Problem Chosen	2025	Team Control Number
C	MCM/ICM Summary Sheet	2509986

Unveiling Medal Trends in the Olympic Games

Summary

As spectators, we are often interested in the Olympic medal table. By analyzing the performance of each team in the previous Olympic Games, we can find the advantages of each team and then predict the number of medals each team will win at the Los Angeles, USA summer Olympics in 2028. At the same time, as some zero medal teams perform better and better, combined with our analysis of the teams that have achieved medals in recent Games, we can predict the probability of these teams to win first medal in the 2028 Olympic Games. Analysis of the "great coach" effect in previous Olympic Games can reveal the vital role of coaches in team promotion. Therefore, it is very important to analyze the performance of each team in the previous Olympic years to predict the performance of each team at the next Olympic Games and to provide a strategy to improve the number of medals and enhance the strength of the team.

Firstly, after data pre-processing, based on the analysis of historical data, we identified the advantages of each country's events. Through the number of medals, host effect, advantage coefficient and the number of participating events of the previous two Olympic Games, a linear regression model was established to fit the number of gold medals, silver medals and bronze medals of the country. The sum of the three medals is the total number of medals. We use this model to predict the medal table of 2028 Olympic Games and get a reasonable result. We found that the United States, as the host country, could do even better at the 2028 Games. We then used logistic regression to predict the countries that achieved the zero breakthrough. By dividing the training set and the testing set, the accuracy of our model reaches 89.83%. The analysis found that those most likely to achieve a zero medal breakthrough were Samoa, Angola, the Republic of Guinea and Mali. At the same time, we found that countries with strong sports tend to do better in the medal table. The more dominant events, the more medals the country wins.

Secondly, we use one way within-group ANOVA to verify whether great coach has an effect on the improvement of national strength. In the F-test, the P-value is less than 0.05, indicating that the data significantly reject the null hypothesis. This proves that a great coach can make a noticeable difference in a country's performance. These results show that hiring a great coach can improve Olympic performance.

Finally, we conduct more in-depth analysis based on models and data to make recommendations for National Olympic Committees that want to achieve breakthroughs. At the same time, countries need to support the development of women's and youth sports and commit to more comprehensive progress in international sporting events.

Keywords: Linear regression, Logistic regression, One-way ANOVA, Olympics

Contents

1 Introduction

1.1 Background

As one of the most influential sporting events in the world, the Olympics will focus on the medal table in addition to all exciting games. Since the revival of the modern Olympic Games in 1896, the Olympic medal table is not only an intuitive embodiment of the sports strength of various countries, but also an important symbol of the country's comprehensive strength, cultural influence, and social development level.

Figure 1: Olympic emblem in 2028

In Olympic history, top countries often get attention for their outstanding sporting achievements. These countries usually perform well in multiple events, especially in track and field, swimming, gymnastics, and diving. For example, the United States, China, Russia (the former Soviet Union), the United Kingdom, Germany, and other countries have long occupied the top of the medal table. Its performance in the Olympic Games not only represents the country's sports level but also reflects its influence and competitiveness on the international sports stage. The sports achievements of these countries not only inspire the love of sports of their own people but also set a benchmark for the development of global sports.

While the countries in the top medal table draw most of the attention, the first medal or gold medal also deserve attention. For example, in the 2024 Paris Olympic Games, Albania (2 medals), Cabo Verde and other countries won Olympic medals for the first time. These achievements are not only the affirmation of athletes' individual efforts, but also encourage more countries and regions to increase their investment in sports. The first-time medal countries often face many challenges, including limited sports resources, lack of professional training facilities and coaches, and relatively weak sports infrastructure. However, through their unremitting efforts and perseverance in the pursuit of the sports cause, these countries finally made a breakthrough in the Olympic Games. These achievements have not only enhanced the international image of the country, but also injected new impetus into the development of domestic sports undertakings.

1.2 Problem Restatement

- By analyzing the number of Olympic medals in each country, we model the number of medals in each country and predict the medal table for the Los Angeles Summer Olympics in 2028. At the same time, the performance of zero-medal teams in previous Olympics predicts the probability of winning medals in the next Olympics and the number of teams with zero-medal breakthrough. Finally, through the analysis of each project and team to find the advantages of each country project.

- The coach plays a crucial role in improving the strength of each team. Through the analysis of Lang Ping, Bela Karolyi and other coaches, we found the role of "great coach" effect in the improvement of team performance, and explained how "great coach" works to them through the examples of three countries.
- Discover more information about the number of Olympic medals and give our insights.

1.3 Our work

In order to predict the medal table of the 2028 Olympic Games and study the great coach effect, we use data for model construction and analysis. In the first question, we constructed Total Advantage Coefficient and used linear regression model and logistic regression model to determine the medal table based on the historical medal table data and the number of participating events. In the second question, we use the One-way within-groups ANOVA. The "great coach" effect is reflected through the team's achievements before and during the coach's coaching. In the third question, we mined more information about the medal table through the data provided by the question. The overview of our work is shown in Figure 2.

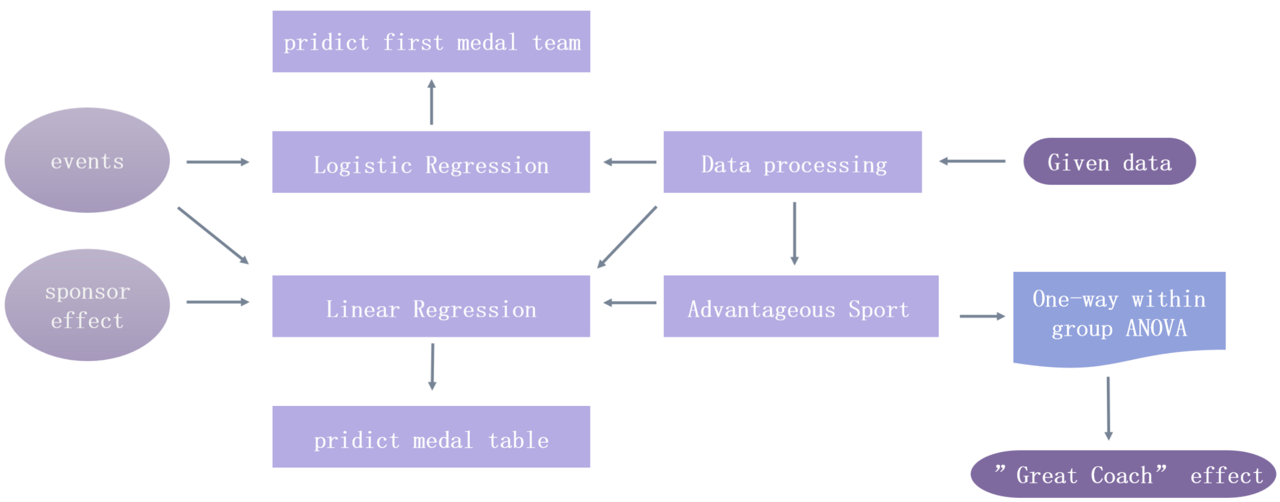

Figure 2: Overview of Our Work

2 Assumptions and Notations

2.1 Assumptions

Assumption 1: Not considering the decreased performance of the athletes due to injury or physical factors.

Assumption 2: Not considering the negative competition or withdrawal or other factors that affecting the competition results due to some unexpected reasons.

Assumption 3: Suppose that the teams for the 2028 Olympics are consistent with the 2024 Olympics.

Assumption 4: The events in the 2028 Olympic game which each teams will participate in are the same as that in the 2024 Olympic game, expect the new events that will be added in 2028.

2.2 Notations

Table 1: Notations used in this literature

Symbol	**Description**
$M_{i,j}$	The total number of medals for the i-th team in the j-th Olympic Game
$G_{i,j}$	The total number of gold medals for the i-th team in the j-th Olympic Game
$S_{i,j}$	The total number of silver medals for the i-th team in the j-th Olympic Game
$B_{i,j}$	The total number of bronze medals for the i-th team in the j-th Olympic Game
$s_{i,j,k,l}$	The score of the number of medals of the k-th sport l-th event for the i-th team in the j-th Olympic Game
$H_{i,j}$	If the i-th team is the home team in the j-th Olympic game, the value is 1, otherwise, the value is 0
$e_{i,j,k}$	The total number of events of the k-th sport for the i-th team in the j-th Olympic Game
$P_{i,j,k}$	The total number of participants of the k-th sport for the i-th team in the j-th Olympic Game
A_i	The collection of the notation of advantageous sports of the i-th team

3 Data Processing

3.1 Data Pre-processing

To prepare the data for analysis, datasets were loaded and processed. The pre-processing steps included harmonizing country names and codes to ensure consistency across different historical contexts. A mapping table was created for country name and code transformations. For example, "Soviet Union" was mapped to "Russian Federation," and "Yugoslavia" was mapped to "Serbia." This step was crucial for aggregating data accurately across years and regions.

In addition, we standardized the country codes in both the medal data and athlete participation data using the ISO 3166-1 standard. This ensured that all country codes were uniform and aligned with international naming conventions. Table 3 provides an overview of the mapping between NOC and their standardized counterparts.

Original Name	Mapped Name
Soviet Union	Russian Federation
East Germany	Germany
Czechoslovakia	Czechia
Chinese Taipei	Taiwan, Province of China

Table 2: Country Name Mapping Examples

NOC	Alpha-3 Code
GER	DEU
GRE	GRC
NED	NLD
TCH	CZE
RSA	ZAF
IRI	IRN

Table 3: Alpha-3 Code Mapping Examples

3.2 Data Cleaning

The cleaned dataset contained columns such as `Year`, `NOC`, `Sport`, `Event`, and `Medal`. To avoid duplicate entries, unique combinations of `Year`, `NOC`, and `Event` were retained, ensuring only one record per medal-awarding event.

Next, a function was implemented to convert the `Medal` column into a tuple format, representing counts of `(Gold, Silver, Bronze)` medals. For instance, a "Gold" medal entry was transformed into `(1, 0, 0)`, while "Silver" was mapped to `(0, 1, 0)`, and "Bronze" was mapped to `(0,0,1)`. Non-medal events were represented as `(0, 0, 0)`.

The dataset was then grouped by `Year`, `NOC`, and `Sport`, and medal counts were aggregated using tuple addition. The results were stored in a new table with the structure shown in Table 4.

Year	NOC	Sport	Medal (Gold, Silver, Bronze)
2024	USA	Athletics	(9, 7, 2)
2024	CHN	Diving	(5, 2, 0)
2024	GBR	Swimming	(1, 1, 0)

Table 4: Processed Data: Medal Counts by Year, NOC, and Sport

4 Models

4.1 Predicting the medal table

4.1.1 Model Preparation

In order to predict the number of medals in the next Olympics, we need to use the corresponding medal data for the previous Olympic games. Because an athlete's career length generally concentrated in 12 to 20 years[1]. With the state and age change, after the previous Olympic athletes also not necessarily attend, and considering for each team, the previous advantage of sports will continue in the short term advantage position. After the testing of MSE and AIC, here we use the data of the last two Olympic medals to fit it. For example, for the gold medals, We use $G_{i,j-1}$, $G_{i,j-2}$ to predict $G_{i,j}$.

At the same time, we observed that the relationship between gold medals, silver medals and bronze medals among the teams was relatively stable. So we fitted the number of gold medals, silver medals and bronze medals respectively, and the sum of the three was the total number of medals.

Then, we need to consider the sponsor effect, and here we set the sponsor effect to the 0-1 variable, $H_{i,j}$. If the i-th team is the home team in the j-th Olympic game, the value is 1, otherwise, the value is 0.

Next, the strength of each country's sports also has an impact on the medal count. An advantage event means that the country is more competitive in the event and has a better chance of winning a medal. And the expected number of medals won by a country in a dominant sport is higher than the expected number of medals won in other sports. We take the Total Advantage Coefficient to show the advantageous sports of a country as the fifth parameter of the model, which we will define in 4.1.2.

Finally, we take the total number of events participated by each country in this Olympic Games as the last parameter of the model. For the i-th team, in the j-th Olympic game, the value is $\sum_k e_{i,j,k}$. This is because the greater the number of events attended, the greater the chance of winning medals and the higher the expected medal count.

4.1.2 Model of Advantageous Sports

In order to analyze the influence of advantageous sports on a country's medal distribution, we build a model incorporating the **Total Advantage Coefficient** to measure a country's performance in specific sports. For each sport, the coefficient reflects the cumulative historical advantage, considering both the medal type and its relevance over time.

The advantage score for sport k in country i during the m-th Olympic Games is calculated as follows:

$$\text{Score} = 4 \cdot G_{i,m,k} + 2 \cdot S_{i,m,k} + 1 \cdot B_{i,m,k}$$

The advantage score is designed to weigh gold medals more heavily, followed by silver and bronze medals. The sports with the highest calculated advantage scores are considered the most advantageous for that country.

Based on the computed advantage scores, sports with higher scores are classified as advantageous for the country, influencing the categorization of sports in this group.

Specifically, for the i-th country in the j-th Olympic game, the advantage coefficient for the k-th sport is defined as:

$$\text{Advantage\_Coefficient}_{i,j,k} = \sum_{m} \exp(k \cdot (Y_m - Y_{\max})) \cdot (4 \cdot G_{i,m,k} + 2 \cdot S_{i,m,k} + 1 \cdot B_{i,m,k})$$

where Y_m is the year of the m-th Olympic game, $Y_{\max}$ is the latest Olympic year, and $G_{i,m,k}, S_{i,m,k}, B_{i,m,k}$ represent the number of gold, silver, and bronze medals won by country i in sport k during the m-th Olympic game. The exponential decay factor $\exp(k \cdot (Y_m - Y_{\max}))$ ensures that recent performances have a stronger influence than older results.

Next, the **Total Advantage Coefficient** for the i-th country is obtained by summing the advantage coefficients across all advantageous sports:

$$\text{Total\_Advantage\_Coefficient}_{i,j} = \sum_{k \in A_i} \text{Advantage\_Coefficient}_{i,j,k}$$

where A_i is the set of advantageous sports for country i. This value captures the overall strength of a country in its most competitive events.

Finally, these parameters are incorporated into the overall medal prediction model, ensuring that both historical performance and event participation are considered for accurate predictions.

4.1.3 Model Establishment

For predicting the medal table

By observing the relationship between the number of medals and various factors, we find that the expected number of medals has a linear relationship with the above parameters. Based on this, a linear model is built. For example, for the gold model, that is,

$$G_{i,j} = \beta_0 + \beta_1 * G_{i,j-1} + \beta_2 * G_{i,j-2} + \beta_3 * H_{i,j} + \beta_4 * \text{Total\_Advantage\_Coefficient}_{i,j} + \beta_5 * \sum_{k} e_{i,j,k}$$

where i is the i-th team, j is the j-th Olympic Game, k is the k-th sport.

The same as the gold model, we can calculate the predicted number of silver medals and bronze medals.

$$S_{i,j} = \beta_0 + \beta_1 * S_{i,j-1} + \beta_2 * S_{i,j-2} + \beta_3 * H_{i,j} + \beta_4 * \text{Total\_Advantage\_Coefficient}_{i,j} + \beta_5 * \sum_{k} e_{i,j,k}$$

$$B_{i,j} = \beta_0 + \beta_1 * B_{i,j-1} + \beta_2 * B_{i,j-2} + \beta_3 * H_{i,j} + \beta_4 * \text{Total\_Advantage\_Coefficient}_{i,j} + \beta_5 * \sum_{k} e_{i,j,k}$$

And the predicted number of the total medals is:

$$M_{i,j} = G_{i,j} + S_{i,j} + B_{i,j}$$

For predicting the first medal teams

Predictions for first-time medal winners are much different from predictions for the medal table, as these teams have not won medals in previous competitions and cannot know their probability of winning a medal from historical medal information.

In this question, we used data other than medals for fitting. Here, we selected the number of participants and the number of participating events, and fitted the teams that achieved the breakthrough of zero medals in previous Olympic Games to establish the logistic regression, that is

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_j X_j)}}$$

where $P(Y = 1|X)$ is the probability of Y being 1 given $X, \beta_0, \beta_1, \beta_2, \ldots, \beta_j$ are the model parameters.

4.1.4 Model Solution and Evaluation

The predicted medal table in 2028 Olympic Game

Based on the above model, we now predict the medal table for the 2028 Olympic Games. The amount of gold, silver and bronze medals in 2024 and 2020 Olympic Games, total advantage coefficient, host effect and events count of each team were fitted. We take 80% of the data as the training set and 20% of the data as the testing set. The coefficient and the mean square error of the testing set are shown in Table 5.

Medal	Previous1	Precious2	Events	Sponsor	Advantage	MSE
Gold	0.4443	0.3199	0.0169	10.5692	0.0114	16.9331
Silver	0.3625	0.2706	0.0270	6.9249	0.0115	10.7679
Bronze	0.3358	0.2193	0.0351	2.1927	0.0115	12.2795

Table 5: The coefficient and MSE

In the 2028 Olympics, the United States will host the games. At the same time, assume that each country participates in the same number of events at the 2028 Games as in 2024, plus Baseball, Crichet, Lacrosse, Squash, Flag Football five new events in 2028[1], which some of the advantageous countries will participate in. The final medal table is shown in the Table 6. Due to space constraints, here we only show the top seven countries by gold medal.

Nation	Gold	Silver	Bronze	Total
United States	48	44	35	131
China	37	27	23	86
Japan	22	14	15	50
Great Britain	18	20	22	62
Australia	18	15	18	51
France	15	19	18	52
Netherlands	13	10	12	34

Table 6: The predicted medal table in 2028

Combined with the actual situation, we can find that this is very close to the actual sports strength of each country, in line with the advantages of each country in sports. At the same time, as the host country, the United States also achieved better results than the 2024 Olympic Games, which is in line with the host effect. Clearly, the US performs better when it has the advantage of being the host country. On the other hand, a French team that loses the advantage of being the host country will not play as well as it did in 2024. The performance of other countries has been relatively stable. Due to changes in sports policy, the emergence of talent in some events, as well as state and performance, they will appear minor changes in the medal table. Therefore, we can consider our model to be reasonable and similar to the actual situation.

The countries that have achieved zero Olympic medal breakthroughs

In this section, we use the data for the corresponding years of countries with zero medal breakthroughs in history as a positive sample (Y=1). Take the years of participation of countries that have not yet won a medal as a plus or minus sample (i.e. Y=0). 155 positive samples and 3175 negative samples were obtained. In order to balance the positive and negative samples and ensure the accuracy of the model, we randomly selected the same number of samples as the positive samples as the data set. Taking 80% of the data as the training set and 20% of the data as the testing set, the final confusion matrix obtained by fitting is as follows:

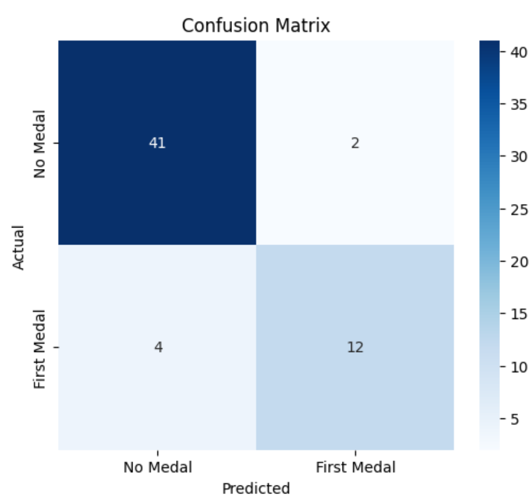

Figure 3: The confusion matrix od testing data

In the test set, the accuracy of the model is 89.93%. It shows that our model is very good at predicting whether a country will achieve a medal-zero breakthrough. Then we use this model to predict the medal-zero breakthrough in 2028 Olympic games. Using data from countries that have not won medals, we finally get the probability of each country achieving a breakthrough of zero medal as shown in the chart below:

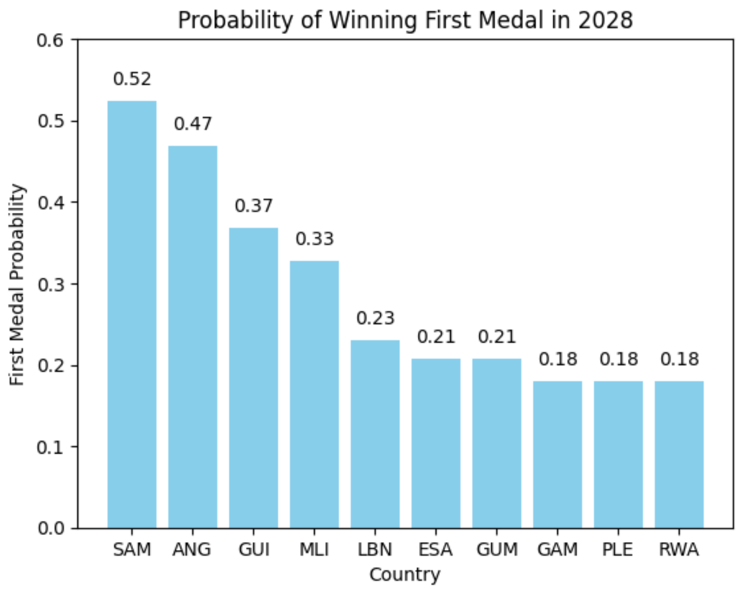

Figure 4: The Probability of Winning first medal in 2028

It shows that Samoa, Angola, Republic of Guinea and Mali have relatively higher probability of making a breakthrough at the 2028 Olympics, possibly claiming the first Olympic medal in history. In the Olympic Games, people of more and more countries have shown their level and spirit, and demonstrated the elegance of themselves and their countries. We believe that these countries that have not yet won a medal will one day make a breakthrough.

The relationship between the events and medals

In this section, we analyze the relationship between advantageous sports and the stability of medal-winning performance for selected countries (CHN, USA, GBR, ITA, DEU). Using data from the past 40 years (1984–2024), we compare the medal counts of advantageous sports and non-advantageous sports within each country. The box plots below visualize the distribution of medal counts, highlighting the stability of performance in advantageous versus non-advantageous sports.

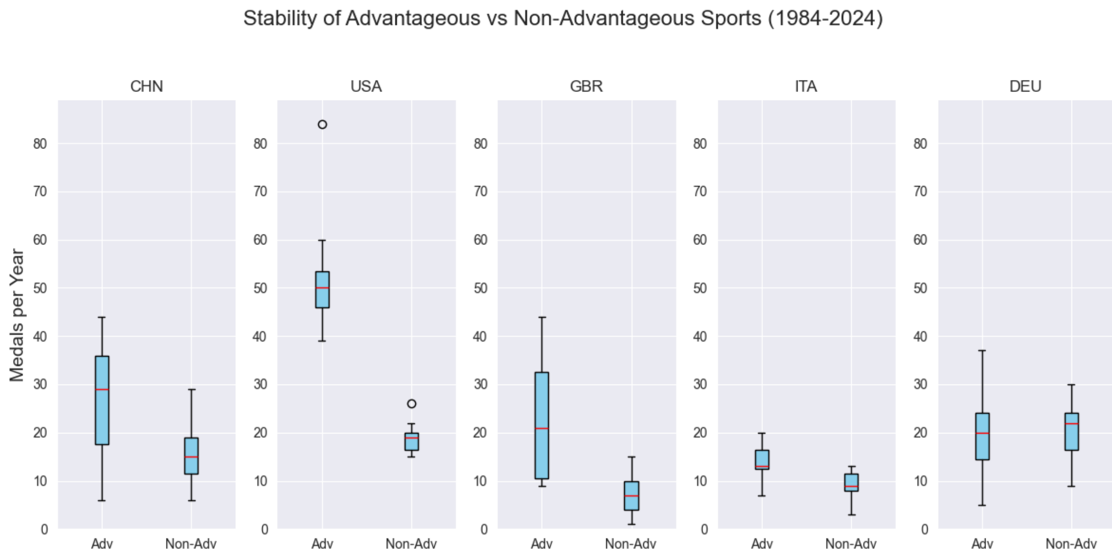

Figure 5: Medal Distribution for Advantageous and Non-Advantageous Sports (1984–2024)

The box plot reveals that advantageous sports generally demonstrate higher peak medal counts compared to non-advantageous sports, highlighting their potential for exceptional performance. For instance, countries like China (CHN) and the United States (USA) exhibit a greater range in medal counts for advantageous sports, indicating their ability to achieve top-tier results despite slightly higher variability. In contrast, non-advantageous sports show relatively lower variability, largely due to many sports consistently achieving zero medals. This suggests that advantageous sports provide a significant opportunity for countries to secure higher medal tallies and excel on the global stage. Overall, the analysis underscores that while non-advantageous sports may appear more stable, this stability often stems from consistently low or zero medal counts. Therefore, strategically focusing on advantageous sports, which offer higher potential for success, remains critical. At the same time, gradual improvements in underperforming areas can further optimize overall performance in future Olympic games.

4.2 The “Great Coach” Effect

4.2.1 Problem Analysis

Because it is difficult to change nationality, it is difficult for athletes to win medals representing multiple countries, but coaches have no such restrictions. Coaches can serve on multiple national teams and win medals for that team. This is the "great coach" effect. These "great coaches" can often bring significant performance improvement to a national team. For example, a team that has not won a medal can be awarded a medal. We are familiar with "great coaches" such as Lang Ping, who brought strength to the Chinese and US women’s volleyball teams, helping both teams win gold medals in Olympics. It is very important to study the role of the "great coach" in order to discover the importance of coaching in improving performance and provide recommendations for countries that need to improve performance. Here we select our "great coach" through the Coach Lifetime Achievement Award awarded by the International Olympic Committee. This award recognizes coaches who have made outstanding achievements in the Olympic Games[2]. The Table 7 shows some of the coaches' events, working teams and working time.

Table 7: The "Great Coach"

Name	**Event**	**Team**	**Time**
Lang Ping	Volleyball	U.S., China	1996,2008,2016
Béla Károlyi	Gymnastics	Romania, U.S.	1976,1996
Jane Figueiredo	Diving	Great Britain	2016,2024
Vitaliy Petrov	Athletics	Brazil	2016
Laura Martinel	Judo	Argentina	2016
Tae-Suk Jang	Fencing	Korea	2020
Myriam Fox-Jerusalmi	Canoeing	Australia	2020
Malcolm Brown	Triathlon	Great Britain	2012-2016
Ulla Koch	Gymnastics	Germany	2000-2008
Malcolm Arnold	Athletics	Great Britain	1996,2012-2020
Katalin Rozsnyói	Canoeing	Hungary	1992-2000
Kaneko Masako	Artistic swimming	Japan	1984-2004
Jon Urbanchek	swimming	U.S.	1988-2012

4.2.2 Model Construction

In this article, we will employ a statistical technique to examine the variations in coach performance before and after their involvement with a team. To achieve this, we will utilize the one-way within-groups ANOVA (Analysis of Variance). This method is particularly useful for comparing the means of the same group of subjects across different conditions or over different time points and will enable us to rigorously evaluate whether there is a measurable difference in the performance of teams before and after a coach's involvement, and to determine the significance of these changes.

The hypothesis testing in one-way ANOVA includes:

- **Null Hypothesis (H_0)**: All group means are equal, i.e., $\mu_1 = \mu_2 = \cdots = \mu_k$.
- **Alternative Hypothesis (H_a)**: At least one group mean is different from the others.

The test statistic F is given by:

$$F = \frac{SSTR/(k-1)}{SSE/(n-k)} = \frac{MSTR}{MSE}$$

where:

- $MSTR$ (Mean Square of Treatments): Between-group mean square.
- MSE (Mean Square Error): Within-group mean square.

In hypothesis testing, we analyze the F value, which is a test statistic derived from the ratio of between-group variance to within-group variance. If this F value is large enough, it suggests that the differences between the group means are significant, and we reject the null hypothesis. This rejection indicates that the observed differences in group means are unlikely due to random chance, implying that the intervention—in this case, the coach's involvement—has had a measurable effect on team performance.

In this problem, we use the score of the number of the medals to measure how much effect the coaches gave to the athletes. We divided the data into two parts: pre-coaching and during coaching. For the pre-coaching group, we selected the corresponding event scores of the team in the previous Olympic Games to be included in the data. For the group of coaching period, if the coach participated in only one Olympic Games in this period, we will use the scores of the corresponding Olympic Games in the data; If the coach has participated in multiple Olympic Games during this period, the average of the scores of the corresponding events in these Olympic Games is taken into the data.

After processing the data, we get two groups with the same data length. Then we perform a one-way ANOVA on this data. The null hypothesis of the testing is that There is no difference in team performance before and during coaching. And the alternative hypothesis is that There is a difference between the team's performance before and during coaching.

4.2.3 Model Solution

Data Preparation and Processing

In order to conduct an analysis of the "great coach" effect, it is first necessary to process the data of the coaches listed in Table 7 and the countries and sports they coach. The processing mainly consists of the following steps:

Based on the coaches and their tenure periods in Table 7, the medal data for each coach's country and sport must be extracted. For each coach, the data will be divided into the "Before Coach" period and the "During Coach" period:

- **Before Coach Periods**: The medal data for the "Before Coach" period is determined by the performance in the Olympics prior to the coach's tenure. For example, Lang Ping began coaching the Chinese women's volleyball team in 1996 and continued through 2008 and 2016. Therefore, for the "Before Coach" period, we need to look up the medal counts for the Chinese volleyball team in the 1992 Olympics, which represents the performance before Lang Ping's tenure.

  For each coach, the "Before Coach" period data is derived by selecting the medal counts for the relevant country and sport from the Olympics preceding the coach's first appointment. For example:

  - Before Lang Ping coached the Chinese women's volleyball team, we examine the 1992 Olympics to find the medal counts for the Chinese volleyball team.
  - Before Béla Károlyi coached the Romanian gymnastics team, we look at the 1976 Olympics to find the medal counts for the Romanian gymnastics team.

- **During Coach Period**: For the "During Coach" period, if the coach was active during multiple Olympic cycles, the medal data for these cycles is averaged. If the coach was involved in only one Olympic Games, the medal data for that particular Olympic cycle is used. For example, Lang Ping coached the Chinese volleyball team during the 2008 and 2016 Olympics, so the medal data for these two Olympics will be averaged to represent her "During Coach" period.

According to the conclusion in 4.1 "The relationship between the events and medals" , advantageous sports may influence fluctuations in medal counts. Therefore, we classify the data into three groups for analysis: the ALL group, the Dominant group, and the Non-Dominant group.

- **ALL Group**: This group includes all sports, regardless of whether they are considered advantageous sports for the country. It serves as the control group.
- **Dominant Group**: These are the sports considered to be advantageous for the country. Based on the formula in the "4.1.2 Model of Advantageous Sports" we calculate the advantage score for each sport for each country across different Olympic cycles, we classify sports events above the threshold as dominant sports.
- **Non-Dominant Group**: These are the sports that are not considered advantageous for the country. By comparing the scores of sports events during the "Before Coach" period with the advantage score criteria emphasized in the "4.1.2 Model of Advantageous Sports", we classify sports events below the threshold as non-dominant sports.

Based on the classification criteria outlined above and Table 7 The "Great Coach", we obtained the medal counts for the corresponding time periods—'During Coach Period'—by country and sport, as well as the medal counts for the previous Olympic cycle—'Before Coach Period'—for each country and sport. For the convenience of display, we have added up the medal counts of the same country, so that two identical countries will not appear in the chart. Below is a heatmap showing the number of medals awarded:

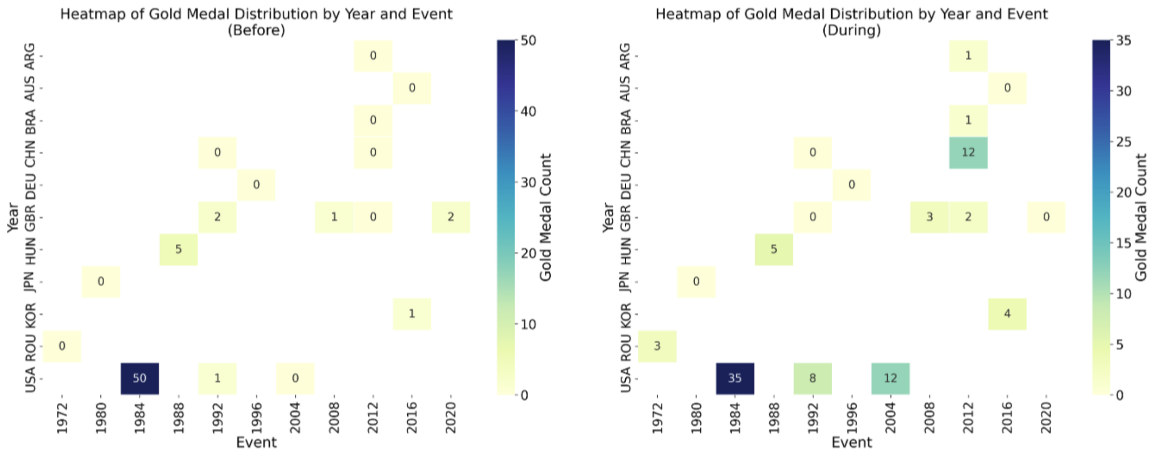

Figure 6: Gold Medal Display

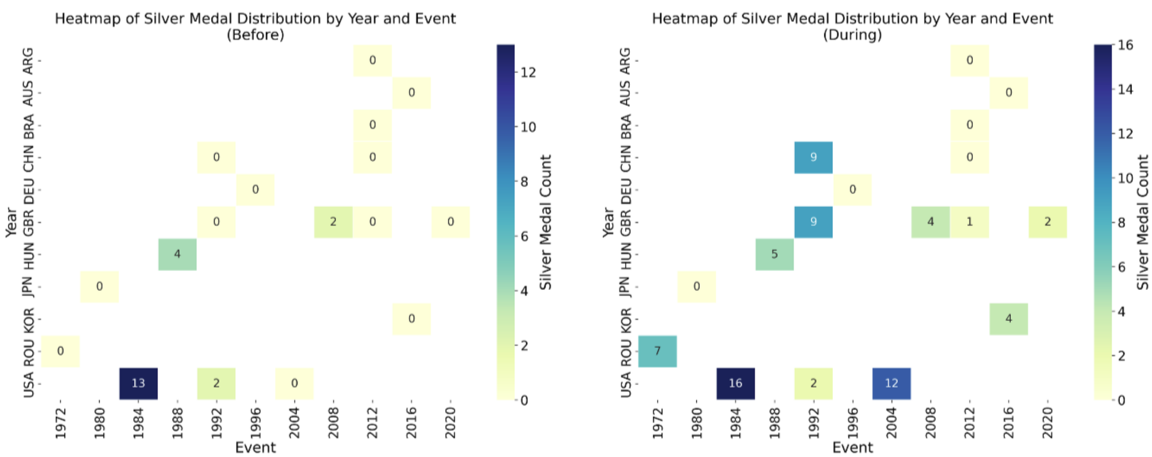

Figure 7: Silver Medal Display

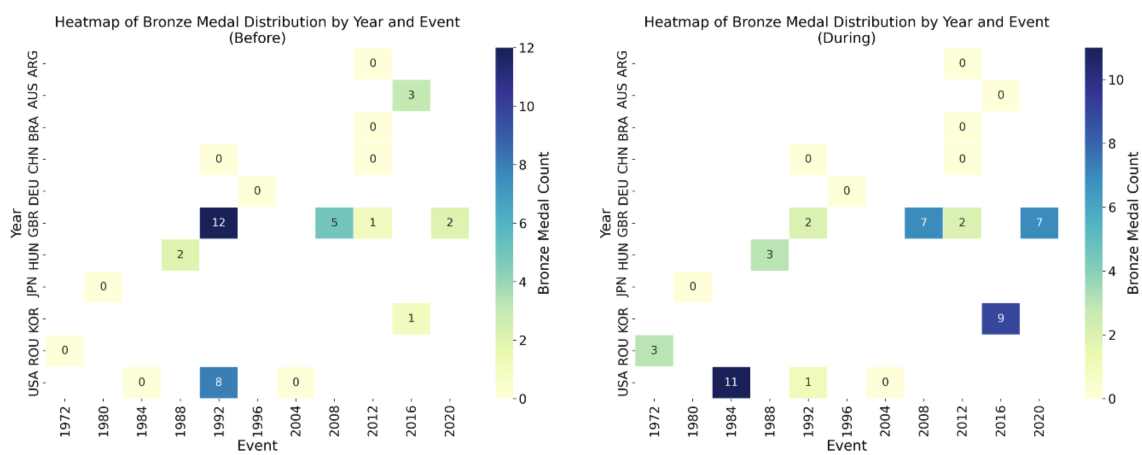

Figure 8: Bronze Medal Display

Based on the number of medals, we also calculated their corresponding scores, Therefore, we have obtained such a table:

Table 8: Advantageous vs. Non-Advantageous Sports Group

Team	**Event**	**Status**	**Time**
Argentina	Judo	Non-dominant	<2016
Australia	Canoeing	Non-dominant	<2020
Brazil	Athletics	Non-dominant	<2016
China	Volleyball	Non-dominant	<2016
China	Volleyball	Non-dominant	<1996
Germany	Gymnastics	Non-dominant	<2000
Great Britain	Athletics	Dominant	<2012
Great Britain	Athletics	Dominant	<1996
Great Britain	Diving	Non-dominant	<2024
Great Britain	Diving	Non-dominant	<2016
Great Britain	Triathlon	Non-dominant	<2012
Hungary	Canoeing	Dominant	<1992
Japan	Artistic swimming	Non-dominant	<1984
Korea	Fencing	Non-dominant	<2020
Romania	Gymnastics	Non-dominant	<1976
U.S.	Gymnastics	Dominant	<1996
U.S.	swimming	Dominant	<1988
U.S.	Volleyball	Non-dominant	<2008

One-way ANOVA Analysis

After identifying the advantageous sports, we visualized the scores and country correspondences used by bar charts and scatter plots, where Score1 represents the score for "Before Coach Period" and Score2 represents the score for "During Coach Period". At the same time, we presented the results of one-way ANOVA to evaluate the impact of the "Great Coach Effect" on team performance before and during coaching. We analyzed the medal counts of three teams: ALL group, Dominant group, and Non-Dominant group. The results of each group, including the F-statistic and p-value, are as follows:

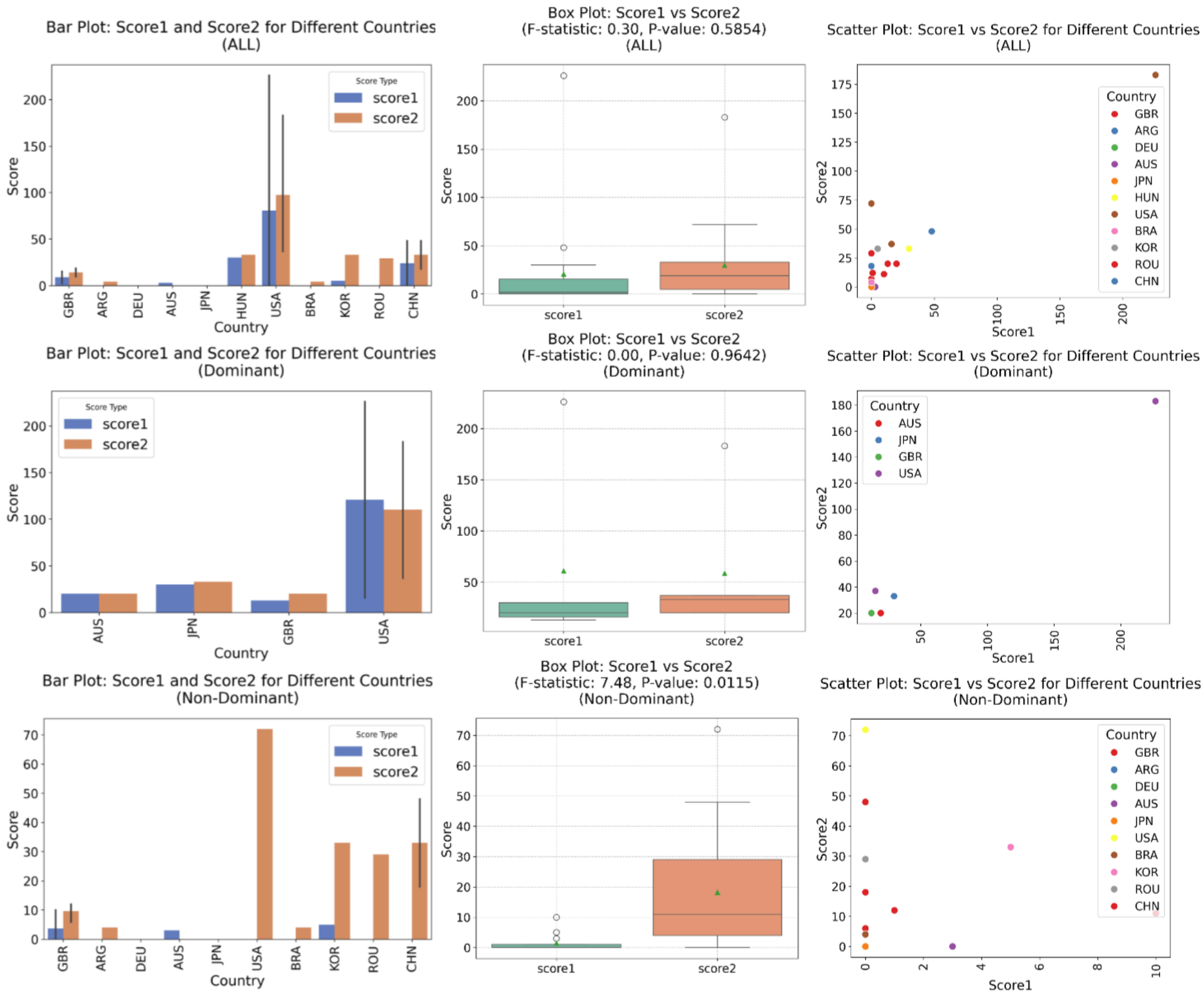

Figure 9: Visualization of Score1 and Score2 Using Bar, Box Plots, and Scatter

Therefore, we have obtained table to display P and F value:

Table 9: Advantageous vs. Non-Advantageous Sports Group

Group type	F	P
ALL Group	0.3033	0.5854
Non-Dominant Group	7.4841	0.0115
Dominant Group	0.0021	0.9642

ALL Group The result indicates that there is no significant difference in the medal performance of the teams before and during the coaching period in the ALL group. Since the p-value (0.5854) is greater than the significance level (0.05), we fail to reject the null hypothesis, meaning that the "Great Coach Effect" is not significant across all sports and countries as a general trend.

Non-Dominant Group For the Non-Dominant group, the analysis shows a significant difference in performance before and during coaching (p-value = 0.0115, which is less than 0.05). This suggests that in non-dominant sports (i.e., sports where the country is not traditionally strong), the "Great Coach Effect" plays a significant role in improving medal performance. The large F-statistic (7.4841) indicates a strong variation between the two periods, supporting the hypothesis that great coaches can notably influence performance in these sports.

Dominant Group In contrast, for the Dominant group, which represents the advantageous sports of the countries, the results indicate no significant difference between the periods (p-value = 0.9642). This high p-value suggests that, in sports where the country already excels, the addition of a great coach does not significantly alter the performance. The F-statistic (0.0021) is extremely low, indicating that the variation between the periods is minimal.

In conclusion, for countries and sports classified as Non-Dominant, the results show a significant improvement in medal performance when a "great coach" is involved. This means that coaching in these sports plays a crucial role in improving results. The **F-statistic** of 7.4841 indicates a substantial variation in medal counts before and during the coaching period. This suggests that the "great coach" effect can lead to a considerable increase in performance, especially in sports where the country was previously not excelling.

"Great Coach" in Three Countries

Based on the given data from the competition, the data predicted in Figure 4: The Probability of Winning first medical in 2028, and the "Great Coach" effect in this section, we have selected three countries that are suitable for investing in "Great Coaches" and provided relevant reasons:

- **China - Volleyball**

  - China has a historical precedent of successfully transforming volleyball into a dominant sport with the influence of great coaches. After the appointment of a renowned coach, the country transitioned from no medals in 2012 (0 gold, 0 silver, 0 bronze) to winning 12 gold medals in 2016 (0 silver, 0 bronze). Additionally, China's performance in 1996 (0 gold, 9 silver, 0 bronze) marked a significant improvement, thus establishing volleyball as a dominant sport during these periods.

  - According to the 2024 data, volleyball currently falls under the category of "Non-Dominant" sports in China. The effect of a "great coach," as confirmed by the data analysis, has a considerable impact on sports in the "Non-Dominant" category, particularly in improving medal performance.

  - China's strong Olympic track record—consistently finishing in the top three in the past five Olympic Games—suggests a deep reservoir of experience and expertise. This positions China in a favorable position to elevate a "Non-Dominant" sport, such as volleyball, to "Dominant" status once again, increasing the likelihood of more gold medals in future Olympic events.

In conclusion, Investing in a "great coach" for volleyball would likely lead China to re-establish volleyball as a dominant sport, with the potential for multiple gold medals, given the proven historical impact of such investments.

- **Samoa - Rugby Sevens**

  - As per the 2024 data, Rugby Sevens in Samoa is classified as a "Non-Dominant" sport. Following the pattern of "great coach" influence observed in other countries, we can expect a substantial effect on improving performance in non-dominant sports like Rugby Sevens.
  - According to Figure 4 in the "4.1.4 Model Solution and Evaluation," predictions indicate a moderate likelihood of Samoa securing its first-ever Olympic medal in Rugby Sevens in the 2028 Olympics. This projection suggests that with the right coach, Samoa could significantly improve its standing.
  - Samoa has shown a strong commitment to developing Rugby Sevens, with 50% (12 out of 24) of its athletes participating in this sport during the 2024 Olympics. This commitment indicates that Rugby Sevens is a major focus for the nation, making it an ideal candidate for investment in a "great coach."

By investing in a "great coach" for Rugby Sevens, Samoa stands a strong chance of winning its first-ever Olympic medal in 2028, marking a significant achievement for the nation.

- **Great Britain - Triathlon**

  - Great Britain has a successful track record of improving its performance in the Triathlon by investing in a "great coach." During the 2012-2016 period, Great Britain achieved a breakthrough, moving from 0 gold, 0 silver, and 0 bronze to 1 gold, 1 silver, and 1 bronze. This indicates the potential for continued success with the right coaching.
  - According to the 2024 data, Triathlon is categorized as a "Non-Dominant" sport for Great Britain. As demonstrated in other cases, "great coach" effects are particularly impactful in sports that are not traditionally dominant for the country, making the Triathlon a prime candidate for improvement.

Investing in a "great coach" for the Triathlon would likely enhance Great Britain's chances of transforming the sport from "Non-Dominant" to "Dominant," potentially leading to an increase in medal counts and improved performances in future Olympics.

4.3 More observations from the data

4.3.1 The original insights

From the figure 10, we can find that for countries with advantageous sports, the number of medals they have won in advantageous events accounts for the vast majority of their total medals. It shows the importance of the advantageous sports for these country. As for these countries, they should focus on maintaining their strengths in the events. Other countries can also learn from relevant experience to further improve their own national strength in this event.

From the coefficient of linear regression model obtained in 4.1.4, we can find that the sponsor effect has a very large impact on the medal count, and can even improve the host country by nearly 20 medals. This reflects the host country's advantage in setting the event and the cheering from the

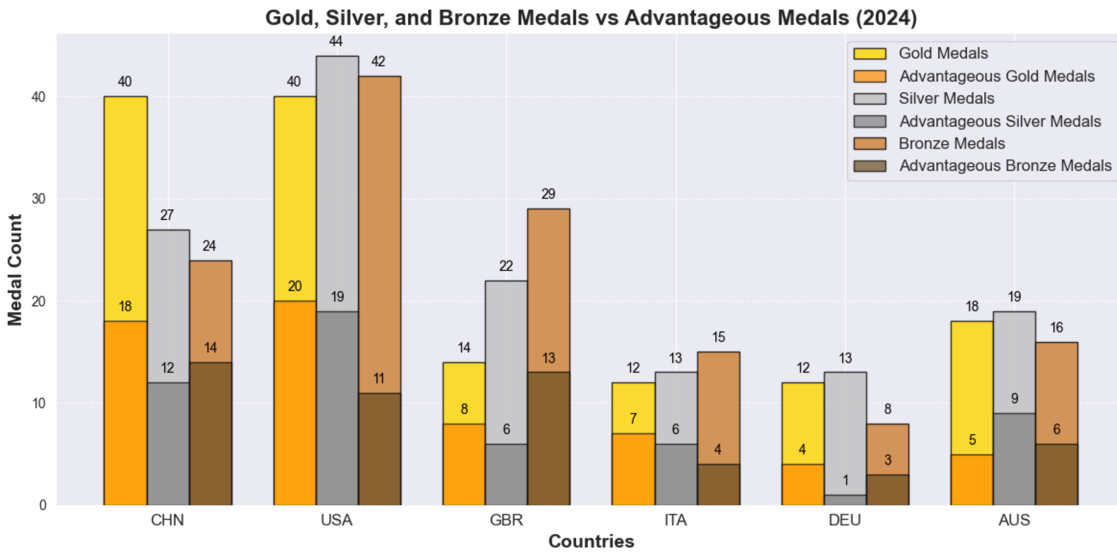

Figure 10: The medal distribution for some countries in 2024

home audience, and also provides great physical and psychological support for the host country players.

From the Figure 11, sports with many events usually have a high popularity in the world and a large number of participants, such as athletes, swimming. These sports have a broad mass base in multiple countries and regions, and typically each requires different training methods and technical requirements. At the same time, these events have a better performance in terms of gender equality and diversity, setting up mixed events or gender equality subcategories. In sports with few events, it is generally difficult to distinguish multiple different games, such as soccer and volleyball, which are only divided into male's and female's groups. These sports are more professional. Some are also limited by competition terrain factors. So it is hard to set more events for these sports.

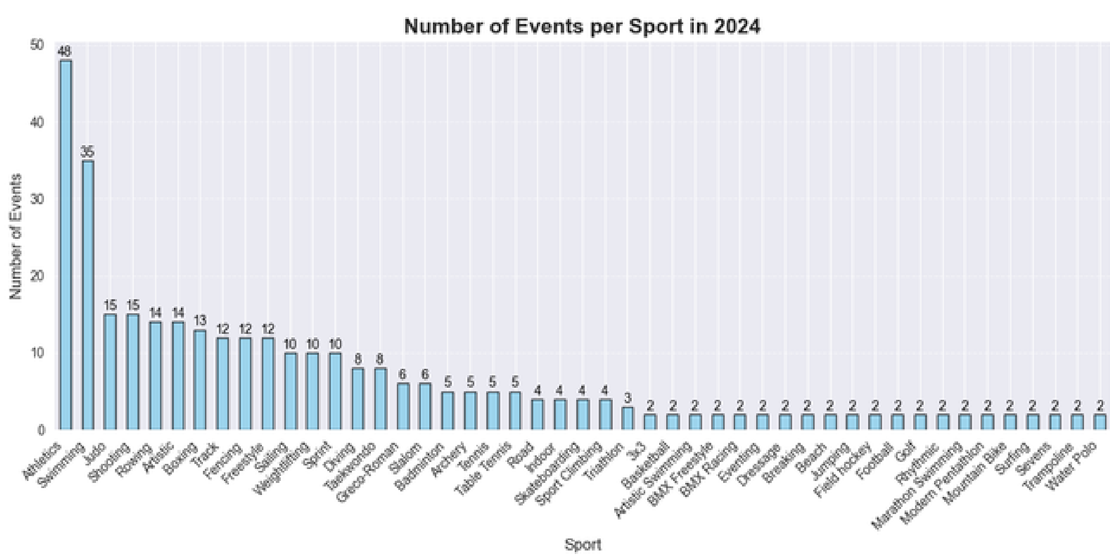

Figure 11: The events distribution in 2024

4.3.2 Suggestions

1. **Assist the sponsor to optimize event distribution:** The sponsor effect has a large impact on the medals. Under the guidance of the International Olympic Committee, the sponsor should adequately take the advantages of the sponsor effect and carefully arrange specific events to make a breakthrough in the number of Olympic medals without disrupting the balance of the event.

2. **Promote zero medal breakthrough in some events for countries that have not won medals:** There are still more than 60 countries that have not won a medal at the Olympics. For these national Olympic committees, they can focus on training athletes to compete for medals in a certain event, so that athletes can achieve better results in the Olympic Games. This can not only show the style of the athletes, but also reflect the improvement of national strength.

3. **Promote transnational coaching of great coaches:** Great coach is an important resource that helps improve your performance. The promotion of transnational coaching makes the comprehensive strength of each country further strengthened. It is very helpful to enhance the competitiveness of countries in various sports and promote more wonderful matches in the Olympic Games.

4. **Support the development of women's and youth sports:** Paris 2024 is the first Olympic Games in history to have an equal number of male and female athletes. This milestone marks an important step forward for gender equality in sports. At the same time, we also find that more and more young athletes have won medals or even gold medals in the Olympic Games. This shows that the development of women's and youth sports plays an important role in sports. In the future, countries should continue to support and promote the development of women's and youth sports.

5 Sensitivity Analysis

5.1 Sensitivity Analysis for the Multiple Linear Regression Model

To assess the robustness of the multiple linear regression model, we conducted a sensitivity analysis by perturbing one or more key features while keeping other variables constant. Specifically, we introduced controlled variations to features such as *Medal_1_prior*, *Medal_2_prior*, and *Total_Advantage_Coefficient* by adding Gaussian noise with varying standard deviations. The goal was to observe how these perturbations influenced the predicted number of gold medals.

The analysis revealed that even under significant perturbations to the selected features, the model's predictions remained stable, demonstrating its robustness. Figures 12 illustrate the minimal fluctuations in prediction outcomes as the features were perturbed. These results underscore the model's ability to maintain reliable performance despite input uncertainties.

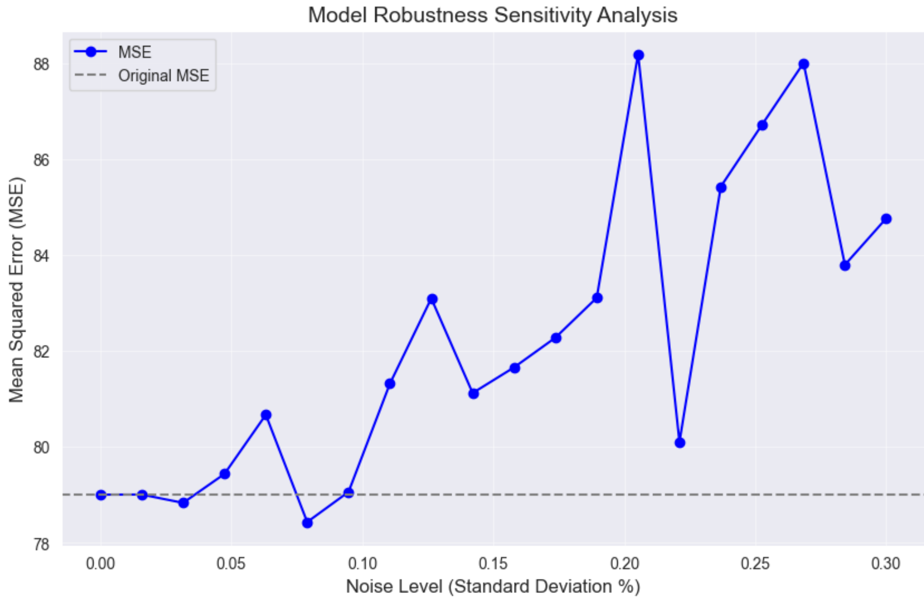

Figure 12: Sensitivity Analysis for the Multiple Linear Regression Model

5.2 Sensitivity Analysis for the Logistic Regression Model

For the logistic regression model predicting the likelihood of a country winning its first medal, a similar sensitivity analysis was performed. We perturbed the features *Events* and *Participants*, which were identified as critical predictors. Controlled noise was added to these features, and the resulting changes in classification outcomes were evaluated.

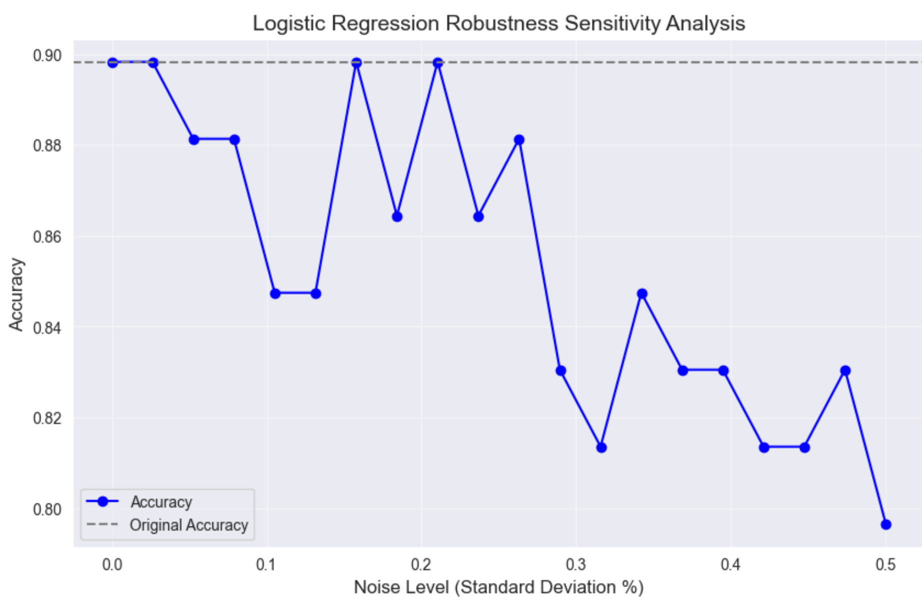

Figure 13: Sensitivity Analysis for the Logistic Regression Model

Figures 13 provide a visual representation of the model's sensitivity analysis, demonstrating its resilience under various conditions.

The results showed that the model's classification accuracy and confusion matrix remained consistent across different levels of perturbation. This indicates that the logistic regression model is robust to variations in its input features, as evidenced by the stability of metrics such as accuracy and the coefficients of the features.

6 Model Evaluation and Promotion

6.1 Model Evaluation

6.1.1 Advantages

- Our model takes into account a variety of data factors, including the historical medal number, the number of participating events, the dominant events and so on, the model can predict the medal number better.
- Our model combines periodic and trend changes in the time dimension, not only considering the current data, but also considering the historical data, making the model more accurate and reasonable.
- Our model gives parameters reasonably to previous events of each country based on time, better studies the advantages of each country.

6.1.2 Limitations

- The types of parameters fitted are few and the prediction accuracy is limited.
- The state of readiness of athletes in specific countries during the Olympic cycle is not taken into account.

6.2 Future Work

6.2.1 Model extension

- Collecting more relevant data such as the GDP of each country, the availability and quality of sports facilities, government sports-related policies, and the level of preparation for each sport. Additionally, recent performance in international competitions could provide valuable insights into current form and competitiveness. These enhancements would allow for more robust and accurate medal predictions.
- The model can be extended to provide a more detailed assessment of the role of coaches in improving athletes' abilities. For example, analyzing how different coaching styles enhance physical fitness, explosive power, or tactical awareness could shed light on the specific contributions of coaching to athletic success.

6.2.2 Model application

• Our model can be applied to other sporting events. Beyond sports, the methodologies used in the model can be applied to other domains, such as arts (e.g., analyzing success in music or dance competitions), finance (e.g., predicting market trends or investment performance), and management (e.g., evaluating team or organizational performance). The versatility of the model highlights its potential for cross-disciplinary applications.

7 Conclusions

Through the study of medal table and other related data, this study has achieved remarkable results. We use linear regression, logistic regression, and One-way within-groups ANOVA to explore the relationship between medal numbers and actual performance. Through the analysis and modeling of the data, we can predict the medal table of the 2028 Olympic Games, discovering the advantages of various countries, studying the countries that may achieve a breakthrough of zero medals, and quantifying the great coaching effect. It plays a role in evaluating each country's ability to win medals, providing data support for each country's sports development, and promoting the rational allocation of sports resources. Although the model still has some limitations in practical application, through future research and improvement, we believe that this model can be more widely used in the sports field. Through better modeling methods and more data support, We can improve the accuracy and applicability of the model, and provide more valuable tools and programs for research in the field of sports.

References

[1] "Country Codes – ISO 3166". International Organization for Standardization (ISO). Archived from the original on 2017-01-23. Retrieved 2018-05-09.

[2] WEI Xin, ZHONG Yaping, ZOU Kening. The More Participation the Longer Sports Career?: Based on the Characteristics of the World Outstanding Track and Field Athletes. Journal of Shanghai University of Sport, 2021, 45(9): 86-98. DOI: 10.16099/j.sus.2021.09.008.

[3] International Olympic Committee. (n.d.). Nominations open for IOC Coaches Lifetime Achievement Awards. Olympics.com. https://olympics.com/ioc/news/nominations-open-for-ioc-coaches-lifetime-achievement-awards [Accessed January 25, 2024].

[4] Los Angeles 2028 Olympic and Paralympic Games. (n.d.). Games Plan. LA28. https://la28.org/en/games-plan.html

[5] Statista. (n.d.). Summer Olympics: share of male and female athletes 1896-2024. Statista. https://www.statista.com/statistics/1090616/olympics-share-athletes-by-gender-since-1896/

Report on Use of AI

1. DeepSeek
   Query1: Where can I find a great coach that has made a huge improvement in a national team from one Olympics to the next, thanks.
   Output: Finding a great coach who has made a significant improvement in a national team from one Olympics to the next involves leveraging various resources and platforms. Here are some specific places and methods to help you identify such coaches:
   1. Official Olympic and Sports Federation Websites
   2. Sports News Outlets and Archives
   3. Academic and Research Institutions
   By utilizing these resources and methods, you can systematically identify and learn about coaches who have made significant improvements in national teams from one Olympics to the next.

2. OpenAI
   Query: Clean and calculate medal counts for advantageous sports
   Output: The model analyzed the concatenated string issue in advantageous sports and suggested using `ast.literal_eval` for proper parsing. It then matched parsed sports with medal data using Python code:

```
import ast

def parse_advantageous_sports(data):
    data['advantageous_sports'] = data['advantageous_sports'].apply(
        lambda x: ast.literal_eval(x.replace("'", "").replace("[", "['").
            ↪ replace("]", "']"))
    )
    return data

# Example medal calculation logic
def calculate_medals(data, medal_data):
    for index, row in data.iterrows():
        sports = row['advantageous_sports']
        medals = sum(medal_data[medal_data['Sport'].isin(sports)]['Total
            ↪ Medals'])
        data.at[index, 'Advantageous Medals'] = medals
```

   This resolved the calculation of medals for advantageous sports, providing both a clean dataset and accurate medal counts.

3. OpenAI
   Query: Parse complex CSV data with tuple-like strings and calculate metrics
   Output: The model identified a parsing issue with CSV columns containing tuple-like strings (e.g., "(1, 2, 3)"). It suggested using `ast.literal_eval` for safe conversion and demonstrated medal computation:

```
import ast

# Parse "Medal" column
```

```
yearly_medals['Medal'] = yearly_medals['Medal'].apply(
    lambda x: ast.literal_eval(x)
)

# Calculate total medals
def calculate_medals(df):
    df['Total_Medals'] = df['Medal'].apply(sum)
    return df

# Example of usage
parsed_data = calculate_medals(yearly_medals)
```

This approach resolved data parsing errors and allowed accurate computation of total medal counts for further analysis.

4. OpenAI
Query: Clean advantageous sports data and match it with medal records
Output: The model addressed issues with concatenated strings in the advantageous sports column. It suggested splitting and cleaning data for accurate mapping:

```
import pandas as pd

# Clean and split advantageous sports
def clean_sports_column(df):
    df['Advantageous_Sports'] = df['Advantageous_Sports'].apply(
        lambda x: [sport.strip() for sport in x.replace('[', '').replace(']
            ↪ ', '').split()]
    )
    return df

# Match and calculate medals
def calculate_advantageous_medals(data, medals_data):
    data['Advantageous_Medals'] = data['Advantageous_Sports'].apply(
        lambda sports: medals_data[medals_data['Sport'].isin(sports)]['
            ↪ Total Medals'].sum()
    )
    return data

# Example usage
cleaned_data = clean_sports_column(advantageous_sports)
result = calculate_advantageous_medals(cleaned_data, yearly_medals)
```

The solution ensured proper matching of advantageous sports with medal records, providing reliable statistics.

5. OpenAI
Query: Kayaking is Canoe Slalom, Which is Canoe Sprint or Canoeing Output: Kayaking belongs to the Canoing category, which can be divided into two main forms of competition:

Canoe Slalom: In this competition, kayaks need to cross a series of water obstacles, usually in turbulent rivers. Athletes need to quickly and accurately maneuver kayaks through these obstacles.

Canoe Sprint: This is a speed competition, usually conducted in calm waters, where athletes need to quickly row kayaks. The distance of the competition is usually 200 meters, 500 meters, or 1000 meters.

So, Kayaking is the core part of the Canoe Slalom and Canoe Sprint competitions.

I hope this helps!