Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Data Science and Marketing Analytics

Winning against the odds: A socio-economic analysis of Olympic success.

Tjedde Peters (511502)

Supervisor:	Professor van de Velden
Second assessor:	Professor Donkers
Date final version:	31st October 2023

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

This thesis examines the relationship between a country's Olympic medal count and its socio-economic factors, addressing four core research questions. Starting with an exploration of influential socio-economic factors, the measurement of the Olympic medals, and methodologies utilized in existing literature. These influential socio-economic factors, including population, GDP/GDP Per Capita, hosting and previously hosting the Olympics, and political regime, are integrated into the analysis. The Olympic medal count is typically assessed based on solely the number of Olympic medals earned by a country and the literature employs various machine learning models, such as Ordinary Least Squares, Tobit, Poisson, and Negative Binomial regression. The second research question focuses on identifying the most effective machine learning model for analyzing the influence of socio-economic factors on Olympic medal counts. In addition to machine learning models from previous literature, this study introduces Random Forest, Gradient Boosting, and Extreme Gradient Boosting. Novel socio-economic factors are considered, including income inequality, healthcare expenditures, technological development, and food supply surplus. The Poisson regression stands out as the preferred machine learning model for its ability to explain and quantify the relationship, primarily due to its interpretability. Conversely, XG-Boosting excels in terms of predictive accuracy, forecasting the number of Olympic medals won with an average deviation of 4.59 in comparison to the actual Olympic medal count. The third research question employs the most explanatory Poisson model to quantify the relationship and identify significant socio-economic factors. The study reveals that population, GDP per capita, hosting the Olympics, autocratic regimes, and healthcare expenditures have positive influences, while income inequality and food supply negatively influence the Olympic medal count. The fourth research question delves into exploring variations in the influence of these socio-economic factors across different contexts. Population has a more significant positive influence on female athletes, while income inequality has a more significant negative influence on them compared to their male counterparts. Population exerts a more significant positive influence on the Summer Olympics than the Winter Olympics, while GDP per capita and autocratic regimes exhibit a stronger positive influence in the Winter Olympics than in the Summer Olympics. Furthermore, the influence of previous hosting and healthcare expenditures was absent in earlier years but has become significant in more recent years.

-

Contents

Chapter 1

Introduction

The Olympic Games, which originated in Athens, Greece, in 1896 (IOC, n.d.), represent one of the most prestigious sporting events worldwide. They bring together diverse cultures and have the power to influence diplomatic relations between nations. For instance, during the Cold War, the Soviet Union and the United States used sports events as a means of expressing their mutual aversion. While sports connections may not carry the same weight as economic or legal relationships, they serve as effective political tools (Kanin, 2019). In 1931, Berlin was controversially chosen to host the 1936 Olympic Games. The National Socialist regime in Germany decided to proceed with the Games, using them to enhance their national image and present themselves as a peaceful and virtuous nation to the world (Mackenzie, 2003). This underscores the inherent connection between the Olympic Games and politics, highlighting their far-reaching impact beyond the scope of the playing field.

The Olympic Games necessitate significant investment in infrastructure and sports facilities by the host city and country. Scandizzo and Pierleoni (2018) categorize the impacts into economic, physical, socio-cultural, psychological, and political aspects. Positive effects include job creation, tourism growth, improved infrastructure, heightened sports interest, and community and national pride. Conversely, negative aspects are underestimated costs, increased taxes and public debt, environmental degradation, social displacement, and post-Olympics underutilization of facilities.

In addition to its impact on the residents of the host nation, the Olympic Games also exert influence on a global audience. For example, the 2020 Olympic Games held in Tokyo garnered a total of 3.05 billion individual viewers via television and digital media platforms, while internet platforms generated approximately 28.0 billion views (IOC, 2021). These figures highlight the immense reach and popularity of the Games, resonating with audiences across the world. Furthermore, the substantial financial investments associated with organizing the 2020 Olympic Games amounted to a total expenditure of 13.0 billion U.S. dollars (IOC, 2020). Such extensive international participation and the significant flow of financial resources within the Olympic process highlight its societal value.

The Dutch government has recently increased its investments in Olympic sports programs. One of the primary motivations behind this increased investment is to enhance national pride, foster cohesion, and increase international prestige. Olympic success leads to a temporary surge in national pride. However, the extent of this effect is somewhat limited because national pride is a relatively stable sentiment influenced by numerous factors (Elling, Van Hilvoorde & Van Den Dool, 2014). Haut, Prohl and Emrich (2016) found similar attitudes in the German population, where adherence to sports values was deemed more important than success, yet the significance of Olympic medals couldn't be dismissed, particularly among younger and less-educated individuals.

Many Western countries base their elite sports development system on the idea that Olympic success contributes, to increased sports participation. The higher demand for sports involvement results in a healthier nation, which provides a larger pool of potential champions for major sporting events. This phenomenon, referred to as the "virtuous cycle of sports," is of interest to the government due to its potential to enhance sports participation and improve the well-being of the population (Grix & Carmichael, 2012).

Olympic success is a valuable asset from a marketing perspective because it enhances collective national pride. Marketing experts could utilize this emotional connection to create favorable brand associations related to succesfull athletes or the Olympics as a whole, enhancing customer engagement. Additionally, the extensive media coverage of the Olympic games presents opportunities for marketing experts to support successful Olympic athletes. This increased visibility provides businesses with chances to elevate their brand exposure (Davis, 2012). Considering the substantial financial resources and attention dedicated to the Olympic games, these marketing opportunities are indeed of significant value.

The governments and other external sponsors face the subsequent question of how they can specifically enhance Olympic performance and what other potential underlying causes of the Olympic success could be. They must consider how to allocate their budgets effectively and identify the opportunities and challenges that arise in this pursuit. These considerations have shaped the primary research goal of this study.

Studying the influence of socio-economic factors on the Olympic medal count of a country.

The construction of the models that determine these influences relies on analyzing historical data of previous performances and exploring the specific circumstances of the countries involved. These two aspects are crucial because historical data allows us to understand the trends and patterns of each country. By considering factors that could potentially affect the rate of success, we can uncover relationships between the number of medals and country-specific variables. To better understand the primary research objective, a series of research questions is formulated.

Olympic achievements and their associated factors have been extensively examined in the literature. What socio-economic factors explain them and how is this relationship captured? This question has led to the exploration of the first research question:

1. Which socio-economic factors are consistently associated with the Olympic medal count, what models are commonly employed to investigate this relationship, and how is the Olympic medal count measured in the existing literature?

The existing literature has explored how socio-economic factors relate to the number of Olympic medals using various models. Each study justifies its choice of a particular model, although there may be one model that performs better than the rest. Additionally, there is a question about whether other socio-economic factors, not previously considered in the literature, might provide valuable insights into the connection between the Olympic medal count and socio-economic factors. Therefore, the second research question is considered:

2. What is the most effective model for analyzing the influence of socio-economic factors on the Olympic medal count, and to what extent do these models differ in statistical significance and predictive accuracy, while also exploring the added value of previously unconsidered socio-economic factors?

After the selection of the most suitable model and determining the composition of socio-economic variables, the next step involves interpreting the model and quantifying the influence of socio-economic factors on Olympic medal counts. This leads to the formulation of the third research question:

3. How can the influence of the socio-economic factors on the Olympic medal count be interpreted, and to which extent do the predictors exert influence?

It is important to acknowledge that the influence of these socio-economic factors may vary depending on the context of the Olympic Games. Hence, the fourth and last research question arises:

4. Do the influences of the socio-economic factors vary across the different genders, seasons and years in relation to the Olympic medal count, and, if so, how?

This thesis holds significance for both the academic and practical domains. From an academic standpoint, this thesis holds scientific value as it provides answers to research questions that contribute to the existing body of knowledge concerning the factors influencing medal-winning outcomes. This paper aims to contribute to the existing literature by examining the influence of a novel combination of independent variables alongside established ones and comparing multiple machine learning models.

In addition to its academic relevance, this research also offers practical implications for decision-makers. Policymakers within national Olympic committees can gain insights into the underlying factors that influence Olympic success. These factors may vary in their influences across different contexts, and understanding these variations can assist national Olympic committees in aligning their investment and training programs effectively. Moreover, this knowledge of the causes of Olympic success can be of interest to governments that invest in Olympic sports. Such investments contribute to the aforementioned ”virtuous cycle of sports,” promoting societal well-being and enhancing athletic performances (Grix & Carmichael, 2012).

Lastly, commercial stakeholders and bookmakers may find this research of interest. Stakeholders can utilize the conditions of a given country to make informed investment decisions and choose which athletes to sponsor. Bookmakers can employ the model and its predictions to adjust their odds in a manner that optimizes their profitability.

The paper is organized as follows. Section 2 provides a comprehensive literature review that examines the current understanding of the association between socio-economic factors within a country and its Olympic medal count. This section presents the existing knowledge on the topic and summarizes the methodological approaches used to investigate these relationships. By reviewing the existing literature, the paper identifies potential gaps in knowledge and highlights opportunities for employing novel methodologies.

In Section 3, an examination of the data used in this study is presented. This section covers the data collection process, including measurement details, followed by an explanation of the data cleaning and merging procedures. Furthermore, it includes a discussion of descriptive statistics related to the data, offering an analysis of the data set.

In Section 4, the paper outlines the chosen models, metrics, and the research approach employed to derive relevant conclusions and answers to the research questions. This section provides insights into the rationale behind selecting specific models and metrics and explains how they are applied to the data set.

Section 5 focuses on presenting the outcomes of the study. It highlights the best-performing models, analyzes the significance of socio-economic factors, and quantifies their influences based on the results obtained. This section aims to provide a clear understanding of the findings and their interpretation within the context of the research questions.

Finally, Section 6 serves as the conclusion and discussion section, summarizing the main findings of the thesis. It addresses the recommendations and insights derived from the results, emphasizing their significance and potential applications. Furthermore, this section acknowledges the limitations of the thesis and highlights areas for future research within this field, suggesting potential approaches to further enhance the understanding of the relationship between socio-economic factors and the Olympic medal count.

Chapter 2

Literature review

2.1 Socio-economic factors

Several studies have been conducted to investigate the influence of socio-economic factors on the Olympic medal count of countries. Within the existing literature, numerous independent variables have been examined, with certain socio-economic factors recurring frequently and demonstrating significant effects. This thesis aims to outline these specific socio-economic factors and to identify areas that warrant further investigation.

2.1.1 Population

The influence of a country's population on its Olympic performance has emerged as a repeated theme in the literature. It is commonly argued that the number of inhabitants should play a significant role in determining the extent of a country's success in the Olympics. The rationale behind this argument is that larger countries possess a larger pool of athletes and talent to choose from, thereby increasing their chances of securing victories (Johnson & Ali, 2004).

An additional perspective in the literature suggests that a larger population size may decrease the likelihood of qualifying for the Olympic Games, attributed to the limited number of available spots for participation. As a result, the selection process becomes highly competitive in countries with larger populations compared to smaller nations. This heightened competition can further enhance the probability of Olympic success (Emrich, Klein, Pitsch, Pierdzioch et al., 2012).

According to Lui and Suen (2008), if athletes worthy of winning medals were distributed randomly across the world, the proportion of medals obtained by a country in the Olympics would be directly proportional to its share of the global population among the participating countries. However, this assumption is limited in scope since it overlooks other significant factors influencing Olympic success. For instance, solely relying on the population as a predictor would lead to the expectation that countries like China, India, Bangladesh, and Indonesia, together representing 43% of the world's population, would have collected more than 6% of total Olympic medals they won in 1996 (Bernard & Busse, 2004).

2.1.2 GDP per capita

A country's population is frequently combined with its GDP per capita, as these two factors emerge as prominent determinants of a country's success in the Olympics (Celik & Gius, 2014). Building upon the example provided by Bernard and Busse (2004), which initially considered population as the sole predictor, the inclusion of GDP per capita further enriches the analysis. The availability of resources for each inhabitant and the government's support significantly impact a country's capacity to invest in training programs, purchase equipment, and ultimately participate in the Olympic Games. Moreover, individuals in wealthier countries have shorter daily working hours to sustain their livelihoods compared to those in poorer countries. Their working time is also expected to decrease as they age, affording them more opportunities to engage in sports activities Emrich et al. (2012).

Historically, wealthier nations have displayed higher participation rates in the Olympics compared to developing countries. Nevertheless, advancements in global travel have led to reduced transportation costs, and improved accessibility to healthcare has contributed to enhanced participation from economically disadvantaged countries (Kuper & Sterken, 2001). These changes have leveled the playing field, allowing developing nations to increase their representation in the Olympic Games.

Johnson and Ali (2000) conducted a study investigating the influence of population and GDP per capita on medal success during the Summer Olympics in the aftermath of the World War. Their findings revealed a noteworthy influence of both population and GDP per capita on medal achievements. Similarly, Andreff (2001) conducted research focusing on the 1996 Atlanta and 2000 Sydney Summer Olympics. In this study, both population and GDP per capita were found to be highly significant, with GDP per capita exhibiting an even greater influence on medal success. These two socio-economic factors emerged as fundamental elements in predicting the Olympic medal count. However, it is essential to acknowledge that other factors may also play a role and should be considered in forecasting the Olympic medal count.

2.1.3 Host country

The act of hosting the Olympic Games can yield favorable outcomes for the Olympic medal count of the participating nation. This potential advantage can be attributed to various factors, including the influence of the home crowd, familiarity with the sporting context, reduced travel fatigue, rule-related factors, refereeing decisions in favor of the home team, and the instinctive sense of territoriality (Legaz-Arrese, Moliner-Urdiales & Munguía-Izquierdo, 2013). During the 2000 Olympic Games held in Sydney, the host nation achieved a remarkable performance, securing nearly 42 medals more than other participating countries exhibiting comparable characteristics (Hoffmann, Ging & Ramasamy, 2004).

Several studies have employed dummy variables to indicate whether a nation hosts the Olympic Games in a given year. Bernard and Busse (2004) found that countries achieved more than 1.8 percent of medals beyond what would be predicted by their GDP alone. In a more comprehensive analysis, Bian et al. (2005) investigated the impact of hosting the Olympics while considering a broader range of socio-economic factors. The findings revealed that being the host nation positively influenced the Olympic performance of athletes from that country.

Hoffmann et al. (2004) reported that hosting the Olympic Games represents a country's affinity with sports, capturing the cultural aspect related to sports participation. The use of a dummy variable for hosting significantly and positively influenced the number of medals won by the host country. Furthermore, the inclusion of lagged dummy variables for previous host countries also yielded significant results, indicating that the benefits of hosting the Olympic Games extend to subsequent events.

Balmer, Nevill and Williams (2003) research findings reveal that the extent of home advantage varies depending on the type of sport. Their study highlighted a significant home advantage in sports that relied on subjective judging by officials or judges. In contrast, sports governed by specific rules with objective judgments showed little to no home advantage. While the advantage of hosting the Olympic Games has been established through various studies, its influence is context-specific and influenced by various factors.

2.1.4 Politics

Politics and the Olympic Games have historically been closely connected and are inseparable(Kanin, 2019). The presence of significant historical incidents such as the 'Nazi Olympics' in 1936, the ideological rivalry between the U.S. and the Soviet Union in 1952, and the Suez invasion in 1956 are not mere coincidences. These occurrences illustrate the evident and enduring link between politics and the Olympic Games (Donald, 1972).

The presence of this relationship is clear but do the political landscape of a country influence its performance? Donald (1972) states that a 'successful nation-state' in the Olympic Games should be: "stable and homogeneous in population, literate, modern and western, with little institutionalized domestic political competition, economically prosperous, characterized by a strong central government staffed by the elite, and probably a member of the Communist Bloc".

The aforementioned assertion was subsequently validated by Grimes, Kelly and Rubin (1974) in their study, where they conducted a regression analysis of the number of medals won to the GDP per capita and population. Notably, the communist countries emerged as outliers in this analysis, as their actual number of medals surpassed the predicted number based on their economic development and population.

In more recent research conducted by Bernard and Busse (2004), it was revealed that the Olympic performance of the Soviet Union and Eastern Bloc countries exceeded their predicted medal share by more than 3 percentage points, considering their GDP and previous performances. Additionally, Johnson and Ali (2004) also observed an over performance of communist and centralized single-party governments in the Olympic Games.

In a communist country with a centralized government, there is a greater emphasis on specialization in sports, and resources are more readily allocated to athletes and training programs compared to societies with open market systems. This prioritization of sports is driven by the significance of Olympic performance in enhancing the national prestige of these communist countries, which holds exceptional importance for them (Bian et al., 2005).

2.2 Gender, season, and year disparities

The importance of various socio-economic factors on the Olympic success has been acknowledged; however, examining whether these influences can be universally applied across all contexts is crucial. Notably, certain influences exhibit variations based on specific circumstances, such as the gender of participants, the season of the Olympic Games, and the particular year in question.

Gender

Many studies exploring the determinants of Olympic success often overlook the distinction between male and female performances. In this regard, the research conducted by Leeds and Leeds (2012) stands out as they observe noteworthy patterns. Specifically, they find that female athletes from Arab countries tend to underperform compared to their counterparts from other countries, while male athletes from Communist countries tend to outperform their peers, whereas female athletes from Communist countries do not exhibit the same advantage. The study seeks to assess the influence of established explanatory variables and incorporates additional gender-related variables into the analysis, namely, fertility rate and the date of women obtaining voting rights.

Rewilak (2021) also researched the determinants of Olympic success, focusing on the separation of data into male and female samples to investigate potential influences specific to each gender. Hosting had a statistically significant positive influence on the performance of both male and female athletes to the same extent. Similarly, the population size also had a statistically significant positive influence on Olympic success for both genders. However, an increase in population size had a twofold influence on female athletes compared to male athletes.

Contrary to the findings of Leeds and Leeds (2012), certain variables such as GDP per capita, political dummy variables, and gender inequality were deemed statistically insignificant in their influence on Olympic success in the study by Rewilak (2021). This discrepancy suggests that the influence of these independent variables might vary depending on the specific research context or sample characteristics. It is important to note that the research in this field may not be extensive enough to draw definitive conclusions, and for a more comprehensive understanding, additional socio-economic factors should be included to assess the differences in their influences on male and female athletes' performance.

Season

The nature of the Summer and Winter Olympics differs significantly from other sporting disciplines in terms of their organization. The Summer Olympics are characterized by a longer duration, a greater number of sporting and non-sporting events, and being scheduled during the holiday season. These factors contribute to a larger economic influence for the Summer Olympics compared to the Winter Olympics, leading to a greater prevalence of research on the Summer Olympics in the current literature (Wood & Meng, 2021).

In a comparative study conducted by Johnson and Ali (2004)), both editions of the Olympics were examined to analyze their respective impacts and the underlying socio-economic factors. The research aimed to investigate two main components: firstly, to explore the relationship between a country's ability to participate in the Olympics and various socio-economic factors, and secondly, to determine the relationship between the Olympic medal count of participating countries and socio-economic factors.

The findings revealed that nations with higher GDP are more likely to send a greater number of athletes to the Olympics, with this effect being more pronounced in the Winter Olympics than in the Summer Olympics. This effect may be because countries with warmer climates are less likely to participate in the Winter Olympics. Additionally, Africa, which has the warmest climate near the equator, is less developed and has limited Winter Olympics participation. Therefore, geography could influence both the GDP and participation rate making it a confounding variable. As a result, the impact of a country's GDP is less significant in the Winter Olympics compared to the Summer Olympics. The Summer Olympics involve more diverse countries, while the Winter Olympics are mainly attended by wealthier nations, reducing the importance of GDP. However, it is essential to note that a higher GDP is generally associated with more Olympic medals in both the Summer and Winter Olympics.

Larger populations also positively influence the number of athletes a country sends, with a stronger effect observed in the Summer Olympics. Interestingly, smaller nations tend to outperform larger nations more prominently in the Winter Olympics, and the population's influence on performance is also less pronounced in the Winter Olympics than in the Summer Olympics. One potential reason for this could be that smaller countries often specialize more to maximize their chances of winning Olympic medals.

Years

The composition of the Olympic Games has evolved significantly over time, particularly in the post-World War II era. The percentage of European athletes participating in the first Olympic Games after the war was as high as 83%, but this figure decreased to 68% during the 2012

Olympic Summer Olympics in London. Furthermore, there has been a noteworthy increase in the representation of women in the Olympics, accounting for nearly half of the athletes in contemporary games, compared to a mere 10% in the postwar games.

Moreover, the distribution of Olympic medals among countries has undergone substantial changes. In the 1980 Olympic Games, the top ten countries accounted for more than 80% of the total medals, but this proportion declined to around 55% for the London Olympics in 2012 (IOC, 2012). Such variations in the composition of the Olympic Games concerning participating countries, gender representation, and medal winners indicate the dynamic nature.

Noland and Stahler (2017) assert that this diversity in Olympic outcomes can be attributed to changes in the underlying correlates, which implies that the socio-economic factors influencing Olympic success have evolved. The significance of certain determinants such as welfare, population size, host advantage, and political structure has been recognized; however, smaller and economically disadvantaged countries have encountered fewer barriers to winning Olympic medals as time has progressed. Hence, the importance of GDP per capita and population size in determining Olympic success has diminished over time.

2.3 Potential socio-economic factors

The importance of socio-economic factors in shaping a country's Olympic performance has been emphasized and will be incorporated into the analysis to account for their influences. Nevertheless, this thesis recognizes the existence of additional potential factors that may also impact a country's Olympic medal count and have not been used in the existing literature. To ensure that no new socio-economic factors strongly correlated with a country's size or wealth are included, correlations will be tested, and socio-economic variables will be chosen based on correlation criteria. The new potential socio-economic factors considered for this study are:

Income inequality

The association between income inequality and a nation's economic condition has been the focus of numerous studies, revealing a negative correlation between higher income inequality and economic growth (Buttrick & Oishi, 2017). Likewise, extensive research has explored the link between income inequality and the overall health of a country, though thus far, evidence demonstrating income inequality as a threat to public health remains non-existent (Subramanian & Kawachi, 2004).

Veal (2016) states that a more equitable distribution of income within a country enhances the well-being of its residents. However, no prior investigations have delved into the potential relationship between income inequality and leisure time and sporting activities. The present study examines this association and finds that countries with lower income inequality tend to have more leisure time and greater participation in cultural and sports activities. Increased sports participation may lead to improved athletes' performance in the Olympic Games.

Healthcare expenditures

GDP per capita is commonly used as an indicator of a country's economic prosperity. However, while this metric provides a useful overall measure, it does not reveal how financial resources are allocated within the country. This paper argues that targeted government investments can offer valuable insights into a nation's specific priorities.

Jakovljevic et al. (2019) emphasize the importance of increased health spending and effective policies in improving public health outcomes and reducing disease prevalence. This finding underscores the significance of health-related investments for countries, particularly fast-growing economies like the BRIC (Brazil, Russia, India, China) governments. Allocating a higher share of the budget to health holds the potential to improve a country's performance in international sporting events like the Olympic Games.

Nutrition

Participating in sports is crucial for Olympic success, but achieving optimal performance relies on proper food intake. Balanced nutrition aids athletes in recovery, effective training, and injury prevention. Ensuring the right nutrients are consumed becomes especially vital to reach the highest levels of the Olympic Games (Maughan, Burke & Coyle, 2004). Nevertheless, the accessibility of nutritious foods for individuals varies across countries, potentially placing some athletes at a disadvantage in their pursuit of Olympic excellence.

Technology

Technology and innovation are playing an increasingly important role in the sports industry. Haake (2009) conducted a study analyzing four distinct disciplines featured in the Olympic Games, investigating their performance improvements throughout history. Although the extent of enhancement varied across disciplines, a significant improvement was observed in all sports, which can be attributed to innovations in equipment. The findings of this research hold applicability from the domain of amateur sports to that of elite athletics.

Also, various information technologies and wearable devices are now available to provide relevant feedback to athletes. It is believed that these technologies can enhance the performance and capabilities of both male and female athletes (Liebermann et al., 2002).

2.4 Employed research methods

Several studies have examined the influence of socio-economic factors, as mentioned in the preceding sections. The question then arises about the methods used and the contexts in which they were applied. To address this, Table 2.1 below provides a summary of these methods employed. In the response variable column, the aggregation of gold, silver, and bronze medals is indicated by (A), while the differentiation between various types of medals is marked as (D).

Table 2.1. Methods overview in existing literature

Date	Author(s)	Predictors	Response variable	Model
1972	Donald	Population GDP/GDP per capita Political structure	Weighed sum of medals	Fisher's exact test
1974	Grimes Kelly Rubin	Population GDP per capita	Medal integer count (A)	OLS
2000	Johnson Ali	Population GDP per capita Host country Neighboring country Political structure	Individual: Probability (D) Country: Medal count (A)	OLS
2001	Andreff	Population GDP per capita	Probability (A)	Logistic regression
2001	Kuper Sterken	Population GDP per capita Host country Political structure	Medal integer count (D)	OLS
2004	Bernard Busse	Population GDP per capita Host country Neighboring country Political structure Climate	Medal integer count (A)	Tobit
2004	Johnson Ali	Population GDP per capita Host country Neighboring country Political structure	Medal share count (A)	OLS
2005	Bian	Population GDP per capita Host country Political structure	Medal integer count (A)	OLS
2008	Lui Sen	Population GDP per capita Host country	Weighed sum of medals	Tobit Poisson
2012	Leeds Leeds	Population GDP per capita Host country Political structure Fertility rate Date woman voting rights	Medal integer count (A & D)	Negative Binomial Regression
2012	Emrich Klein Pitsch Pierdzjoch	Population GDP per capita	Medal integer count (A)	OLS
2014	Celik Gius	Population GDP per capita Lagged medals won	Medal integer count (A)	OLS
2017	Nohland Stahler	Population GDP per capita Host country / Post host country Political structure Education	Medal share count(A)	Tobit
2021	Rewilak	Population GDP per capita Host country Political structure	Medal share count (A)	Tobit

Donald (1972) investigates the link between Olympic performance and national socio-economic indicators. It categorizes countries' Olympic outcomes as high or low scores based on a weighted sum of medals, and national socio-economic indicators are similarly classified as high or low using thresholds. Fisher's exact test is employed to assess differences between these categories, examining whether high-scoring countries differ from low-scoring countries in terms of the socio-economic indicators, and vice versa. This approach, however, conservatively establishes the relationship, relying on thresholds that limit the representation of a precise relationship.

Ordinary Least Squares (OLS) regression is a widely employed technique in statistical analysis. Nonetheless, when the dependent variable exhibits many values of zero, conventional statistical methods can introduce downward biases into the estimates. In such cases, Tobit regression, as applied by Lui and Suen (2008), Noland and Stahler (2017), and Rewilak (2021), offers a comparable alternative to OLS regression. The Tobit regression method is designed to accommodate and address the limitations posed by the distribution of the dependent variable (McBee, 2010).

The Generalized Linear Model (GLM) transforms non-linear problems into linear ones. Andreff (2001) utilized logistic regression, a type of GLM, to predict binary outcomes, where the dependent variable is categorical, indicating the presence or absence of an outcome (Walsh, 1987)

Lui and Suen (2008) study uses Poisson regression, a common GLM for count outcomes, modeling the number of events within a fixed time frame. The dependent variable follows a Poisson distribution, determined by the average event rate, and is transformed into a natural logarithm. it is important to note that Poisson regression assumes equidispersion, where the variance equals the mean. If this assumption is violated, Coxe, West and Aiken (2009) suggests Negative Binomial Regression (NGB) as an alternative solution, which relaxes the equidispersion assumption.

The assessment of Olympic performance is influenced by the models utilized, as the choice of the dependent variable determines the appropriateness of the analytical approach. The study conducted by Donald (1972) was initiated by computing a weighted sum of medals, categorizing them into two groups based on this criterion. Similarly, Lui and Suen (2008) employed a similar approach to measure medal performance, with gold medals receiving the highest weight on a scale of 1 to 3. In a different context, Andreff (2001) assessed the probability of winning any Olympic medal, whereas Johnson and Ali (2000) investigated the likelihood of winning specific types of Olympic medals for individuals.

Bernard and Busse (2004), Noland and Stahler (2017) and Rewilak (2021) employed a ratio-based count, which considers the number of medals won to the total number of medals available. In contrast, Grimes et al. (1974), Kuper and Sterken (2001), Johnson and Ali (2004), Bian et al. (2005), Emrich et al. (2012), Leeds and Leeds (2012) and Celik and Gius (2014) adopted an integer count approach, focusing on the total number of medals secured by a particular country. The latter method is more commonly employed and often does not differentiate between gold, silver, or bronze medals, except for Kuper and Sterken (2001) and Leeds and Leeds (2012).

Chapter 3

Data

This research investigates the relationship between socio-economic factors and a country's Olympic medal count. The utilization of data is essential in this investigation as it facilitates the quantification of this relationship. This section outlines the sources of data, the reasons for considering this data representative, and the specifications associated with it. Finally, the process of cleaning and merging the separate data sets is explained.

3.1 Data collection

3.1.1 Olympic performance

The data set employed in this study referred to as '120 years of Olympic history: athletes and results,' has been sourced from Kaggle. This data set contains extensive information about both Summer and Winter Olympic games achievements, covering events from the first Olympics in 1896 to the 2016 edition. It was compiled by Rgriffin (2018), who collected the data by scraping information from www.sports-reference.com and combining it into one comprehensive data set. The data set consists of 271,116 rows and 15 columns, with each row representing an individual athlete taking part in an Olympic discipline, accompanied by their respective personal information contained within the columns. The columns of the data set are displayed in Table A.1.

3.1.2 Socio-economic factors

All the socio-economic data used in this study was obtained from www.ourworldindata.org, a trusted open-source website that compiles information from around the world. This website focuses on key global challenges, including poverty, disease, hunger, climate change, war, existential risks, and inequality, all of which are closely related to the socio-economic state of our planet. The website serves as a third-party platform by aggregating data from numerous official databases (e.g. World Bank and United Nations) maintained by trusted institutions. The socio-economic factors have been obtained individually as separate data sets, with each data set primarily focused on a specific socio-economic variable.

GDP per capita

Gross Domestic Product (GDP) per capita serves as a metric for assessing the average standard of living and financial resources available to individuals within a given country. To facilitate the cross-country comparisons, GDP per capita values are standardized by converting them into international dollars, a process based on Purchasing Power Parity (PPP) rates. These PPP rates enable a relative assessment of the cost of living across different countries, making GDP per capita data applicable and meaningful within a global context (OurWorldinData, 2021b).

Population

The population data set includes data points that record historical global population figures for past decades and provide projections for future decades (OurWorldinData, 2022b).

Political regime

The state of democracy in a country is a measure of how much political freedom and participation its citizens enjoy. In the data set the global political regimes are classified into four distinct categories, which serve as a categorical variable (OurWorldinData, 2022a).

Closed Autocracies: In these political systems, citizens do not have the right to elect their political leaders or participate in multi-party elections. The level of political freedom is severely restricted.

Electoral Autocracies: Within electoral autocracies, citizens do possess the formal right to participate in elections and vote for their leaders. However, these rights are often constrained, leading to elections that are less free and fair, and limitations on broader political freedoms.

Electoral Democracies: These political systems grant citizens the right to participate in legitimate multi-party elections. While citizens have a significant degree of political agency, the level of political rights and freedoms may still vary within this category.

Liberal Democracies: Representing the most advanced form of democracy, liberal democracies not only provide individuals with extensive political rights but also ensure equality under the law. Furthermore, they establish legal mechanisms to constrain the authority of elected leaders.

Income inequality

The GDP per capita serves as an indicator of a country's average wealth; however, it does not provide insights into how financial resources are distributed within the society. In contrast, the Gini index, a coefficient that ranges from 0 to 1, offers a measure of income inequality, with a higher coefficient signifying a greater degree of income inequality (OurWorldinData, 2021a).

Healthcare expenditures

The state of a country's healthcare system reflects its capacity to provide for its citizens' wellbeing. The allocation of funds to healthcare offers an insight into its prioritization. However, absolute public healthcare investments are influenced by a country's financial resources. Therefore, healthcare expenditures are measured as a percentage of the GDP. Analyzing the connection between this socio-economic indicator and the Olympic medal count may yield insights into the importance of emphasizing such investments (OurWorldinData, 2019).

Technology

Technology and the internet have become deeply integrated into global society, with widespread usage being the norm. Nevertheless, the percentage of the population that has accessed the internet within the last three months remains a significant indicator of a country's technological development, reflecting the extent of active users and overall technological advancement (OurWorldinData, 2021d).

Nutrition

To assess global access to nutritious food, one data set compared the minimum costs of a nutritious diet to average food expenditures. However, this data is only available from 2017 onwards, which does not align with the 2016 cutoff of the Olympic performance data set. As an alternative, data on daily caloric food supplies, including macronutrient composition like plant protein, animal protein, fat, and carbohydrates, is used (OurWorldinData, 2020). Additionally, a separate data set outlines daily minimum caloric requirements for each (OurWorldinData, 2021c).

The socio-economic data sets obtained for this thesis are summarized in Table A.3.

3.2 Data cleaning

3.2.1 Olympic data

Firstly, the Olympic data set is examined and only the relevant variables that align with the thesis objectives are retained. The variables ID, Name, Age, and Height are personal but this study is focused on the relationship on a country-specific level thus these are removed from the data set. Additionally, the redundant variables NOC and Games which duplicate information found in Team, Year and Season are excluded. Also, the variable Team is renamed to Country for better clarity in describing the values of this variable.

The Olympic data set has been limited to the years between 2000 and 2016 because data for variables related to nutrition and healthcare expenditures is only available from the year 2000 onwards. The updated data set now includes events from the 2000 Olympic Games in Sydney through the 2016 Olympic Games in Rio de Janeiro, encompassing a total of five Summer Olympics and four Winter Olympics.

The variable of interest, *Medal*, denotes whether an individual athlete has achieved a gold, silver, bronze, or no medal. It is standardized such that all three types of medals are represented as 1, while the absence of a medal is coded as 0. Furthermore, one year is subtracted from the *Year* variable for each observation. This adjustment facilitates the merging of Olympic performance data with socio-economic predictors from the year immediately preceding the Olympic Games. This transformation is implemented to account for the fact that when merging data for identical years, it assumes a retroactive impact during the 5- and 9-month intervals in the aftermath of the respective Summer and Winter Olympics, which is not relevant. Additionally, any unnecessary numerical suffixes in the *Country* variable are eliminated.

Finally, the Olympic medal count is aggregated and summed for each year and country, resulting in the creation of the *Total Medals* variable. This variable represents the Olympic medal count for every country in a specific year and edition of the Olympic Games. Olympic team winners are considered as a single medal winner for the Olympic rankings, contrasting with the initial data set where each team member is categorized as a medal winner and, thus, individual athletes are aggregated with their respective teams. To conclude, the *Maximum Medals* variable is created to show the total of all medals won for the involved edition of the Olympic Games.

The updated Olympic data set includes a total of 1377 observations and is composed of 6 variables: *Year, Country, Season, City, Total Medals*, and *Maximum Medals.*

3.2.2 Socio-economic data

The socio-economic data sets are loaded individually and undergo minor adjustments. Variable names for all the socio-economic indicators are slightly modified to enhance clarity. The *Political Regime* variable is assigned numerical codes ranging from 0 to 3, with each number representing a distinct political context. These numerical codes are subsequently transformed into corresponding strings that correspond to their respective numbers. However, these string representations are not suitable for further analysis. Consequently, four dummy variables are created, with each one denoting the presence (1) or absence (0) of a specific political regime for the country in a given year.

The daily caloric food supply, referred to as *Total Supply*, for each country should be determined by aggregating the caloric contributions of individual macronutrients. Subsequently, the separate data set containing the minimum daily caloric requirements should be integrated with the caloric supply data. Following this integration, a new variable termed *Food Supply Surplus* can be computed by dividing the daily caloric supply by the daily minimum requirement. Lastly, the absolute caloric content attributable to each macronutrient is transformed into a ratio by dividing it by the *Total Supply.* In summary, the concept of *Food Supply Surplus* explains the capacity of each country to meet its needs. Additionally, variables such as *Plant Protein, Animal Protein, Fat*, and *Carbohydrates* contextualize these caloric values by showing their proportions relative to the *Total Supply.*

3.2.3 Merging, handling NA values and multicollinearity

The initial step involved merging the previously cleaned Olympics data set with the separate sets of socio-economic variables. After this merging process, an adjustment was made to the *Year* variable by incrementing it by 1, thereby restoring its original state. However, a noteworthy point to mention is the absence of data for the *Nutrition* and *Healthcare Expenditures* variables in the year 1999, which directly preceded the 2000 Sydney Olympics. To address this data gap, the values for these variables were substituted with data from the year 2000. Additionally, a novel independent variable was introduced to indicate whether a country hosted the Olympic Games. Consequently, a lagged host variable was incorporated to denote hosting one of the two preceding editions. The data set contains information from 101 countries across 9 Olympic Games, giving a total of 1377 rows and 23 columns.

The merged data set is not yet suitable for analysis, primarily due to the presence of 4156 missing values that require attention. To solve this issue, missing values in continuous socio-economic variables will be replaced with the mean value of that variable across all available years of data for the respective country. This imputation process reduces the number of missing values to 3653. Following imputation, rows containing these missing values are removed from the data set. The data set has been reduced from its initial 1377 rows to 759 rows. This reduction also resulted in a decrease in the total number of medals from 5458 to 4885. However, this decrease in medal count is relatively minor compared to the removal of rows, indicating that the removed rows mainly represent small countries without socio-economic data that do not win many medals. It is important to note that the sum of all medal winners is accounted for in the initial phase before the cleaning, resulting in the creation of the *Maximum Medals* variable. When this variable is incorporated into the analysis, it considers the potential total number of Olympic medals that could be won, thus offering insight into athlete achievements compared to the other Olympic Games editions.

Ultimately, it is critical to evaluate multicollinearity, as it reveals strong correlations among independent variables. This can result in enlarged standard errors for coefficients in regression models, introducing uncertainty regarding the actual influence of each independent variable on the dependent variable. The correlation matrix among the independent variables reveals their relationships. Ratner (2009) defines a correlation coefficient exceeding an absolute value of 0.7 as a strong relationship. This criterion is consistently applied to determine whether to keep or exclude independent variables. The correlation matrix is presented in Table A.2 where *Animal Protein, Plant Protein, Fat*, and *Carbohydrates*, *Share Internet Users* and *Liberal Democracy* also exceed this threshold concerning the *GDP Per Capita.* If all these independent variables are retained, they may collectively capture some of the influence of *GDP Per Capita* obscuring its true influences. Therefore only *GDP Per Capita* is kept and all other strongly correlated independent variables are eliminated. The ultimate data set, aligned with the research goals of this thesis, comprises 759 rows and 19 columns.

3.3 Data descriptives

The data sets have been sourced, merged, and cleaned to make them ready for analysis. Before executing the analyses, an examination of the statistical properties of the data set's variables has been conducted. This examination includes measures of central tendency (mean and median), measures of dispersion (range and variance), and measures of shape (quantiles and skewness). These descriptive statistics of the numerical variables are summarized in Table 3.1.

Table 3.1. Descriptive statistics

Variable	Mean	Median	Variance (SD)	Range	1st Quantile	3rd Quantile	Skewness
Total Medals	6.44	1.00	188.26 (13.72)	0-110.00	0.00	6.00	3.83
Maximum Medals	685.00	879.00	99691 (316.73)	207.00-947.00	265.00	917.00	-0.72
GDP Per Capita	21359	14474	366,849,551 (19153)	698-120,648	6078	34788.8	1.39
Population	61,018,574	10,464,537	3.75×10^{16} (1.94×10^{8})	78,848-1,393,715,456	4,521,640	38,601,774	5.68
Gini Coefficient	0.37	0.35	0.0065 (0.08)	0.24-0.65	0.31	0.41	1.07
Food Supply Surplus	1.580	1.589	0.043 (0.21)	1.062-2.011	1.425	1.745	-0.13
Healthcare Expenditures	6.67	6.56	6.14 (2.48)	1.85-20.41	4.80	8.33	0.69

Table 3.1 highlights notable aspects of the dependent variable *Total Medals*. A key observation is the relatively low variance when considering the range of values, suggesting that a significant portion of the data points cluster around the low-valued mean, having high-valued outliers. This notion is further substantiated by the substantial skewness evident in the last column, which measures the distribution's asymmetry. Specifically, when the coefficient is positive, it signifies right-skewed data, indicating that the tail and the minority of the distribution are located on the right-hand side, the high-valued end of the distribution. Conversely, a negative coefficient suggests left-skewed data, where this relationship is reversed. Data is considered highly skewed if it falls below -1 or exceeds 1 (Groeneveld & Meeden, 1984).

In this context, *GDP Per Capita* and particularly the *Gini Coefficient* slightly exceed this threshold, while *Total Medals* and *Population* exhibit extremely right-skewed distributions. The right-skewed nature of the dependent variable *Total Medals* is visually represented in Figure A.1.

The descriptive statistics associated with categorical variables adhere to distinct standards and are grounded in the fundamental concept of frequency analysis. Consequently, these categorical variables are summarized in Table A.4. The occurrence of the three most tolerant regimes seems to be fairly evenly distributed, while the *Closed Autocracy* is less common. The number of hosts in the data set aligns with expectations since each edition could only be hosted by one country.

Chapter 4

Methodology

The primary aim of this study is to investigate how the number of Olympic medals relates to various socio-economic factors. The first section will describe the research type used to establish this relationship. The second section will explain the research approach used to assess the study's reproducibility. Finally, the last section will provide a detailed examination of the technical aspects of the models used

4.1 Research type

This paper employs a quantitative approach because it deals with variables that can be measured numerically, particularly counting variables. Unlike experimental research, where variables are intentionally changed in different groups, this study merely observes these variables (Kamil, 2004). Therefore, this research can be defined as explanatory since it tries to explain the relationship between the dependent variable and independent variables without altering the data. Additionally, the study also explores predictive aspects by assessing how well the models can make predictions. However, it is crucial to distinguish between explanatory and predictive modeling when it comes to evaluating their performance, potential issues, and overall objectives (Sainani, 2014).

A primary concern regarding the model validity is the risk of overfitting to the training data. Overfitting occurs when the model becomes excessively fitted to the data set, capturing noise and random fluctuations instead of the genuine underlying patterns or relationships. This overfitting issue significantly worsens the model's capability to provide accurate predictions for unseen data and apply these predictions effectively in real-world scenarios. This problem becomes more pronounced when dealing with a relatively small data set, as the limited information available makes it harder to accurately represent the true underlying patterns. (Sainani, 2014).

4.2 Research approach

The models utilized in this research align with the models employed in the various studies summarized in Table 2.1. These models include Ordinary Least Squares (OLS), Tobit, Negative

Binomial (NGB), and Poisson regression, which are consistent with the literature.

In addition to the regression models, other methods will also be applied. Ensemble methods, which combine multiple models to improve results will be used (Dietterich, 2000). Specifically, ensemble methods based on decision trees, including Random Forest, Gradient Boosting, and XGBoosting, will be employed. Detailed technical specifications for these models will be provided in Section 4.3.

Before proceeding with the analysis, the data set will be randomly divided into an 80% training sample and a 20% test sample. The training sample will be used to train the model, with subsequent predictions made on the unseen test set. Given the limited data set size, the allocation of data points to these subsets may substantially influence the analysis outcome. Therefore, the training set is divided into 10 equal folds, enabling cross-validation within the model training process. In a repeated procedure of 10 iterations, each fold serves as the validation set once, while the remaining nine folds constitute the training set. This technique proves particularly valuable in cases with a limited number of observations, as it ensures that every data point is utilized for both training and validation purposes. The performance of each fold out of the 10 will be investigated, aiming for a narrower range of performance metrics across these folds to enhance consistency and increase generalizability to real-world scenarios. Furthermore, the average of these performance metrics obtained through the 10-fold cross-validation on the observed training data will be compared to the performance metrics when applying the cross-validated model to unobserved test data in the out-of-sample predictons. When the performance metrics are nearly equal, it suggests that the model can generalize its predictions effectively, thereby guaranteeing external validity and confirming the potential consistency in the 10-fold cross-validation.(Berrar et al., 2019).

The Root Mean Squared Error (RMSE) serves as a commonly employed statistical measure for assessing a model's error rate and predictive efficacy (Chai & Draxler, 2014). The mathematical expression of this metric involves the summation of errors across all data points, their squared values, and subsequently, the extraction of the square root (4.5).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.5}$$

In this context, y_i represents the observed value for the i-th data point, $\hat{y}_i$ is the predicted value for the i-th data point, and n is the total number of data points in the data set. In this study, this metric serves as an indicator of the extent to which the average deviation between the predicted and actual medal counts can be observed. Furthermore, apart from assessing the models' predictive capabilities, additional metrics are employed to shed light on their explanatory aspects.

One such metric is the R-squared (R^2) coefficient. R^2 is utilized to assess the goodness of fit, offering insights into how effectively the independent variables explain the variation in the dependent variable. On a scale that ranges from 0 to 1, a higher R^2 value indicates a more robust fit, meaning that the model is better at handling outliers, providing more accurate estimates for its parameters, and ultimately enhancing its ability to explain the observed variations in the data. In simpler terms, a higher R^2 suggests that the model does a better job of capturing the underlying patterns in the data (Miles, 2005). The formula for the R^2 is presented in (4.6).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.6}$$

In this formula, y_i represents the observed value for the i-th data point, $\hat{y}_i$ is the predicted value for the i-th data point, $\bar{y}$ is the mean of the dependent variable and n is the total number of data points in the data set. Both RMSE and R^2 will be computed for both cross-validation and out-of-sample predictions.

The specific relationship between the independent variables and the dependent variables is established by analyzing the magnitude of coefficients associated with socio-economic predictors. However, it is important to note that the magnitude of coefficients alone does not provide insights into the statistical significance of the predictors. To assess the significance of each seperate predictor, null hypothesis testing is employed.

$$H_0 : \text{The coefficient is equal to zero}$$

$$H_1 : \text{The coefficient is not equal to zero}$$

The null hypothesis (H_0) asserts that the coefficient is equal to zero, indicating it does not influence the dependent variable. If the null hypothesis is rejected, the alternative hypothesis (H_1) is confirmed, signifying that the coefficient is not equal to zero and indeed exerts an influence on the dependent variable.

The null hypothesis is evaluated using the t-value for each coefficient, calculated by dividing the estimated coefficient by its standard error. A high t-value indicates a significant coefficient estimate, while a t-value close to zero implies insignificance. The t-value also aids in computing the associated p-value, using the fixed t-distribution and the degrees of freedom specific to the analysis. Degrees of freedom indicate the number of independently varying parameters, with a higher value indicating greater analysis stability (James, Witten, Hastie, Tibshirani et al., 2013).

The p-value reflects the probability of obtaining a t-value as extreme as or more extreme than the calculated t-value, assuming the null hypothesis is true. If a coefficient of zero could yield a larger t-value, it implies the coefficient estimate is not statistically significant. The significance threshold, typically set at $p < 0.05$, determines the maximum allowable probability for insignificance. When the p-value is below this threshold, it indicates rejecting the null hypothesis in favor of the alternative hypothesis, signifying the coefficient's significance. A stricter threshold at $p < 0.01$ signifies a higher level of significance, while the highest significance is at $p < 0.001$ (Schervish, 1996).

The model employs Sequential Backward Elimination (SBE) to select independent variables. It begins with the full set of variables and assesses their statistical significance based on a predefined p-value threshold of $p < 0.05$. The least significant variable is then iteratively removed until all remaining independent variables are statistically significant. This approach simplifies the model, enhances interpretability, and improves performance and efficiency by eliminating irrelevant features (Mao, 2004). It is important to note that this method is not suitable for ensemble models since ensemble models do not yield coefficients. Instead, ensemble models depend on variable importance scores, which quantify how much these models utilize independent variables to formulate predictions and explain the variance in the dependent variable. These scores are typically expressed on a scale spanning from 0 to 100. (Grömping, 2009).

Skewed socio-economic predictors in Table 3.1 will undergo a natural logarithmic transformation for multiple reasons. This transformation normalizes their distribution, aligning them with other independent variables. It also stabilizes variance, reduces outlier influence, and enhances model performance. Furthermore, it simplifies the interpretation of regression coefficients, particularly for the *Population* variable with a wide range of values. Ensemble models, while handling complex relationships, may not substantially benefit from this transformation, but it will not adversely affect their results. Thus, the natural logarithm is applied consistently to these variables in both regression and ensemble models.

For each individual model, SBE is performed, and the model is selected based on the criterion that all independent variables must be significant at $p < 0.05$. Subsequently, a comparison of the different models is carried out. The RMSE for cross-validation and out-of-sample predictions will help identify the best predictive model, while the R^2 for cross-validation and out-of-sample predictions will indicate the best-explaining model. However, to ensure the internal and external validity of the results, it is essential to consider whether the assumptions of each specific model are satisfied or violated. Finally, the best model will be chosen, and the entire data set will be used to explain the relationship between the Olympic medal count and the socio-economic factors.

The preceding steps are replicated for the research question that examines variations in socio-economic influence across gender, season, and year, where the best explanatory model is selected. The pre-processed data set will be divided into subsets based on these categories, and distinct analyses will be conducted for each, aiming to demonstrate their respective causal relationships within their specific contexts. The coefficients of the socio-economic predictors are extracted for the regression models and compared to assess whether there exist differences between them.
A confidence interval is established for the coefficient of a socio-economic predictor within different contexts. This interval is associated with a confidence level, reflecting the likelihood that if a series of confidence intervals were constructed from different random samples drawn from the same population, a certain percentage of these intervals would encompass the true value of the parameter, and thus, the coefficient. Hence, it signifies the probability that the provided confidence interval contains the true coefficient parameter. The confidence interval (CI) is calculated

as follows (4.7) (Hazra, 2017).

$$\text{CI} = \text{Coefficient estimate} \pm \text{Critical value (z)} \times \text{SE of coefficient estimate} \tag{4.7}$$

The z-value is determined by the chosen confidence level, where higher z-values correspond to increased confidence but also wider confidence intervals. Critical z-values are consistent for each confidence level. In this study, a 95% confidence interval is used, resulting in z ≈ 1.96 (Hazra, 2017). Two separate 95% confidence intervals for the same socio-economic predictor will be constructed for different contexts, and the presence of any overlap between these intervals will be assessed. Keeping in mind that there is a 95% probability for the true coefficient to fall within each individual confidence interval, it can be inferred that the likelihood of the true coefficients differing is $95\% \cdot 95\% \approx 90\%$ for two non-overlapping 95% confidence intervals.

4.3 Method specifications

4.3.1 Regression models

OLS

The Ordinary Least Squares (OLS) method is a frequently utilized technique employed for estimating the coefficients within a linear regression equation, and it is grounded in the principle of minimizing the Residuals Sum of Squares (RSS). The RSS is computed as the sum of the squares of the differences between the predicted and observed values for all data points. This minimization process yields a fitted regression equation, represented as (4.8).

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{i1} + \hat{\beta}_2 \cdot x_{i2} + \ldots + \hat{\beta}_p \cdot x_{ip} + e_i \tag{4.8}$$

In this equation $\hat{y}$ signifies the predicted dependent variable for i-th data point, $\hat{\beta}$ represents the fitted coefficients for the p-th independent variable, x denotes the value of the p-th independent variables for the i-th data point, and e accounts for the error term of the i-th data point, representing the disparities between predictions and observations (Montgomery, Peck & Vining, 2021).

The OLS regression method is underpinned by a set of fundamental assumptions. Violation of these assumptions can potentially introduce bias into parameter estimates, thereby impacting the validity of subsequent statistical inferences. These critical assumptions are outlined by Long (2008) as follows:

- **Linearity**: The relationship between the dependent and independent variables is linear. The change in the dependent variable remains constant with a one-unit change in the independent variable.
- **Independence of errors**: Residuals, representing the gaps between predicted and actual values, are independent across observations. The residual of an observation should not influence the prediction of another observation.

- **Homoscedasticity**: The variance of the residuals remains consistent, regardless of the predicted values of the dependent variable.
- **No multicollinearity**: Independent variables are not highly correlated with each other.

Tobit

Tobit regression is much like OLS regression and adheres to most of the same assumptions. However, there is a key distinction: Tobit regression is specifically designed for situations where the dependent variable is censored and hence only partially observed based on predefined thresholds. Tobit regression is particularly useful when dealing with data that has a significant number of censored observations, allowing researchers to model and analyze relationships between variables while accounting for the censoring process. While Tobit retains the assumptions from OLS, it introduces an extra assumption to address this censoring.

- **Censoring mechnanism**: The censoring mechanism and hence the threshold at which censoring occurs is assumed to be known

The threshold for censorship is indeed specified in this study, and the constraint is defined as (4.9).

$$Total\ Medals = \begin{cases} Total\ Medals^* & if\ Total\ Medals^* > 0 \\ 0 & if\ Total\ Medals^* \leq 0 \end{cases} \tag{4.9}$$

The variable *Total Medals* represents the number of medals won, and Total Medals* represents a latent variable describing the number of medals won. When a country has earned more than zero medals, their data is fully observed, along with their underlying socio-economic factors. Conversely, for countries with zero medals or fewer, their data is censored and less weight is attributed to these data points when estimating the model parameters, as noted by Rewilak (2021).

In the study conducted by Noland and Stahler (2017), the same decision was made to censor the countries that did not win any medals. This decision was based on the observation that there was a relatively high occurrence of these non-winning countries in the data set. Employing the Tobit regression model enabled the researchers to assign greater emphasis to the explanatory capacity of countries that achieved Olympic medal success.

Censoring in the data limits the feasibility of calculating R^2, as the variance explained by the independent variables in the dependent variables cannot be assessed through the censored data points. In the context of Tobit regression, the log-likelihood is employed as a metric, estimating the probability that the model effectively describes the observed data, given the model's parameters.

Poisson and Negative Binomial

The subsequent models employed are Poisson and Negative Binomial (NGB) regression, each introducing distinct assumptions compared to the two previous models discussed. The key assumptions of Poisson regression, as outlined by James et al. (2013), include:

- **Count dependent variable**: The dependent variable signifies counts, such as the number of occurrences happening within a designated period.
- **Non-negative**: The dependent count variable should be whole numbers, non-negative, and mutually independent.
- **Poisson distribution**: The dependent count variable follows a Poisson distribution.
- **Equidispersion**: The mean and variance of the dependent count variable should be equal to each other.
- **Log-linear**: The natural logarithm of the dependent count variable is a linear function of the predictor variables.

The dependent count variable undergoes a logarithmic transformation following the log-linear assumption. This leads to the formulation of a Poisson regression model (4.10). (James et al., 2013).

$$\ln(\mu_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{i1} + \hat{\beta}_2 \cdot x_{i2} + \ldots + \hat{\beta}_p \cdot x_{ip} \tag{4.10}$$

$\ln(\mu)$ represents the natural logarithm of the number of Olympic medals for the i-th data point. β corresponds to the coefficients associated with the p-th predictor variable, while x denotes the values of the p-th predictor variable for the i-th data point.

The crucial assumption of Poisson regression is the equidispersion, where the mean and variance of the dependent variable are equal. When this assumption is violated, the NGB regression, an extension of Poisson regression, is a suitable alternative. The NGB regression permits a more flexible modeling approach and relaxes the assumption of equidispersion.

Both Poisson regression and NGB regression fall within the category of Generalized Linear Models (GLM), which represents an expansion of conventional linear regression models designed to accommodate a wider spectrum of data types and distributions, where non-linear problems are transformed into linear ones (James et al., 2013).

4.3.2 Ensemble models

Decision trees

Ensemble models, which are explored next, are based on the foundation of decision trees. Decision trees known for their interpretability, start at the root with the entire data set. At each internal node, a predictor and split point are chosen to divide the data into two subsets, minimizing the difference between actual and predicted values (RSS). The predicted value is typically the mean of observations in a node. This process continues until specific stopping conditions, like reaching maximum tree depth or minimal RSS improvement, are met. Terminal nodes, or leaves, emerge when these conditions are satisfied, holding the predicted value—either the mean of the data points in that leaf or a specific observation's value. Although decision trees are user-friendly and transparent for tracing predictions, they can overfit with excessive depth and complexity. Therefore, ensemble models are introduced to mitigate these issues and enhance overall model performance (Loh, 2011).

Random Forest

The random forest algorithm is based on bagging, or bootstrap aggregating. In this technique, the original training data set is randomly resampled to create B subsets. Each of these B subsets is used to build a decision tree, introducing increased diversity among the resulting trees. The final predictive outcome is obtained by averaging the predictions from all the constructed decision trees, which helps reduce variance. Notably, the assessment of predictive performance during the training of the bagged model does not have to rely solely on cross-validation; it can also be accomplished through out-of-bag error estimation. In this process, each decision tree is constructed using a bootstrapped subset containing most of the original observations. The remaining data, not used in the bootstrapped subset, can be used to make predictions for the decision tree on which it was not trained. Subsequently, these predictions are compared to the actual values to evaluate the accuracy of the model's training.

This model provides a valuable feature: assessing variable importance by analyzing the mean decrease in RSS for each variable's split. The higher the average decrease, the more crucial the predictor is considered. Compared to decision trees, Random Forests offer a subtle improvement. Traditional decision trees often place the most influential predictors near the tree's top, leading to high similarity among the constructed trees. This similarity can undermine the goal of combining diverse models to reduce variance. To address this, Random Forests use a strategy where only a subset of predictors is used at each split, increasing dissimilarity among the B decision trees and ultimately enhancing predictive performance. A notable advantage of Random Forest is its flexibility, not being bound by strict statistical assumptions as traditional regression models (Loh, 2011).

Gradient Boosting and Extreme Gradient Boosting

Boosting algorithms work by building decision trees one after another, where each new tree aims to correct the mistakes of the previous one. This involves creating multiple decision trees, each using the errors from the predictions of the preceding tree.

Gradient boosting initiates with an initial decision tree, where each data point's initial prediction is the average of the response variable. Subsequent decision trees follow as weak learners, adjusting their predictions to minimize residuals. This process involves a specific loss function, often the Mean Absolute Error (MAE), and a learning rate that scales adjustments. A lower learning rate provides a slower learning process, necessitating more decision trees for reduced residuals but offering computational efficiency. Conversely, a higher learning rate demands fewer trees for efficient performance. The iterative process of diminishing residuals and growing new decision trees continues until the loss function and residuals become negligible or predefined stopping conditions are met. The final prediction combines outputs from all decision trees. One remarkable feature of Gradient Boosting is its flexibility, not confined by strict statistical assumptions about data distribution or relationships. Nonetheless, achieving optimal performance necessitates tuning specific hyperparameters, such as the learning rate, decision tree depth, and minimum node observations.(Natekin & Knoll, 2013).

Extreme Gradient Boosting, often abbreviated as XGBoosting, is an advanced variant of Gradient Boosting. While the core concepts are similar, XGBoost introduces subtle adjustments that impact its operational behavior. Initially, all data points receive identical predictions, set to the average value. However, the formation of decision trees is regularized by the introduction of the parameter λ. Its presence leads to the pruning of the decision tree, which means that the decision tree becomes less complex and deep by removing the leaves from the bottom to the top. The higher the value of λ, the quicker the decision trees get pruned. Regularization here thus functions as a mechanism to reduce tree complexity, making it less sensitive to outliers and mitigating the risk of overfitting. By default, λ is set to 1, a value consistent with this research. XGBoosting does not adhere to strict assumptions but involves tuning several hyperparameters, including: the learning rate, maximum depth of decision trees, γ the minimum reduction in the loss function, fraction of columns subsampled at each level, and the minimum number of observations in a leaf.

Both boosting algorithms, XGBoosting and Gradient Boosting, are employed in this study. While XGBoosting is known for its robustness in handling outliers, there remains the possibility of excessive pruning in the decision tree. Hence, it is of interest to empirically evaluate its superior performance. The optimization of hyperparameters for both boosting algorithms is carried out via random search within predefined parameter value ranges (Chen & Guestrin, 2016).

Chapter 5

Results

5.1 Regression results

The SBE process starts with the OLS regression model, aiming to determine the optimal set of independent variables for evaluating regression outcomes and performance metrics. Initially, all socio-economic predictors are included in the analysis. Through cross-validation, coefficients of predictors with p-values exceeding the 0.05 threshold are systematically removed. The first to be eliminated is the variable *Electoral Democracy*, with a p-value of 0.520. Subsequently, the model undergoes another round of cross-validation, resulting in the exclusion of *ln(Gini Coefficient)* with a p-value of 0.413. The process is reiterated with the remaining variables, leading to the removal of *Food Supply Surplus* at a p-value of 0.128. Following another cycle of cross-validation, all socio-economic predictors exhibit p-values lower than 0.05. Predictors that exceed the threshold are excluded from the analysis and performance evaluation, as they fail to reject the null hypothesis, indicating that their coefficients are statistically equal to zero.

The second model employed is the Tobit regression, and like the OLS regression, it begins with the inclusion of all socio-economic predictors. Through cross-validation, the model provides coefficient estimates, with *Electoral Democracy* having the highest p-value of 0.613 for its coefficient. Upon cross-validating the Tobit regression with the reduced set of socio-economic predictors, all these remaining predictors are found to be statistically significant at the 0.05 significance level. In other words, every coefficient, except that of *Electoral Democracy* is found to be different from zero, thus rejecting their null hypothesis. Consequently, these significant coefficients and hence predictors are included in the final model used for assessing the relationship and performance metrics.

Then the Poisson regression model is cross-validated using a similar approach as the previous models. In this process, *Electoral Democracy* is the first predictor to be eliminated, as it has a p-value of 0.110. After this removal, the Poisson regression model is cross-validated again, and it is found that all the remaining socio-economic predictors are statistically significant at the 0.05 significance level. Therefore, in the final model used for assessment, *Electoral Democracy* is the only socio-economic predictor that cannot reject the null hypothesis that its coefficient is equal to zero.

The last regression model, NGB regression undergoes cross-validation, initially with all socio-economic predictors. *Past Host* is removed first due to its p-value of 0.538. Subsequently, *Electoral Democracy* is eliminated with a p-value of 0.357 in the next cross-validation round. In the following iteration, *Closed Autocracy* is removed, having a p-value of 0.211. When cross-validating the reduced model, all remaining coefficients are statistically significant at the 0.05 level. As a result, the socio-economic predictors that exceeded the significance threshold of 0.05 cannot reject the null hypothesis that their coefficient is equal to zero and are therefore excluded from the final model.

The SBE process has been carried out for various regression models, leading to the identification of the significant socio-economic predictors included in their respective final models. The regression outcomes of these models and the corresponding composition of independent variables are summarized in Table 5.1.

Table 5.1. Regression results

Variable	OLS	Tobit	Poisson	NGB
Intercept	-114.30***	-264.10***	-22.86***	-23.49***
	(7.10)	(14.50)	(0.46)	(1.09)
ln(GDP Per Capita)	4.28***	11.86***	1.16***	1.05***
	(0.52)	(13.67)	(0.04)	(0.10)
ln(Population)	3.73***	7.50***	0.64***	0.71***
	(0.28)	(0.51)	(0.01)	(0.04)
ln(Gini Coefficient)		-12.54**	-1.32***	-2.22***
		(3.91)	(0.11)	(0.32)
Closed Autocracy	6.16**	10.52**	0.83***	
	(2.29)	(3.79)	(0.09)	
Electoral Autocracy	3.96***	7.06***	0.75***	0.40*
	(1.17)	(2.04)	(0.05)	(0.16)
Electoral Democracy				
Healthcare Expenditures	1.46***	1.78***	0.03***	0.06*
	(0.21)	(0.34)	(0.008)	(0.03)
Food Supply Surplus		-11.90*	0.59***	1.00*
		(5.34)	(0.14)	(0.43)
Host	23.00***	30.59***	0.82***	0.95*
	(4.13)	(6.15)	(0.07)	(0.45)
Past Host	10.65***	8.22*	0.22***	
	(3.11)	(4.02)	(0.06)	
Maximum Medals	0.01***	0.03***	0.002***	0.002***
	(0.001)	(0.04)	(0.001)	(0.001)
N	609	609	609	609
R^2 (Cross-validation)	0.45		0.70	0.67
RMSE (Cross-validation)	10.82	16.71	7.38	8.92
R^2 (Out-of-sample)	0.40		0.61	0.31
RMSE (Out-of-sample)	8.10	15.77	6.51	8.67
Log-likelihood		-1254		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

In both the OLS and NGB regressions, three independent variables were excluded, while in the Tobit and Poisson regressions, only one independent variable was removed. The variable *Electoral Democracy* showed no significance in any of the models and was excluded from all.

The RMSE values observed during cross-validation showed the lowest for the Poisson model (= 7.38), followed by the NGB model (= 8.92), the OLS model (= 10.82), and the Tobit model (= 16.71). RMSE for out-of-sample predictions serves as an indicator of the model's generalization performance. For all models, the RMSE decreased, with the lowest for the Poisson model (= 6.51), followed by the OLS model (= 8.10), the NGB model (= 8.67), and lastly, the Tobit model (= 15.77). In the cross-validation, the R^2 was highest for the Poisson model (= 0.70), followed by the NGB model (= 0.67), and lastly, the OLS model (= 0.45). For the out-of-sample predictions, the Poisson model also had the highest R^2 (= 0.61), followed by the OLS model (= 0.40), with the NGB model explaining the least well (= 0.31).

It is important to note that the Tobit model could not provide an R^2 value due to the censored nature of the model, which resulted in the log-likelihood metric remaining at -1254. This metric is challenging to interpret and cannot be directly compared to the R^2. However, the very high RMSE in both cross-validation and out-of-sample predictions for the Tobit regression, indicates a lack of explanatory power and accuracy in the Tobit model. Interestingly, the OLS and NGB models switched their ranks for both the RMSE and R^2 in cross-validation when compared to out-of-sample predictions. A closer examination of the cross-validation results presented in Table B.1 reveals notable differences between the performance of various models. Specifically, when we focus on the R^2 values, the OLS model demonstrates a more consistent range (= [0.22-0.58]) in comparison to the NGB model, which exhibits a wider and less consistent range (= [0.38-0.92]). Similarly, for RMSE, OLS shows a narrower and more consistent range (= [8.45-14.80]) compared to the broader and less consistent range of RMSE values for NGB (= [4.92-16.29]). While, on average, the NGB model yields better performance metrics, it is essential to consider the reliability of these metrics. The broader range of NGB performance metrics indicates lower reliability. This is evident from the decline in R^2 from 0.67 during cross-validation to less than half of 0.31 in out-of-sample predictions, highlighting the model's limited generalizability.

In contrast, the Poisson regression model exhibits a more consistent range for both R^2 and RMSE in Table B.1. The Poisson regression model retains the most socio-economic predictors, all of which are highly significant. Moreover, it achieves the most favorable values for both R^2 and RMSE in both cross-validation and out-of-sample predictions. The detailed analysis of the 10-fold cross-validation suggests that the Poisson regression model demonstrates a higher reliability of cross-validation metrics compared to the second-best performing NGB regression.

5.2 Ensemble results

Ensemble models automatically determine the importance of independent variables and do not necessitate manual selection. The variable importance for Random Forest, Gradient Boosting, and Extreme Gradient Boosting, along with their corresponding performance metrics and tuned hyperparameters, are presented in Table 5.2.

Table 5.2. Ensemble results

Variable	Random Forest	Gradient Boosting	XGBoosting
lm(Population)	100.00	100.00	100.00
Healthcare Expenditures	48.62	28.47	36.21
ln(GDP Per Capita)	46.89	37.48	43.75
ln(Gini Coefficient)	16.43	10.50	33.42
Food Supply Surplus	18.89	17.87	14.79
Host	7.39	9.36	2.90
Past Host	2.00	1.35	0.00
Maximum Medals	41.10	36.90	61.85
Electoral Democracy	1.70	1.04	5.79
Closed Autocracy	1.25	0.69	2.78
Electoral Autocracy	0.00	0.00	0.34
N	609	609	609
R^2 (Cros-validation)	0.71	0.75	0.80
RMSE (Cross-validation)	7.67	6.82	6.18
R^2 (Out-of-sample)	0.84	0.79	0.81
RMSE (Out-of-sample)	4.20	4.77	4.59
Number of decision trees		175	863
Tree depth		9	6
Learning rate		0.07	0.16
Minimum observations in node		5	4
Subsample of predictors	1.00		0.85
Subsample of observations			0.98
Gamma			0.56

The variable importance analysis for all three ensemble models highlights that *ln(Population)* is the most influential in constructing the models and explaining the Olympic medal count, serving as the baseline score of 100.00. *ln(GDP Per Capita)* is the second most important independent variable in Gradient Boosting (= 37.48) and XGBoosting (= 43.75) models, while it takes the third position in the Random Forest model (= 46.89). *Healthcare Expenditures* is the second most influential independent variable in the Random Forest model (= 48.62) but ranks third in both the Gradient Boosting (= 28.47) and XGBoosting models (= 36.21). Notably, the *Maximum Medals* variable is excluded from this ranking as it functions as a control variable in the analysis. However, its relatively high importance in all ensemble models underscores the importance of controlling for the number of Olympic medals that could potentially be won in each edition. It is worth highlighting that in all ensemble models, the regime of *Electoral Autocracy* consistently emerges as the least important, while *Electoral Democracy* stands out as the most significant among the regime categories. In contrast, all regression models eliminate the *Electoral Democracy.*

In the cross-validation metrics for the ensemble models, Random Forest achieved the lowest R^2 (= 0.72) and the highest RMSE (= 7.67). In contrast, Gradient Boosting performed better for the R^2 (= 0.75) and RMSE (= 6.78), while the XGBoosting model excelled in cross-validation for the R^2 (= 0.80) and the RMSE (= 6.18). However, in the out-of-sample predictions, a different picture emerges. The R^2 of the Random Forest was the highest (= 0.84), and the RMSE was the lowest (= 4.20). Gradient Boosting, on the other hand, performed the least well in both R^2 (= 0.75) and RMSE (= 5.23). XGBoosting improved its performance in the R^2 (= 0.81) and RMSE (= 4.59) compared to cross-validation. When observing Table B.1, it becomes evident that the Random Forest model exhibits a wider range of R^2 values (= [0.44-0.90]) in comparison to XGBoosting (= [0.59-0.94]). However, the ranges of RMSE values are nearly equal, with [4.21-12.18] for Random Forest and [3.10-11.46] for XGBoosting. Therefore, the slightly higher out-of-sample performance of Random Forest compared to XGBoosting is not conclusive. This lack of conclusiveness is attributed to the differing nature of cross-validation consistency and hence generalizeability, particularly in terms of R^2.

In the Random Forest model, the only tuning hyperparameter involved adjusting the number of predictors considered at each split. This resulted in the use of the full set of predictors at every split. In the case of the Gradient Boosting model, it utilized fewer decision trees (= 175) compared to the XGBoosting model (= 863). However, the decision trees within the Gradient Boosting model were deeper (= 9) than those in the XGBoosting model (= 6). This difference in tree depth may be attributed to the regularization and pruning characteristics of the XGBoosting model. The learning rate in the Gradient Boosting model (= 0.07) is less than that in the XGBoosting model (= 0.16), a notable observation. This difference is noteworthy because a lower learning rate typically suggests a requirement for a greater number of weak learners to construct the model effectively. However, in this case, the Gradient Boosting model employs fewer decision trees compared to the XGBoosting model.

5.3 Assumption testing

The regression and ensemble models were trained, and predictions along with performance metrics were obtained. However, the validity of these findings depends on whether the assumptions are met or violated. In this section, the focus is solely on testing the assumptions of the regression models since the ensemble models are not constrained by these statistical assumptions.

5.3.1 OLS and Tobit

The OLS and Tobit regression models adhere to the same assumptions, except for the censoring mechanism that is added to the Tobit model.

- **Censoring mechanism**: The censoring mechanism and hence the threshold at which censoring occurs is assumed to be known.

This assumption is met in the Tobit regression model since the threshold is determined based on censoring for observations where the Olympic medal count is zero.

- **Linearity**: The relationship between the dependent and independent variables is linear.

The change in the dependent variable remains constant with a one-unit change in the independent variable.

- **Homoscedasticity**: The variance of the residuals remains consistent, regardless of the predicted values of the dependent variable.

Residual plots illustrate the relationship between residuals (y-axis) and fitted values (x-axis) in regression models. Ideally, they should display a straight, horizontal line at y=0, indicating a well-fitted model. However, curved lines in the residuals suggest non-linearity (Tsai, Cai & Wu, 1998). To assess homoscedasticity, it is essential to ensure that residuals' variance is randomly distributed across fitted values. Deviations from this random pattern or funnel shapes may indicate heteroscedasticity. Residual plots are valuable for testing linearity and homoscedasticity assumptions in regression. Figure B.1 and Figure B.2 display the residual plots for OLS and Tobit regressions, respectively.

In both the OLS and Tobit models, the trend lines, highlighted in red, exhibit a curved shape, which suggests non-linearity in both regression models. Additionally, the variance in these models varies with respect to the fitted values. Notably, negative fitted values correspond to high positive residuals, while fitted values around zero yield residuals close to zero. This observation aligns with the constraint that the observed Olympic medal count cannot be negative, causing predictions below zero to result in highly positive residuals. In summary, the linearity and homoscedasticity assumptions are violated for the OLS and Tobit regression models.

- **No multicollinearity**: Independent variables are not highly correlated with each other.

In Section 3.2.3, the exclusion of socio-economic predictors in the data set was based on their high correlations with each other, as observed in Table A.2. This action was taken to ensure the absence of highly correlated variables which satisfies the assumption.

- **Independence of errors**: Residuals, representing the gaps between predicted and actual values, are independent across observations. The error term of an observation should not influence the prediction of another observation

This represents the last assumption tested for both regression models. Autocorrelation evaluates the extent to which preceding residuals influence succeeding residuals. Autocorrelation among residuals can be determined by constructing an autocorrelation plot. Figure B.3 and Figure B.4 present the autocorrelation plots for the OLS and Tobit regression models, respectively.

The dashed line in the plot represents the autocorrelation threshold of 0.05, serving as the confidence interval indicating autocorrelation in the model at a specific lag. The first lag holds particular significance as it reveals whether the current residual is correlated with the preceding one. In the case of the OLS regression model, it marginally exceeds the 0.05 treshold, while for the Tobit regression model, it falls between the threshold and zero for the first lag. Consequently, the independence error assumption is met for the Tobit regression but not for the OLS regression.

5.3.2 Poisson and NGB

Both the Poisson and NGB models share the same assumptions, although there is a relaxation of one assumption in the NGB model.

- **Count dependent variable**: The dependent variable signifies counts, such as the number of occurrences happening within a designated period

Each specific edition and year of the Olympic Games indeed represents a count, and as such, this assumption is met for both the Poisson and Negative Binomial (NGB) models.

- **Non-negative**: The dependent count variable should be whole numbers and non-negative.

The count of Olympic medals is inherently composed of whole numbers, and a country can not win a negative number of Olympic medals. Therefore, this assumption is also valid for both the Poisson and NGB models.

- **Poisson distribution**: The dependent count variable follows a Poisson distribution

The observed counts of the dependent variables are compared to the expected count of the variables according to the Poisson distribution and the given mean of the number of Olympic medals won, which is equal to 6.44 in this study. The observed versus the expected counts are visualized in Figure B.5.

In an ideal scenario, a perfectly matching Poisson distribution would result in a plot where the y and x-axes share the same scale, forming a straight diagonal line, signifying equivalence between observed and expected counts based on the distribution. However, the Poisson distribution is constrained, assuming a maximum of 16 Olympic medals based on the average medal count, while the observed data set records more than 100 Olympic medals awarded. Consequently, the scaling of both axes differs significantly. This discrepancy can be attributed to the distinctive nature of the observed Olympic medal distribution, characterized by numerous zero-medal winners and a few outliers with exceptionally high medal counts.

The Poisson distribution also reveals extreme outliers, with the majority of Olympic medal recipients clustered toward the distribution's beginning and center. Consequently, the Poisson distribution exhibits right-skewness, a characteristic shared with the visualized distribution of Olympic medal recipients in Figure A.1 and confirmed by the positive skewness measure in Table 3.1. The Poisson distribution assumption is not satisfied based on Figure B.5, indicating a disparity. However, it is noteworthy that both the observed and Poisson distributions exhibit right-skewness with long tails, implying some similarity in this regard.

- **Equidispersion**: The mean and variance of the dependent count variable should be equal to each other.

Table 3.1 presents the statistics for the number of Olympic medals won, revealing a mean of 6.44 and a variance of 188.26. These statistics highlight a notable difference between the mean and variance, suggesting that the Poisson regression assumption of equidispersion is not met. The NGB regression, by relaxing this assumption, ensures that such differences in mean and variance do not exert any influence on the interpretation of the model.

- **Linearity**: The natural logarithm of the dependent count variable is a linear function of the predictor variable.

The relationship between the natural logarithm of the dependent count variables and the socio-economic predictors is revealed through residual plots. Residual plots for both Poisson and NGB regression models are generated to illustrate the distribution of residual values along the range of fitted values. These residual plots are visualized in Figure B.6 for the Poisson regression and in Figure B.7 for the NGB regression.

For both the Poisson and NGB regression models, the trend line appears relatively flat, positioned close to a straight horizontal line starting at y=0. The lower fitted values exhibit a fairly random distribution in both regression models, with a mix of high positive and negative values. On the other hand, higher fitted values tend to have smaller residuals, being closer to zero. This suggests that the models effectively capture and fit outliers, particularly related to high Olympic medal winners. Notably, the range of residuals is more extensive for the Poisson regression compared to the NGB regression. In summary, the consistent flatness of the trend line indicates that the natural logarithm of the count is indeed a linear function of the predictor variable, satisfying this assumption for both the Poisson and NGB regression models.

5.4 Final model selection and interpretation

In the preceding sections, various models underwent cross-validation, followed by predictions on the test data set. The coefficients, importance of independent variables, and performance metrics for models were obtained. Furthermore, the assumptions for the specific regression models were tested to provide context for the model's output. All these steps contribute to the selection of the final machine learning model used to explain the relationship between the Olympic medal count and socio-economic factors, as well as to make predictions for the future.

The Tobit regression produced the lowest RMSE for both cross-validation and out-of-sample predictions and used the log-likelihood metric as a measure of goodness of fit, which may not be the preferred choice over R^2. The OLS regression displayed the lowest R^2 and the second lowest RMSE in cross-validation. Both the Tobit and OLS models failed to meet the assumptions of linearity and homoscedasticity. Additionally, the OLS model could not satisfy the independence of errors assumption.

In contrast, Poisson regression excelled in terms of performance metrics, achieving the highest R^2 and the lowest RMSE for both cross-validation and out-of-sample predictions while maintaining consistency within the cross-validation process. However, it faced challenges in adhering to the Poisson distribution assumption and maintaining equality of mean and variance. Similarly, the NGB model did not adhere to the Poisson distribution assumption but allowed for the relaxation of the mean and variance equality assumption. Nonetheless, the reduced generalizability of the NGB model, as evidenced by the decline in R^2 from cross-validation (= 0.67) to out-of-sample predictions (= 0.31), outweighed the benefits of relaxing the mean and variance equality assumption. Consequently, among the regression models, the Poisson model emerged as the most favorable choice.

The XGBoosting model exhibited the highest R^2 and the lowest RMSE during cross-validation, whereas the Random Forest model displayed the least favorable performance metrics, manifesting the lowest R^2 and RMSE. Notably, in out-of-sample predictions, the Random Forest model outperformed both the Gradient Boosting and XGBoosting models. Nonetheless, the disparity in out-of-sample performance between Random Forest and XGBoosting was less pronounced than the notable consistency advantage of XGBoosting during cross-validation. As a result, XGBoosting is the preferred choice for generalizability and real-world scenario predictions. Collectively, the ensemble models demonstrated superior performance metrics in comparison to the best-performing regression model, the Poisson. However, it is essential to note that the assessment of relative variable importance in the construction of these models failed to explain the precise relationship between Olympic medal counts and socio-economic factors. This limitation results in a lack of interpretive and explanatory value crucial for a more comprehensive understanding of this relationship. Consequently, the Poisson regression model is preferred for quantifying these influences, leveraging the retrieved coefficients of the socio-economic predictors. Conversely, the XGBoosting model is the recommended choice for making predictions concerning unseen data in real-world scenarios, where it exhibits the ability to predict with an average deviation of 4.59 Olympic medals from the actual count.

The Poisson regression model is employed to explain the association between the Olympic medal count and the socio-economic determinants. This interpretation encompasses all gender categories and encompasses both the Summer and Winter Olympics spanning the years 2000 to 2016. The Poisson regression undergoes cross-validation on the entire data set, yielding the outcomes outlined in Table 5.3.

Table 5.3. Poisson regression

Variable	
Intercept	-22.260***
	(0.410)
ln(GDP Per Capita)	1.113***
	(0.035)
ln(Population)	0.627***
	(0.012)
ln(Gini Coefficient)	-1.383***
	(0.107)
Closed Autocracy	0.728***
	(0.075)
Electoral Autocracy	0.712***
	(0.048)
Healthcare Expenditures	0.035***
	(0.007)
Food Supply Surplus	-0.412***
	(0.118)
Host	0.844***
	(0.059)
Past Host	0.260***
	(0.053)
Maximum Medals	0.002***
	(0.001)
N	759
R^2	0.73
RMSE	6.95

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

In the hypothetical scenario where all independent variables are set to zero, the projected number of Olympic medals won approximates -22.260. However, this scenario is highly implausible given the nature of the independent variables involved. Specifically, for each 1% increase in a country's *GDP Per Capita*, the expected Olympic medal count rises by approximately 1.113%, all other variables being held constant (SE = 0.035). Likewise, a 1% growth in a country's *Population* corresponds to a 0.627% increase in the expected Olympic medal count, while maintaining other variables at constant levels (SE = 0.012). Conversely, a 1% increase in the *Gini Coefficient*, indicative of greater income equality, results in a 1.383% reduction in the expected Olympic medal count, assuming all other factors remain unchanged (SE = 0.107).

In the context of this analysis, it is observed that the presence of a *Closed Autocracy* within a country leads to an increase in the number of Olympic medals obtained by a factor of $e^{0.728} - 1 = 2.0709 - 1 = 1.0709 \approx 107\%$ compared to both the *Liberal Democracy* and *Electoral Democracy*, all other factors being held constant (SE = 0.075). Similarly, the presence of an *Electoral Autocracy* within a nation results in an expected increase of $e^{0.712} - 1 = 2.0381 - 1 = 1.0381 \approx 104\%$ in the number of Olympic medals secured compared to both the *Liberal Democracy* and *Electoral Democracy*, while controlling for all other relevant variables (SE = 0.048).

When *Healthcare Expenditures* increase by 1% as a share of the GDP, the estimated number of Olympic medals won experiences a rise of $e^{0.035} - 1 = 1.0356 - 1 = 0.0356 \approx 3.6\%$, under the condition that all other variables remain constant (SE = 0.007). Conversely, when a country's *Food Supply Surplus* increases by one unit, there is a decrease of $e^{-0.412} - 1 = 0.6623 - 1 = -0.3378 \approx 34\%$, in the projected number of Olympic medals obtained, holding all other factors constant (SE = 0.118). Additionally, hosting the Olympics is associated with an increase of $e^{0.844} - 1 = 2.3257 - 1 = 1.3257 \approx 133\%$ in the expected number of Olympic medals won while controlling for all other factors (SE = 0.059). Furthermore, having organized the Olympics in one of the four preceding years is linked to an expected increase of $e^{0.260} - 1 = 1.2969 - 1 = 0.2969 \approx 30\%$ in the number of Olympic medals won while keeping all other factors constant (SE = 0.058). Lastly, for each additional medal that can be earned in the entire Olympic Games, the expected number of Olympic medals won by a country increases by $e^{0.002} - 1 = 1.0020 - 1 = 0.0020 \approx 0.2\%$, under the assumption that all other factors remain constant (SE = 0.001).

5.5 Gender, season and year interpretation

The preceding section has explained the influences of socio-economic predictors on the total number of Olympic medals won. It is important to note that this model is applicable across various genders, seasons, and years. Consequently, the same Poisson regression model is employed to assess this relationship within different gender categories, across various seasons, and over different years, enabling the evaluation of variations in the extent of these influences. 95% confidence intervals are established for the coefficients in all Poisson models, utilizing the standard errors, enabling a comparison of differences in coefficients.

5.5.1 Gender

In this context, the analysis focuses on investigating whether the influences of specific socio-economic factors differs between male and female athletes. The results are summarized in Table 5.4.

Table 5.4. Poisson regression: Male and Female

Variable	Male	Female	95% CI Male	95% CI Female
Intercept	-21.530***	-23.990***	[-25.572,-20.508]	[-25.322,-22.700]
	(0.527)	(0.669)		
ln(GDP Per Capita)	1.072***	1.097***	[0.982,1.162]	[0.991,1.204]
	(0.046)	(0.54)		
ln(Population)	0.563***	0.679***	[0.532,0.594]	[0.641,0.716]
	(0.016)	(0.029)		
ln(Gini Coefficient)	-1.162***	-1.787***	[-1.434,-0.893]	[-2.138,-1.440]
	(0.138)	(0.178)		
Closed Autocracy	0.555***	0.909***	[0.342,0.764]	[0.694,1.125]
	(0.108)	(0.110)		
Electoral Autocracy	0.789***	0.633***	[0.663,0.914]	[0.483,0.784]
	(0.064)	(0.077)		
Healthcare Expenditures	0.040***	0.046***	[0.021,0.058]	[0.024,0.068]
	(0.001)	(0.011)		
Food Supply Surplus	-0.280.	-0.649***	[-0.586,0.027]	[-1.004,-0.294]
	(0.156)	(0.181)		
Host	0.927***	0.746***	[0.769,1.079]	[0.565,0.919]
	(0.079)	(0.090)		
Past Host	0.265***	0.276***	[0.121,0.405]	[0.121,0.426]
	(0.072)	(0.078)		
Maximum Medals	0.003***	0.004***	[0.0029, 0.0034]	[0.0031,0.0039]
	(0.001)	(0.001)		
N	746	705	746	705

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Through an examination of the 95% confidence intervals, it becomes evident that the *Intercept, ln(Population)*, and *ln(Gini Coefficient)* exhibit non-overlapping intervals. Consequently, there exists a distinction in the influence of these socio-economic factors on the Olympic medal count when considering male and female athletes. The true coefficients for these independent variables are not equivalent between the two genders, with a level of confidence estimated to be approximately 90%, derived from the product of two 95% confidence intervals.

In the hypothetical scenario wherein all independent variables are set to zero, the projected number of Olympic medals achieved is estimated to be approximately -21.53 for male athletes and -23.99 for female athletes. For every 1% increase in a country's *Population*, there is a corresponding 0.563% increase in Olympic medals for male athletes, and a higher increase of 0.679% for female athletes, with all other factors held constant. Moreover, an increase in the *Gini Coefficient* by 1% is associated with a decrease in the number of Olympic medals won by 1.162% for male athletes and 1.787% for female athletes, under the condition of all other factors remaining constant.

5.5.2 Season

In this context, the analysis examines whether the influence of particular socio-economic factors varies between the Summer and Winter Olympics. The findings are presented in Table 5.5.

Table 5.5. Poisson regression: Season

Variable	Summer	Winter	95% CI Summer	95% CI Winter
Intercept	-15.370***	-35.730***	[-16.691,-14.042]	[-39.482,-32.156]
	(0.676)	(1.868)		
ln(GDP Per Capita)	1.087***	2.414***	[1.012,1.163]	[2.144,2.706]
	(0.037)	(0.141)		
ln(Population)	0.650***	0.530***	[0.626,0.687]	[0.469,0.603]
	(0.013)	(0.036)		
ln(Gini Coefficient)	-1.136***	-3.164***	[-1.364,-0.91]	[-3.872, -2.472]
	(0.114)	(0.359)		
Closed Autocracy	0.708***	3.344***	[0.551,0.877]	[2.736,3.965]
	(0.082)	(0.312)		
Electoral Autocracy	0.804***	1.841***	[0.701,0.917]	[1.482,2.216]
	(0.053)	(0.185)		
Healthcare Expenditures	0.042***	0.055***	[0.026,0.057]	[0.014,0.096]
	(0.008)	(0.020)		
Food Supply Surplus	-0.505***	0.035	[-0.768,-0.253]	[-0.601,0.679]
	(0.130)	(0.323)		
Host	0.880***	0.664***	[0.741,1.027]	[0.419,0.914]
	(0.071)	(0.126)		
Past Host	0.297***	0.009	[0.185,0.419]	[-0.283,0.266]
	(0.058)	(0.138)		
Maximum Medals	-0.006***	-0.006***	[-0.007,-0.005]	[-0.009,-0.003]
	(0.001)	(0.002)		
N	510	249	510	249

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The *Intercept, ln(GDP Per Capita), ln(Population), ln(Gini Coefficient), Closed Autocracy*, and *Electoral Autocracy* all manifest non-overlapping 95% confidence intervals. Consequently, the true coefficients for these independent variables vary between the two seasons of the Olympic games, with a confidence level estimated to be approximately 90%.

An 1% increase in *GDP Per Capita* results in a 1.087% increase in the count of Olympic medals for the Summer Olympics and a 2.414% increase for the Winter Olympics, with all other factors held constant. Furthermore, a 1% increase in the *Population* is expected to yield a 0.650% increase in the count of Summer Olympic medals and a 0.530% increase in the count of Winter Olympic medals, while holding all other factors constant. Participating in the Summer Olympics under a *Closed Autocracy* leads to an $e^{0.708} - 1 = 2.0299 - 1 = 1.0299 \approx 103\%$ increase in Olympic medal count compared to *Liberal Democracy* and *Electoral Democracy*, while in the Winter Olympics, this increase is higher at $e^{3.344} - 1 = 28.3322 - 1 = 27.3322 \approx 2733\%$, holding all factors constant. For an *Electoral Autocracy*, the expected number of Summer Olympic medals won is estimated to increase by $e^{0.804} - 1 = 2.2345 - 1 = 1.2345 \approx 123\%$ and the Winter Olympic medal count is projected to increase by $e^{2.058} - 1 = 6.3028 - 1 = 5.3028 \approx 530\%$ compared to both the *Liberal Democracy* and *Electoral Democracy*, with all factors held constant.

5.5.3 Year

This section explores variations in the influences of socio-economic factors over time. It focuses on the analysis of the first and last two available Olympic Games editions in the data set, encompassing both the Summer and Winter Olympics. The findings are outlined in Table 5.6.

Table 5.6. Poisson regression: Year

Variable	2000 & 2002	2014 & 2016	95% CI 00 & 02	95% CI 14 & 16
Intercept	-22.930***	-22.230***	[-24.763,-21.186]	[-24.239,-20.312]
	(0.912)	(0.998)		
ln(GDP Per Capita)	1.082**	1.079***	[0.923,1.247]	[0.925,1.244]
	(0.081)	(0.083)		
ln(Population)	0.636***	0.548***	[0.583,0.698]	[0.492,0.609]
	(0.027)	(0.028)		
ln(Gini Coefficient)	-1.970***	-1.756***	[-2.475,-1.483]	[-2.289, -1.254]
	(0.252)	(0.263)		
Closed Autocracy	0.997***	0.991***	[0.616,1.385]	[0.663,1.327]
	(0.197)	(0.170)		
Electoral Autocracy	0.910***	0.501***	[0.691, 1.132]	[0.295,0.727]
	(0.113)	(0.110)		
Healthcare Expenditures	0.022	0.080***	[-0.019, 0.062]	[0.049,0.111]
	(0.021)	(0.016)		
Food Supply Surplus	-0.329	-0.100	[-0.843,0.185]	[-0.622,0.438]
	(0.262)	(0.267)		
Host	0.831***	1.385***	[0.565,1.086]	[1.043,1.719]
	(0.131)	(0.170)		
Past Host	-0.693	0.818***	[-2.492,0.451]	[0.624,1.012]
	(0.713)	(0.099)		
Maximum Medals	0.002***	0.002***	[0.0018,0.0023]	[0.0015,0.0019]
	(0.001)	(0.001)		
N	158	169	158	169

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Upon examining the 95% confidence intervals, only *Past Host* exhibits non-overlapping intervals. However, it's important to note that its coefficient is not statistically significant in the context of the Poisson regression for the years 2000 and 2002, with a significance level of $p < 0.05$. Consequently, the null hypothesis cannot be rejected, indicating that the coefficient of *Past Host* for the years 2000 and 2002 is equal to zero, exerting no influence on the number of Olympic medals won. Conversely, being a *Past Host* in the years 2014 and 2016 resulted in an $e^{0.818} - 1 = 2.2660 - 1 = 1.2660 \approx 127\%$ increase in the number of Olympic medals won, holding all other factors constant. The *Healthcare Expenditures* does show overlap in the 95% confidence intervals; however, the coefficient estimate in the years 2000 and 2002 is not significant at the level of $p < 0.05$, making the constructed 95% confidence interval insignificant too. In these years, the null hypothesis cannot be rejected, indicating that the influence of *Healthcare Expenditures* is effectively zero. Conversely, in the years 2014 and 2016, for every 1% increase in *Healthcare Expenditures*, the Olympic medal count is projected to increase by $e^{0.080} - 1 = 1.0833 - 1 = 0.0833 \approx 8.3\%$, while holding all other factors constant.

Chapter 6

Conclusion & Discussion

6.1 Conclusion

This thesis investigates the relationship between a country's Olympic medal count and its socio-economic factors, addressing four research questions. The first research question explores the influential socio-economic factors, the measurement of the Olympic medals, and the methodologies employed in existing literature. The literature review confirms the favorable influences of socio-economic factors such as population, GDP/GDP per capita, hosting the Olympics, and political regime on the Olympic medals won. These factors have been integrated into the analysis, with population representing the number of residents, GDP per capita measured in international dollars (2017), and regime classification as either closed/electoral autocracy or electoral/liberal democracy. Furthermore, the analysis takes into account the influence of past Olympic hosting. In the existing literature, Olympic success is typically assessed based on the total number of Olympic medals earned, with the majority of studies favoring the Olympic medals measured as integers rather than shares. In this approach, gold, silver, and bronze medals are frequently aggregated. To explore this relationship, the literature employed a range of machine learning models, including Ordinary Least Squares (OLS), Tobit, Poisson, and Negative Binomial (NGB) regression.

The second research question delved into identifying the most effective machine learning model for analyzing the influences of socio-economic factors on the Olympic medal count. In addition to the machine learning models found in the existing literature, this study also incorporated the Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoosting). Furthermore, novel socio-economic factors were introduced, encompassing income inequality, healthcare expenditures, technological development, and nutrition. These factors were evaluated through metrics such as the Gini coefficient, healthcare expenditures as a percentage of GDP, internet user proportions, and the supply versus demand of food in a country.

Among the machine learning models utilized in prior research, the Poisson regression outperformed the others in explaining the variance of the Olympic medal count, achieving the highest R-squared (R^2) values of 0.70 in cross-validation and 0.61 in out-of-sample scenarios. The R^2 values for the OLS regression were 0.45 in cross-validation and 0.40 in the out-of-sample scenario, while the NGB regression achieved R^2 values of 0.67 in cross-validation and 0.31 in out-of-sample predictions. Furthermore, the Poisson regression exhibited the lowest Root Mean Square Error (RMSE) and the highest predictive accuracy, with an RMSE of 7.38 in cross-validation and 6.51 for out-of-sample predictions. In contrast, the OLS, Tobit, and NGB regressions yielded RMSE values of 10.82, 16.71, and 8.92, respectively, in cross-validation, and 8.10, 15.77, and 8.67 for out-of-sample predictions. The OLS and Tobit regression models failed to meet their linearity and homoscedasticity assumptions, and in the case of OLS, the independence of errors assumption was also not satisfied. Both the Poisson and NGB regression models did not conform to the Poisson distribution assumption, and the equality of mean and variance was not met for the Poisson regression either. However, the advantage of relaxing the equality of mean and variance for the NGB regression is considered less significant than the lack of generalizability, as demonstrated by the decrease in R^2 between cross-validation and the out-of-sample scenario, and the detailed examination of the cross-validation. Therefore, the Poisson regression is the preferred choice among the regression models. In the Poisson regression, only the *Electoral Democracy* variable was removed based on Sequential Backward Elimination (SBE). Consequently, all the coefficients of the other socio-economic factors, including those previously unconsidered, were found to be statistically significant at the level of $p < 0.05$.

The ensemble models, which were not taken into account in the prior literature, demonstrated more favorable performance metrics compared to the earlier regression models. XGBoosting exhibited the highest R^2 and the lowest RMSE in cross-validation, with values of 0.80 and 6.18, respectively. Among the ensemble models, Random Forest performed the least effectively during cross-validation, achieving an R^2 of 0.71 and an RMSE of 7.67. However, in the out-of-sample scenario, Random Forest excelled with an R^2 of 0.84 and the lowest RMSE of 4.20. In contrast, XGBoosting and Gradient Boosting yielded R^2 values of 0.81 and 0.79, with RMSE values of 4.59 and 4.77, respectively. Despite the superior performance metrics of Random Forest in out-of-sample predictions, the consistency advantage of XGBoosting in terms of R^2 within the cross-validation was notably more substantial. As a result, XGBoosting emerged as the most effective model for predicting Olympic medal counts based on socio-economic factors, with an average deviation of 4.59 Olympic medals from the actual achievements. Notably, the variable *Healthcare Expenditures*, which had not been considered in previous literature, ranked as the second most influential independent variable in the construction of the Random Forest model and the third most influential in both Gradient Boosting and XGBoosting. However, the relative importance of independent variables does not provide a precise understanding of the relationship between socio-economic factors and Olympic medal counts. Consequently, the Poisson regression model is preferred for explaining this relationship due to its ability to provide interpretable coefficients.

The third research question focuses on interpreting the influence of socio-economic factors on the Olympic medal count and evaluating the extent of the predictors' influences. This was achieved using the Poisson regression model on the full set of observations. The model's interpretation applies to athletes of both genders, across Summer and Winter Olympics, spanning all editions from 2000 to 2016.

Almost all socio-economic variables, except for *Electoral democracy*, displayed a significant influences on the Olympic medal count, yielding coefficients with p-values lower than 0.05. For instance, a 1% increase in *GDP Per Capita* and *Population* corresponded to approximately 1.113% and 0.627% increases in Olympic medals. Conversely, a 1% increase in the *Gini Coefficient*, representing income inequality, resulted in about a 1.383% reduction in the expected number of Olympic medals. Being in a *Closed Autocracy* or *Electoral Autocracy* correlated with roughly 107% and 104% increases in the number of Olympic medals won, compared to living in both the *Liberal Democracy* and *Electoral Democracy.* Furthermore, a 1% increase in *Healthcare Expenditures* led to a 3.6% increase in Olympic medals for a country. For every unit increase in the *Food Supply Surplus*, the number of Olympic medals won decreased by 34%. Hosting the Olympic Games had a substantial positive influence, resulting in a significant increase of 133% in the number of Olympic medals won. Additionally, having hosted the Olympic Games in the previous four years was associated with an increase of 30% in the expected number of Olympic medals won. These individual interpretations of the independent variables assume that all other factors in the analysis remain constant.

The last research question aimed to distinguish the influence of socio-economic variables with respect to gender, season, and year, employing a Poisson regression model. To achieve this, 95% confidence intervals were established for the coefficients within the regression. These intervals facilitated the identification of overlapping or non-overlapping 95% confidence intervals for the same independent variable in different contexts. This analysis enabled to infer that when two 95% confidence intervals do not overlap, there is approximately a 90% probability that the true coefficients of the independent variable differ in another context.

Female athletes exhibited stronger influences, both positive and negative. For instance, a 1% increase in the *Population* led to a 0.563% increase in expected Olympic medals for male athletes, while female athletes saw a 0.679% increase. When the *Gini Coefficient* increased by 1%, indicating higher income inequality, male athletes experienced a 1.162% decrease in Olympic medals won, while female athletes faced an even greater 1.787% reduction. These individual interpretations of the independent variables are made under the assumption that all other factors in the analysis remain unchanged.

A 1% increase in *GDP Per Capita* corresponds to a 1.087% increase in expected Olympic medals for the Summer Olympics and a 2.414% increase for the Winter Olympics. A 1% *Population* increase leads to a 0.650% rise in Olympic medals for the Summer Olympics and a 0.530% increase for the Winter Olympics. *Closed Autocracies* boost Olympic medal counts by 103% for the Summer Olympics and a staggering 2733% for the Winter Olympics compared to both the *Liberal Democracy* and *Electoral Democracy. Electoral Autocracies* also have significant influences, increasing medal counts by 123% for the Summer Olympics and 530% for the Winter Olympics, compared to inhabitants from both the *Liberal Democracy* and *Electoral Democracy.* These individual interpretations of the independent variables are based on the assumption that all other factors in the analysis remain unchanged.

When contrasting the early Olympic editions (2000 and 2002) with the more recent ones (2014 and 2016), a significant contrast becomes apparent regarding the influences of *Past Host* status and the role of *Healthcare Expenditures.* In the earlier years, *Past Host* status and *Healthcare Expenditures* had no influence, as evidenced by the coefficients lacking statistical significance at the p ¡ 0.05 level. However, in recent editions, this influence became pronounced, resulting in a 127% increase in Olympic medal count for countries that had previously hosted the Olympic Games, assuming all other factors remained constant. Furthermore, for every 1% increase in *Healthcare Expenditures*, the number of Olympic medals won saw an 8.3% boost in the more recent editions, holding all other factors constant.

6.2 Discussion

6.2.1 Implications

The study's outcomes align with its research objective. Established socio-economic factors showed the expected positive influence, and previously unexamined variables revealed significant influences. By comparing different models, this research effectively identified these influences and evaluated their additional contributions.

The study had both similarities and differences compared to prior research. The study confirmed Poisson regression's superiority over Tobit and Negative Binomial regression in the goodness of fit, aligning with (Lui & Suen, 2008). Like Rewilak (2021) his study, this research examined the influence of socio-economic factors on both male and female athletes. Both studies revealed similar findings, indicating that the host influence affects both genders similarly and emphasizing the greater influence of population size on female athletes' Olympic medal counts. Furthermore, this study introduced a distinctive perspective by uncovering a more substantial negative influence of income inequality on female athletes' Olympic medal counts, enriching the existing literature.

Johnson and Ali (2004) stressed the financial aspect of the Winter Olympics and the population for the Summer Olympics, aligning with this study. This research, however, introduced a unique perspective by highlighting a stronger host advantage in the Summer Olympics, not discussed in Johnson and Ali (2004) their work. Additionally, this study uncovered the heightened influence of residing in an autocracy as opposed to a democracy in the context of the Winter Olympics, providing new insights. Analyzing changes over time, it found that the positive influences of the population did not differ over time, contrary to the diminishing positive influences found by Noland and Stahler (2017). Furthermore, the influence of prior Olympic Games hosting and a higher allocation of GDP to healthcare expenditures on Olympic medal counts did not manifest in earlier years but became evident in more recent years.

The enhanced Poisson regression model, with its inclusion of new variables, provides valuable insights for National Olympic Committees and governments. It deepens the understanding of the link between Olympic succes and socio-economic factors, enabling informed actions. This model quantifies the necessary adjustments to achieve specific increases in Olympic medal counts. It also guides the emphasis on socio-economic variables for those targeting factors like gender or season, making it a useful tool. Improving the relative Olympic performance of female athletes compared to male athletes could be achieved by mitigating income inequality and promoting population growth. Over time, the role of healthcare expenditures has become increasingly apparent, emerging as an influential independent variable in ensemble models. This underscores the importance of prioritizing healthcare improvement for enhancing Olympic performance for athletes of both genders and across various Olympic seasons. This research also benefits sports organizations, bookmakers, and gamblers alike. The selection of the XGBoosting model, driven by its predictive capabilities, allows bookmakers and gamblers to make more accurate predictions, enhancing their odds and betting strategies. Furthermore, marketing professionals have the opportunity to adjust their marketing campaign and target audience following the predicted successful nations.

6.2.2 Limitations

While this study has provided valuable insights into the relationship between socio-economic factors and Olympic medal success, it is essential to recognize its limitations. Access to additional data would be preferable, but it is not always available. Currently, healthcare expenditures serve as an indicator of a country's health prioritization. However, having a variable reflecting targeted sports investments would be more meaningful. Unfortunately, such data is often complex to obtain due to many external subsidies from different parties. Moreover, using the ease of maintaining a nutritious diet may better represent a country's food status compared to dividing the food supply by the minimum requirement, as it does not account for nutritional quality. Expanding the sample size over a longer period could enhance reliability, but socio-economic data availability is limited to a specific time frame.

The data cleaning process had some undesirable consequences. Removing countries with no available socio-economic data also meant removing some Olympic medal winners, albeit a small number due to the relatively small size of those countries. In an ideal scenario, it would be preferable to retain all Olympic medal winners. Excluding liberal democracy, which was strongly correlated with GDP Per Capita, affected the interpretation of the remaining regime categories. In the Sequential Backward Elimination (SBE) process, electoral democracy was consistently removed from all regression models. In this instance, liberal and electoral democracy were manually designated as the reference category, rather than allowing the model to select it. This manual selection might explain the notably high coefficient for the closed autocracy regime in Table 5.5 for the Winter Olympics.

In the Tobit regression, obtaining R^2 values was not possible due to the censored nature of the data. This limitation made it more challenging to make consistent comparisons across all methods. However, it was evident that the RMSE for both cross-validation and out-of-sample predictions was significantly higher for Tobit compared to Poisson, clearly distinguishing their performance. The interpretation of the Poisson regression model has several limitations. First, the assumptions of Poisson distribution and equal mean variance were not fully met, despite some resemblance in the distribution of Olympic medal winners. Although the performance metrics for the Poisson regression were the most favorable and all coefficients exhibited high significance levels, not all assumptions were satisfied, potentially leading to biased parameter estimates. Furthermore, the Poisson models were applied to different contexts using distinct data sets with varying sample sizes. Smaller sample sizes in some cases could introduce biases into the parameter estimates. Consequently, the comparison across different subsets was not always as straightforward as it would have been with consistent sample sizes. Finally, the assessment of model accuracy was relative, and the best-performing models were chosen. However, it is worth noting that the average deviation of 4.59 Olympic medals from the actual achievements in XGBoosting, while promising, falls short of perfection and could potentially be further improved for greater precision.

6.2.3 Further research

In terms of future research, there is room to expand the scope and deepen the understanding of the factors influencing the Olympic medal count. The analysis could benefit from the inclusion of additional socio-economic variables that might have an influences on the number of Olympic medals. Moreover, it may be valuable to explore how altering the measurement of specific socio-economic factors affects the extent of their influence. Furthermore, the application of alternative machine learning models could offer valuable insights, enhancing both explanatory and predictive modeling in this context.

In addition to exploring variations in socio-economic influence based on gender, season, and year, it would be worthwhile to investigate the influence of an athlete's specific sport or other more specific contexts. This analysis can reveal how socio-economic factors may differ based on the unique characteristics of each sport or context.

Finally, the XGBoosting model developed in this study could find practical application in making predictions for the 2024 Olympic Games in Paris. To achieve this, socio-economic data from the year 2023 would be required, even though the year is ongoing. The availability of such data would allow the model to generate predictions that could be valuable for external stakeholders and decision-makers.

References

Andreff, W. (2001). The correlation between economic underdevelopment and sport. *European Sport Management Quarterly*, *1*(4), 251–279.

Balmer, N. J., Nevill, A. M. & Williams, A. M. (2003). Modelling home advantage in the summer olympic games. *Journal of sports sciences*, *21*(6), 469–478.

Bernard, A. B. & Busse, M. R. (2004). Who wins the olympic games: Economic resources and medal totals. *Review of economics and statistics*, *86*(1), 413–417.

Berrar, D. et al. (2019). *Cross-validation.*

Bian, X. et al. (2005). Predicting olympic medal counts: The effects of economic development on olympic performance. *The park place economist*, *13*(1), 37–44.

Buttrick, N. R. & Oishi, S. (2017). The psychological consequences of income inequality. *Social and Personality Psychology Compass*, *11*(3), e12304.

Celik, O. B. & Gius, M. (2014). Estimating the determinants of summer olympic game performance. *International Journal of Applied Economics*, *11*(1), 39–47.

Chai, T. & Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae). *Geoscientific model development discussions*, *7*(1), 1525–1534.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Coxe, S., West, S. G. & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of personality assessment*, *91*(2), 121–136.

Davis, J. A. (2012). *The olympic games effect: How sports marketing builds strong brands.* John Wiley & Sons.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1–15).

Donald, W. B. (1972). Olympic games competition: structural correlates of national success. *International Journal of Comparative Sociology*, *13*, 186.

Elling, A., Van Hilvoorde, I. & Van Den Dool, R. (2014). Creating or awakening national pride through sporting success: A longitudinal study on macro effects in the netherlands. *International review for the sociology of sport*, *49*(2), 129–151.

Emrich, E., Klein, M., Pitsch, W., Pierdzioch, C. et al. (2012). On the determinants of sporting success–a note on the olympic games. *Economics Bulletin*, *32*(3), 1890–1901.

Grimes, A. R., Kelly, W. J. & Rubin, P. H. (1974). A socioeconomic model of national olympic performance. *Social science quarterly*, 777–783.

Grix, J. & Carmichael, F. (2012). Why do governments invest in elite sport? a polemic. *International journal of sport policy and politics*, *4*(1), 73–90.

Groeneveld, R. A. & Meeden, G. (1984). Measuring skewness and kurtosis. *Journal of the Royal Statistical Society Series D: The Statistician*, *33*(4), 391–399.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, *63*(4), 308–319.

Haake, S. J. (2009). The impact of technology on sporting performance in olympic sports. *Journal of Sports Sciences*, *27*(13), 1421–1431.

Haut, J., Prohl, R. & Emrich, E. (2016). Nothing but medals? attitudes towards the importance of olympic success. *International review for the sociology of sport*, *51*(3), 332–348.

Hazra, A. (2017). Using the confidence interval confidently. *Journal of thoracic disease*, *9*(10), 4125.

Hoffmann, R., Ging, L. C. & Ramasamy, B. (2004). Olympic success and asean countries: Economic analysis and policy implications. *Journal of Sports Economics*, *5*(3), 262–276.

IOC. (n.d.). *History of the ioc.* Retrieved from `https://olympics.com/ioc/history` (Accessed: October 13, 2023)

IOC. (2012). *London 2012 facts figures.* Retrieved from `https://stillmed.olympic.org/Documents/Reference_documents_Factsheets/London_2012_Facts_and_Figures-eng.pdf` (Accessed: October 13, 2023)

IOC. (2020). *Tokyo 2020 organising committee publishes final balanced budget.* Retrieved from `https://olympics.com/ioc/news/tokyo-2020-organising-committee-publishes-final-balanced-budget` (Accessed: October 13, 2023)

IOC. (2021). *Olympic games tokyo 2020 watched by more than 3 billion people.* Retrieved from `https://olympics.com/ioc/news/olympic-games-tokyo-2020-watched-by-more-than-3-billion-people` (Accessed: October 13, 2023)

Jakovljevic, M., Timofeyev, Y., Ekkert, N. V., Fedorova, J. V., Skvirskaya, G., Bolevich, S. & Reshetnikov, V. A. (2019). The impact of health expenditures on public health in brics nations. *Journal of sport and health science*, *8*(6), 516.

James, G., Witten, D., Hastie, T., Tibshirani, R. et al. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Johnson, D. K. & Ali, A. (2000). Coming to play or coming to win: Participation and success at the olympic games. *Wellesley College Dept. of Economics Working Paper*(2000-10).

Johnson, D. K. & Ali, A. (2004). A tale of two seasons: participation and medal counts at the summer and winter olympic games. *Social science quarterly*, *85*(4), 974–993.

Kamil, M. L. (2004). The current state of quantitative research. *Reading Research Quarterly*, *39*(1), 100–107.

Kanin, D. B. (2019). *A political history of the olympic games.* Routledge.

Kuper, G. H. & Sterken, E. (2001). Olympic participation and performance since 1896. *Available at SSRN 274295*.

Leeds, E. M. & Leeds, M. A. (2012). Gold, silver, and bronze: Determining national success in men's and women's summer olympic events. *Jahrbücher für Nationalökonomie und Statistik*, *232*(3), 279–292.

Legaz-Arrese, A., Moliner-Urdiales, D. & Munguía-Izquierdo, D. (2013). Home advantage and sports performance: evidence, causes and psychological implications. *Universitas Psychologica*, *12*(3), 933–943.

Liebermann, D. G., Katz, L., Hughes, M. D., Bartlett, R. M., McClements, J. & Franks, I. M. (2002). Advances in the application of information technology to sport performance. *Journal of sports sciences*, *20*(10), 755–769.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), 14–23.

Long, R. G. (2008). The crux of the method: assumptions in ordinary least squares and logistic regression. *Psychological reports*, *103*(2), 431–434.

Lui, H.-K. & Suen, W. (2008). Men, money, and medals: An econometric analysis of the olympic games. *Pacific Economic Review*, *13*(1), 1–16.

Mackenzie, M. (2003). From athens to berlin: The 1936 olympics and leni riefenstahl's olympia. *Critical Inquiry*, *29*(2), 302–336.

Mao, K. Z. (2004). Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *34*(1), 629–634.

Maughan, R., Burke, L. M. & Coyle, E. F. (2004). *Food, nutrition and sports performance ii: the international olympic committee consensus on sports nutrition.* Routledge.

McBee, M. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the tobit model. *Gifted Child Quarterly*, *54*(4), 314–320.

Miles, J. (2005). R-squared, adjusted r-squared. *Encyclopedia of statistics in behavioral science.*

Montgomery, D. C., Peck, E. A. & Vining, G. G. (2021). *Introduction to linear regression analysis.* John Wiley & Sons.

Natekin, A. & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.

Noland, M. & Stahler, K. (2017). An old boys club no more: pluralism in participation and performance at the olympic games. *Journal of Sports Economics*, *18*(5), 506–536.

OurWorldinData. (2019). *Total healthcare expenditure as a share of gdp, 2019.* Retrieved from `https://ourworldindata.org/grapher/total-healthcare-expenditure-gdp` (Accessed: October 13, 2023)

OurWorldinData. (2020). *Daily caloric supply derived from carbohydrates, protein and fat.* Retrieved from `https://ourworldindata.org/grapher/daily-caloric-supply-derived-from-carbohydrates-protein-and-fat,` (Accessed: October 13, 2023)

OurWorldinData. (2021a). *Economic inequality - gini index.* Retrieved from `https://ourworldindata.org/grapher/economic%20inequality-gini-index,` (Accessed: October 13, 2023)

OurWorldinData. (2021b). *Gdp per capita.* Retrieved from `https://ourworldindata.org/grapher/gdp-per-capita-worldbank?tab=table,` (Accessed: October 13, 2023)

OurWorldinData. (2021c). *Minimum daily requirement of calories.* Retrieved from `https://ourworldindata.org/grapher/minimum-requirement-calories?time=earliest,` (Accessed: October 13, 2023)

OurWorldinData. (2021d). *Share of the population using the internet.* Retrieved from `https://ourworldindata.org/grapher/share-of-individuals-using-the-internet,` (Accessed: October 13, 2023)

OurWorldinData. (2022a). *Political regime.* Retrieved from `https://ourworldindata.org/grapher/political-regime?time=1871,` (Accessed: October 13, 2023)

OurWorldinData. (2022b). *Population.* Retrieved from `https://ourworldindata.org/grapher/population-with-un-projections,` (Accessed: October 13, 2023)

Ratner, B. (2009). The correlation coefficient: Its values range between+ 1/- 1, or do they? *Journal of targeting, measurement and analysis for marketing*, *17*(2), 139–142.

Rewilak, J. (2021). The (non) determinants of olympic success. *Journal of sports economics*, *22*(5), 546–570.

Rgriffin. (2018). *120 years of olympic history: Athletes and results. kaggle.* Retrieved from `https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-andresults` (Accessed: October 13, 2023)

Sainani, K. L. (2014). Explanatory versus predictive modeling. *PM&R*, *6*(9), 841–844.

Scandizzo, P. L. & Pierleoni, M. R. (2018). Assessing the olympic games: The economic impact and beyond. *Journal of economic surveys*, *32*(3), 649–682.

Schervish, M. J. (1996). P values: what they are and what they are not. *The American Statistician*, *50*(3), 203–206.

Subramanian, S. V. & Kawachi, I. (2004). Income inequality and health: what have we learned so far? *Epidemiologic reviews*, *26*(1), 78–91.

Tsai, C.-L., Cai, Z. & Wu, X. (1998). The examination of residual plots. *Statistica Sinica*, 445–465.

Veal, A. J. (2016). Leisure, income inequality and the veblen effect: Cross-national analysis of leisure time and sport and cultural activity. *Leisure Studies*, *35*(2), 215–240.

Walsh, A. (1987). Teaching understanding and interpretation of logit regression. *Teaching sociology*, 178–183.

Wood, J. & Meng, S. (2021). The economic impacts of the 2018 winter olympics. *Tourism Economics*, *27*(7), 1303–1322.

Appendix A

Appendix A: Data

Table A.1. Variable overview of Olympic performance data set

Variable	**Type**	**Description**
ID	Integer	Unique number for each athlete
Name	String	Full name of athlete
Sex	String	Gender of athlete (Male or Female)
Age	Integer	Age of athlete
Height	Integer	Height of athlete
Weight	Integer	Weight of athlete
Team	String	Country represented by athlete
NOC	String	National Olympic Committee 3-letter code
Games	Integer + String	Year and season
Year	Integer	Year of Olympic Games
Season	String	Season of organizing (Winter or Summer)
City	String	Host city
Sport	String	The type of sport the athlete participated in
Event	String	Specific discipline within the sport
Medal	String	Medal won (Gold, Silver, Bronze, or NA)

Table A.2. Correlation matrix

	Total Medals	GDP Per Capita	Population	Closed Autocracy	Electoral Autocracy	Liberal Democracy	Electoral Democracy	Share Internet Users	Gini Coefficient	Healthcare Expenditures	Animal Protein	Plant Protein	Fat	Carbohydrates	Supply Excess	Host
Total Medals	X															
GDP Per Capita	0.30	X														
Population	0.29	-0.12	X													
Closed Autocracy	0.07	-0.19	0.34	X												
Electoral Autocracy	-0.09	-0.42	-0.08	-0.14	X											
Liberal Democracy	0.23	0.76	-0.13	-0.18	-0.50	X										
Electoral Democracy	-0.19	-0.33	0.06	-0.14	-0.38	-0.50	X									
Share Internet Users	0.26	0.77	-0.10	-0.13	0.37	0.62	-0.25	X								
Gini Coefficient	-0.10	-0.40	0.05	0.05	-0.04	-0.27	0.32	-0.39	X							
Healthcare Expenditures	0.32	0.57	-0.13	-0.21	-0.35	0,57	-0.17	0.55	-0.16	X						
Animal Protein	0.25	0.75	-0.16	-0.23	-0.43	0.68	-0.21	0.68	-0.37	0.51	X					
Plant Protein	-0.22	-0.68	0.17	0.26	0.46	-0.60	0.06	-0.59	0.20	-0.50	-0.79	X				
Fat	0.29	0.76	-0.12	-0.25	-0.42	0.68	-0.20	0.66	-0.35	0.59	0.82	-0.80	X			
Carbohydrates	-0.29	-0.78	0.12	0.24	0.42	-0.70	0.23	-0.68	0.38	-0.58	-0.88	0.76	-0.99	X		
Food Supply Surplus	0.30	0.63	-0.05	-0.08	-0.29	0.52	-0.24	0.58	-0.31	0.46	0.54	-0.41	0.58	-0.60	X	
Host	0.24	0.07	0.10	0.03	-0.04	0.06	-0.04	0.06	0.01	0.08	0.07	-0.05	0.09	-0.09	0.09	X
Past Host	0.24	0.12	0.14	0.06	-0.07	0.12	-0.09	0.12	-0.02	0.13	0.10	-0.06	0.14	-0.14	0.14	-0.16

Table A.3. Variable overview of socio-economic factors

Socio-economic Variable	**Measure**	**Observations**	**Time Span**
GDP Per Capita	GDP (in 2017 international dollars) / Population	6364	1990-2021
Population	Number of inhabitants	38.355	1950-2100
Political regime	Closed autocracy	30.766	1789-2022
	Electoral autocracy		
	Electoral democracy		
	Liberal democracy		
Income inequality	Gini index (0-1)	2125	1967-2021
Healthcare expenditures	% of GDP	3974	2000-2019
Technology	% of internet users in last three months	6570	1960-2021
Nutrition	Daily calorie supply (including nutrient composition)	3596	1961-2020
	Daily minimum calorie requirement	4972	2000-2021

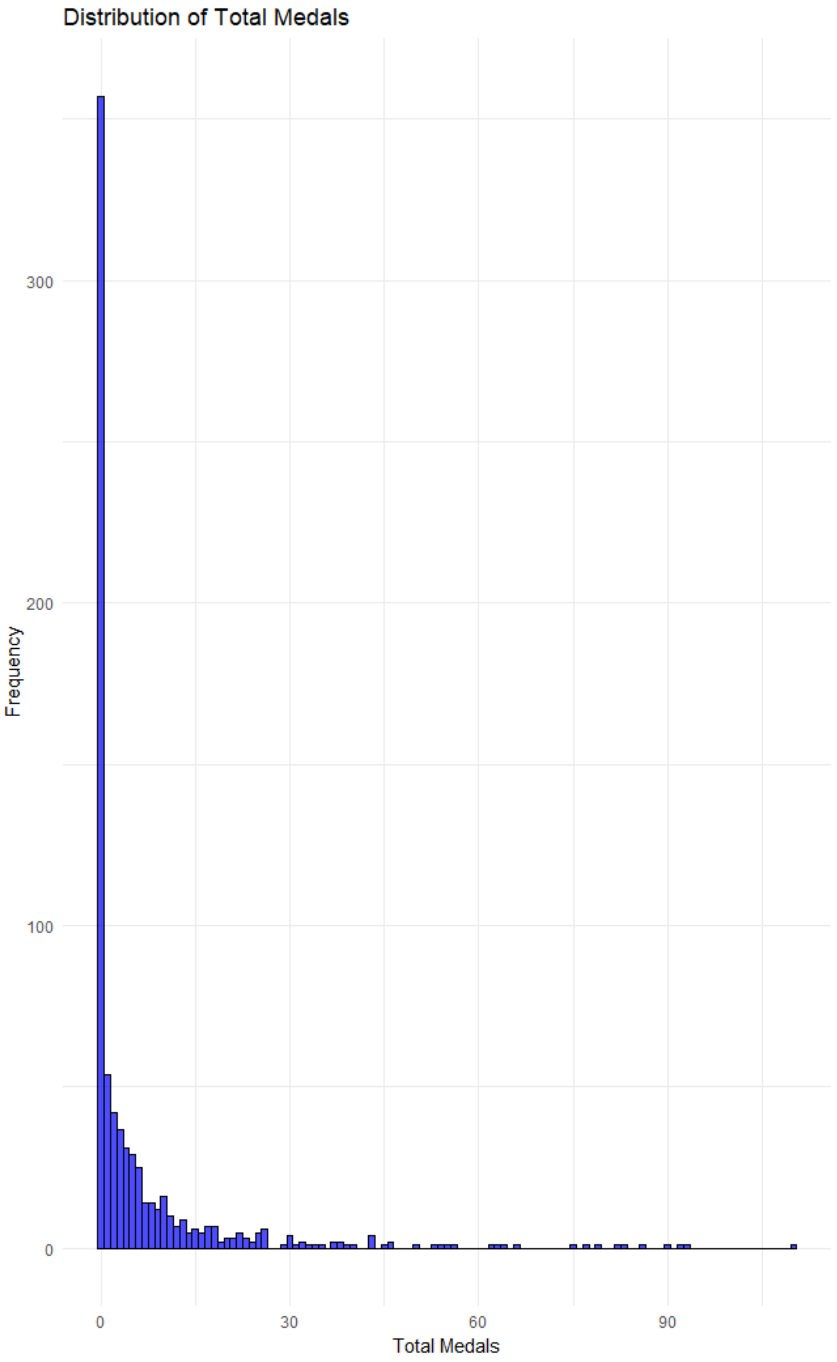

Figure A.1. Olympic medal distribution

Table A.4. Categorical variables frequency table

Variables	**Frequency**
Regime	
Closed Autocracy	36 (4.8%)
Electoral Autocracy	210 (28.0%)
Electoral Democracy	211 (28.1%)
Liberal Democracy	293 (39.1%)
Host / Past Host	
Yes	8 (1.1%) / 13 (1.7%)
No	742 (98.9%) / 737 (98.3%)

Appendix B

Appendix B: Results

Table B.1. Cross-validation results

Fold	OLS		Tobit		Poisson		NGB	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
1	0.58	9.14	NA	16.51	0.59	8.18	0.64	10.35
2	0.55	14.80	NA	18.78	0.76	7.35	0.69	10.35
3	0.38	11.29	NA	14.56	0.91	6.42	0.54	8.11
4	0.48	10.19	NA	17.81	0.55	5.02	0.65	10.31
5	0.22	13.25	NA	17.21	0.44	7.45	0.38	16.29
6	0.50	11.80	NA	16.31	0.67	6.75	0.60	7.95
7	0.42	10.45	NA	17.85	0.84	6.43	0.92	6.97
8	0.51	8.77	NA	15.51	0.71	10.64	0.71	4.92
9	0.48	10.02	NA	14.52	0.75	8.63	0.80	11.30
10	0.37	8.45	NA	17.47	0.79	6.94	0.81	5.99
Range	[0.22-0.58]	[8.45-14.80]	NA	[14.52-18.78]	[0.44-0.91]	[6.42-10.64]	[0.38-0.92]	[4.92-16.29]

Fold	Random Forest		Gradient Boosting		XGBoosting	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
1	0.64	7.66	0.47	11.18	0.59	5.56
2	0.44	12.18	0.79	4.60	0.72	11.46
3	0.74	9.14	0.73	4.57	0.70	5.77
4	0.64	6.95	0.89	7.28	0.94	3.10
5	0.61	7.14	0.86	8.57	0.83	4.61
6	0.90	7.25	0.85	5.14	0.85	6.73
7	0.77	7.21	0.90	4.29	0.72	8.71
8	0.68	9.97	0.68	4.81	0.90	3.58
9	0.92	4.21	0.64	10.05	0.81	6.49
10	0.74	5.25	0.47	7.28	0.87	3.75
Range	[0.44-0.90]	[4.21-12.18]	[0.47-0.90]	[4.29-11.18]	[0.59-0.94]	[3.10-11.46]

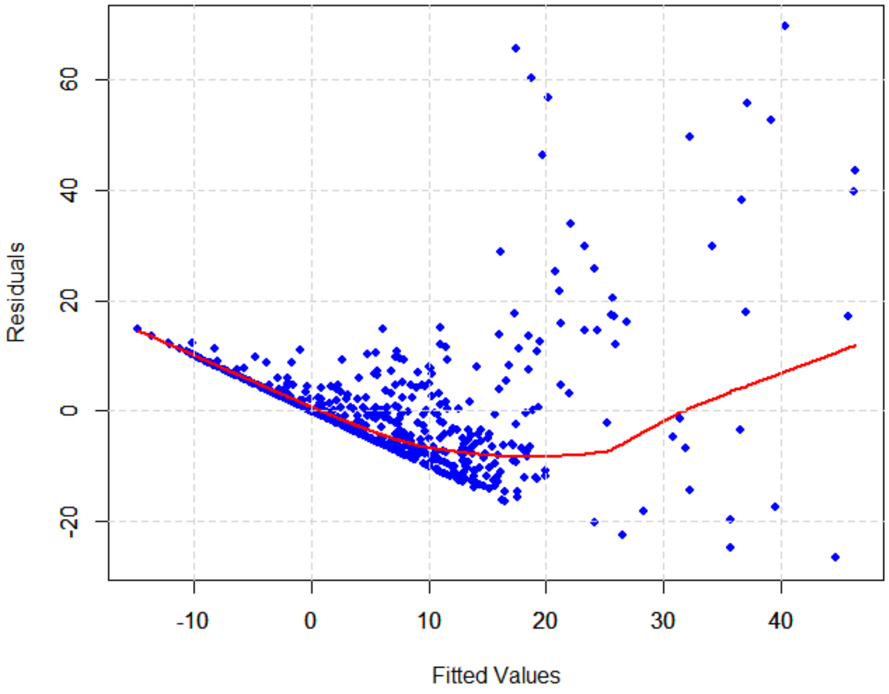

Figure B.1. Residual plot: OLS regression

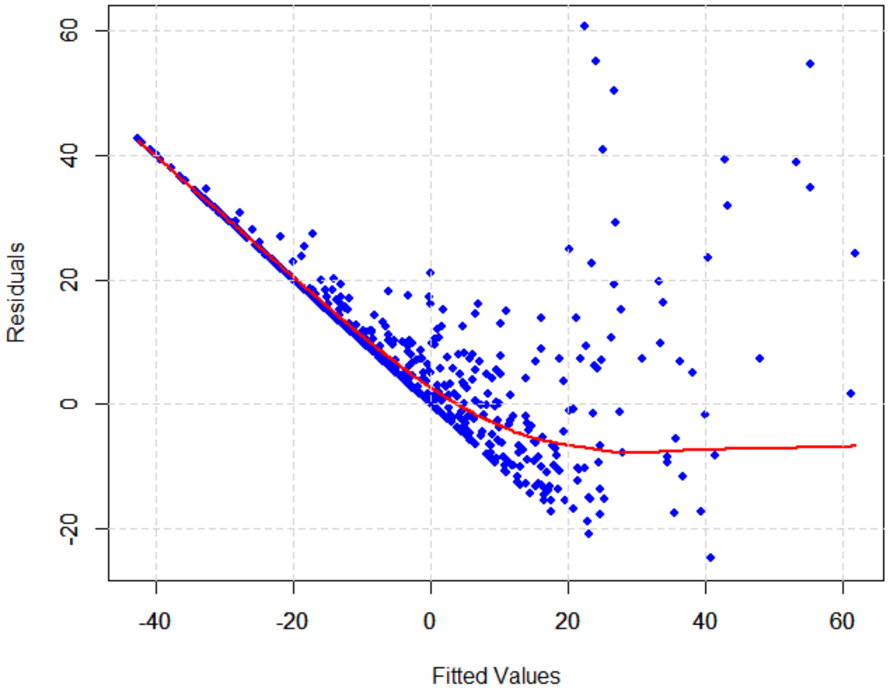

Figure B.2. Residual plot: Tobit regression

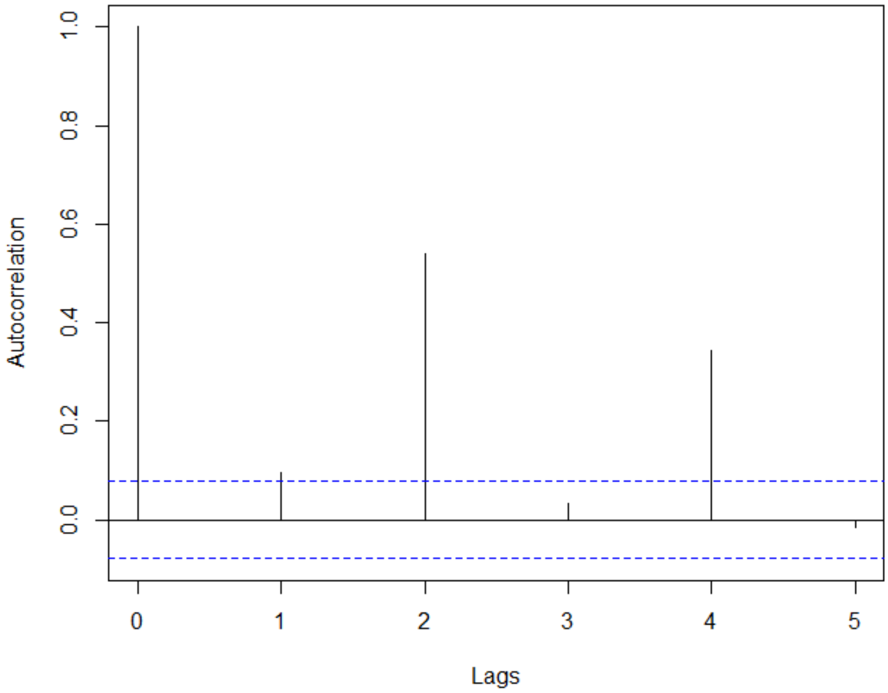

Figure B.3. Autocorrelation plot: OLS regression

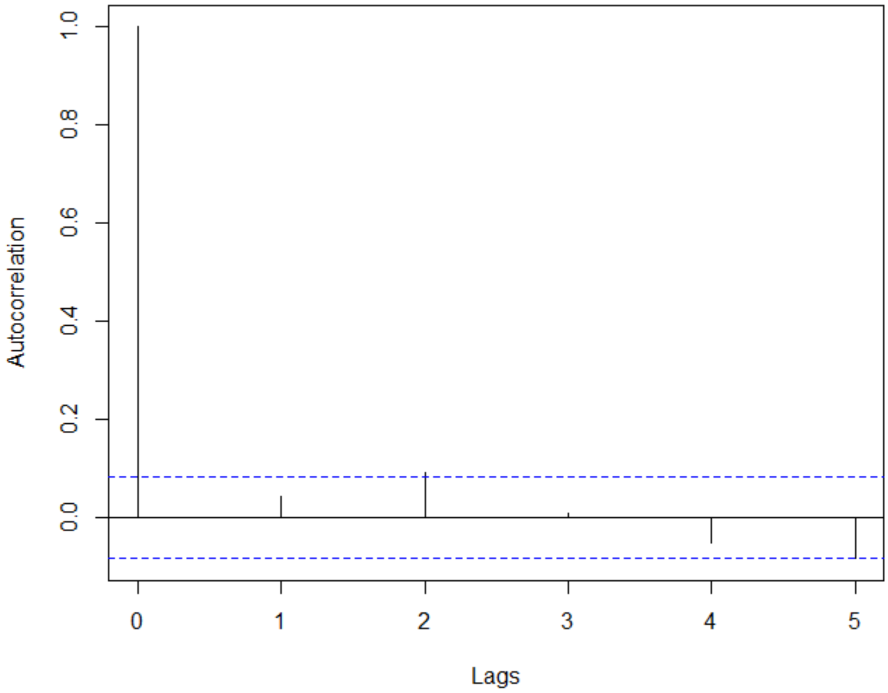

Figure B.4. Autocorrelation plot: Tobit regression

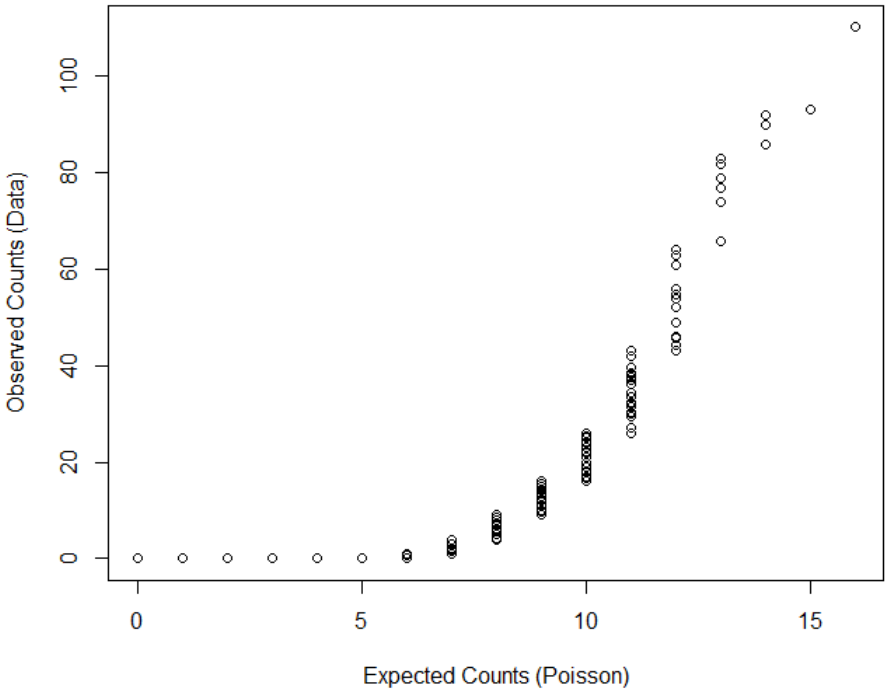

Figure B.5. Data distribution vs Poisson distribution

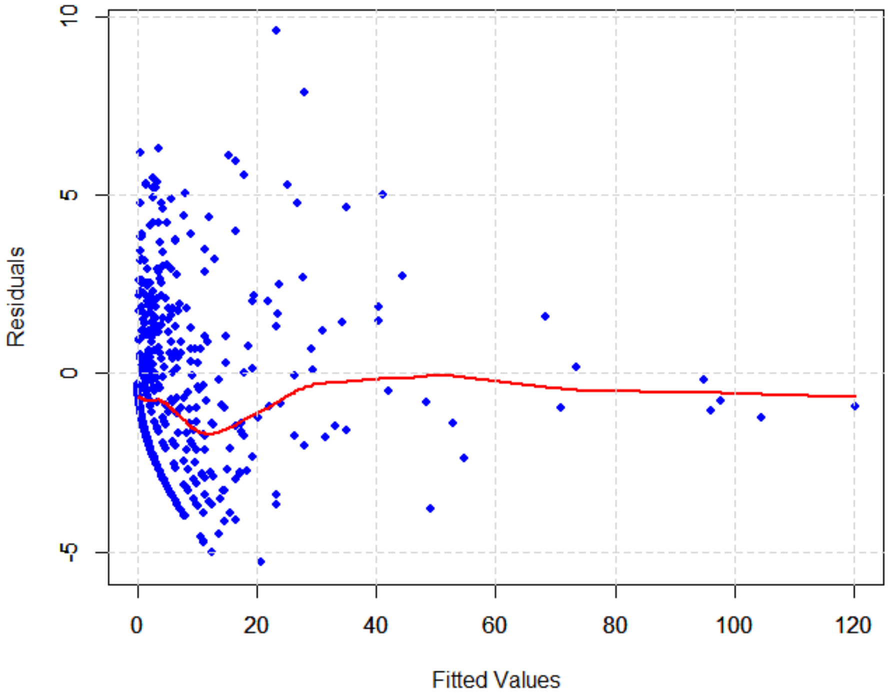

Figure B.6. Residual plot: Poisson regression

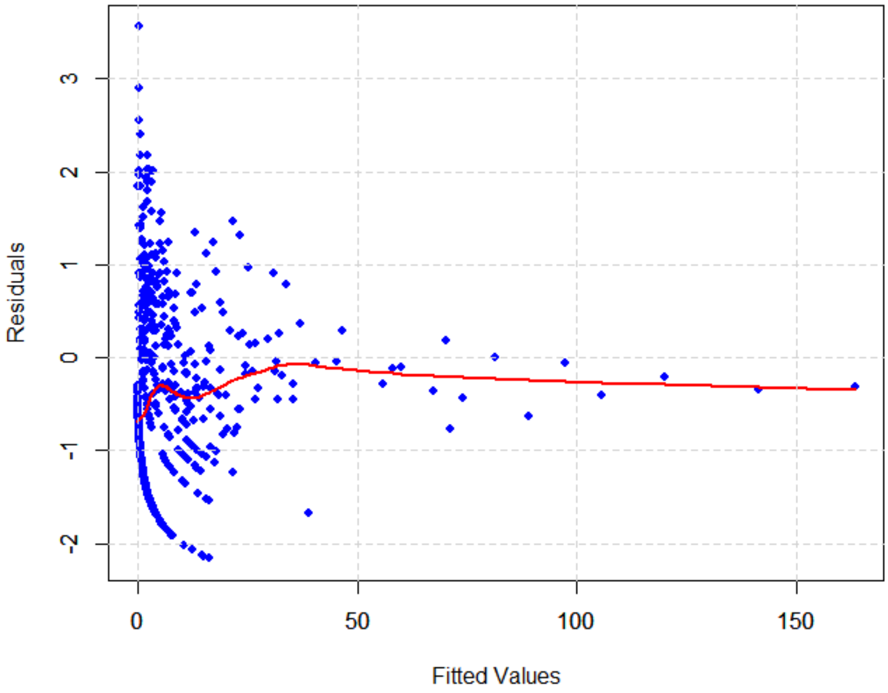

Figure B.7. Residual plot: NGB regression