Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2526908

The MCM Thesis of Team 2526908

Summary

This study focuses on using the **Prophet** model to predict the medal table for the 2028 Los Angeles Olympics, along with the corresponding prediction intervals. We then apply **KMeans** clustering to identify countries that show significant changes in their medal counts. Additionally, we use **Negative Binomial Regression** to examine the relationship between the number of events a country participates in and the number of medals they win, and assess this relationship using **Pearson Correlation** coefficients. We also apply **Network Analysis** to identify key sports for prominent countries in various Olympic Games. The study includes visualizations to uncover insights such as the impact of the host country, gender equality, and Japan's investment in the Olympics. We also utilize **Randomized Inference** to test and estimate the contribution of the "great coach" effect on medal counts. Finally, we compiled our findings into a memorandum to send to the International Olympic Committee (IOC), highlighting the key insights from our research.

Keywords: Olympic, Prophet, KMeans, Negative Binomial Regression, Pearson Correlation, Network Analysis, Randomize Inference, IOC

Contents

1 Introduction

1.1 Problem Background

For sports enthusiasts, predicting outcomes before a match is a common and enjoyable pastime. This excitement reaches new heights during the Summer Olympics, the world's largest sporting event, held every four years and bringing together nations from across the globe. By analyzing data from previous Olympic Games, we can predict the "medal table" for the upcoming event, identify each country's strengths in specific sports, and explore the impact of the "great coach" effect. These insights not only enhance our understanding of global sports dynamics but also uncover fascinating trends and stories behind the competition.

1.2 Problem Restatement

The following outlines key objectives to analyze factors influencing Olympic medal outcomes, focusing on predictive modeling, performance evaluation, and actionable insights for national Olympic committees:

1. **Medals count prediction:** Develop a model to predict medal counts, estimate uncertainty/precision, and evaluate its performance. Include confidence intervals to assess reliability.

2. **Performance evaluation:** Analyze predictions to identify countries performing better or worse than in 2024. Highlight potential first-time medalists.

3. **Events and medal counts:** Investigate the relationship between sports events and medal outcomes. Identify key sports for countries and assess how event choices impact results.

4. **"Great coach" effect:** Quantify the influence of exceptional coaching on results with examples from three countries.

5. **Insights for national committees:** Extract actionable insights from the model and prepare a memorandum for national Olympic committees.

1.3 Our Work

To tackle this challenge, our approach combines statistical analysis, hypothesis testing, and the Prophet time series model to develop a comprehensive framework for predicting Olympic outcomes. By thoroughly analyzing historical performance data of athletes and nations, we ensure a robust validation process to guarantee the reliability and applicability of the model across diverse Olympic events. Through this innovative perspective, we aim to provide valuable insights for athletes, coaches, and analysts, supporting strategic preparation and decision-making in the high-pressure context of the Olympic Games. The process and method of solving the problem are presented in Figure 1.

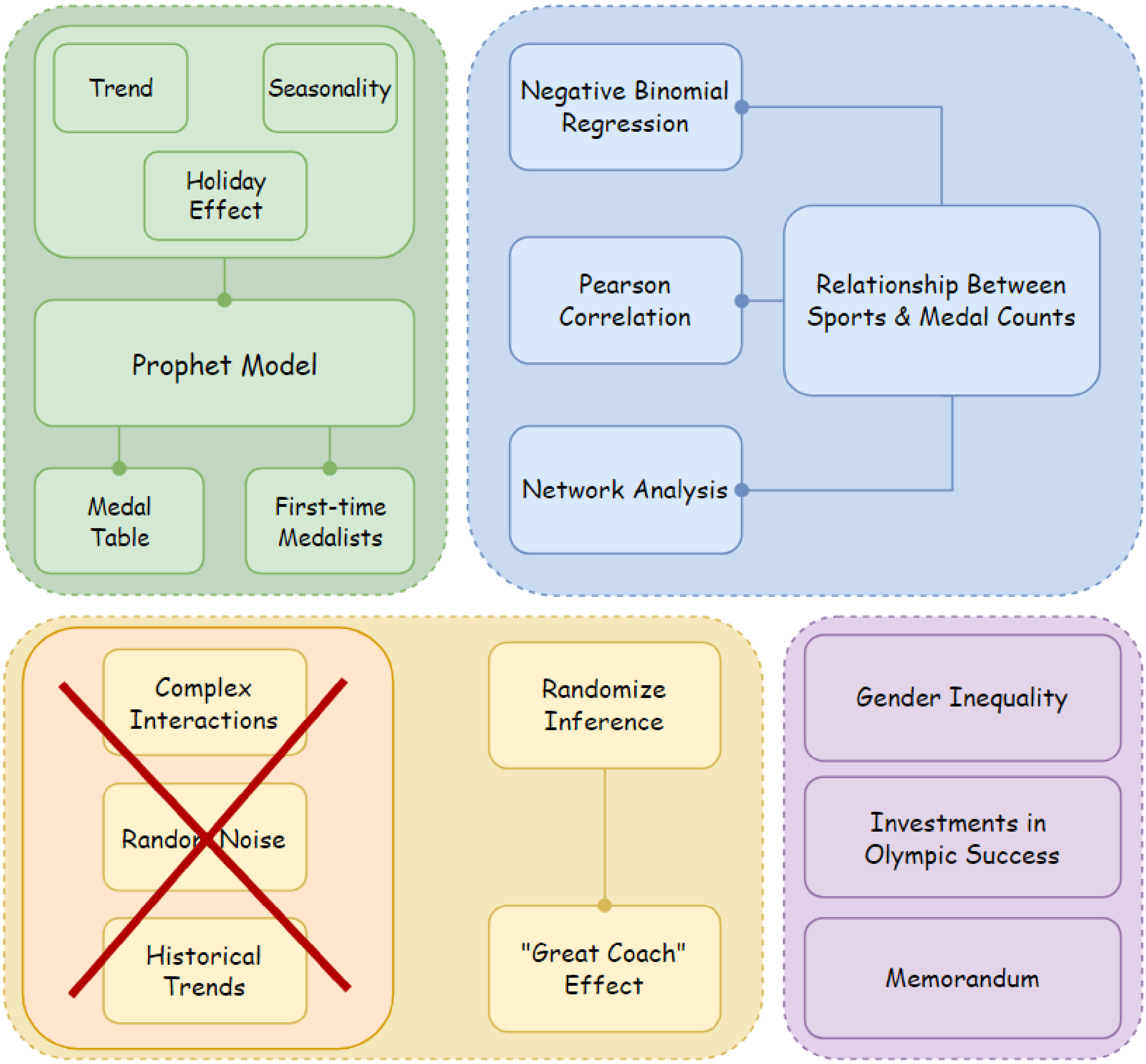

Figure 1: Framework of Our Work

2 Preparation for Modeling

2.1 Assumptions

- **Structural Consistency in Olympic Framework**
  We assume that the number of participating countries, the number of sports, the number of athletes per sport, the medal distribution ratio, and the competition format remain unchanged from 2024.

- **Coach Tenure Stability**

  We assume that coaches remain in charge throughout a single Olympic cycle and that each coach holds their position for at least 4 years.

- **Independent Long-Term Medal Trends**

We assume that the number of medals won by each country follows a long-term trend, and the performance of one country is not directly influenced by another.

- **Unbiased Coaching Efficacy**

  We assume that all coaches demonstrate their full capabilities and do not intentionally influence match outcomes.

2.2 Notations

Table 1: Notations Table

Symbol	Definition
$y(t)$	Predicted value at time t
$g(t)$	Trend component in Prophet model
$s(t)$	Seasonal component in Prophet model
$h(t)$	Holiday effect, describe the impact of "holiday" effect
ϵ_t	Residual, the error term between the predicted value and the actual value at time t
S	The number of changepoints at time s_j
$\boldsymbol{\delta}$	A vector of rate adjustents
k	Base rate
m	Offset parameter to connect the endpoints of the segments
$\mathbf{a}(t)$	A event indicator vector
$\boldsymbol{\gamma}$	A parameter describing changepoint correction,
P	The regular period we expect the time series to have
$\boldsymbol{\beta}$	A seasonality parameter vector.
$X(t)$	A matrix of seasonality
$Z(t)$	Describe the effect of the host country
D_{Host}	A set of host
$\boldsymbol{\kappa}$	A parameter vector adjusting the impact of hosting on the forecast
H_0	Null hypothesis
H_1	Alternative hypothesis
μ	Expected value of the total medals
β_0	Intercept of the Events variable.
β_1	Coefficient of the Events variable.
ρ	Population correlation coefficient
r	Pearson correlation coefficient
df	Degrees of freedom of the t-Student.
Φ_t	Cumulative probability of the t-Student distribution.
$P(n)$	Points through medals
$n_G,\ n_S,\ n_B$	Number of *Gold*, *Silver*, *Bronze* medals

2.3 Data Preprocessing

We edited the file `summerOly_programs.csv` by replacing bullet points with 0 (since these years were only for demonstration purposes and can be considered as having no medal events). Then, we corrected the country names with character errors in the file `summerOly_medal_counts.csv`, retrieved the variable *NOC_CODE* to use as the identifier for country names, and used this file as the main file. Additionally, we merged countries that were previously separated (e.g., *East Germany* and *West Germany*). For the other files, we integrated them into the main file as follows:

- `summerOly_hosts.csv`: Retrieved the binary variable *Host*, which indicates the host country.
- `summerOly_programs.csv`: Retrieved the variable *Events*, which represents the total number of events that the country participated in during that Olympic Games.
- `summerOly_athletes.csv`: Queried to extract specific sports columns, representing the number of medals that the country won in each sport, and added countries that had never won medals before (since our main file only contains countries with medals).

The final dataset consists of 80 features and 3216 samples, as shown in Figure 2.

	Rank	NOC_CODE	Gold	Silver	Bronze	Total	Year	Events	Host	3x3 Basketball	...	Taekwondo	Tennis	Trampoline Gymnastics	Trampolining	Triathlon	Tug-Of-War
0	1	USA	11	7	2	20	1896	16	0	0	...	0	0	0	0	0	0
1	2	GRE	10	18	19	47	1896	39	1	0	...	0	4	0	0	0	0
2	3	GER	6	5	2	13	1896	27	0	0	...	0	1	0	0	0	0
3	4	FRA	5	4	2	11	1896	18	0	0	...	0	0	0	0	0	0
4	5	GBR	2	3	2	7	1896	19	0	0	...	0	3	0	0	0	0

Figure 2: The first five rows of preprocessed data

3 Olympic Medal Insights: Who Will Shine in 2028?

3.1 Projection for the medal table

In sports competitions, results are often influenced by various factors such as athlete performance, the quality of recruits, team rotation cycles, and non-cyclical influences like a country's political situation. This study employs the **Prophet algorithm**[1], which decomposes forecasts into three components: a trend component $g(t)$, a seasonality component $s(t)$, and a holiday effect $h(t)$. The forecasted value $y(t)$ is expressed as:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t, \tag{1}$$

where ϵ_t represents the residual or unexplained variation. Each of these components uses time t as a regressor.

The trend $g(t)$ models long-term changes and incorporates changepoints to handle shifts in growth rates. For example, the 2008 Olympics, hosted by China, led to a significant increase in medals, while short-term disruptions like the COVID-19 pandemic affected the Tokyo 2020 Olympics.

Suppose there are S changepoints at times $s_j,\ j = 1, \ldots, S$. We define a vector of rate adjustments $\boldsymbol{\delta} \in \mathbb{R}^S$, where δ_j represents the change in growth rate at time s_j. The growth rate at time t is given by $k + \sum_{j:t>s_j} \delta_j$, where k is the base rate. To express this compactly, let $\mathbf{a}(t) \in \{0,1\}^S$ be a vector defined as:

$$a_j(t) = \begin{cases} 1, & \text{if } t \geq s_j, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Thus, the growth rate at time t becomes $k + \mathbf{a}(t)^\top \boldsymbol{\delta}$. To ensure continuity of the trend at each changepoint, the offset parameter m is adjusted using $\gamma_j = -s_j \delta_j$. The trend component is then:

$$g(t) = (k + \mathbf{a}(t)^\top \boldsymbol{\delta})t + \big(m + \mathbf{a}(t)^\top \boldsymbol{\gamma}\big). \tag{3}$$

Seasonality $s(t)$ captures recurring patterns, such as the four-year cycle of the Summer Olympics. To model these periodic effects, we use a Fourier series[2]. Let P represent the period (in our case, $P = 365.25 \times 4 = 1461$ days over 4 years). The seasonality is approximated as:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right) \right), \tag{4}$$

where N determines the complexity of the series. Estimation involves fitting $2N$ parameters $\boldsymbol{\beta} = [a_1, b_1, \ldots, a_N, b_N]^\top$. For our case, with $P = 1461$ and optimal $N = 5$, the seasonal matrix is constructed as:

$$X(t) = \left[\cos\left(\frac{2\pi(1)t}{1461}\right), \ldots, \sin\left(\frac{2\pi(5)t}{1461}\right) \right]. \tag{5}$$

The seasonal component is expressed as:

$$s(t) = X(t)\boldsymbol{\beta}, \tag{6}$$

where $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2)$ imposes a smoothing prior to regularize the seasonality.

The holiday effect $h(t)$ accounts for special influences, such as the advantage a host country gains during an Olympic event. Let D_{Host} denote the set of host countries. Each host country is assigned a parameter κ_i, representing its impact. This effect is modeled as:

$$Z(t) = [\mathbf{1}(t \in D_{\text{Host}})], \tag{7}$$

and the holiday component becomes:

$$h(t) = Z(t)\boldsymbol{\kappa}, \tag{8}$$

where $\boldsymbol{\kappa} \sim \mathcal{N}(0, \nu^2)$ is a prior, and ν controls the flexibility of the host effect.

To calculate the uncertainty interval with a confidence level of 95%, we set $\alpha = 0.05$, which gives $Z_{\alpha/2} = Z_{0.025} \approx 1.96$. Since the components are uncorrelated, the total variance is the sum of the individual variances:

$$\text{Var}[\hat{y}(t)] = \text{Var}[g(t)] + \text{Var}[s(t)] + \text{Var}[h(t)] + \sigma^2 \tag{9}$$

The prediction interval is then given by $\hat{y}(t) \pm Z_{0.025}\sqrt{\text{Var}[\hat{y}(t)]}$ or equivalently:

$$\left[\hat{y}(t) - 1.96\sqrt{\text{Var}[\hat{y}(t)]}\ ,\ \hat{y}(t) + 1.96\sqrt{\text{Var}[\hat{y}(t)]} \right] \tag{10}$$

3.2 Model Evaluation and Forecasting of Medal Table Results

We trained the model to predict the number of *Gold*, *Silver*, and *Bronze* medals for each country, and then summed them to get the *Total* number of medals. The ranking of countries was also determined based on these medal counts.

To evaluate the model's performance, we used data on the medal achievements of countries from the Olympic Games up to and including 2016 for training and prediction and then evaluated the results for the 2020 and 2024 Olympic Games. The model was assessed by comparing its predictions to the actual results from the 2024 Olympic Games, specifically the number of gold, silver, and bronze medals won by each country. We also experimented with Random Forest and XGBoost and compared the results, as shown in Table 2.

Table 2: Metrics for Gold, Silver, and Bronze Medals Prediction

Model	**Gold**			**Silver**			**Bronze**		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
XGBoost	2.390	0.773	0.764	1.932	0.684	0.825	1.937	0.673	0.826
Random Forest	2.122	0.696	0.814	1.747	0.62	0.857	1.721	0.627	0.863
Prophet	**1.527**	**0.664**	**0.904**	**1.331**	**0.619**	**0.917**	**1.639**	**0.678**	**0.875**

From the MAE, RMSE, and R^2 metrics, the Prophet model outperformed other models such as XGBoost and Random Forest due to its characteristics that are well-suited to time-series data with seasonality and clear trends. Prophet, developed to address time-series forecasting problems, is particularly strong in handling seasonal factors and special events, which are important for predicting sports outcomes like Olympic medals, where there are seasonal fluctuations and external factors that impact results.

The parameters of the Prophet model were optimized to fit the data, with the ability to flexibly adjust for trends, seasonality, and the "holiday" effect, helping the model produce accurate and reliable predictions. Although models like XGBoost and Random Forest can yield accurate results, they are not as well-optimized for time-series characteristics such as seasonality and trends, making Prophet the ideal choice in this case.

Additionally, Prophet is easy to adjust and interpret, making analyses clearer and more understandable. This not only makes Prophet a powerful tool for predicting medals but also adds value in strategic decision-making, supporting deeper analyses of the potential success of countries in the 2028 Summer Olympics.

Next, we used Prophet to predict the results for the 2028 Summer Olympics for various countries. The predicted results are shown in Figure 3, which includes the predicted values for the top 10 countries. Based on the predicted rankings of countries for the 2024 and 2028 Olympic Games, there are some significant changes. The United States maintains its leading position. China and Japan remain in second and third positions, respectively, showing the stability of countries with strong traditional performances. Great Britain sees a significant jump from 7 to 3. Germany is out of the top 10, making room for Hungary, which climbs from 14 to 9.

Rank	Country	Gold	Silver	Bronze	Total
1	United States	45	43	37	125
2	China	38	29	27	94
3	Japan	21	14	16	51
4	United Kingdom	19	21	26	66
5	Australia	19	18	18	55
6	France	14	25	25	64
7	South Korea	13	8	12	33
8	Netherlands	12	10	14	36
9	Hungary	9	5	4	18
10	Italy	8	12	13	33

Figure 3: Los Angeles Olympic 2028 Predicted Medal Table - Top 10 Countries

We also present the 95% prediction intervals with lower and upper bounds, applying interpolation to the denoised data, as shown in Figure 4. Prophet helps generate these prediction intervals to indicate the level of uncertainty in the predictions, providing deeper insights into the likelihood of success for countries in the 2028 Olympic Games. The intervals reflect the model's certainty while accounting for natural fluctuations in the data, such as changes in athlete performance, strategic shifts, or even unforeseen factors like injuries or competition conditions. In the predicted chart for total medals, the prediction interval appears narrower due to the stretching of the medal counts of leading countries, causing the prediction interval to appear more condensed.

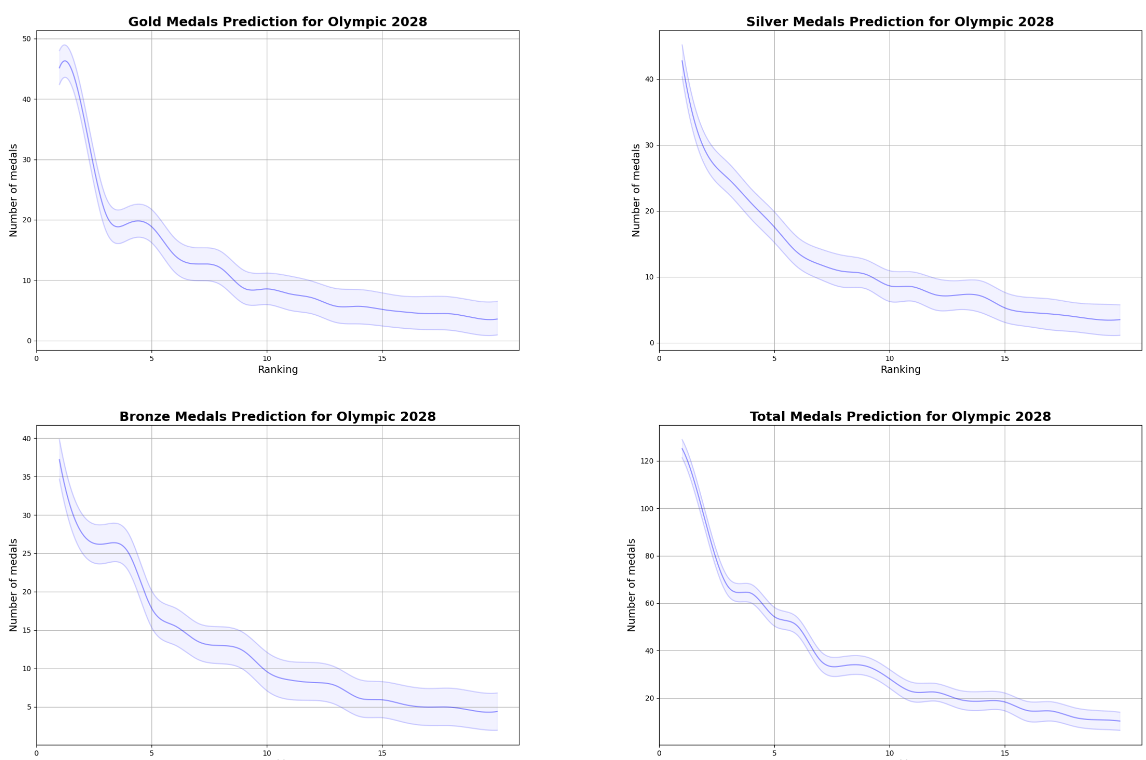

Figure 4: Prediction Interval for Olympic 2028

Overall, the model has shown promising results in predicting Olympic outcomes, but improvements are still needed to handle unexpected fluctuations in large-scale competitions. External factors, such as changes in competition strategies or the emergence of new athletes, could be factors that need to be further considered during the training process.

3.3 Nation Performance Variability and First-time Medalists

To further analyze the potential improvement or decline of countries, we applied the **KMeans** clustering algorithm to the 2024 results and 2028 predicted results to group countries into 3 clusters:

- Participant: Countries with few or no medals.
- Good: Countries with a certain number of medals.
- Outstanding: Countries with exceptional performance.

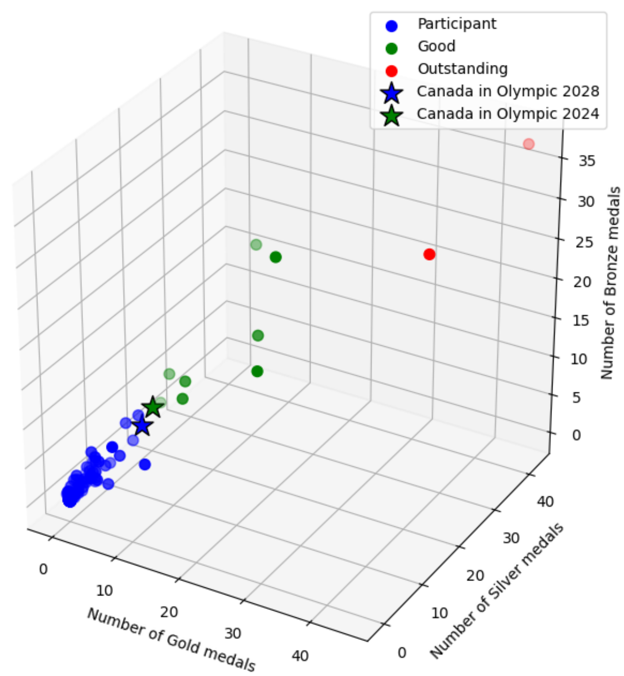

Figure 5: Clustering of 2028 Olympic Results

In Figure 5, by clustering countries into separate groups based on the predicted results, we can track whether any country moves from a low-performance group to a high-performance group, or vice versa, from a high group to a low group. The clustering results show no significant changes between the groups, except that **Canada shifted from the *Good* group in the 2024 Olympics to the *Participant* group in the 2028 Olympics**. This suggests that there are no major performance fluctuations for any country, or it may also indicate that the current model is not strong enough to capture large performance variations across countries.

Based on the predicted value of the Prophet model, we have calculated for the countries that have never won medals, the most promising nations to potentially win medals in the 2028 Olympics are Samoa (SAM), Vanuatu (VAN), and Rwanda (RWA), with the highest predicted value of 0.125, 0.059, and 0.052, respectively. These nations have shown the greatest potential for success, suggesting that their performances in upcoming competitions may improve significantly. Other countries, including Palestine (PLE), Yemen (YEM), and Vanuatu (VIN), also exhibit notable probabilities, though at a lower level compared to the leading countries.

4 Relationship between sports events and medal counts

4.1 Negative Binomial Regression

The number of medals is a count variable, non-negative, and often exhibits overdispersion (variance greater than the mean), as some countries may participate in many events but win few medals, while other countries may participate in fewer events but win many medals.

Thus, we use **Negative Binomial Regression**[3] to test the following hypotheses:

- Null hypothesis (H_0): There is no relationship between the number of events participated in and the number of medals.
- Alternative hypothesis (H_1): The more events a country participates in, the more medals it wins.

The Negative Binomial Regression model is defined as follows:

$$\ln(\mu) = \beta_0 + \beta_1 \cdot \text{Events} \tag{11}$$

where μ is the expected number of total medals, β_0 is the intercept, and β_1 is the coefficient for the number of events.

From the regression results, we obtained $\beta_1 = 0.0158$ and $\text{SE}(\beta_1) = 0.000623$, yielding $\frac{\beta_1}{\text{SE}(\beta_1)} \approx 25.393$. Therefore, we compute the p-value as:

$$p\text{-value} = 2\left(1 - \Phi\left(\left|\frac{\beta_1}{\text{SE}(\beta_1)}\right|\right)\right) \approx 2(1 - \Phi(25.393)) \approx 0. \tag{12}$$

Since the p-value is approximately 0, the Negative Binomial Regression provides strong statistical evidence that **participating in more events increases the number of medals**.

4.2 Pearson Correlation

Additionally, we test whether participating in more events, particularly in the strongest sports categories for each country, increases the likelihood of winning medals for that country.

We create a new variable *Top3_Events* representing the total number of medals won by each country in their top 3 strongest sports. We then analyze the correlation between the *Top3_Events* and the total number of medals for the top 10 countries with the highest total medal counts across all Olympics. The scatter plot showing the correlation between *Top3_Events* and total medals is shown in Figure 6.

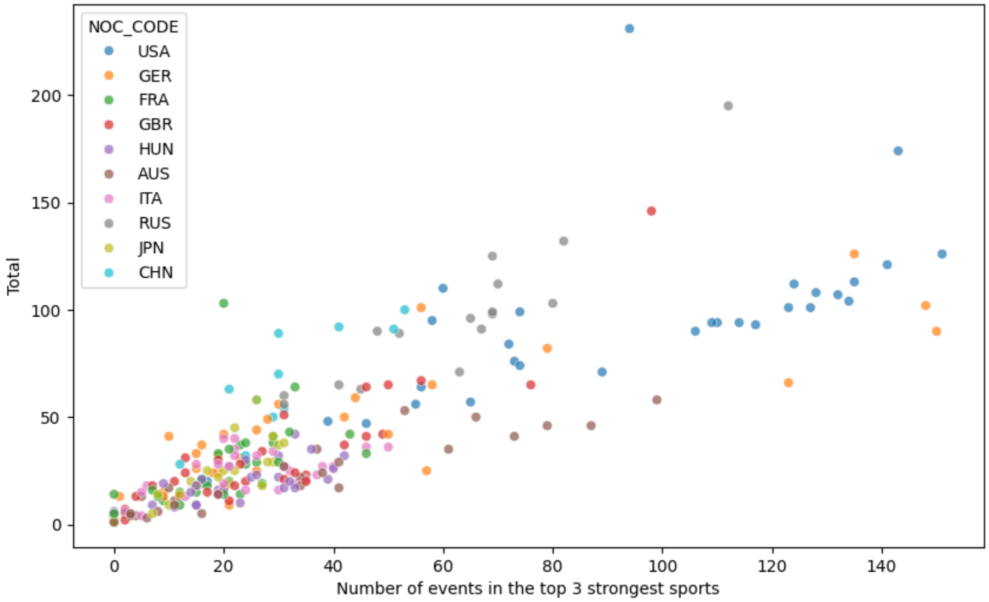

Figure 6: Correlation: Top 3 Events vs. Total Medals (Top 10 Countries)

Each point on the plot represents a specific country in a given year (only considering the 10 countries with the highest medal counts across all Olympic Games). The x-axis represents the number of events that countries participate in from their top 3 strongest sports, and the y-axis represents the total number of medals won by each country in these top 3 events.

To determine the existence of a linear correlation between the two continuous variables, with ρ as the population correlation coefficient between *Top3_Events* and the total number of medals, we perform a Pearson correlation test to test the following hypotheses:

$$\begin{cases} H_0 : \rho = 0 & \text{(No linear correlation)} \\ H_1 : \rho \neq 0 & \text{(There is a linear correlation)} \end{cases}$$

We calculated the Pearson correlation coefficient $r = 0.822$, with $n = 252$ observations. The Pearson test uses the Student's t-distribution to assess the significance of the Pearson correlation coefficient, which is calculated as:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.822 \cdot \sqrt{250}}{\sqrt{1-0.822^2}} \approx 22.826. \tag{13}$$

With degrees of freedom $df = n - 2 = 250$, the two-tailed p-value is determined by:

$$p\text{-value} = 2 \cdot \mathbb{P}\left(T \geq |t|\right) = 2\left(1 - \Phi_t(22.826)\right) \approx 0 \tag{14}$$

where Φ_t is the cumulative distribution function of the Student's t-distribution with $df = 250$.

Since the p-value is approximately 0, we reject the null hypothesis H_0 and accept the alternative hypothesis $H_1 : \rho \neq 0$ for a strong and statistically significant linear correlation.

Therefore, there is strong statistical evidence of a relationship between participating in more events in the top 3 sports and the likelihood of winning medals, although further models are needed to confirm **since causal relationships have not been verified yet**.

4.3 Important Sports Mapping by Network Analysis

In this section, we use Network Analysis[4] to analyze the relationship between countries and sports to assess their strengths and weaknesses in specific sports. We illustrate Network Analysis in Figure 7, focusing on five nations with outstanding achievements throughout Olympic history and five sports with the greatest influence in this competition.

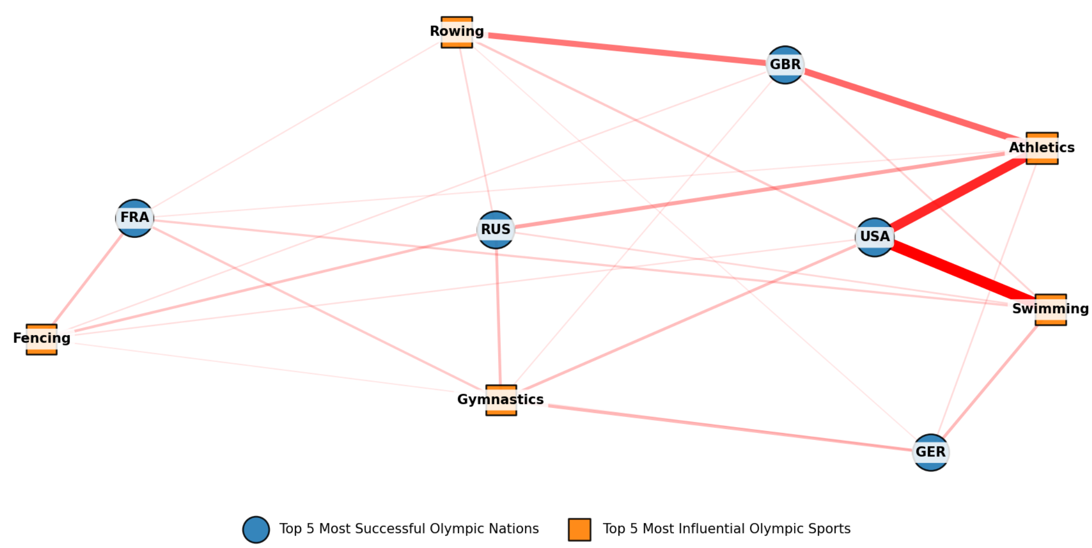

Figure 7: Network of Olympic Nations and Sports Influence

The nodes are categorized into two groups: blue nodes represent countries (e.g., USA, GBR, RUS, FRA, GER), and orange nodes represent sports (e.g., Swimming, Athletics, Gymnastics, Rowing, Fencing). The edges between the nodes represent the relationship or influence between a country and a sport, with the edge thickness reflecting the importance or frequency of medals won by that country in the respective sport. This network structure visualizes the relationships and mutual influences between countries and sports.

Some countries exhibit a strong focus on specific sports, such as the USA in Swimming and Athletics, or Great Britain in Rowing. These countries might consider strategic investments in sports where they already have strong connections or growth potential. For example, Germany could invest more in Fencing, or Russia in Gymnastics. Additionally, sports with thick edges, such as Swimming, Gymnastics, and Athletics, are often susceptible to the "great coach" effect, where the presence of an exceptional coach can lead to significant improvements in medal performance.

Moreover, we analyzed the top three most successful sports for each of the top 10 most successful countries in Olympic history in Figure 8.

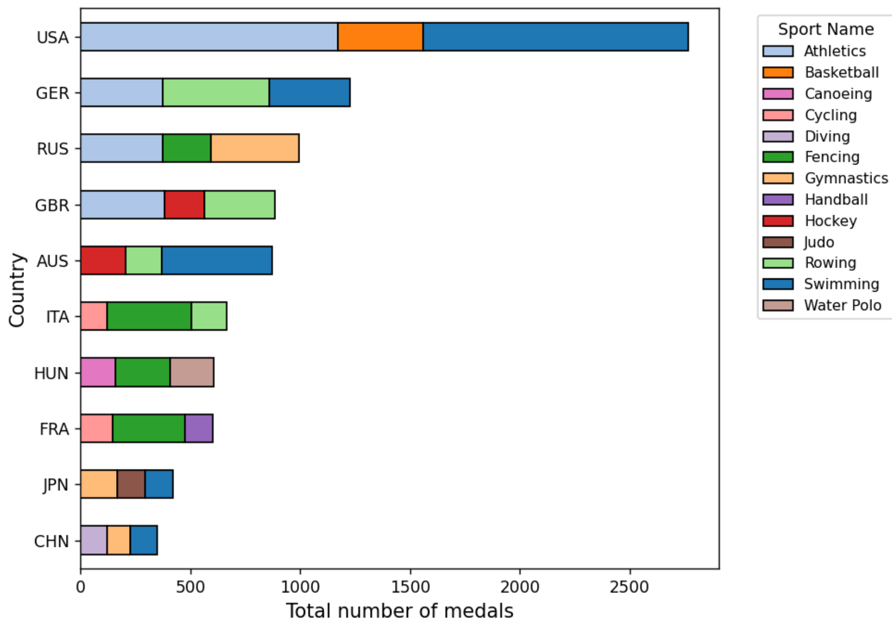

Figure 8: Top 3 Most Successful Sports for Top 10 Countries

From the above chart, we can observe which sports have a long history of success for each country. Notably, the USA has more than twice the number of medals as Germany, which ranks second, with two iconic Olympic sports (Athletics and Swimming) and one sport where the USA is considered a global leader (Basketball).

4.4 Impact Analysis of Hosting

Since 2020, the International Olympic Committee (IOC) has allowed host countries of the Olympic Games to propose the inclusion of one or more sports in the official competition program[5]. This policy not only aims to promote diversity in the events but also provides a distinct advantage for the host country by giving them additional opportunities to win medals.

A notable example is the 2020 Tokyo Olympic Games, where the Japanese Olympic Committee chose to introduce five new sports: baseball and softball, karate, sport climbing, surfing, and skateboarding. These sports were particularly popular in Japan and showcased the strengths of Japanese athletes. Similarly, for the upcoming 2024 Olympics, France has proposed and successfully included breaking – an innovative and increasingly popular street dance – into the program.

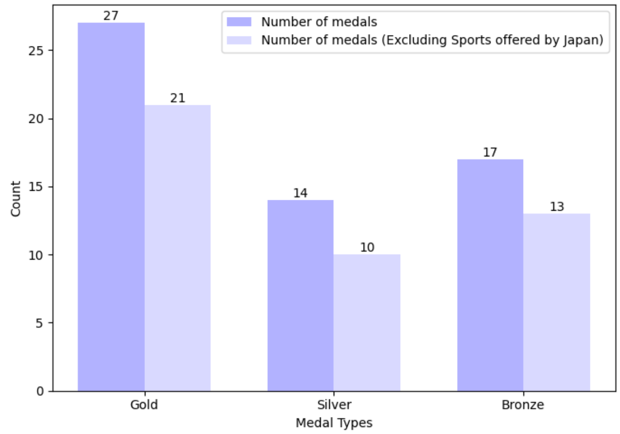

Figure 9: Comparison of Medal Counts of Japan in 2020

In practice, the host nation has consistently won medals in at least one of the sports they proposed. This demonstrates that newly added sports not only expand the host nation's competitive scope but also have a direct positive impact on their overall performance in that Olympic Games. However, since there have been only a limited number of Olympics held under this policy, we have not yet been able to comprehensively evaluate the extent of the influence these additional sports have on the host nation's overall achievements.

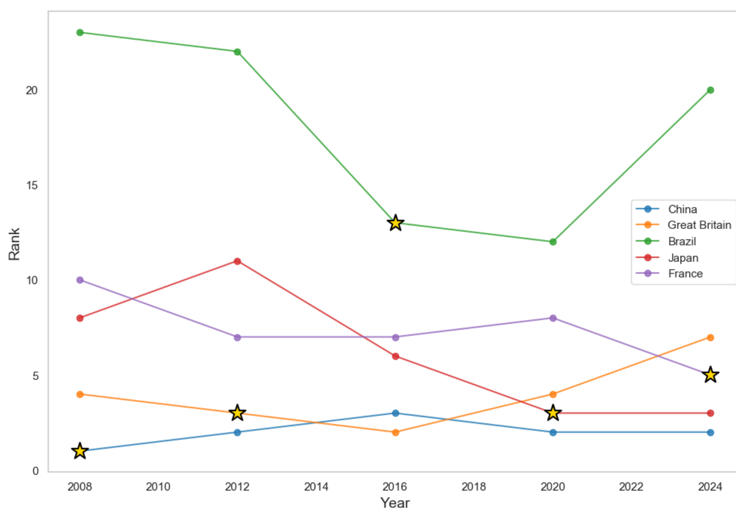

Figure 10: Host Country Rank (2008 - 2024)

In general, Figure 10 shows the rankings of different countries in the Olympic Games from 2008 to 2024, with stars marking the years in which each country hosted the Olympics. Overall, the years when countries hosted the event generally saw higher rankings compared to their previous Olympics. This can be explained by various factors, including the thorough preparation of the host countries, the support in terms of infrastructure and morale for the athletes, as well as the strong public and media encouragement during the years the event takes place. Therefore, with the limited number of Olympic Games since this policy was implemented, we cannot determine the exact impact of the host country's chosen sports on their performance.

5 The "Great Coach" Effect: A Case Study of Medal Impact

5.1 Examining Béla Károlyi's Contribution to Olympic Medals

Béla Károlyi[6], renowned for coaching the Romanian and later the U.S. women's gymnastics teams, significantly impacted Olympic medal counts. In the 1976 Olympics, as the coach of Romania, he guided the legendary gymnast Nadia Comaneci, who won three gold medals, one silver, and one bronze. Notably, she became the first gymnast in Olympic history to achieve a perfect 10.0 score.

In the 1984 Olympics, Károlyi served as the individual coach for Mary Lou Retton, who won the all-around championship, and Julianne McNamara, who secured the gold medal for uneven bars. To evaluate the "great coach" effect, we compare the performances of these athletes with their U.S. teammates not coached by Károlyi during the same year.

We define a scoring function $P(n)$ for individual athletes based on their medal counts, where medal weights increase arithmetically: a bronze medal is weighted as 1, silver as 3, and gold as 6. Let n_G, n_S, and n_B represent the counts of gold, silver, and bronze medals, respectively. The total weighted score is given by:

$$\text{Weighted\_Total} = 6 \times n_G + 3 \times n_S + n_B \tag{15}$$

We rank athletes based on their scores and visualize the results, plotting the athletes on the x-axis and their scores on the y-axis, as shown in Figure 11.

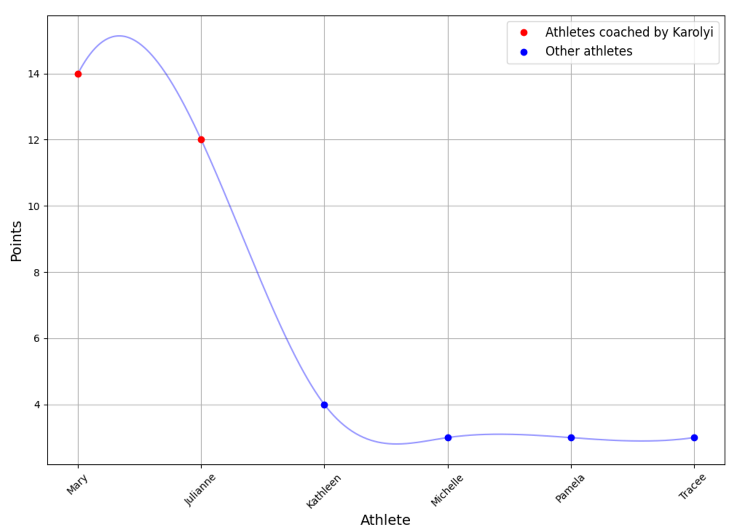

Figure 11: Scores of U.S. Gymnastics Team Members in the 1984 Olympics

The results clearly demonstrate that **the two athletes coached by Károlyi significantly outperformed their U.S. teammates in 1984**. The elbow-shaped curve at the highlighted blue point emphasizes the sharp difference in performance.

5.2 Analyzing the "Great Coach" Effect

Assessing the influence of coaches on Olympic performance is challenging due to several factors:

- **Random noise:** Performance is influenced by unpredictable factors such as injuries or competition conditions.
- **Historical trends:** Some nations have traditionally excelled in specific sports, independent of coaching impact.
- **Complex interactions:** Athlete quality, investment levels, and national sports policies interplay in intricate ways.

Following the framework of Randomization Inference[7], we test the hypothesis:

$$\begin{cases} H_0 : \text{All coaches are equally effective.} \\ H_1 : \text{Coaching has a significant impact.} \end{cases} \tag{16}$$

While Randomization Inference identifies whether coaching impacts performance overall, it does not estimate the individual effect of each coach due to random noise. To address this, we define the following independent variables and regression framework:

- **Host:** A binary variable (1 if the country is the host nation, 0 otherwise). Hosts often enjoy advantages such as familiarity, focused investment, and the ability to include sports favorable to their athletes.
- **Weighted_Total_lag:** Weighted medal performance in the previous Olympics. This reflects a country's existing capabilities.
- **Events_y:** The number of events in the year. More events provide more medal opportunities.
- **NOC_CODE:** Fixed effects for countries to account for inherent differences, such as traditional strength in specific sports or consistent investment strategies.
- **Year:** Fixed effects for years to control for global factors, such as rule changes or environmental conditions, that may influence all nations equally.

The regression model is as follows:

$$\text{Weighted\_Total}_{i,t} = \beta_0 + \beta_1 \text{Host}_{i,t} + \beta_2 \text{Weighted\_Total\_lag}_{i,t} + \beta_3 \text{Events}_{i,t} + \mathbf{FE} + \epsilon_{i,t} \tag{17}$$

Here, i represents the country code (e.g., NOC_CODE), and t represents the year (e.g., 2024). We calculate residuals as the difference between actual and predicted performance:

$$\text{Residuals}_{i,t} = \text{Weighted\_Total}_{i,t} - \widehat{\text{Weighted\_Total}}_{i,t} \tag{18}$$

To evaluate coach-specific effects, we employ a Block Permutation Test. Since Olympic cycles last four years, we assume that coaching tenure spans at least one cycle. A "coach block" (coach_id) is defined as the uninterrupted participation of a country in a sport over consecutive Olympics. If a nation skips an Olympic cycle, a new "coach block" is created. This definition preserves the temporal structure and controls for serial correlation.

The regression model for residuals is:

$$\text{Residual}_i = \beta_1 \cdot \text{Coach}_{1,i} + \beta_2 \cdot \text{Coach}_{2,i} + \cdots + \beta_{K-1} \cdot \text{Coach}_{K-1,i} + \epsilon_i \tag{19}$$

where β_j represents the fixed effect of coach j, indicating the average difference in residuals compared to the baseline coach.

Using 500 block permutations, we compute the observed and permuted R^2 values, calculate p-values, and define effect size and potential gain as follows:

$$R^2 = 1 - \frac{\sum(\text{residuals}_i - \hat{\text{residuals}_i})^2}{\sum(\text{residuals}_i - \bar{\text{residuals}})^2} \tag{20}$$

$$\text{p-value} = \frac{1}{N_{\text{perm}}} \sum_{k=1}^{N_{\text{perm}}} \mathbf{I}(R^2_{\text{perm}_k} \geq R^2_{\text{obs}}) \tag{21}$$

$$\text{Effect Size} = R^2_{\text{obs}} - \frac{1}{N_{\text{perm}}} \sum_{k=1}^{N_{\text{perm}}} R^2_{\text{perm}_k} \tag{22}$$

$$\text{Potential Gain}_c = \text{Effect Size}_c \times \overline{\text{Weighted Total}}_c \tag{23}$$

Figure 12 below are the results based on the countries where the p-value is most statistically significant

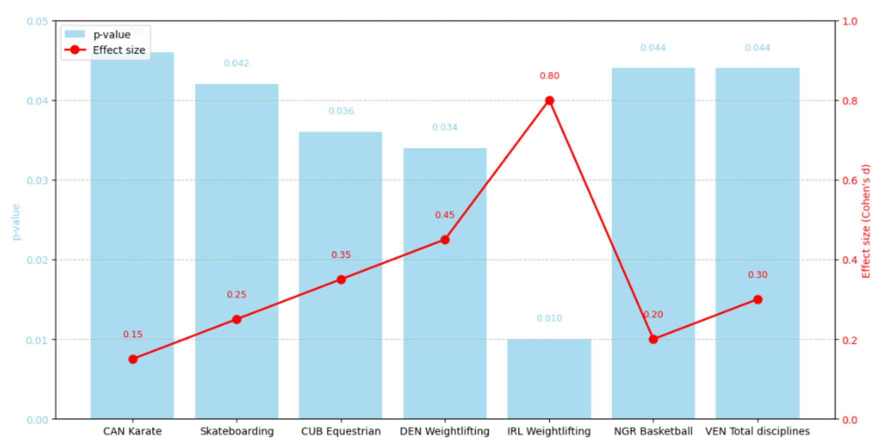

Figure 12: Significance and Effect Size Across Sports

From the above analysis, we conclude that **the "great coach" effect significantly influences Olympic performance**. For example, our model identifies that **Canada should invest in Karate coaches, with a projected medal-weighted score increase of 32 points,**

Cuba in Equestrian could lead to an increase of 14 points, and Denmark in Weightlifting may see an increase of 19 points, as these represent the top three potential gains.

6 Insights into Olympic Trends and Performance

6.1 Gender Equality in the Olympic Games

During the process of selecting variables to train the Prophet model, we identified an interesting trend related to gender participation in the Olympic Games. This trend is illustrated in Figure 13 below.

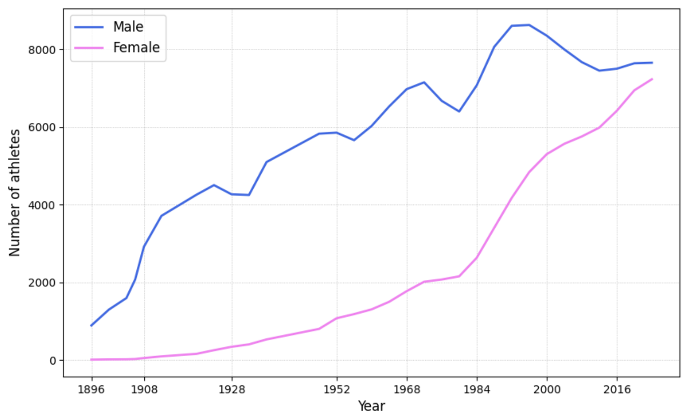

Figure 13: Number of males and females by year

Historically, the number of male athletes has always outnumbered female athletes. However, this gap has been narrowing over time, reflecting a **positive trend toward gender equality**. This highlights the increasing participation of women in the Olympics, partly due to global efforts to reduce gender disparities in sports. By promoting equal opportunities and implementing more inclusive policies, the Olympic Games have become not only a stage for showcasing athletic talent but also a symbol of progress in ensuring rights and opportunities for both men and women.

6.2 Examining Japan Investments in Olympic Success

Countries with a strong tradition in certain sports often excel in those sports or at least secure medals consistently in their areas of expertise. A specific example is Japan, as shown in Figure 14.

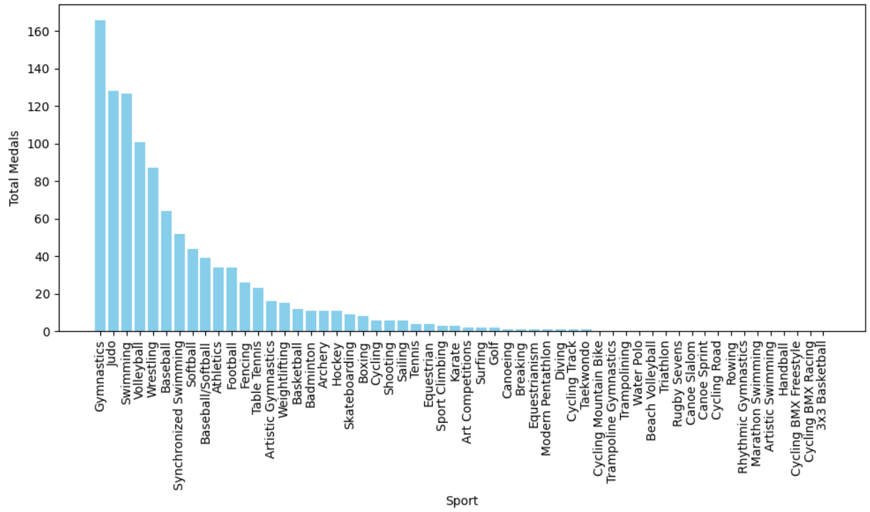

Figure 14: Medal Total of Japan by Sport

In addition to excelling in sports such as Gymnastics and Judo, Japan leveraged its role as the host nation for the 2020 Olympics to introduce its favored sports, such as Baseball and Karate, into the official competition. Although there is insufficient data to conclusively determine the impact of these sports on Japan's performance, they are areas worth continued investment. Furthermore, Japan's performance across various sports in different Olympic Games can be visualized by calculating the score $P(n)$ based on the number of medals, as detailed earlier, and visualized in Figure 15.

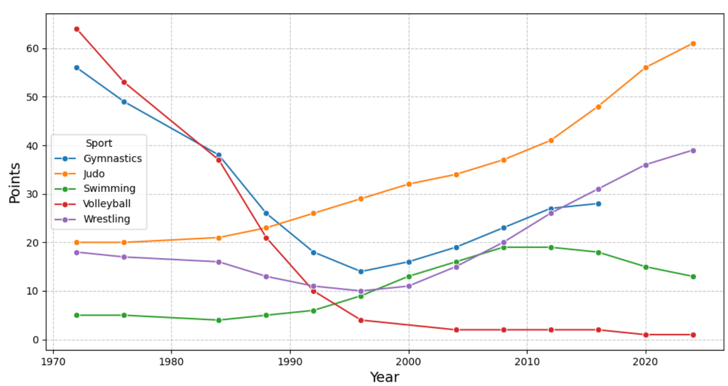

Figure 15: Trends in Japan's Olympic Sports Performance (1972-2024)

It is evident that Japan has transitioned from excelling in Gymnastics and Volleyball to focusing on Wrestling, Swimming, and Judo. However, Gymnastics has shown a resurgence after a period of decline. Regarding Volleyball, Japan has historically been the strongest team in Asia[8]. A hypothesis might be that Volleyball has seen reduced interest domestically, or perhaps other countries, benefiting from physical advantages, have started focusing more on Volleyball, potentially impacting Japan's performance in this sport.

In the case of Wrestling and Judo, these are widely recognized as emblematic sports associated with Japan. Since Japan's performance in these two disciplines has been improving, maintaining investments in these areas could lead to continued success and an increase in medal counts.

7 Conclusion

7.1 Strength and Weakness

Strengths:

Effective use of the Prophet model: The Prophet model has been proven to be effective in predicting Olympic medal counts, especially with data that exhibits clear seasonal patterns and trends. It outperforms other models such as XGBoost and Random Forest in predicting the number of gold, silver, and bronze medals. The model is also easy to adjust and interpret, making the analysis clearer and more understandable.

Multidimensional analysis: The study not only focuses on predicting the number of medals but also analyzes the relationship between sports and medal counts, the influence of the host country, and the effect of great coaches ("great coach" effect). Various analysis methods such as Negative Binomial Regression, Pearson Correlation, and Network Analysis are used to provide insights.

Careful data preprocessing: The data is meticulously preprocessed, including correcting character errors, merging countries that were previously split, and integrating different data sources to create a complete dataset.

Analysis of the "Great Coach" Effect: The study used Randomization Inference to assess the impact of great coaches, helping eliminate random noise and provide statistically meaningful conclusions.

Weaknesses:

Causality not proven: Although the study has shown correlations between participation in many events and the number of medals, it is not yet convincing enough to prove a causal relationship between these factors.

Limitations in evaluating the coach effect: Although the study used Randomization Inference to assess the coach effect, this method does not estimate the individual effectiveness of each coach but only identifies the existence of overall differences.

Lack of data on social and economic factors: The study mainly focuses on sports and Olympic history factors, but lacks data on social and economic factors that could influence a country's performance, such as investment in sports infrastructure or national policies.

KMeans clustering does not reflect significant fluctuations: KMeans clustering was used to group countries based on performance, but the results showed no significant changes between groups, which may indicate that the model is not strong enough to capture major fluctuations in a country's performance.

7.2 Potential Improvement

The improved Randomize Inference framework would: Track athlete migrations as proxies for coach movement linked to medal surges. Shuffle migrations to simulate no coach impact.Compare real vs. permuted medal counts to isolate coach effects. Quantify the impact of hiring a "great coach" on medals.

8 Memorandum

To: International Olympic Committee

From: Team #2526908

Subject: Insights from Olympic Data Analysis and Predictions for Los Angeles 2028

Date: January 27, 2025

Dear Members of the International Olympic Committee,

We are excited to share key insights derived from our analysis of Olympic data spanning up to the 2024 Games. Our research focuses on predicting the medal table for the Los Angeles 2028 Olympics, analyzing relationships between sports, events, and medal counts, assessing the impact of host nations, and examining the "great coach" effect, among other insights.

Prediction of the Medal Table for Los Angeles 2028: With the United States hosting the 2028 Games, our projections indicate a slight increase in the gap between the U.S. and the second-ranking team, China, compared to 2024. Notably, Great Britain is expected to see a significant rise, moving from 7th to 3rd place. Meanwhile, Germany falls out of the top 10, making way for Hungary, which climbs from 14th to 9th. These shifts offer intriguing narratives to explore ahead of the Games.

Relationship Between Events and Medal Counts: Our findings confirm a strong linear correlation between participation in more events and an increase in medal counts. The top three most popular sports categories show a clear link to total medal achievements. Additionally, the host nation effect plays a significant role in influencing overall performance, as does strategic participation in certain sports.

The "Great Coach" Effect: The "great coach" effect emerges as a critical factor in Olympic success. For instance, our model suggests that Canada should invest in Karate coaches, with a projected medal-weighted score increase of 32 points. Similarly, Cuba would benefit from focusing on Equestrian (increase of 14 points), and Denmark on Weightlifting (increase of 19 points). These insights highlight key opportunities for maximizing medal outcomes.

Gender Balance in Participation: Our analysis also shows that the gender gap in athlete participation has narrowed significantly. Whereas male athletes previously outnumbered female athletes substantially, the numbers are now nearly equal, reflecting a positive trend toward gender parity in the Games.

These insights provide a valuable perspective on Olympic trends and opportunities. We hope our findings will help inform planning and strategies to make the Games even more competitive and engaging.

Yours sincerely,
Team #2526908

References

[1] S. J. Taylor and B. Letham, "Forecasting at scale," *PeerJ Preprints*, 2017.

[2] A. C. Harvey and N. Shephard, "Structural time series models," in *Handbook of Statistics*, (edited by G.S. Maddala, C.R. Rao and H.D. Vinod). Amsterdam: North Holland, 1993, vol. Vol. 11:Econometrics, pp. 261–302.

[3] V. Nijimbere, "Hypothesis test and confidence interval for the negative binomial distribution via coincidence: A case for rare events," *The International Journal of Contemporary Mathematical Sciences*, vol. 12, pp. 243–253, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:67788987.

[4] X. Bi, Z. Sun, and B. Hu, "The network of sports: Using network analysis to understand the relationship between sports and socio-physiological factors in contemporary china," *Applied Network Science*, vol. 9, no. 1, pp. 1–17, 2024. DOI: 10.1007/s41109-024-00628-0. [Online]. Available: https://doi.org/10.1007/s41109-024-00628-0.

[5] A. Mambo. "How are new sports chosen for the olympics?" NPR. (Jul. 26, 2024), [Online]. Available: https://www.npr.org/2024/07/26/nx-s1-5051616/how-are-new-sports-chosen-for-the-olympics.

[6] Wikipedia contributors, *Béla károlyi — Wikipedia, the free encyclopedia*, [Online; accessed 27-January-2025], 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=B%C3%A9la_K%C3%A1rolyi&oldid=1270080690.

[7] C. R. Berry and A. Fowler, "How much do coaches matter ?," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202615628.

[8] Wikipedia contributors, *Japan men's national volleyball team — Wikipedia, the free encyclopedia*, [Online; accessed 27-January-2025], 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Japan_men%27s_national_volleyball_team&oldid=1270435110.

Report On Use of AI

We primarily use AI tools like ChatGPT for translation, enabling quick and accurate conversion of text while preserving context and meaning. This has greatly improved the efficiency and quality of my work. We also use AI to ask specialized questions like the ones below.

OpenAI ChatGPT

Query 1: Brief describe about Randomization Inference

Answer: Randomization Inference (RI) is a statistical method used to test hypotheses by comparing observed data to what would be expected under a null hypothesis, based on randomization. It's often employed in experimental settings to analyze treatment effects without relying heavily on assumptions about the underlying distribution of data. Key Points: Focus on Random Assignment:RI leverages the fact that treatment or intervention was assigned randomly in an experiment.It uses the randomization mechanism itself to generate a reference distribution for the test statistic.Test Statistic:A test statistic (e.g., difference in means) is calculated from the observed data. The same statistic is calculated for many hypothetical randomizations (permutations) of the treatment assignment.Null Hypothesis Testing:RI tests whether the observed effect could have occurred by chance under the null hypothesis (typically that the treatment has no effect).

Non-Parametric: Unlike traditional statistical tests, RI does not assume a specific data distribution (e.g., normality). This makes it robust, especially with small sample sizes. Applications: commonly used in fields like economics, medicine, and social sciences to evaluate causal effects.Often applied in randomized controlled trials or studies with well-defined randomization protocols.By using the actual randomization process to create the reference distribution, RI ensures that the p-values and conclusions are valid under the given experimental design.

Query 2: Why Lang Ping is a successful coach for both USA and China?

Answer: Lang Ping is a successful coach for both the USA and China due to her deep volleyball expertise, global experience, and exceptional leadership. As a legendary player turned coach, she adapted strategies to fit each team's strengths, leading the USA to Olympic silver in 2008 and China to gold in 2016. Her ability to navigate cultural differences, inspire players, and modernize tactics made her a unifying and innovative figure in volleyball, respected worldwide.

Query 3: What are the factors that affect the number of medals a country wins.

Amswer: The number of medals a country wins is influenced by factors like economic resources, population size, and sports infrastructure, which enable better training and talent development. Government support, cultural emphasis on sports, and talent identification also play key roles. Regular participation in international events, climate suitability for specific sports, and private sponsorship further enhance performance. Host countries often excel due to increased investment and motivation.