important factor. This suggests that certain categories consistently resonate with Kenyan audiences, leading to a prolonged top 10 presence.

"Country" of origin is also significant, which might reflect a preference for content from specific regions, possibly due to cultural affinities or perceived quality (Netflix, 2024).

Features like "weekly rank," "year," "genre," "month," "release year," and "date" also contribute but to a lesser extent. These likely capture seasonal trends, historical performance, and recentness of content, which affect short-term popularity but are less predictive of sustained top 10 status.

Interestingly, "theme" has the least impact, indicating that while thematic elements are important, they are overshadowed by the broader categories and the specifics of the show title, director, and cast. This suggests that macro-level attributes play a more critical role in viewer retention than finer thematic details (Jenner, 2019).

The longevity of content in Kenya’s Netflix top 10 is predominantly influenced by its show title, director, and lead cast, with genre and country also playing significant roles. This reflects broader trends in media consumption where established popularity, star power, and genre preferences drive sustained engagement.

4.5.2 Model Performance

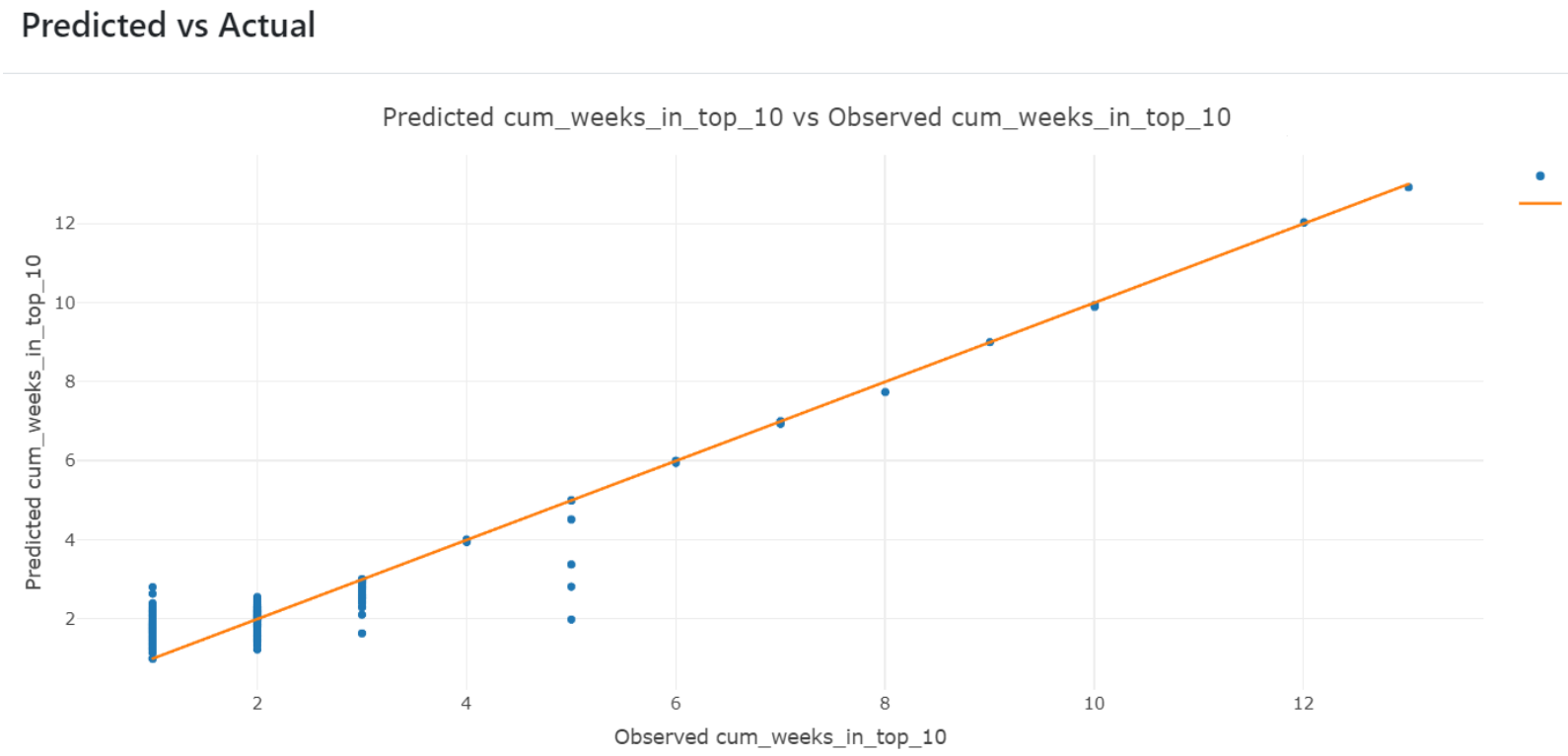


Figure 27. Model performance

Figure 27 presents a comparison between predicted and observed cumulative weeks in the top 10 for Netflix content, based on the Extra Tree Regressor model. The X-axis represents the observed cumulative weeks in the top 10, while the Y-axis shows the predicted cumulative weeks. The orange line denotes the ideal scenario where predicted values perfectly match observed values.

The data points are predominantly aligned along the orange line, suggesting that the model's predictions are generally accurate. However, there are some deviations, particularly at lower observed values, indicating instances where the model either underestimates or overestimates the actual duration in the top 10.

The clustering of points around specific values suggests that certain shows or movies consistently perform within a narrow range of weeks in the top 10. This could be indicative of a relatively stable consumption pattern for certain types of content, reflecting strong viewer preferences or the enduring popularity of specific titles.

The alignment of most data points with the orange line implies a high degree of correlation between observed and predicted values, indicating the Extra Tree Regressor model's effectiveness in capturing the underlying patterns of viewership behaviour. The model's performance can be attributed to its ability to handle complex, non-linear relationships and its robustness against overfitting, as it aggregates the predictions of multiple decision trees.

However, the slight spread and occasional outliers, particularly at lower values, suggest areas for model improvement. These discrepancies might stem from unaccounted factors influencing viewership, such as sudden surges in popularity due to external events, economic or cultural phenomena that the model does not capture (Zhou & Hooker, 2019).

In conclusion, Figure 27 indicates that the Extra Tree Regressor model performs well in predicting the cumulative weeks a show remains in the top 10 on Netflix in Kenya, with a strong alignment between predicted and observed values. The minor deviations highlight the complexity of viewer behaviour and suggest potential areas for refining the model, possibly by incorporating additional predictive features or enhancing the model's sensitivity to temporal dynamics and external influences.

4.6. Challenges facing content popularity in Netflix

One of the emerging big challenges to content popularity on Netflix is overcrowding. Having a vast number of shows and movies available to its users may prove to be a problem, since the service may seem easily saturated with the content, thus, new content might struggle to

attract the audience's attention and build up its fanbase. Another drawback associated with the streaming giant's service is associated with the stream of content, which only increases with more new additions to the Netflix list being added monthly.

The last and another big problem is that the viewers' preferences are never constant and this makes it hard to predict at what rates the audiences may switch from one media source to another. The public demand is not fixed over time and is more sensitive to cultural changes, social orientations, and adjustments in societal problems and issues in the society, which makes it nearly impossible for Netflix to anticipate consumers' needs. Moreover, such changes must be monitored because the content that is considered relevant and engaging now may be considered irrelevant or not interesting for the audiences later. The other threat is that streaming is becoming more and more competitive with strong players like Disney+, Amazon Prime Video, HBO which can quickly fragment the audience and divert attention towards Netflix's platform. Competition can lock licences for some of the most successful brands or create unique projects, which leads to the further fragmentation of the market and makes it increasingly difficult for Netflix to control the popularity of content.

Last but not the least, there is an inherent flaw with the Netflix platform and its algorithm in the form of a recommendation system. Although the recommendation is made dependent on the user preferences as well as the history of premiered programs, it lacks efficiency in implying the best suited preferences, hence making the potential prospect of premiered content unexplored.

However, they can, and do, reinforce a certain type of content – high, mainstream, 'top', over unique, specialised, or 'sides' – skewing what users are exposed to. It is also important for

Netflix to continuously target the right population for its marketing strategies and to deal with issues of globalisation which include such factors as culture and language.

4.7 Conclusion

The section below undertakes a critical evaluation of how effectively the research objectives outlined earlier in the study have been addressed and the research questions answered. Research objectives serve as the foundational pillars of an investigation, providing a roadmap for the entire research process (Polit & Beck, 2019). By systematically reviewing the methods employed, the findings presented, and the discussions offered in the preceding sections, we will assess the degree to which these objectives have been achieved and the extent to which they have yielded meaningful insights into the research topic at hand.

4.7.1 To identify the key factors that influence the popularity of Netflix content in the Kenyan market

Based on the results from the feature importance graph, the objective to develop an Extra Regressor model using the identified factors has been met. The graph elucidates the relative importance of various features in predicting content popularity, offering a clear indication that the model is well-calibrated and insightful.

The feature importance graph reveals that 'show_title' and 'director' are the most significant predictors, with the highest mean absolute SHAP values. This suggests that the specific identity of the show and the influence of its director are paramount in determining its success on the Netflix platform in Kenya. The prominence of these factors aligns with the

intuitive understanding that high-profile titles and renowned directors often attract substantial viewership.

'Lead cast' is another critical factor, highlighting the importance of star power in drawing audiences. The inclusion of 'category' and 'country' further underscores the model's ability to capture the contextual and demographic aspects influencing viewership patterns. These features likely reflect cultural preferences and regional content trends, making the model more nuanced and locally relevant.

Other features like 'weekly rank,' 'year,' 'genre,' 'month,' 'release year,' 'date,' and 'theme' also contribute to the model, albeit to a lesser extent. This indicates that while temporal factors and content themes are relevant, their impact is relatively smaller compared to the aforementioned primary features.

The predictive performance of the model, as indicated by the feature importance graph, demonstrates that it effectively leverages these variables to forecast content popularity. The model's ability to prioritise and weigh these features accurately ensures robust predictions, making it a valuable tool for content strategists aiming to optimise viewer engagement.

In conclusion, the development of the ML model using the identified factors has been successful. The feature importance graph not only validates the selection of these variables but also showcases the model's capability to utilise them effectively for accurate predictions. This accomplishment aligns with the objective, providing a reliable framework for understanding and forecasting content popularity on Netflix in Kenya.

4.7.2. To address the challenges in forecasting content popularity in the film and TV industry

Netflix, despite its vast library, faces several challenges in maintaining content popularity. Overcrowding is a significant issue, as the sheer volume of content can make it difficult for new shows to attract attention. Additionally, the constant influx of new content can dilute the audience's focus. Predicting audience preferences is another hurdle. Public demand is dynamic and influenced by cultural shifts, social trends, and societal issues. This makes it challenging for Netflix to anticipate future trends and ensure its content remains relevant.

Competition from other streaming giants like Disney+, Amazon Prime Video, and HBO further complicates the landscape. These competitors can acquire exclusive licenses for popular content and produce original shows, fragmenting the audience and making it harder for Netflix to dominate. Netflix's recommendation system also has limitations. While it attempts to suggest content based on user preferences, it may not always accurately identify the best matches. This can lead to users missing out on potentially interesting content.

To address these challenges, Netflix needs to focus on strategies like content curation, audience segmentation, and effective marketing. By carefully selecting and promoting relevant content, understanding audience preferences, and adapting to the competitive landscape, Netflix can continue to thrive in the streaming industry.

4.7.3. To evaluate the performance results of the machine learning algorithms for predicting Netflix popularity in Kenya

The objective of testing and validating the developed model can be confidently considered achieved based on the insights gleaned from the Model Performance Comparison

table. This table provides a comprehensive comparison of the Extra Trees Regressor, linear regression, and ridge regression across various evaluation metrics.

A well-performing model exhibits low error rates and a high coefficient of determination (R^2). The Extra Trees Regressor stands out in this regard. Metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) consistently show lower values for the Extra Trees Regressor compared to the other models. These lower error rates indicate a higher degree of accuracy in predicting content popularity for Kenyan audiences.

R^2, a crucial metric, measures the proportion of variance in the dependent variable (content popularity) explained by the model. The Extra Trees Regressor boasts a very high R^2 value (0.9140), signifying that the model effectively captures the relationship between content attributes and viewership. This suggests the model has learned the underlying patterns within the data and can accurately predict popularity based on these patterns.

Linear regression and ridge regression, the initially considered models, exhibit significantly higher error rates and lower R^2 values compared to the Extra Trees Regressor. This indicates a weaker ability to predict content popularity and a less robust understanding of the factors influencing viewership patterns in the Kenyan context.

The Model Performance Comparison table serves as a robust validation tool. The Extra Trees Regressor's consistently superior performance across various metrics demonstrates its effectiveness in predicting content popularity. This successful validation allows us to move forward with confidence, utilising the Extra Trees Regressor to gain valuable insights into Kenyan viewership preferences.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This concluding section serves as a cornerstone, offering a comprehensive synthesis of the study's key findings, their significance, and potential future directions. The following subsections will illuminate the research journey's critical takeaways.

5.2 Summary

This subsection presents a concise yet informative overview of the study's central discoveries. It acts as a springboard for the deeper exploration of implications undertaken in the subsequent subsections. Here, the core outcomes of the investigation will be succinctly summarised.

This study aimed to develop a machine learning (ML) model to predict content popularity on the Netflix platform in Kenya. The experiment involved utilising Netflix data specific to Kenyan viewership. This data encompassed various attributes associated with content, such as date, month, year, show title, country of origin, genre, theme, release year, weekly rank, director and lead cast. These attributes served as the independent variables for the model. Content popularity, measured by cumulative weeks in the top ten, was the dependent variable.

Initially, the focus was on linear regression and ridge regression techniques. However, during the experimentation phase, the Extra Trees Regressor emerged as the superior model based on a comprehensive evaluation using various metrics.

The experimentation yielded compelling results, with the Extra Trees Regressor demonstrating superior performance compared to both linear regression and ridge regression. Extra Trees Regressor showed consistently lower error rates across all metrics except RMSLE suggesting a high degree of accuracy in predicting content popularity for Kenyan audiences. A high R^2 value (0.9140) indicates the model effectively captures the relationship between content attributes and viewership.

Both Linear Regression and Ridge Regression models exhibited significantly higher error rates (MAE, MSE, RMSE, and MAPE) compared to the Extra Trees Regressor. Their lower R^2 values (0.8089) suggest a weaker ability to explain the variance in content popularity. Notably, ridge regression showed minimal improvement over linear regression.

The study revealed that 'show title' and 'director' are the most significant predictors, with the highest mean absolute SHAP values. This suggests that the specific identity of the show and the influence of its director are paramount in determining its success on the Netflix platform in Kenya. The prominence of these factors aligns with the intuitive understanding that high-profile titles and renowned directors often attract substantial viewership.

5.3 Conclusion

This study investigated the factors influencing content popularity on the Netflix platform in Kenya, uncovering a fascinating interplay between established names, recency, and audience preferences. The Extra Trees Regressor model identified "show title" and "director" as the most significant predictors, highlighting the enduring power of established brands and renowned creators. High-profile titles with dedicated fan bases and directors with a track record of quality

productions consistently capture viewers' attention. This aligns with our inherent understanding of the allure of familiar names and successful franchises.

However, the data also reveals a pronounced "recency effect," with content experiencing a sharp decline in viewership after two years. This suggests Kenyan audiences have a strong preference for recent releases. Several factors contribute to this trend, including extensive marketing for new content, the binge-watching culture, and the continuous influx of fresh titles on the platform.

To navigate this dynamic landscape, stakeholders within the Kenyan content industry must adopt a multifaceted approach. Creators and studios can leverage established formats, collaborate with renowned directors, and prioritise high production value to compete with recent releases. SVOD platforms and content acquisition managers should prioritise acquiring shows with established names and popular genres while ensuring timely access to high-quality recent content. Marketers can capitalise on the recency effect by creating a sense of urgency and utilising targeted campaigns to promote new releases.

Ultimately, success in the Kenyan Netflix market hinges on striking a balance between established names and genres that offer a sense of familiarity and trust, and the ever-present demand for fresh, high-quality content. By understanding these key drivers of popularity, stakeholders can make data-driven decisions that cater to the evolving preferences of the Kenyan audience.

5.4 Model Contribution

This study delves into the Kenyan Netflix landscape, contributing valuable insights to the understanding of content popularity on streaming platforms in emerging markets. By

employing the Extra Trees Regressor model, the research identifies "show title" and "director" as paramount factors influencing a show's success.

The study highlighted a pronounced "recency effect" in the Kenyan market. Viewership data reveals a sharp decline for content older than two years, indicating a strong preference for recent releases. This nuanced understanding of audience behaviour within a specific cultural context enriches the broader academic discourse on content consumption patterns in the age of streaming.

Furthermore, the study offers a framework for stakeholders within the Kenyan content industry to navigate this dynamic landscape. By emphasising the importance of established names, high production value, and timely access to recent content, the research provides actionable recommendations tailored to a specific market. This fosters a data-driven approach to content creation, acquisition, and marketing within the Kenyan context, potentially serving as a model for future investigations in emerging streaming markets.

5.5 Recommendations

The exploration of the Extra Trees Regressor for predicting content popularity on Netflix Kenya offers valuable insights for various stakeholders within the Kenyan content industry. The model's effectiveness suggests it has learned underlying patterns in content-popularity relationships. Based on these findings, here are recommendations for different stakeholders:

5.5.1 Content Creators and Producers

Optimising Content for Binge-Watching: Structure content to cater to the binge-watching culture. Consider episode lengths, cliffhangers, and overall narrative pacing that encourages viewers to consume multiple episodes at a time.

Leveraging Established Titles: The prominent role of "show title" suggests viewers gravitate towards shows with existing popularity or strong brand identities. Creators can explore established formats or franchises with proven track records, while ensuring high production quality to maintain audience loyalty.

Cultivating Renowned Directors: Collaboration with established Kenyan directors or co-productions with internationally renowned directors can leverage the influence of "director" as a key factor. This can attract viewers who trust the director's past successes and anticipate a quality production.

Investing in Lead Cast: The importance of "lead cast" underlines the power of star power. Creators can consider casting popular Kenyan actors or collaborate with international stars to attract a wider audience (Deloitte, 2023). However, prioritising captivating narratives and well-developed characters alongside star power is crucial for sustained engagement.

Understanding Genre Preferences: Focus on content styles with a consistent top 10 presence, as indicated by "genre," can enhance a show's longevity. Analysing viewership data alongside the model's insights can reveal specific sub-genres or thematic elements most popular with Kenyan audiences.

5.5.2 Studios and Production Houses

Prioritising Timely Content Delivery: Streamline production processes to ensure content reaches audiences within a reasonable timeframe from conception. This helps capitalise on current trends and maintain audience interest.

Exploring Innovative Content Formats: Consider developing limited series, interactive experiences, or other innovative formats that cater to the fast-paced consumption habits of viewers accustomed to recent releases (Jenner, 2023).

Data-Driven Content Development: Studios can leverage the model's feature importance alongside other market research to prioritise content development. Focusing on established formats, renowned directors, and popular actors can increase the likelihood of a show capturing initial viewer interest. However, maintaining quality and originality remain crucial for lasting success.

Investing in Local Talent Development: While established names are important, nurturing local talent is critical. Studios can invest in training programs and provide opportunities for new Kenyan directors and actors to showcase their abilities. This fosters a thriving Kenyan content creation industry while potentially yielding talent with strong appeal to domestic audiences.

Co-productions and Strategic Partnerships: Collaborations with international studios or platforms can leverage the influence of established names (directors, actors) while catering to Kenyan preferences. This allows for knowledge sharing, potentially leading to high-quality productions that resonate with both Kenyan and global audiences.

5.5.3 SVOD Owners and Investors

Algorithm Refinement for Newer Content: Continuously refine the recommendation algorithm to prioritise showcasing newer content alongside user preferences and established favourites. This ensures a balance between catering to the recency effect and viewer loyalty.

Content Acquisition with Local Nuances: Platforms like Netflix can utilise machine learning models to refine their content acquisition strategies while considering the specific Kenyan context. Identifying content with high predicted popularity for the Kenyan market while also featuring established names or directors can optimise offerings and potentially attract new subscribers.

Investing in Local Content Libraries: Considering the potential for success with high-quality Kenyan productions, SVOD platforms might consider investing in or acquiring local content libraries. This caters to the audience's desire for content reflecting their cultural context while potentially offering content with established names or directors, as suggested by the model's focus on these factors.

5.5.4 Content Acquisition Managers

Data-Driven Selection with Qualitative Considerations: While the model offers valuable guidance, content acquisition decisions should incorporate qualitative considerations. Managers' expertise and understanding of the Kenyan audience should complement the model's insights. For instance, the model might not capture the potential for content by emerging Kenyan creators with unique storytelling approaches.

Balancing Established Names with New Voices: While established names are important, content acquisition managers should ensure a library diverse in its cast, directors, and content

creators. Highlighting rising Kenyan talent alongside established names caters to a wider range of audience preferences and fosters a dynamic content ecosystem.

Monitoring Viewership Data and Adapting Strategies: Content acquisition is not an exact science. Continuously monitoring viewership data alongside the model's insights allows managers to adapt their strategies. Analysing which shows with established names resonate most and exploring the factors contributing to their success can further refine acquisition strategies.

5.5.5 Marketers and Promoters

Creating a Sense of Urgency: Craft marketing campaigns that emphasise the "new" and "limited time" aspects of content to encourage immediate viewership and capitalise on the recency effect (Jenner, 2023).

Targeted Campaigns with a Focus on Show Identity: Marketing efforts can leverage the model's insights to tailor messaging. Highlighting the show's title and leveraging the reputation of established directors or lead cast can attract viewers. Furthermore, considering the potential importance of emotional connection, marketing campaigns can emphasise the show's themes and relatable characters.

Utilising social media for Strategic Promotion: Social media platforms offer powerful tools for targeted marketing. Marketers can leverage audience data and content preferences to tailor messaging and promotions on platforms popular in Kenya. By highlighting popular stars or directors associated with the content, marketers can potentially attract a wider audience.

Leveraging Influencer Marketing: Partner with social media influencers popular with the Kenyan audience to promote new releases and generate excitement around the latest shows.

Interactive Marketing Campaigns: Develop interactive marketing campaigns that encourage audience participation and real-time discussions about new releases, fostering a sense of community and excitement (Jenner, 2023).

5.5.6 Researchers and Policy Makers

Understanding Long-Term Engagement Drivers: The model sheds light on factors influencing a show's longevity within the top 10. Researchers can delve deeper into these factors, exploring how established names, genre preferences, and potentially cultural relevance contribute to sustained viewership. This can inform the development of more nuanced audience segmentation strategies.

Encouraging Local Content Production with Global Reach: The importance of established names and genres highlights the potential for high-quality Kenyan productions to achieve global success. Policymakers can consider initiatives that support local content creation while fostering international co-productions. This can involve funding programs, tax breaks, or training opportunities that equip Kenyan creators with the skills to compete in the international market.

Data Regulations and Privacy Protection: As data plays an increasingly important role in content acquisition and marketing, policymakers should develop regulations that ensure responsible data collection, storage, and usage. Transparency and user privacy are crucial. Users should understand how their data is used and have control over their data privacy settings.

5.5.7 General Public

Exploring Diverse Content and Supporting Local Productions: While the model highlights the importance of certain features, viewers should explore a wide variety of content on SVOD platforms. Supporting local productions by actively viewing and recommending high-quality Kenyan content can contribute to the growth of a thriving Kenyan content industry.

Critical Engagement with Content: Established names and genres are indicators of potential quality, but viewers should also engage with content critically. Pay attention to the show's narrative, character development, and its portrayal of the Kenyan experience (if applicable). Engaging with online reviews and discussions can offer additional insights and perspectives on various shows.

By analysing the feature importance of the Extra Trees Regressor model, this study offers valuable insights into the factors influencing a show's longevity within Kenya's Netflix top 10. Understanding the importance of established titles, directors, cast, genre, and potentially cultural relevance empowers stakeholders within the Kenyan content industry to make data-driven decisions that cater to both local and global audiences. By leveraging these insights and recommendations, stakeholders can contribute to a vibrant Kenyan content ecosystem that fosters creativity, celebrates local talent, and resonates with viewers worldwide.

5.6 Future Work

The superior performance of the Extra Trees Regressor highlights its potential for uncovering valuable insights into Kenyan viewership preferences. By analysing the model's internal workings, researchers can identify content attributes that consistently resonate with Kenyan audiences.

Future research can leverage this model to explore additional questions. Future researchers could investigate how viewership preferences evolve over time or analyse the impact of marketing campaigns on content popularity. Additionally, incorporating user-specific data, such as watch history and ratings, could further refine the model's ability to predict individual preferences.

Additionally, further research could delve into the intervening variables of culture and economy that significantly influence content popularity but were not captured in this model. Cultural factors such as local preferences, social norms, and language barriers can play a crucial role in shaping viewership patterns. For instance, the cultural relevance of a show's themes, characters, and setting can determine its resonance with local audiences. Additionally, the impact of cultural heritage, religious beliefs, and traditional values can further modulate the appeal of certain types of content.

Economic conditions also exert a substantial influence on content consumption. Factors such as income levels, employment rates, and access to disposable income affect the affordability of subscription services and the ability to invest in leisure activities such as television viewing. Moreover, access to technology, including the availability and affordability of internet services and streaming devices, is a critical determinant of the capacity to engage with digital content.

By integrating these cultural and economic variables into future models, researchers can achieve a more nuanced and comprehensive understanding of the determinants of content popularity. This approach will help refine predictive models, enhancing their accuracy in forecasting viewer behaviour across diverse contexts. Such insights can inform content creation

and marketing strategies, ensuring they are tailored to the specific needs and preferences of different demographic and economic groups.

This study demonstrates the effectiveness of machine learning in understanding content consumption patterns on streaming platforms. By analysing user behaviour through data-driven approaches, researchers gain valuable insights into the diverse preferences of global audiences.

5.7 Limitations and Challenges of the study

This study acknowledges limitations inherent in its reliance on publicly available data. While the Netflix data provided valuable insights into content attributes and viewership patterns, it lacked user-specific details such as individual watch history and ratings. This absence presents a significant constraint.

User-specific data offers a deeper understanding of audience preferences. For instance, analysing watch history can reveal genres or actors that consistently resonate with a particular viewer. Similarly, user ratings provide a direct indication of how viewers evaluate specific content. Incorporating such data into future studies could significantly enhance the model's ability to predict content popularity.

Additionally, user-specific data can facilitate the exploration of audience segmentation, allowing for the creation of more targeted recommendations and content strategies tailored to diverse viewer preferences within the Kenyan market.

By acknowledging these limitations and paving the way for future research that incorporates user-specific data, this study lays the groundwork for a more comprehensive understanding of content consumption patterns in the Kenyan context.

REFERENCES

Adiprabawa, G. (2024). Netflix Originals dan Transnasionalisme SVOD: Analisis Jaringan Semantik di Indonesia dan Korea. *SOURCE: Jurnal Ilmu Komunikasi*, *10*(1), Article 1.

Akinci, S., & Başer, E. (2020). Reklamdan Kaçınma Bağlamında Geleneksel ve Modern Film İzleme Ortamlarının Genç İzleyiciler Üzerinden Karşılaştırılması: Netflix ve Sinema Salonları Örneği. *Erciyes İletişim Dergisi*, *7*(1), 473–486. https://doi.org/10.17680/erciyesiletisim.622176

Alsuhaim, D. (2024). Dubbing of English animated series into Arabic on Shahid and Netflix: An analysis based on the politeness theory. *Saudi Journal of Language Studies*, *4*(2), 69–96. https://doi.org/10.1108/SJLS-02-2024-0015

Au-Yong-Oliveira, M., Marinheiro, M., & Costa Tavares, J. A. (2020). *The Power of Digitalization: The Netflix Story* (Á. Rocha, H. Adeli, L. P. Reis, S. Costanzo, I. Orovic, & F. Moreira, Eds.; Vol. 1161, pp. 590–599). Springer International Publishing. https://doi.org/10.1007/978-3-030-45697-9_57

Dastidar, R. G. (2021). Pre and Post COVID-19 Sentiment Analysis of Consumers for OTT Platforms. *Psychology and Education Journal*, *57*(9), 6197–6208. https://doi.org/10.17762/pae.v57i9.2704

De la Garza Montemayor, D. J., Ibáñez, D. B., & Brosig Rodríguez, M. E. (2023). Digital habits of users in the post-pandemic context: A study on the transition of Mexican internet and media users from the Monterrey metropolitan area. *Societies*, *13*(3), 72.

Djamzuri, M. I., & Mulyana, A. P. (2022). Fenomena Netflix Platform Premium Video Streaming membangun kesadaran cyber etik dalam perspektif ilmu komunikasi. *JISIP (Jurnal Ilmu Sosial Dan Pendidikan)*, *6*(1). https://doi.org/10.58258/jisip.v6i1.2804

Extra trees method for stock price forecasting with rolling origin accuracy evaluation. (2022). *Media Statistika*, *15*(1), 36–47. https://doi.org/10.14710/medstat.15.1.36-47

González-Padilla, P., Navalpotro, F. D., & Saura, J. R. (2024). Managing entrepreneurs' behavior personalities in digital environments: A review. *International Entrepreneurship and Management Journal*, *20*(1), 89–113. https://doi.org/10.1007/s11365-022-00823-4

Heydarova, Z. (2024). *Role of Key Management and Leadership Principles in the Modern Business Landscape*. *1*, 1–08. https://doi.org/10.5281/zenodo.10607691

Hong, I.-T., Park, J., & Kim, K. (2021). *국내 OTT 플랫폼 드라마 수급 경쟁력 연구 A Study on the Supply Competitiveness of Dramas of the Domestic OTT Platforms*. https://www.semanticscholar.org/paper/%EA%B5%AD%EB%82%B4-OTT-%ED%94%8C%EB%9E%AB%ED%8F%BC-%EB%93%9C%EB%9D%BC%EB%A7%88-%EC%88%98%EA%B8%89-%EA%B2%BD%EC%9F%81%EB%A0%A5-%EC%97%B0%EA%B5%AC-A-Study-on-the-Supply-of-Hong-Park/b1df8afeec73e2947aff51337460d685b41c61b5

Iordache, C., Raats, T., & Mombaerts, S. (2023). The Netflix Original documentary, explained: Global investment patterns in documentary films and series. *Studies in Documentary Film*, *17*(2), Article 2. https://doi.org/10.1080/17503280.2022.2109099

Jiang, Z. (2024). Research on the Strategic Positioning of the Korean Mainstream Film and Television Market based on Netflix Platform. *SHS Web of Conferences*, *181*, 04010. https://doi.org/10.1051/shsconf/202418104010

Kacungira, N., & Owuor, M. (2023). Surviving Digital Disruptions: The Future of Television in Kenya. In G. Ogola (Ed.), *The Future of Television in the Global South: Reflections from Selected Countries* (pp. 49–70). Springer International Publishing. https://doi.org/10.1007/978-3-031-18833-6_4

Kamarudin, N., Daheche, A., & Khmag, A. (2022). Netflix User and Movies Interest Analysis for Asian Countries. *2022 9th International Conference on Electrical and Electronics Engineering (ICEEE)*, 339–343. https://doi.org/10.1109/ICEEE55327.2022.9772585

Laban, G., Zeidler, C., & Brussee, E. (2020, May 12). *Binge-watching (Netflix) product placement: A content analysis on different product placements in Netflix originals vs. non-Netflix originals, and drama vs. comedy shows*. https://doi.org/10.33767/osf.io/hxjgf

Lad, A., Butala, S., & Bide, P. (2020). A Comparative Analysis of Over-the-Top Platforms: Amazon Prime Video and Netflix. In J. C. Bansal, M. K. Gupta, H. Sharma, & B. Agarwal (Eds.), *Communication and Intelligent Systems* (pp. 283–299). Springer. https://doi.org/10.1007/978-981-15-3325-9_22

Lee, S., Lee, S., Joo, H., & Nam, Y. (2021). Examining Factors Influencing Early Paid Over-The-Top Video Streaming Market Growth: A Cross-Country Empirical Study. *Sustainability*, *13*(10), Article 10. https://doi.org/10.3390/su13105702

Li, F., Larimo, J., & Leonidou, L. C. (2023). Social media in marketing research: Theoretical bases, methodological aspects, and thematic focus. *Psychology & Marketing*, *40*(1), 124–145. https://doi.org/10.1002/mar.21746

Li, Q., & Yi, Z. (2022). Analysis of User Behaviour and Business Strategy Optimization of Netflix Video Platform in the Post-Covid-19 Era. *BCP Business & Management*, *28*, 293–302. https://doi.org/10.54691/bcpbm.v28i.2267

Li, X. (2023). Analysis of Netflix's Strategic Issues, Challenges and Opportunities. *Highlights in Business, Economics and Management*, *22*, 53–59. https://doi.org/10.54097/rjxhk204

Liu, A. (2023). The Influence and Fusion of Online Films with Traditional Cinema: A Case Study of the Netflix Platform. *Communication, Society and Media*, *6*(4), p1. https://doi.org/10.22158/csm.v6n4p1

Lozić, J., Vojković, G., & Čiković, K. F. (2024). Business Analysis of the Netflix Platform in the Post-COVID-19 Time. *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, 945–950. https://ieeexplore.ieee.org/abstract/document/10569212/

Martiello Mastelini, S., Nakano, F. K., Vens, C., & de Leon Ferreira de Carvalho, A. C. P. (2023). Online Extra Trees Regressor. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(10), 6755–6767. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2022.3212859

Neira, E., Clares-Gavilán, J., & Sánchez-Navarro, J. (2021). New audience dimensions in streaming platforms: The second life of Money heist on Netflix as a case study. *El Profesional de La Información*, e300113. https://doi.org/10.3145/epi.2021.ene.13

Netflix in Europe: Four Markets, Four Platforms? A Comparative Analysis of Audio-Visual Offerings and Investment Strategies in Four EU States—Catalina Iordache, 2022. (2024, September 18). https://journals.sagepub.com/doi/abs/10.1177/15274764211014580

Netflix: The Disruptor Faces Headwinds - The Challenges of Penetrating the Indian Market - ProQuest. (n.d.). Retrieved September 20, 2024, from https://www.proquest.com/openview/2db3a1fd2e98d844e74f0f7bdea831dc/1?pq-origsite=gscholar&cbl=2026366&diss=y

Panda, B., Kishore, Dr. K. N., & Baid, M. (2023). Over-The-Top (OTT) Platforms in the Indian Entertainment Industry: A Comprehensive Study of Digital Streaming Services. *International Journal of Research Publication and Reviews*, *4*(12), 2418–2422. https://doi.org/10.55248/gengpi.4.1223.123450

Park, J. H. (2022). *Netflix and Platform Imperialism: How Netflix Alters the Ecology of the Korean TV Drama Industry*. https://www.semanticscholar.org/paper/Netflix-and-Platform-Imperialism%3A-How-Netflix-the-Park/a44e7485e265f6ce518a1358beaee93436fcee2a

Park, S., & Lee, J. (2022). Effects of OTT Platform Original Content on Platform Brand—Focused on Netflix Original Contents -. *JOURNAL OF THE KOREA CONTENTS ASSOCIATION*, *22*(11), 548–560. https://doi.org/10.5392/JKCA.2022.22.11.548

Penmatcha, A. (2022). Netflix the World's Most Entertaining Platform—Behind the Scenes. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4298925

Performance Comparison K-Nearest Neighbor, Naive Bayes, and Decision Tree Algorithms for Netflix Rating Classification (Vol. 1, Issue 1, pp. 16–22). (2024, January 10). [Video recording]. https://doi.org/10.57152/ijatis.v1i1.1104

Phillo, C., & Ruchimat, T. (2022). *The Authority of Indonesian Broadcasting Commission in Selecting Content That is Suspected to Contain the Pornographic Elements in Netflix Streaming Platform in Indonesia:* 3rd Tarumanagara International Conference on the Applications of Social Sciences and Humanities (TICASH 2021). https://doi.org/10.2991/assehr.k.220404.046

Pluta, M., & Siuda, P. (2022). Cancer entertainment education and Netflix – an exploratory study. *Educational Media International*, *59*(1), 80–93. https://doi.org/10.1080/09523987.2022.2054115

Putri, A. S., & Kleden, K. L. (2022). PENYIARAN FILM TANPA SENSOR DI PLATFORM NETFLIX. *Yudishtira Journal : Indonesian Journal of Finance and Strategy Inside*, *2*(1), 131–143. https://doi.org/10.53363/yud.v2i1.31

Sánchez-Mompeán, S. (2021). Netflix likes it dubbed: Taking on the challenge of dubbing into English. *Language & Communication*, *80*, 180–190. https://doi.org/10.1016/j.langcom.2021.07.001

Shim, D., Lee, C., & Oh, I. (2022). Analysis of OTT Users' Watching Behavior for Identifying a Profitable Niche: Latent Class Regression Approach. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(4), Article 4. https://doi.org/10.3390/jtaer17040079

Subscribers Forecasting of Netflix Based on Multiple Linear Models (Vol. 34, pp. 229–236). (2022, December 14). [Video recording]. https://doi.org/10.54691/bcpbm.v34i.3018

Susilo, D. & Harliantara. (2023). The Digital Promotion of Japanese and Korean Movie in OTT Platform by Netflix. *Indonesian Journal of Business Analytics*, *3*(5), 1979–1994. https://doi.org/10.55927/ijba.v3i5.6418

Türkmen, B. (2020). Utilising Digital Media as a Second Language (L2) Support: A Case Study on Netflix with Translation Applications. *Interdisciplinary Description of Complex Systems*, *18*(4), 459–470. https://doi.org/10.7906/indecs.18.4.6

Varela, D., & Kaun, A. (n.d.). *A User-focused Approach to the Netflix Recommendation Algorithm*.

Wang, Y., Xiang, Z., & Zhang, X. (2022). Research on the Current Situation and Marketing Strategies of Netflix Platform Marketing. *BCP Business & Management*, *33*, 50–56. https://doi.org/10.54691/bcpbm.v33i.2719

Wayne, M. L. (2022a). Netflix audience data, streaming industry discourse, and the emerging realities of 'popular' television. *Media, Culture & Society*, *44*(2), Article 2. https://doi.org/10.1177/01634437211022723

Wayne, M. L. (2022b). Netflix audience data, streaming industry discourse, and the emerging realities of 'popular' television. *Media, Culture & Society*, *44*(2), 193–209. https://doi.org/10.1177/01634437211022723

Wei, Y. (2024). Analysis of the Success of The Netflix Korean Drama the Glory from The Perspective of Cross-cultural Communication. *International Journal of Education and Humanities*, *13*(1), Article 1.

Yadav, D., & Jain, A. (2024). Factors Influencing The Success Of OTT Platforms: A Literature Review. *INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS*, *12*, a672–a679.

Yao, Y. (2023). An Investigation on the Streaming Industry: With the Case of Netflix. *SHS Web of Conferences*, *165*, 01001. https://doi.org/10.1051/shsconf/202316501001

Yilmaz, E. S., & Erdem, A. (2022). Dijital Platform Üyeliklerinin Devamlılığına Etki Eden Faktörler: Netflix Örneği. *İktisadi İdari ve Siyasal Araştırmalar Dergisi*, *7*(17), 47–67. https://doi.org/10.25204/iktisad.970186

중앙대학교 일반대학원 문화예술경영학과 석사과정, Auh, Y., & Limb, S.-J. (2022). A Comparative Study on the Compositional Factors of OTT Platform Dramas and TV Dramas: Focusing on Netflix Original Korean Dramas and Simultaneous Airing Korean TV Dramas. *Journal of Arts and Cultural Management*, *15*(3), 57–87. https://doi.org/10.15333/ACM.2022.12.30.57

Appendix 1: Schedule

	12 WEEKS	**4 WEEKS**	**4 WEEKS**	**4 WEEKS**
Title				
Introduction				
Problem Statement				
Research Objectives				
Literature Review				

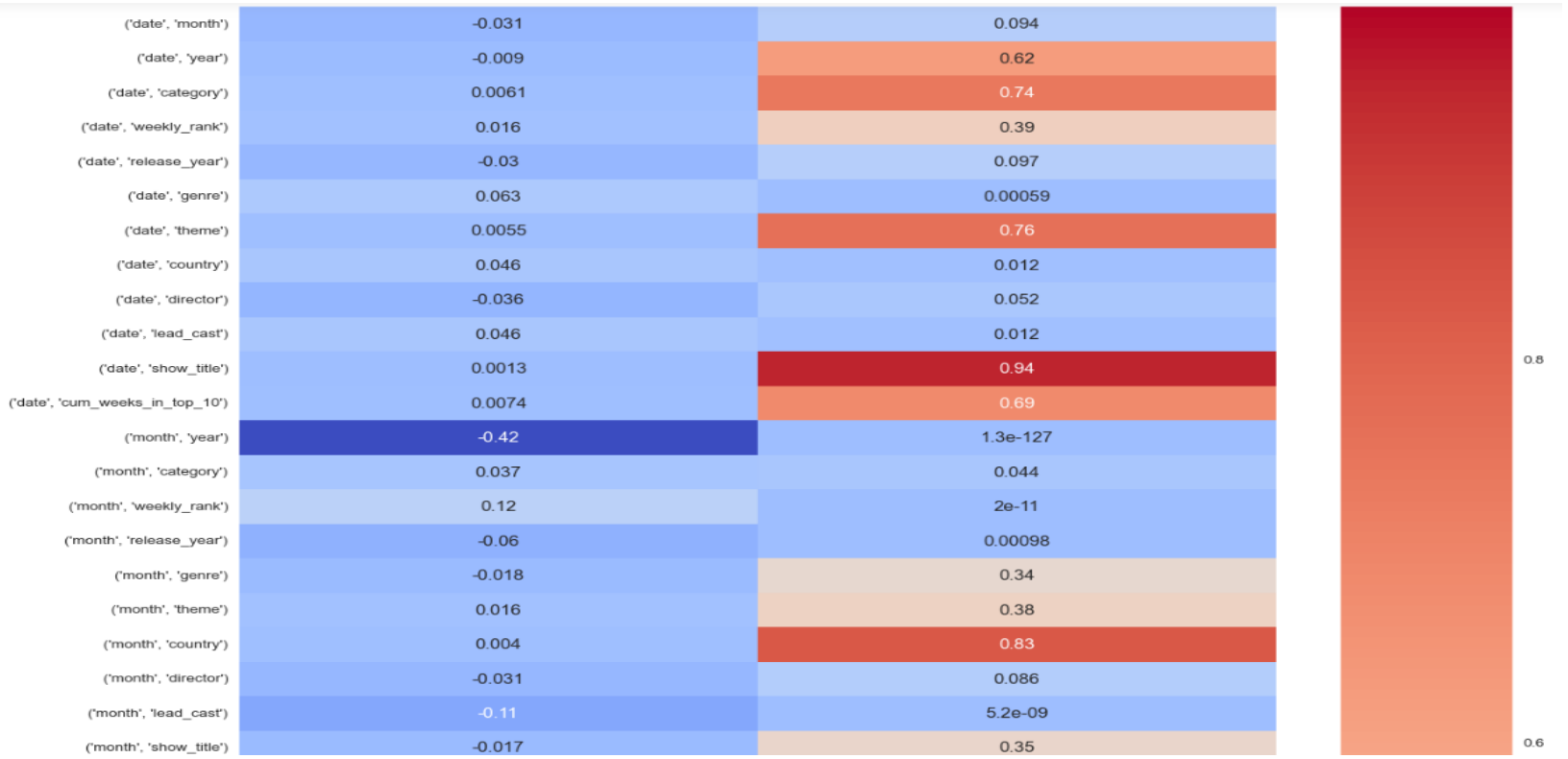
Research Methodology				
Analysis				
Discussion				
Conclusion and Recommendation				
Final Reviewing				

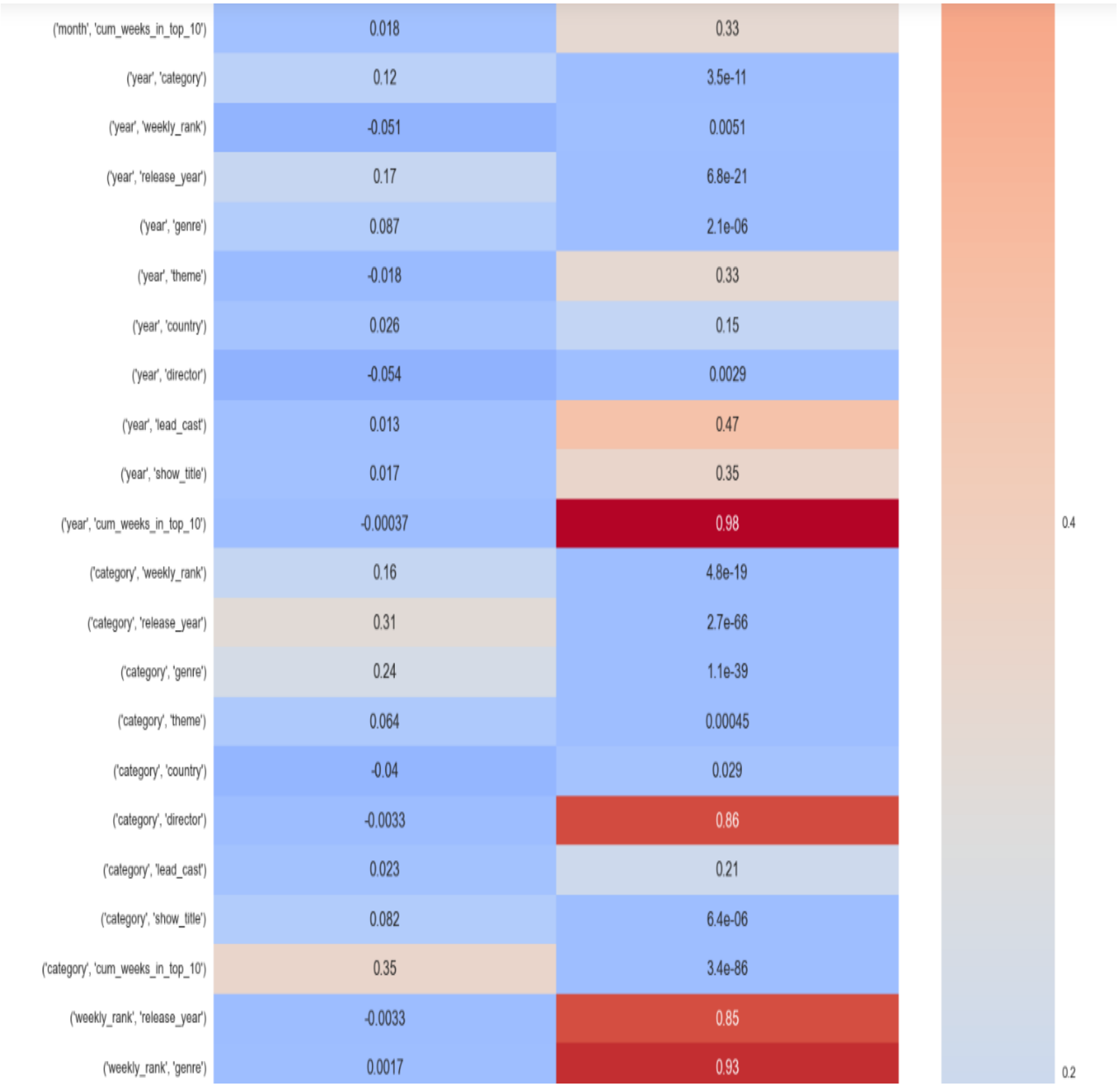
Appendix 2: Resources and Budget

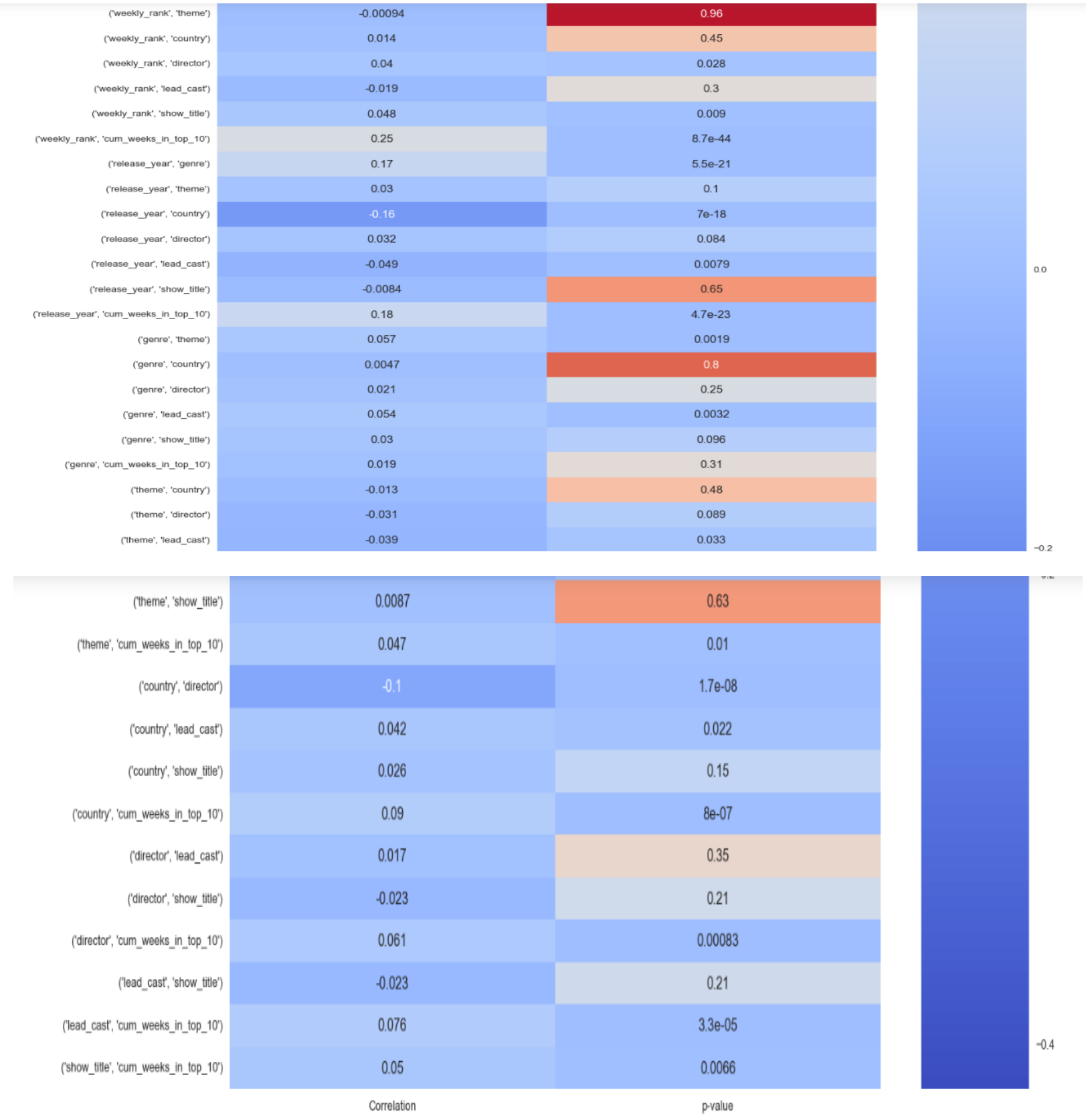
	Item	**Unit**	**Price (Kes)**	**Total (Kes)**
1	Internet Connection	12 Months		
	- Research		5,000	60,000
	- Downloads		19,000	228,000
	- Zoom calls			
	- Email communications		5,000	60,000

2	Final Dissertation Preparation - Printing - Binding	150 pages * 3 copies	20	9,000
4	Miscellaneous expenses	1		50,000
5	**Total**			**407,000**

Appendix 3: Correlation and P-value Chart




('month', 'cum_weeks_in_top_10') 0.018 0.33
('year', 'category') 0.12 3.5e-11
('year', 'weekly_rank') -0.051 0.0051
('year', 'release_year') 0.17 6.8e-21
('year', 'genre') 0.087 2.1e-06
('year', 'theme') -0.018 0.33
('year', 'country') 0.026 0.15
('year', 'director') -0.054 0.0029
('year', 'lead_cast') 0.013 0.47
('year', 'show_title') 0.017 0.35
('year', 'cum_weeks_in_top_10') -0.00037 0.98
('category', 'weekly_rank') 0.16 4.8e-19
('category', 'release_year') 0.31 2.7e-66
('category', 'genre') 0.24 1.1e-39
('category', 'theme') 0.064 0.00045
('category', 'country') -0.04 0.029
('category', 'director') -0.0033 0.86
('category', 'lead_cast') 0.023 0.21
('category', 'show_title') 0.082 6.4e-06
('category', 'cum_weeks_in_top_10') 0.35 3.4e-86
('weekly_rank', 'release_year') -0.0033 0.85
('weekly_rank', 'genre') 0.0017 0.93
0.4
0.2


('weekly_rank', 'theme') -0.00094 0.96
('weekly_rank', 'country') 0.014 0.45
('weekly_rank', 'director') 0.04 0.028
('weekly_rank', 'lead_cast') -0.019 0.3
('weekly_rank', 'show_title') 0.048 0.009
('weekly_rank', 'cum_weeks_in_top_10') 0.25 8.7e-44
('release_year', 'genre') 0.17 5.5e-21
('release_year', 'theme') 0.03 0.1
('release_year', 'country') -0.16 7e-18
('release_year', 'director') 0.032 0.084
('release_year', 'lead_cast') -0.049 0.0079
('release_year', 'show_title') -0.0084 0.65
('release_year', 'cum_weeks_in_top_10') 0.18 4.7e-23
('genre', 'theme') 0.057 0.0019
('genre', 'country') 0.0047 0.8
('genre', 'director') 0.021 0.25
('genre', 'lead_cast') 0.054 0.0032
('genre', 'show_title') 0.03 0.096
('genre', 'cum_weeks_in_top_10') 0.019 0.31
('theme', 'country') -0.013 0.48
('theme', 'director') -0.031 0.089
('theme', 'lead_cast') -0.039 0.033
0.0
−0.2
('theme', 'show_title') 0.0087 0.63
('theme', 'cum_weeks_in_top_10') 0.047 0.01
('country', 'director') -0.1 1.7e-08
('country', 'lead_cast') 0.042 0.022
('country', 'show_title') 0.026 0.15
('country', 'cum_weeks_in_top_10') 0.09 8e-07
('director', 'lead_cast') 0.017 0.35
('director', 'show_title') -0.023 0.21
('director', 'cum_weeks_in_top_10') 0.061 0.00083
('lead_cast', 'show_title') -0.023 0.21
('lead_cast', 'cum_weeks_in_top_10') 0.076 3.3e-05
('show_title', 'cum_weeks_in_top_10') 0.05 0.0066
−0.4
Correlation
p-value