
From Data to Gold: Tier-Based Modeling for Olympic Medal Prediction

Summary

Our research develops a robust framework for predicting medal counts at the 2028 Los Angeles Summer Olympics, grounded in an extensive analysis of historical data, and addresses three key tasks: forecasting medal distribution, evaluating the influence of elite coaching, and uncovering novel insights into Olympic medal dynamics.

Task 1 focuses on medal prediction through a **tier-based modeling approach**. Using **K-means clustering**, we classified countries into five tiers based on recent historical performance, tailoring prediction models to the unique dynamics of each group. With a **metaheuristic** algorithm, we optimized the feature selection process. **Regression techniques**, including LSBoost, XGBoost, random forests, etc., as well as optimizing algorithms, including SSA [8], DBO [9], SCA [10], etc., were implemented with the **TOPSIS** method used to identify the most effective model and algorithm. We also introduced **K-Fold algorithm** to split datasets into training, validating, and testing sets. Additionally, an analysis of the relationship between sports' global importance and medal distribution, using **Spearman's rank correlation coefficient**, confirmed the substantial benefits host countries derive, such as increased medal counts and event-specific advantages.

Task 2 explores the "Great Coach Effect" wherein elite coaches significantly influence national Olympic success. By analyzing **coaching transfers** and **performance trends**, we quantified the impact of high-caliber coaching on medal counts. **Regression analysis** revealed that this effect is particularly pronounced in **lower-tier/underperforming countries**, where elite coaching can lead to substantial improvements. Based on these findings, we recommend prioritizing the recruitment of top-tier coaches in countries with the potential for notable performance gains.

Task 3 delves into the original insights revealed from our model. Firstly, we explored the concept of **Global Sports Equity**, highlighting the significant impact of elite coaching on lower-tier countries and recommending the sharing of coaching expertise to foster a more equitable global sports landscape. Secondly, we analyzed the **Host Country Effect**, observing substantial increases in medal counts for host nations and suggesting regulations to mitigate this advantage. Thirdly, we identified a **Time-Sensitive Element** in our prediction model, emphasizing the importance of recent match data in optimizing performance. Lastly, we discussed the **Granularity Effect** in prediction models, recommending a diversified participation strategy to maximize medal-winning opportunities. Our **sensitivity analysis** further confirmed the model's stability with respect to the number of estimators, ensuring robust prediction performance.

Keywords: Olympics Medal Prediction; Regression Models; K-Fold; Optimizing Algorithm; Metaheuristic Algorithm

Contents

1	Introduction	3
1.1	Problem Background	3
1.2	Literature Review	3
1.3	Our Work	3
2	Assumptions and Justifications	3
2.1	Historical Consistency	3
2.2	Comprehensive national power's great impact on medal count	3
2.3	Host Country Advantage	4
2.4	Time Proximity	4
2.5	Sport as minimal granularity	4
3	Notations and Definitions	4
4	Data Preprocessing	5
4.1	Early Attempts: Athlete-based Prediction	5
4.2	Current Method: Sport/Event Oriented Data Preprocessing	5
4.3	Cumulative Medal-Winning Countries by Olympic Year	6
5	Establishing the Model	6
5.1	Data Loading & Transformation	7
5.2	Feature Engineering	7
5.3	Data Splitting	7
5.4	Model Training	7
5.5	Model Evaluation	7
5.6	Model Selection	8
5.7	Model Optimization	8
5.8	Uncertainty Quantification and Prediction Intervals	9
5.9	Conclusion	9
6	Task1 — Employing the model for prediction	10
6.1	Subtask 1 — Predicting Medal Counts for the 2028 Los Angeles Olympics	10
6.1.1	Prediction Results	10
6.1.2	Analysis of Prediction Results	10
6.2	Subtask 2 — Predicting the Number of Countries Winning Their First Medal	10
6.2.1	Analysis	10
6.2.2	Validation	10

6.2.3	Results	13
6.3	Subtask 3.1 — Exploring the Relationship Between Event Count and Medal Count	13
6.3.1	Visualization	13
6.3.2	Interpretation of Results	14
6.4	Subtask 3.2 — Identifying the Most Important Sports for Various Countries . . .	14
6.4.1	Methodology	14
6.4.2	Results and Insights	15
6.5	Subtask 3.3 — Impact of Events Chosen by Host Countries on Olympic Results .	15
6.5.1	Results and Comparative Insights	16
6.5.2	Conclusion	16
7	Task2 — The Great Coach Effect: its identification and impact	16
7.1	Identification	16
7.1.1	Methodology	16
7.1.2	Implementation	17
7.2	Quantification	18
7.2.1	Methodology	18
7.2.2	Implementation	18
7.3	Exemplification	19
7.3.1	Synchronized Swimming Women’s Team: Ana Tarré	19
7.3.2	Gymnastics Women’s Team All-Around: Béla Károlyi	19
7.3.3	Women’s Volleyball: Lang Ping	20
7.4	Recommendation	20
8	Task3 — Original Insights revealed from the model	20
8.1	Global Sports Equity	20
8.2	Host Country Effect and Regulation	21
8.3	Time-Sensitive Element in Prediction	22
8.4	Granularity Effect in Prediction Models	22
9	Sensitivity Analysis	22
10	Conclusion	23

1 Introduction

1.1 Problem Background

The Olympic medal table is a key focus for nations and fans, reflecting athletic success and national pride. Predicting medal counts is challenging due to complex factors like event types, host country advantages, and new competitors. This problem requires models based on provided datasets, including historical medal tables, event breakdowns, and athlete performance.

Traditional forecasting methods, like OLS regression and Poisson models, often lack accuracy, especially when the precise data of participants and events is absent. This problem emphasizes predicting medal breakthroughs for the introduction of new algorithms and projection strategies based on historical trends.

Meanwhile, the dynamic nature of sports, with evolving rules, emerging talents, and varying levels of investment in different countries, adds another layer of complexity. Our approach incorporates regression model techniques and tier-based modeling to address these challenges, aiming to provide more accurate and insightful predictions. Our model strives to capture the multifaceted influences on Olympic performance, offering a comprehensive tool for forecasting future medal counts.

1.2 Literature Review

Predicting Olympic medal counts has been a topic of interest for researchers across various fields, including economics, sociology, and sports science. Early studies, such as those by Ball (1972)[1], focused on identifying fundamental socioeconomic and demographic factors that influence a nation's medal count. Subsequent research, including work by Bernard and Busse (2004) and Xun Bian (2005) [2, 3], further explored the impact of political systems and hosting advantages on medal counts, finding that hosting the Games and having a centrally-planned economy can significantly enhance a country's performance. Recent advancements in statistical and machine learning techniques have led to more sophisticated models for predicting Olympic medals. Forrest et al. (2010) [4] enhanced traditional regression-based models by including additional regressors such as public spending on recreation and the effects of future hosts.

Despite significant progress, predicting Olympic medals remains a complex task due to the interplay of numerous factors. Overall, the field has evolved from simple correlation-based models to sophisticated machine learning algorithms, with traditional socioeconomic factors remaining significant predictors of Olympic success.

1.3 Our Work

Our entire workflow is demonstrated in *Figure 1*.

2 Assumptions and Justifications

2.1 Historical Consistency

We assume that historical trends in medal counts are consistent with future performance. This assumption is at the heart of our model and is supported by previous research on the predictability of Olympic medal counts [4, 5, 7].

2.2 Comprehensive national power's great impact on medal count

It's been widely acknowledged that a country's comprehensive national power, including economic, political, and social factors, has a significant impact on its Olympic performance

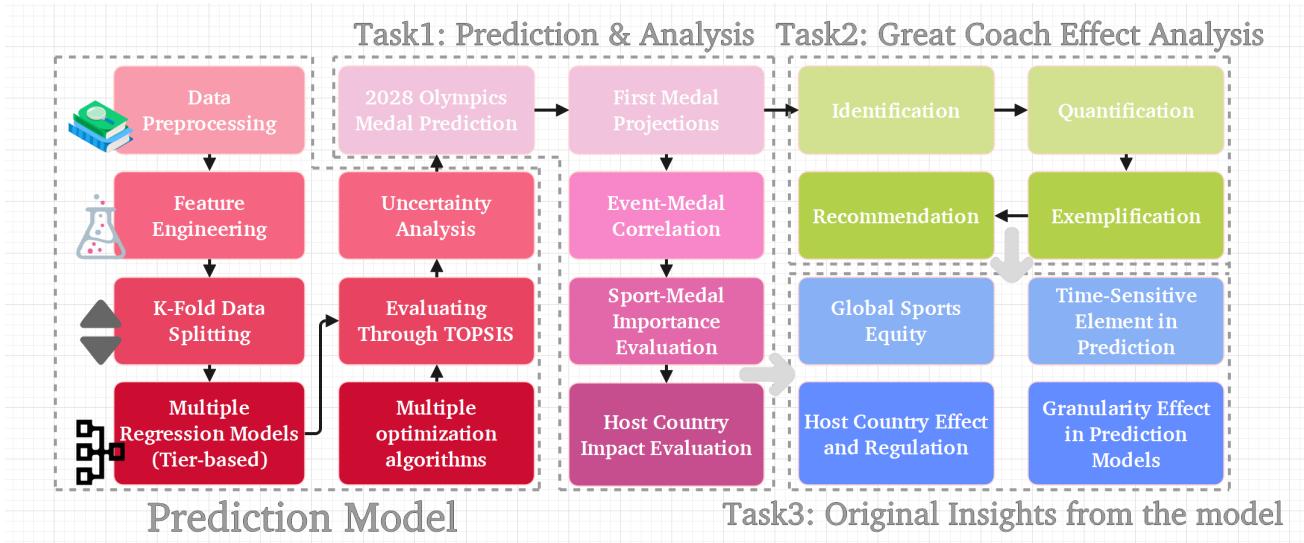


Figure 1: Our entire workflow

[1, 2, 3, 7]. Because the provided dataset doesn't include the detailed information of these factors like GDP, we manually categorize countries into different tiers based on their historical medal counts, which is a good proxy for their comprehensive national power. For more detail, see Data Preprocessing part.

2.3 Host Country Advantage

We assume that host countries have a competitive advantage due to factors such as home crowd support and familiarity with the venues.

2.4 Time Proximity

We assume that only the recent Olympic data is relevant for predicting future medal counts. This assumption is based on the idea that recent performance trends and current athlete conditions are more indicative of future outcomes than older data. Therefore, our model primarily focuses on the last few Olympic Games to ensure the predictions are based on the most relevant and up-to-date information.

2.5 Sport as minimal granularity

We assume that using sport as the minimal granularity of data is the best for predicting medal counts, namely, we predict the medal outcome of each sport and sum them up to obtain the projected medal count for each country. This is justified by our trial-and-error process, which found that using event or individual athlete as the minimal granularity of data leads to overfitting and poor generalization. Granularity too small leads to higher randomness and fewer samples for each unit, and neither of them are beneficial for the accurate prediction of medal counts in the next Olympic Games. Therefore, we use sport as the minimal granularity of data in our model.

3 Notations and Definitions

Our model employs a range of notations and definitions to facilitate the analysis and prediction of Olympic medal counts. These notations are summarized in *Table 1*.

Notation	Definition
C_i	Country i , $i \in \{1, 2, \dots, N\}$; N : total number of countries.
E_j	Event j , $j \in \{1, 2, \dots, M\}$; M : total number of events.
S_k	Sport k , composed of related events E_j , $k \in \{1, 2, \dots, L\}$; L : total number of sports.
$m_{i,j}$	Medals won by country C_i in event E_j from 2012 to 2024.
M_i	Total medals won by country C_i , $M_i = \sum_{j=1}^M m_{i,j}$.
$w_{i,j}$	Importance of event E_j for country C_i , $w_{i,j} = \frac{m_{i,j}}{M_i}$.
W_j	Global importance of event E_j , $W_j = \sum_{i=1}^N w_{i,j}$.
I_k	Global importance of sport S_k , $I_k = \sum_{j \in S_k} W_j$.
S_{xx}	Sum of square deviation of x , $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$.

Table 1: Summary of notations and definitions

4 Data Preprocessing

4.1 Early Attempts: Athlete-based Prediction

First, we ruled out the approach of predicting medal counts based solely on the historical total medal data of countries, as this method is both logically flawed and overly simplistic. Given that the most detailed dataset provided includes information on individual Olympic competitors, their respective sports, events, and results, it is a natural and more precise approach to base medal predictions on the performance of individual athletes participating in specific events.

Furthermore, we aimed to incorporate the host country as a training feature by matching it to the competing country, while the dataset only provided the host city. Recognizing the National Olympic Committee (NOC) code as a standardized and uniform identifier for countries, we created a dataset named "City_NOC.csv" to serve as a dictionary containing all Olympic host cities and their corresponding NOC codes. Using this dictionary, we developed a Python script to add a new column, "host_NOC", to the dataset while removing the original "City" column. Additionally, we excluded all data from the 1906 Olympics, as it was an informal event and disrupted the four-year cycle traditionally followed by the Olympic Games.

To further enhance data quality, we normalized medal counts by dividing them by their corresponding team sizes [AI: 2] using a Python script. This normalization step ensured that the data accounted for the nature of team events, eliminating biases caused by repeated medal counts. We subsequently created a training dataset consisting of individual participants as rows and their corresponding medals earned and host NOC for each year as columns.

4.2 Current Method: Sport/Event Oriented Data Preprocessing

Despite these efforts, the athlete-based prediction results were significantly lower than the actual medal counts. This discrepancy likely stems from the instability of individual performances and the reality that most athletes do not participate in multiple Olympic Games, let alone win multiple medals. Consequently, we shifted to a more aggregated and macroscopic approach, categorizing data by country(NOC) and sport. We anticipated this method would yield more stable and reliable predictions.

To refine the dataset further, we addressed an issue where substitutes were included in

the list of athletes, rendering team size unsuitable for normalizing medal counts. To resolve this, we removed columns such as Name, Sex, Team, and Event, which were irrelevant to the country- and sport-based approach. We then eliminated duplicate rows to create a cleaned dataset, "summerOly_medal_norepeat.csv", which contained all medals without repetition or omission.

Moreover, we recognized "No Medal" as a potential indicator of a country's capability in specific events. Countries participating in a sport are likely to perform better than those that do not. Therefore, we introduced a "No Medal" column as a parallel feature alongside the Gold, Silver, and Bronze medal counts for model training.

To enhance prediction accuracy, we employed the K-means clustering algorithm to classify all participating countries into 5 distinct tiers based on their historical medal counts. As a result, Tier 5, the highest tier, comprised only the United States (USA) and China (CHN), which dominated in gold and total medal counts. Tier 4 and Tier 3 were identified based on a significant gap in gold medal counts, with Italy (38) marking the boundary between the two tiers and New Zealand (27) serving as the lower boundary of Tier 3. Tier 2 and Tier 1 were marked by total medal counts of 14 and 6, respectively. This tier-based categorization allowed for a more structured and nuanced analysis of countries' performance trends.

4.3 Cumulative Medal-Winning Countries by Olympic Year

Problem 1.2 requires projecting how many countries will earn their first medal in the next Olympics. To address this, we utilized the "summerOly_medal_norepeat.csv" dataset that was prepared earlier. A new column, "Medal", was added to the dataset, representing the sum of the "Gold", "Silver", and "Bronze" columns to indicate whether a given row corresponds to a medal-winning entry.

Subsequently, we developed a Python script to calculate the year in which each country earned its first medal. For countries that have not yet won any medals, the "Year" column was assigned the value 2028 to reflect the next Olympic cycle. Following this, another Python script was created to calculate the cumulative number of countries that had won medals up to each Olympic year. This cumulative count served as the foundational data for time series projection.

5 Establishing the Model

In this section, we outline the process of developing a regression-based model to predict medal counts (Gold, Silver and Bronze medals) for each country in the Los Angeles 2028 Summer Olympics. Our methodology accounts for historical performance, and hosting effects. The proposed approach includes estimates of uncertainty, prediction intervals, and measures of model performance.

Thanks to the nature of this problem being inherently correlated to chronological projection and the provided data being complete with regard of countries, sports, events and individual athletes, we adopt time series analysis and regression models to predict medal counts for the 2028 Los Angeles Olympics.

To specify, we adopt a tier-based modeling strategy, aligning with the classification to be established in Task 1 - Subtask 3. Countries are grouped into five tiers based on their historical performance. Separate regression models are trained for each tier to capture the unique dynamics of countries within the same tier. The methodology involves the following steps:

5.1 Data Loading & Transformation

- **Loading:** We use data from the most recent eight Olympic Games, including information on hosting nations and medal counts (Gold, Silver, and Bronze) for each participating country.
- **Transformation:** Hosting nation information is one-hot encoded to introduce categorical variables. To address class imbalance, the SMOTE (Synthetic Minority Oversampling Technique) algorithm is applied, generating synthetic samples for underrepresented classes and balancing the dataset.

5.2 Feature Engineering

Features are selected and optimized using metaheuristic algorithm to determine the most effective combinations. Like the image showed, this process involves transforming the feature selection problem into a binary optimization problem, where:

- A value of 0 indicates the exclusion of a feature.
- A value of 1 indicates the inclusion of a feature.

A regression model is trained iteratively with different feature combinations, and the performance of each combination is evaluated.

5.3 Data Splitting

Given the relatively small dataset size, we adopt k-fold cross-validation to split the data into training, validation, and testing subsets. This approach ensures that the model is evaluated robustly across multiple splits of the data.

5.4 Model Training

We train several regression models on the training set, including: LSBoost, XGBoost, Multivariate Linear Regression, Gaussian Regression, Decision Tree, Random Forest. Each model is trained using the training set and evaluated on the validation set to determine the optimal hyperparameters.

5.5 Model Evaluation

The performance of each model is assessed on the test set using the following metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2)

The evaluation result is shown in *Figure 3*. From the result we can see that each medal type can have very different best models. This phenomenon also vouches for the veracity of our tier-based modeling strategy.

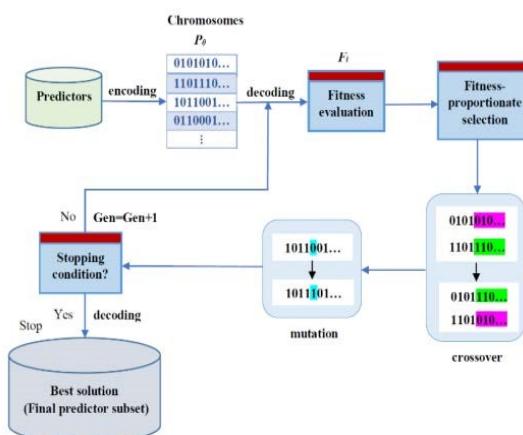


Figure 2: Mechanism of metaheuristic algorithm

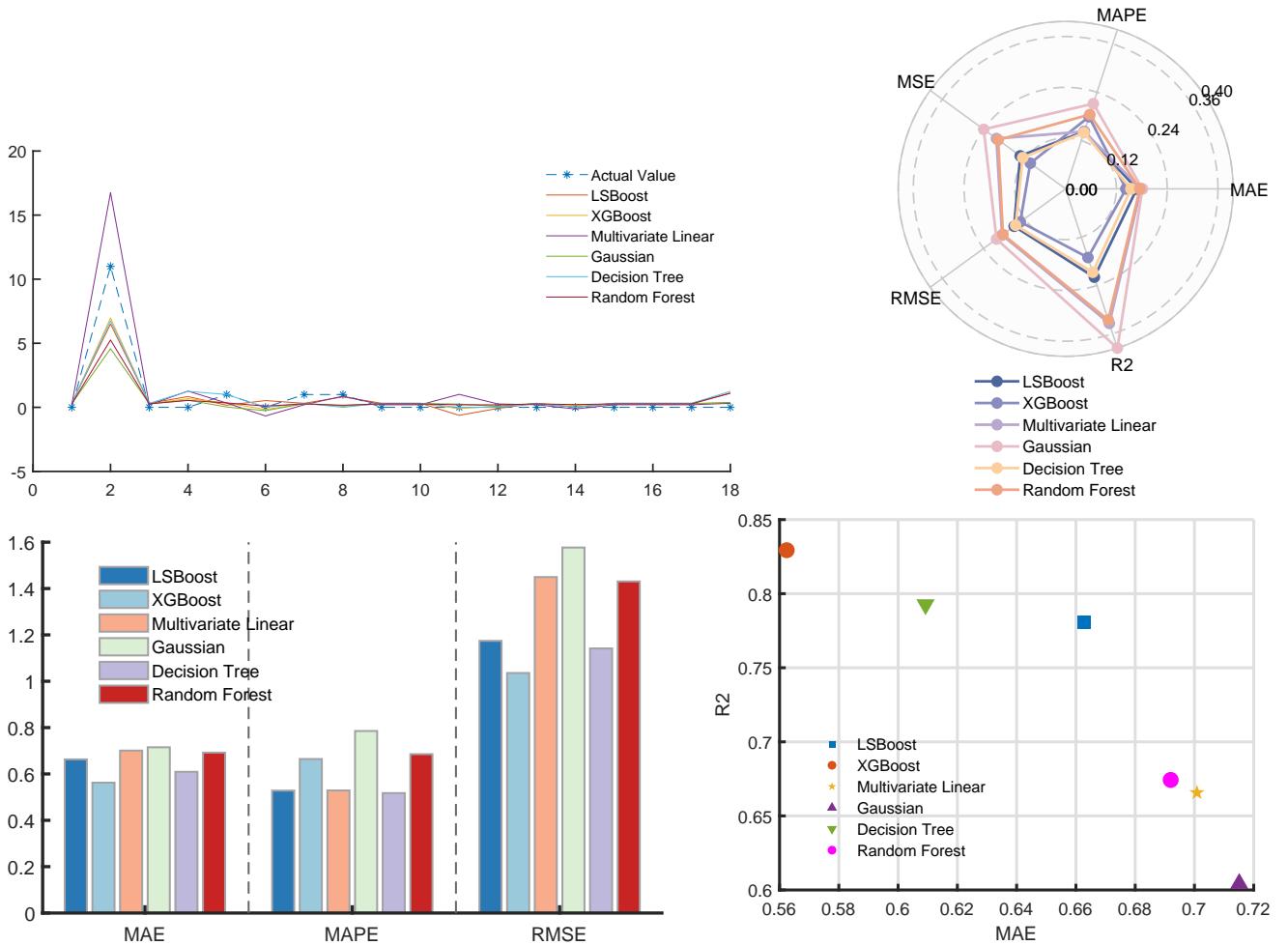


Figure 3: Performance comparison of different regression models for predicting Tier 1 Gold medal prediction model. Subfigures illustrate different comparisons across models.

5.6 Model Selection

The best-performing model is selected using the TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) method. This approach ensures a balanced evaluation by simultaneously considering models with:

- Smaller values of MAE, MAPE, MSE, and RMSE.
- Larger values of R^2 (closer to 1 indicates better fit).

Among the Tier 1 Gold medal prediction models, for instance, XGBoost exhibited the best performance after evaluation. It's also shown later in practice that each medal type can have very different best models.

5.7 Model Optimization

We apply various optimization algorithms to enhance model performance further. The algorithms tested include: SSA [8], DBO [9], SCA [10], SA [11], PSO [12], SO [13], POA [14], GWO [15].

As an example, SSA algorithm demonstrated the best results for optimizing the Tier 1 Gold

medal prediction model, and the optimization outcome is shown in *Figure 4*.

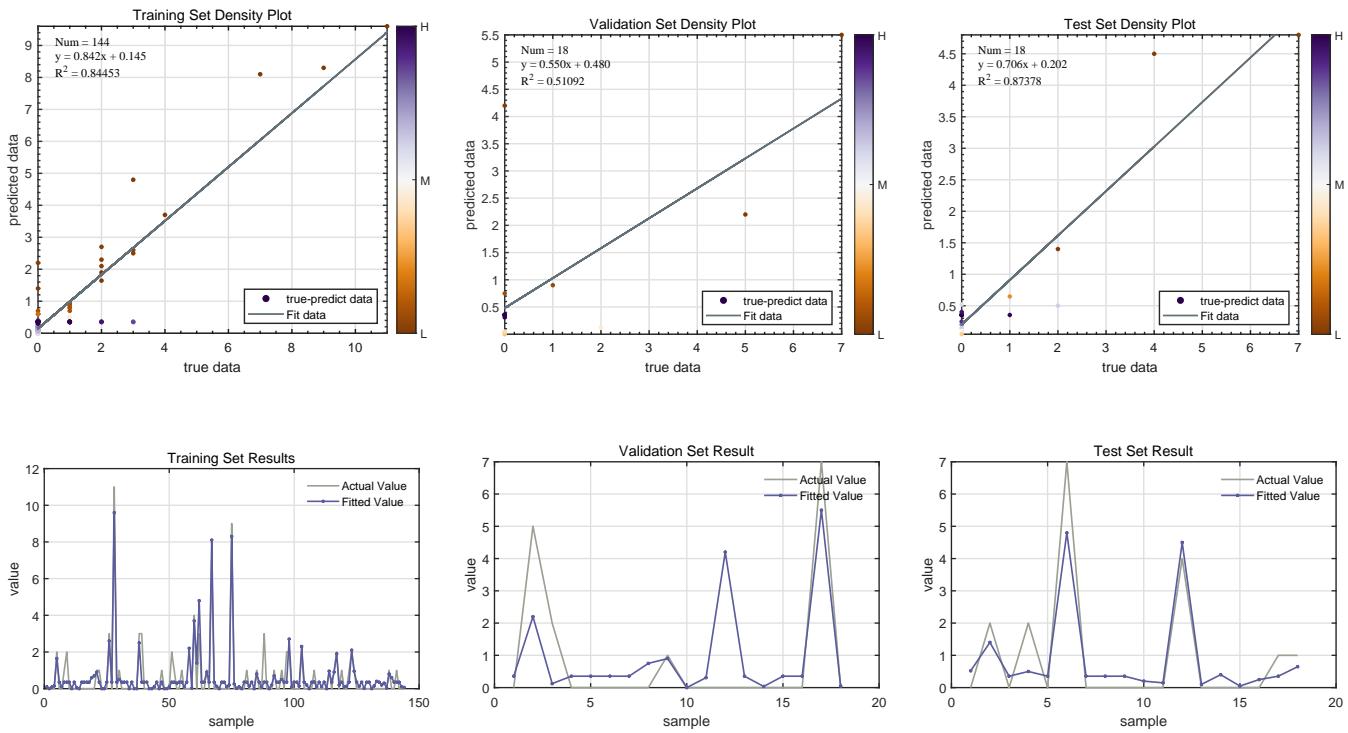


Figure 4: Density plot and result of training set, validation set, and test set for Tier 1 Gold medal prediction model.

5.8 Uncertainty Quantification and Prediction Intervals

To quantify the uncertainty in model predictions, we use Gaussian Probability Interval Prediction. This method calculates prediction intervals, providing upper and lower bounds for the estimated medal counts. The intervals represent the range within which the actual medal counts are expected to fall within at a specified confidence level, as shown in *Figure 5*.

5.9 Conclusion

The tier-based regression modeling approach provides a structured and efficient framework for predicting Olympic medal counts. By incorporating historical performance, host nation effects, and optimization techniques, the model achieves robust performance and reliable projections. The results offer actionable insights for countries aiming to improve their Olympic performance.

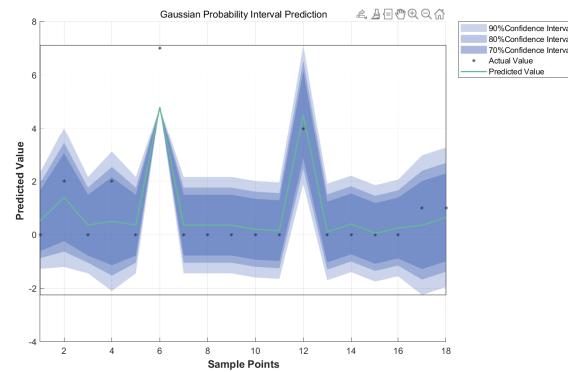


Figure 5: Gaussian Probability Interval Prediction.

6 Task1 — Employing the model for prediction

6.1 Subtask 1 — Predicting Medal Counts for the 2028 Los Angeles Olympics

6.1.1 Prediction Results

We tested the established model to predict medal counts for the 2028 Los Angeles Olympics. The resulting medal table is presented as *Table 2*: (results rounded to one decimal place, and the prediction intervals attached are calculated at a 95% confidence level)

6.1.2 Analysis of Prediction Results

In this section, we analyze the improvements or declines in medal counts. We define "*improvement*" and "*decline*" as follows in order to separate them from normal fluctuations:

- **Improvement:** If the predicted medal count for a country in 2028 is higher than its medal count from the previous Olympics more than a constant X percent (**Condition 1**) or for every type of medal, the count doesn't decrease and at least one type of medal shows an increase in count (**Condition 2**), we consider it an improvement.
- **Decline:** If the predicted medal count for a country in 2028 is lower than its medal count from the previous Olympics more than a constant X percent (**Condition 1**) or for every type of medal, the count doesn't increase and at least one type of medal shows a decrease in count (**Condition 2**), we consider it a decline.

With X set to 15%, we identified the following countries that are expected to improve or decline in medal counts at the 2028 Los Angeles Olympics shown in *Table 3*.

6.2 Subtask 2 — Predicting the Number of Countries Winning Their First Medal

In this subtask, we aim to predict the number of countries that will win their first Olympic medal at the 2028 Los Angeles Olympics. This involves analyzing historical data to identify trends and factors that contribute to countries winning their first medals and applying these insights to make predictions for the upcoming Olympics.

6.2.1 Analysis

Firstly, we calculated historical data on countries that have won their first Olympic medals in past Games, which is done in Data Preprocessing. The visualization of this dataset is shown in *Figure 6* [AI: 1].

Looking at this figure we discovered that the data points of recent Olympics exhibit a strong linear relationship. Therefore, we applied linear regression to the data point of recent 8 Olympics (y represents the number of countries that have won their first Olympic medals in past Games, x represents year), the result was as follows.

6.2.2 Validation

Theorem 6.1. *The confidence interval for the regression function $\mu(x_0) = \hat{a}x_0 + \hat{b}$ at a confidence level of $(1 - \alpha)$ is*

$$\left(\hat{a}x_0 + \hat{b} \pm t_{\alpha/2, n-2} \cdot \sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

It's notable that $R^2 = 0.9935$, indicating a very high fit to the data, meaning the linear model is highly accurate for this dataset.

NOC	B	S	G	BL	BU	SL	SU	GL	GU	Tot
USA	48.8	45.3	45.9	46.3	51.3	43.9	46.8	43.3	48.6	140.1
CHN	27.3	30.1	46.7	25	29.6	28.8	31.4	44.3	49.1	104.1
GBR	38.8	24.6	20.5	36.5	41.1	23.3	25.9	18.1	22.9	83.9
JPN	19.1	22	26.3	16.8	21.4	20.7	23.3	23.9	28.7	67.4
FRA	30.4	12.1	21.3	28.1	32.7	10.8	13.4	18.9	23.7	63.8
AUS	18.3	15.4	21.7	16	20.6	14.1	16.7	19.3	24.1	55.4
ITA	15.4	11.5	13.6	13.1	17.7	10.2	12.9	11.2	16	40.6
NED	17	10.1	11.7	14.7	19.3	8.8	11.4	9.3	14.1	38.9
GER	6.4	18	10.8	4.1	8.7	16.6	19.3	8.4	13.2	35.2
CAN	11.5	5.5	10.5	9.2	13.7	4.2	6.8	8.1	12.9	27.4
KOR	10.7	7.9	7.8	8.4	13	6.6	9.2	5.4	10.2	26.4
BRA	5.6	8.2	6.8	3.3	7.9	6.9	9.5	4.4	9.2	20.6
NZL	1.7	5.5	12.5	0	4	4.2	6.8	10.1	14.9	19.8
UKR	7.9	5.2	5.7	5.6	10.2	3.9	6.5	3.3	8.1	18.8
ESP	7.5	5.7	2.6	5.2	9.8	4.4	7	0.2	5	15.8
HUN	3.9	7.5	3	1.6	6.2	6.2	8.8	0.6	5.4	14.3
AZE	5	5.2	3	2.7	7.3	3.9	6.5	0.6	5.4	13.2
DEN	6.6	3.8	2.3	4.3	8.9	2.5	5.1	0	4.8	12.7
BEL	5.5	2	5.2	3.2	7.8	0.7	3.3	2.8	7.6	12.7
IRI	4.9	3.1	3.5	2.6	7.2	1.8	4.4	1.1	5.9	11.5
JAM	2.5	6.6	2.3	0.2	4.8	5.3	7.9	0	4.7	11.4
KAZ	4.6	3.8	2.8	2.3	6.9	2.5	5.2	0.4	5.2	11.3
CUB	5.7	3.2	2.1	3.4	8	1.9	4.5	0	4.5	11.1
KEN	3.2	5.4	2.4	0.9	5.5	4	6.7	0	4.8	11
SWE	3.2	3.6	3.9	0.9	5.5	2.3	4.9	1.5	6.3	10.7
SUI	5.7	2.5	2.4	3.4	8	1.2	3.8	0	4.8	10.6
POL	4.6	2.2	3.2	2.3	6.9	0.9	3.5	0.8	5.6	9.9
GRE	6.8	1.2	0.7	4.5	9.1	0	2.6	0	3.1	8.7
UZB	3.2	3.1	2.4	0.9	5.5	1.8	4.4	0	4.8	8.7
GEO	3.2	2.4	2.7	0.9	5.5	1.1	3.8	0.3	5.1	8.4
LTU	3.2	1.6	2.5	0.9	5.5	0.3	2.9	0.1	4.9	7.3
CRO	3.1	3.2	0.9	0.8	5.4	1.9	4.5	0	3.3	7.3
TUR	4.8	2.6	0	2.5	7.1	1.3	3.9	0	2.1	7.1
ROU	3.3	0.9	2.9	1	5.6	0	2.2	0.5	5.3	7.1
BUL	2.5	1.4	3	0.2	4.8	0.1	2.7	0.6	5.4	6.9
NOR	3.8	1.4	1.4	1.5	6.1	0.1	2.7	0	3.8	6.6
THA	2.6	2.3	1.5	0.3	4.9	1	3.6	0	3.9	6.4
CZE	2	2.7	1.4	0	4.2	1.4	4.1	0	3.8	6.1
RSA	2.1	3.4	0.6	0	4.4	2	4.7	0	3	6.1
ETH	1	2.8	1.2	0	3.3	1.5	4.1	0	3.6	5
TPE	2.2	1.3	1.5	0	4.5	0	2.6	0	3.9	5
IND	1.4	1.9	1.6	0	3.7	0.5	3.2	0	4	4.9
MEX	2.2	2.7	0	0	4.5	1.4	4	0	2.2	4.8
...

Table 2: Predicted medal counts for 2028 Los Angeles Olympics. B: Bronze, S: Silver, G: Gold, BL: Bronze Lower Bound, BU: Bronze Upper Bound, SL: Silver Lower Bound, SU: Silver Upper Bound, GL: Gold Lower Bound, GU: Gold Upper Bound. Due to space constraints, only the top 43 countries are shown.

NOC	Change	Conditions Met	NOC	Change	Conditions Met
ARG	↑	1, 2	KAZ	↑	1, 2
AZE	↑	1, 2	SUI	↑	1, 2
BEL	↑	1	SVK	↑	1
CHN	↑	2	UKR	↑	1, 2
CUB	↑	1	USA	↑	2
DEN	↑	1, 2	HUN	↓	1
ETH	↑	1	IND	↓	1
GBR	↑	1, 2	KOR	↓	1
GEO	↑	1	MAR	↓	1
INA	↑	1	NOR	↓	1
JAM	↑	1, 2	POR	↓	1
JPN	↑	1, 2	ROU	↓	1

Table 3: Countries with Predicted Changes in Medal Counts for the 2028 Los Angeles Olympics

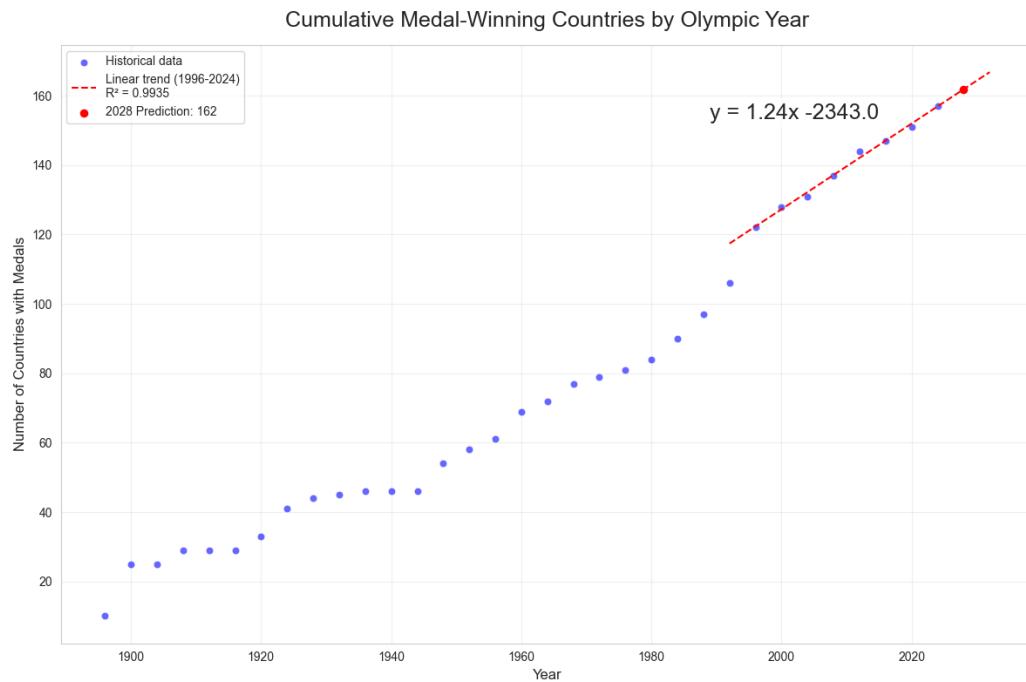


Figure 6: Cumulative Medal-Winning Countries by Olympic Year

Then the projection for number of countries that have won medals till 2028 ($x=2028$) $\hat{y}_{2028} = \hat{a}x + \hat{b} = 161.86$. Therefore, the projection for new medal-winning countries in the next Olympics is $\Delta y = \hat{y}_{2028} - y_{2024} = 4.86$.

Based on our calculations, the residual standard deviation $\sigma = 0.917$, and $S_{xx} = \sum_{k=1}^8 (x_k - \bar{x})^2 = 672$. For a 70% confidence level, we refer to the t-distribution table and find $t_{0.15,6} = 1.134$.

6.2.3 Results

Based on our analysis and model predictions, we estimate that approximately 5 countries will win their first Olympic medal at the 2028 Los Angeles Olympics. This prediction is derived from the linear regression analysis of historical data, which shows a strong correlation ($R^2 = 0.9935$) between the year and the cumulative number of countries winning their first medal. The confidence interval for this prediction, at a 70% confidence level, is between 4.05 and 5.67 new medal-winning countries.

6.3 Subtask 3.1 — Exploring the Relationship Between Event Count and Medal Count

In this subtask, we investigate the relationship between the number and types of events a country participates in and the number of medals it earns. The analysis is performed using data from the 2024 Olympics, focusing on three primary aspects: visualization, correlation analysis, and interpretation of results.

6.3.1 Visualization

To better understand the relationship between event count and medal count, we employed two visualization methods:

- **Scatter Plot:** Displays the relationship between event count and medal count.
- **Box Plot:** Analyzes the distribution of medal counts across different event count intervals.

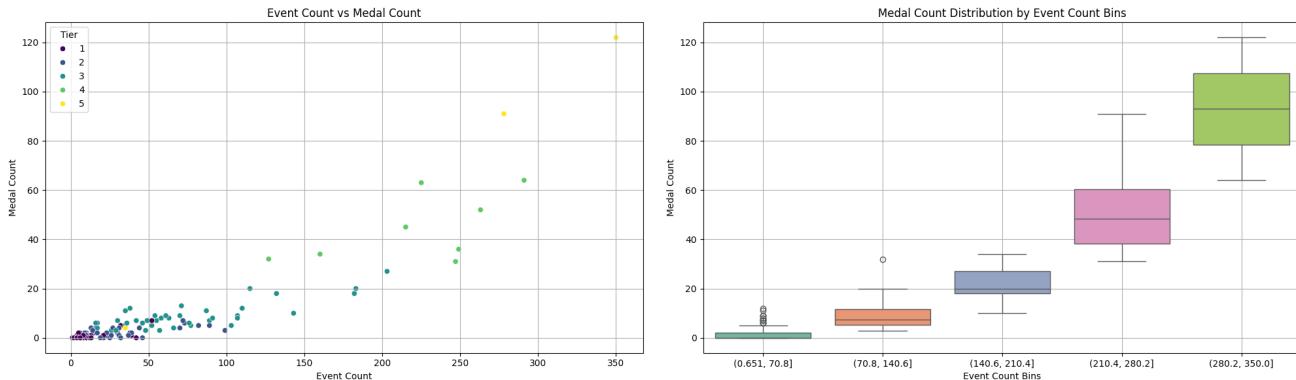


Figure 7: Visualizations: (Left) Scatter plot showing the correlation; (Right) Box plot depicting the distributions.

To better understand the relationship between event count and medal count, countries were grouped into five distinct tiers based on their competitive performance using the K-means clustering algorithm.

The decision to use K-means clustering was motivated by the scatter plot observations, which suggested heterogeneous patterns in the event count and medal count relationship across different types of countries. The clustering process was informed by previous medal achievements(first gold medals, then total medals), ensuring that the grouping reflected meaningful competitive distinctions.

Once the countries were assigned to tiers, we conducted separate correlation analyses within each tier. Spearman's correlation coefficient was used to measure the strength and direction

of the relationship, as it is well-suited for non-parametric data and can handle potential non-linearity. The results are shown in following Table 4.

6.3.2 Interpretation of Results

The analysis indicates that for most tiers (1, 2, 4, and 5), there is a statistically significant positive relationship between event count and medal count ($p < 0.05$). This suggests that the number of events a country participates in positively impacts its medal count. For Tier 3, the p -value is greater than 0.05, indicating no statistically significant correlation between event count and medal count within this group. This anomaly could be due to specific factors such as high variability in event outcomes or unmeasured country-specific attributes. However, the overall analysis aligns with the expectation that more competitive countries (higher tiers) are better equipped to convert event participation into medals due to superior resources, training, and athletes.

6.4 Subtask 3.2 — Identifying the Most Important Sports for Various Countries

6.4.1 Methodology

To answer the question of which sports are most important for various countries, we proceed with the following steps:

Step 1 Compute the event importance value $w_{i,j} = \frac{m_{i,j}}{M_i}$, ensuring that the sum of importance values for all events for a given country satisfies $\sum_{j=1}^M w_{i,j} = 1$.

Step 2 For each event E_j , we calculate its global importance value $W_j = \sum_{i=1}^N w_{i,j}$ as the sum of its importance values across all countries, which captures the relative importance of the event across all countries, regardless of individual country performance.

Step 3 For each sport S_k , we compute its global importance $I_k = \sum_{j \in S_k} W_j$. It is completed through aggregating the global importance values W_j of all events E_j under the sport, providing an overall measure of the importance of each sport for all countries collectively.

Step 4 The sports are then ranked based on their global importance values I_k . A visualization of the rankings illustrates that athletics, boxing, and wrestling are the top three most important sports globally, where athletics has the absolute highest (more than twice of the second place) importance value. Result is shown in *Figure 8*.

Tier	Spearman Correlation	p-value
1	0.38	2.16×10^{-5}
2	0.39	0.01
3	0.61	4.99×10^{-5}
4	0.57	0.13
5	0.98	0.02

Table 4: Spearman correlation and p-values across different tiers.

6.4.2 Results and Insights

The analysis reveals that the importance of sports varies significantly across countries. The ranking of sports based on global importance highlights sports such as **athletics, swimming, and gymnastics** as the most influential globally. These sports typically feature a larger number of events and offer more medal opportunities, which contributes to their higher global importance.

6.5 Subtask 3.3 — Impact of Events Chosen by Host Countries on Olympic Results

Following these steps, we analyze the impact of events chosen by some major host countries on their Olympic results.

Categorization of Events for a Country Events are categorized into three types:

- **New Events:** Events that the country participated in for the first time during the given Olympic Games. These are identified as events not entered by the country in the previous five cycles.
- **Dominant Events:** Events in which the country earned medals in two consecutive previous Olympic Games.
- **Normal Events:** All remaining events that do not fall into the above categories.

Analysis of Event Performance For each event type:

- Calculate the total number of events (X) and medals won (Y) in each year across a 16-year period.
- Compute the yearly ratio as $R = Y/X$.
- Derive the Z-score for the host year compared to neighboring years using: $Z = \frac{R_{\text{host year}} - \mu}{\sigma}$ where μ stands for mean ratio of neighboring years ($\mu = \frac{\sum_{i=1}^{16} R_i}{16}$), and σ stands for standard deviation ($\sigma = \sqrt{\frac{\sum_{i=1}^{16} (R_i - \mu)^2}{16}}$).
- Determine if $Z > 1.96$, indicating statistical significance.

Application to Multiple Host Countries Applied the algorithm above to various host countries. Z-scores for new, dominant, and normal events are calculated for each host country to identify patterns in performance.

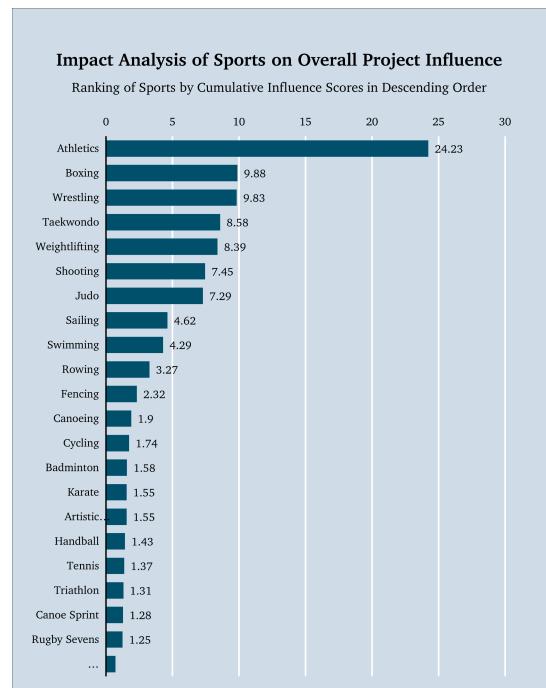


Figure 8: Ranking of sports by global importance based on aggregated event importance values across all countries.

6.5.1 Results and Comparative Insights

Overall Trends Across Hosts Across 16 Olympic Games, the proportion of Z-scores exceeding 1.96 for each event category is:

- **New Events:** 9 occurrences (56.25%),
- **Dominant Events:** 8 occurrences (50.00%),
- **Normal Events:** 11 occurrences (68.75%).

6.5.2 Conclusion

The findings demonstrate that host countries rely most heavily on normal events to enhance their medal counts, with new events providing additional opportunities. Dominant events, while important, tend to show smaller incremental impacts, likely due to the high level of competition.

7 Task2 — The Great Coach Effect: its identification and impact

The Great Coach Effect is a phenomenon where the presence of a highly skilled coach significantly improves the performance of athletes. Due to the fact that coaches are not bound to a specific country, they can have a significant impact on the medal counts of multiple nations. Identifying the Great Coach Effect and quantifying its impact on medal counts is crucial for predicting Olympic success.

7.1 Identification

7.1.1 Methodology

We reckon that the Great Coach Effect can be identified by comparing the performance of athletes under the same coach across different countries. It should, by nature, cause one country to have a sharp increase in their medal count in specific sports (otherwise this coach wouldn't be that great), and one other country to go through a gradual-to-sharp decrease in their medal count in the same sports. This is because the coach basically can only focus on coaching one country simultaneously, and the athletes from the other country would not receive the same level of training and support.

Again, due to the inherent limitation of usable dataset, we can only seek to identify the Great Coach Effect from this characteristic, rather than other methods such as relying on coach information to deduce the effect.

Therefore, we propose a two-step approach to identify and quantify the Great Coach Effect:

1. Identify such pattern across the dataset.
2. Refer to outside sources to interpret and cross-verify the result, finding out which specific coach is causing the effect each time.

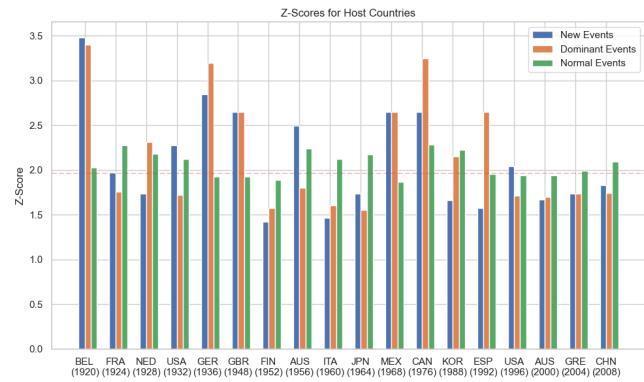


Figure 9: Comparison of Z-scores for new, dominant, and normal events across host countries.

7.1.2 Implementation

We implemented the data filtering logic in Python with 0.23k LOC. The program reads the dataset, groups the data by country, sport, and year, and calculates the weighted medal count for each combination(gold medals with weight 3, silver with 2 and bronze with 1). It then uses linear regression on the most recent 12 years of data to calculate the slope of the medal trend over time. A positive slope exceeding a defined threshold (POS_K_LOWER) indicates a sharp increase, while a negative slope below another threshold (NEG_K_UPPER) indicates a gradual-to-sharp decrease. Countries with significant trends are recorded as having possible increases or decreases.

The program then pairs countries with overlapping trends in the same sport, checking the overlap period to ensure it spans at least four years and does not exceed a defined gap (OVERLAP_GAP_UPPER). Paired trends suggest the transfer of coaching expertise from one country to another.

By carefully calibrating the thresholds and gaps, we can effectively identify the Great Coach Effect and quantify its impact on national performances. We present the visual representation of the potential occurrences of the Great Coach Effect from both sport level and event level in *Figure 10*.

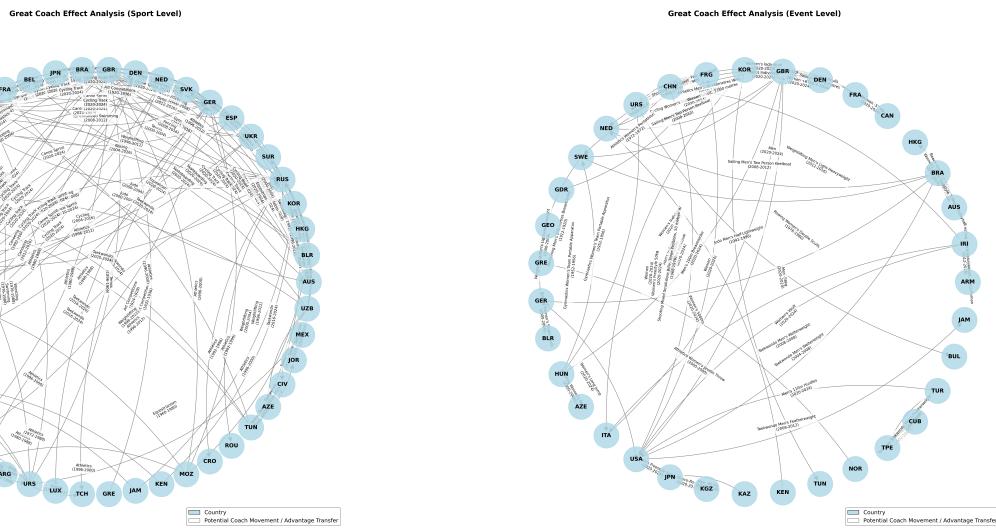


Figure 10: Potential occurrences of The Great Coach Effects from sport level and event level

However elaborate we may calibrate the parameters, there still exists the issue of false positives and false negatives, with false positives accounting more because other factors may also contribute to this same characteristic. We recommend cross-referencing the results with external sources, such as coaching records and athlete interviews, to validate the findings and pinpoint the specific coaches responsible for these shifts. This data-driven approach allows for systematic identification of the Great Coach Effect, providing actionable insights into the strategic importance of coaching in achieving Olympic success.

7.2 Quantification

7.2.1 Methodology

We will quantify two main aspects of the Great Coach Effect: the magnitude of the performance change in both medal counts changes and the percentage of that change to the total medal count of the country.

The magnitude of the performance change can be calculated by comparing the average medal count before and after the coach transfer, while the percentage of the change can be calculated by dividing the change in medal count by the total medal count of the country in the respective sport.

These two aspects, especially the percentage of medal change to the total medal count of the country, have inextricable correlations with the overall strength of the country. Therefore, we will employ the same strategy as in our model establishment to cluster countries based on their historical medal counts, and then quantify the Great Coach Effect within each cluster.

7.2.2 Implementation

We further polish the program used in the identification step to calculate the magnitude and percentage of the Great Coach Effect, adding an additional 0.23k LOC. The program categorizes countries based on their tier value, calculates the average increase/decrease in medal count, and computes the average percentage of this change relative to the total medal count of the country. These metrics provide a comprehensive view of the Great Coach Effect's potential impact on different countries and sports.

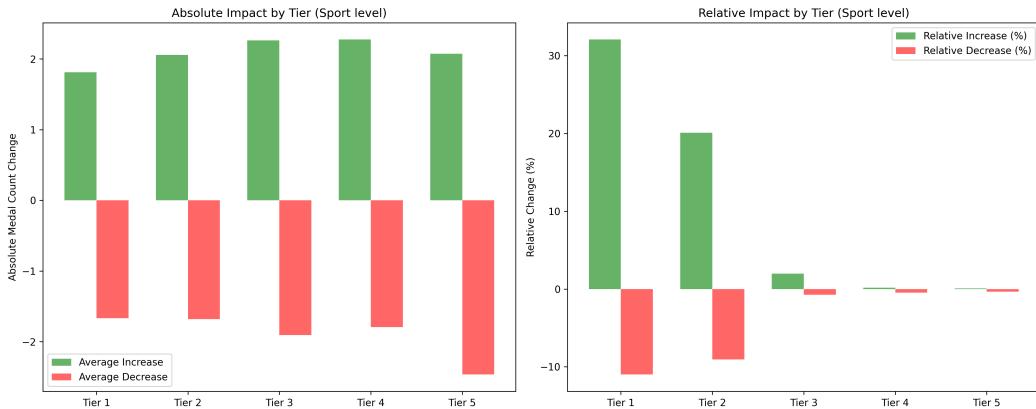


Figure 11: Quantification of The Great Coach Effect

Tier	Avg. Increase	Avg. Decrease	Relative Increase(%)	Relative Decrease(%)
1	1.8125	-1.6667	32.0772	-10.9944
2	2.0562	-1.6795	20.1028	-9.0563
3	2.2623	-1.9041	1.9766	-0.7317
4	2.2743	-1.7923	0.1780	-0.4518
5	2.0746	-2.4638	0.0660	-0.3472

Table 5: Impact of The Great Coach Effect on Different Tiers of Countries

Figure 11 and Table 5 clearly shows the impact of the Great Coach Effect on different tiers of countries, highlighting the potential for significant performance improvements under the

guidance of elite coaches. We can see that the positive and negative impact of the Great Coach Effect(namely, the moving in and out of great coaches) on absolute medal changes has remained constant among different tiers of countries, while its impact on relative medal changes decreases as the tier goes up. This is an expected phenomenon because countries with better overall power already do well in sports across a wide range of fields, so the arrival or departure of one specific coach won't have much effect relative to the total medal count. By quantifying the magnitude and percentage of the effect, we can better understand the strategic value of investing in high-caliber coaching talent and its implications for national Olympic success.

7.3 Exemplification

By analyzing the performance trends in specific sports, we identified notable cases where the transfer of a coach coincided with a sharp increase in performance of one country and a decline or stagnation for another, allowing for the cross-verification of our algorithms. Below are three exemplifications of the Great Coach Effect based on historical Olympic data and their correlations with our results.

It is worth mentioning that the data generated by the program above provides pairs of trends where one country's performance in a certain event increases significantly while another decreases over the same period of time. To spot such pattern, we assign each kind of medal with a value and apply linear regression to the data over a certain of range to get the slope, where a positive slope indicates improvement and vice versa.

7.3.1 Synchronized Swimming Women's Team: Ana Tarré

- 1 - Event: Synchronized Swimming Women's Team, Overlap Years [2008-2012], ↪
→Increase: CHN(slope=0.125) / Decrease: ESP(slope=-0.250)

This piece of data generated by the program reveals a significant overlap in performance trends between China (CHN) and Spain (ESP) during the period from 2008 to 2012. China's positive slope (0.125) and Spain's negative slope (-0.250) aligns well with Ana Tarré's transition from coaching Spain to China in 2012.

During Ana Tarré's tenure with Spain (1996–2012), the team gradually improved, earning higher medals, peaking with a Silver medal in 2008 and a Bronze in 2012. However, after she transitioned to coaching China in 2012, Spain had dropped out of the podium positions since then until 2024, while China experienced a gradual rise, achieving a series of Silver medals and finally a Gold medal in 2024, strongly aligning with the Great Coach Effect.

7.3.2 Gymnastics Women's Team All-Around: Béla Károlyi

- 1 - Event: Gymnastics Women's Team All-Around, Overlap Years [2000-2008], ↪
→Increase: USA(slope=0.125) / Decrease: ROU(slope=-0.250)

According to the program data, the overlap period between Romania (ROU) and the United States (USA) spans from 2000 to 2008. During this time, Romania exhibited a steep negative slope (-0.250), whereas the United States showed a positive one (0.125), which matches the historical narrative of Béla Károlyi's influence.

Béla Károlyi's legendary coaching career demonstrates a clear pattern of the Great Coach Effect. While coaching Romania (1974–1981), he elevated the team to global prominence, culminating in Silver medals in 1976 and 1980. After moving to coach the United States (1981–

2016), Romania's performance declined, while the U.S. women's gymnastics team emerged as a dominant force.

7.3.3 Women's Volleyball: Lang Ping

- 1 - Event: Volleyball Women's Volleyball, Overlap Years [1992–2004], Increase: ↗
→CHN(slope=0.175) / Decrease: USA(slope=-0.075)
- 2 - Event: Volleyball Women's Volleyball, Overlap Years [2004–2012], Increase: ↗
→USA(slope=0.250) / Decrease: CHN(slope=-0.375)

The program identifies two overlapping periods: 1992-2004 and 2004-2012. In both periods, one of CHN/USA rises and the other falls, which closely corresponds to Lang Ping's coaching career.

Looking at Lang Ping's coaching career spanning over 20 years, we can also discover the distinct influence of a capable coach. During her first tenure with China (1995-2005), the team achieved notable successes, including a Gold medal in 2004. After transitioning to coach the U.S. team (2005-2013) which had won few medals in history, U.S. quickly surpassed China and won 2 consecutive Silver medals. When Lang Ping returned to coach China in 2013, the Chinese team quickly regained its dominance with a Gold in 2016.

7.4 Recommendation

In the Quantification part, we represented that the Great Coach Effect has a consistent impact on the absolute medal count changes across different tiers of countries, while its impact on relative medal count changes decreases as the tier goes up. This suggests that investing in elite coaching talent can yield substantial performance improvements, particularly for countries with lower historical medal counts(represented by low tier).

Therefore, we recommend that national Olympic committees prioritize recruiting and retaining top-tier coaches for tier 2 countries. Tier 1 countries usually have too tight a budget to afford the cost of coaches with great expertise, while it won't be really necessary for tier 3 5 countries to do so because of the low relative effect.

To determine which three tier 2 countries can benefit the most from the Great Coach Effect, we use the python code from Quantification part and modifies it to find the 3 countries with the highest potential for performance improvement.

As is shown in *Figure 12*, we recommend the top 3 countries with the highest potential for performance improvement: **ISR (Israel)**, **BRN (Bahrain)**, and **VEN (Venezuela)**, with respective projected medal gain of 5.5, 3.9 and 3.9.

8 Task3 — Original Insights revealed from the model

8.1 Global Sports Equity

In our analysis of the Great Coach Effect, we've shown that the impact of elite coaching talent can exert great influence on low-tier countries, assisting in the establishment of a more level playing field in the global sports arena and contributing to a more equitable global sports landscape.

We recommend that countries with elite coaches consider sharing their expertise with nations that have historically underperformed in those sports. This collaborative approach can elevate the global standard of competition, foster international goodwill, and ensure a more equitable distribution of knowledge and skills across the sporting world. By promoting a more

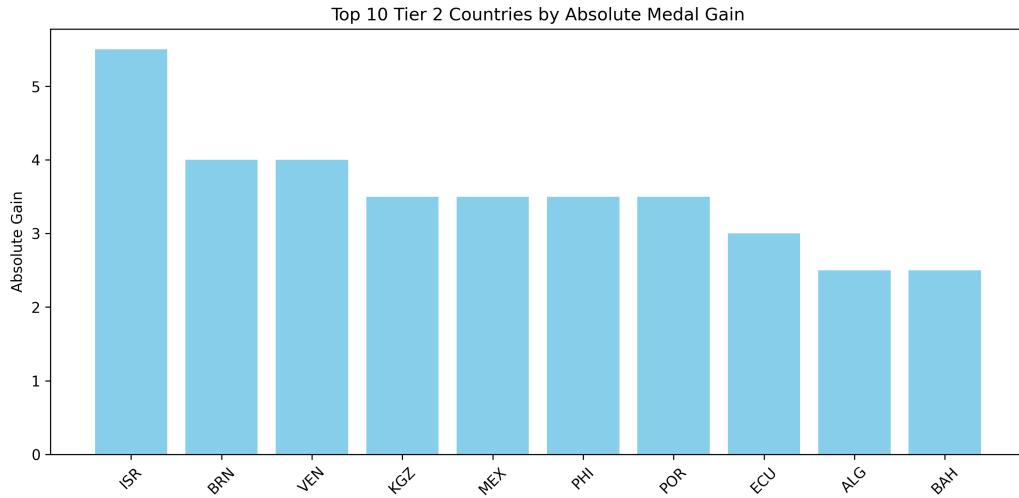


Figure 12: Quantification of The Great Coach Effect on Tier 2 Countries

inclusive and diverse competitive environment, we can enhance the global appeal and accessibility of various sports.

8.2 Host Country Effect and Regulation

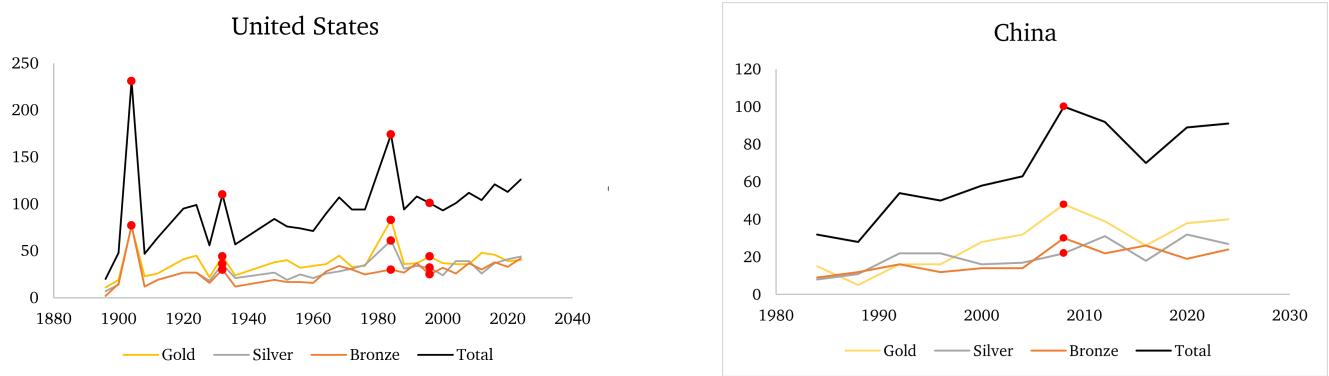


Figure 13: The Host-Country Effect on Medal Counts

Analyzing the provided data, we observe a significant increase in medal counts for the United States during the years it hosted the Olympics: 1904, 1932, 1984, and 1996. For instance, in 1904, the US won 231 medals, a stark contrast to the 48 medals won in 1900. Similarly, China, as the host in 2008, saw a substantial increase in its medal count, winning 100 medals compared to 63 in 2004. This pattern suggests a strong host effect, where the host country benefits from factors such as home crowd support, familiarity with venues, and possibly more favorable conditions.

To curb such undesired host effects, regulations could be implemented to ensure a more level playing field. One approach could be to standardize training facilities and access for all participating countries well before the Games. Additionally, introducing stricter oversight on judging and scoring to prevent any potential bias towards the host nation can help. Another measure could be to limit the number of new events introduced by the host country, which

might be tailored to their strengths. These regulations can help mitigate the host advantage, ensuring fairer competition and maintaining the integrity of the Olympic Games.

8.3 Time-Sensitive Element in Prediction

In our model evaluation part, we found that the number of matches immediately preceding the target year used to predict the outcome had a significant impact on the prediction performance. This suggests that the model's predictive power is highly sensitive to the number of matches played in the year leading up to the Olympics, strongly indicating that there is a time-sensitive element to the data.

Hence, we suggest that national Olympic committees and sports organizations pay close attention to the recent matches played in the year leading up to the Olympics. By monitoring and analyzing this data, they can gain valuable insights into the performance trends of athletes and teams, identify potential areas for improvement, and make informed decisions on training and preparation strategies. This time-sensitive element underscores the importance of adapting to ever-changing conditions and leveraging real-time data to optimize performance and maximize medal-winning opportunities.

8.4 Granularity Effect in Prediction Models

Through our initial attempt at using individual athletes as the smallest granularity, we found that the model's prediction performance was significantly affected by the high randomness and variability in individual athlete performance and is thus defective.

In our initial attempt at using individual athletes as the smallest granularity, we found that the model's prediction performance was significantly affected by the high randomness and variability in individual athlete performance, making it defective.

This also reflects the real-world scenario, where allocating training resources to individual athletes is challenging due to the high uncertainty and variability in individual performance. Over-investing in a single project or athlete does not necessarily guarantee a return on investment, but in the meantime, taking part in more sports, as shown in the model, can steadily increase the chances of winning medals.

Therefore, we recommend that national Olympic committees focus on diversifying their participation in various sports to maximize their chances of winning medals. This strategy can help mitigate the risks associated with investing heavily in individual athletes and improve the overall performance of the country in the Olympics.

9 Sensitivity Analysis

In order to evaluate how changes in a key hyperparameter affect the overall prediction performance, we perform a sensitivity analysis that varies the number of estimators for all the regression models we used. Specifically, we adjust the `n_estimators` parameter (if the model supports it) in the range of 25 to 175 with the step of 5, while maintaining consistent data and procedural steps otherwise.

We use 3-fold cross-validation on the same features and labels, and for each medal type (*Gold, Silver, Bronze*), we select the best performing model just as in the Model Evaluation part. We compare each tested value against a baseline model where `n_estimators=100`, calculating two proportions:

1. Inner_param_change_proportion: how much `n_estimators` changes relative to the

baseline, i.e. $\frac{n_value - 100}{100}$.

2. **Result_change_proportion:** how much the average MAE changes relative to the MAE of the baseline, i.e. $\frac{\text{MAE}(n_value) - \text{MAE}(100)}{\text{MAE}(100)}$.

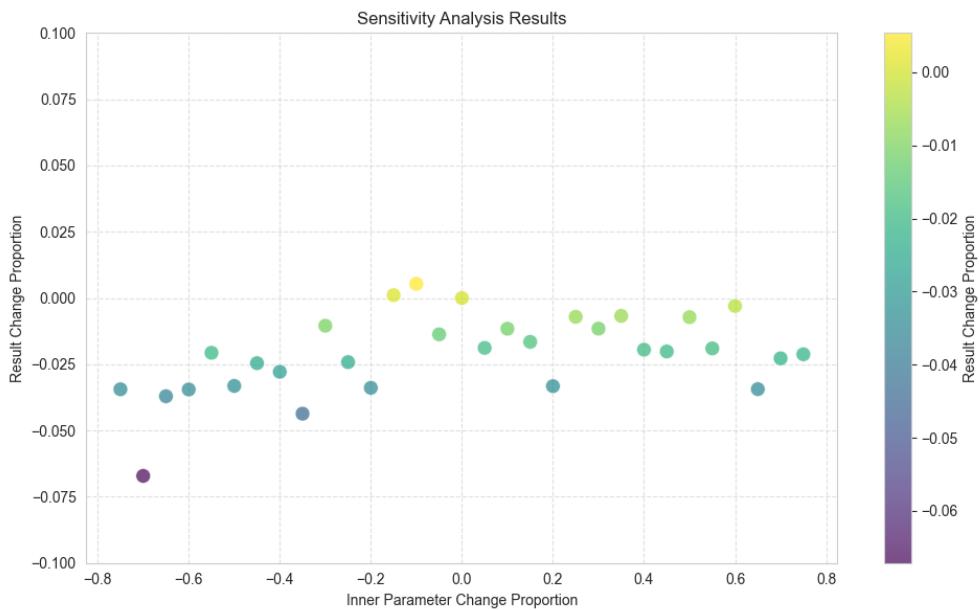


Figure 14: Sensitivity Analysis of the Number of Estimators

We can see from *Figure 14* that the change in the number of estimators has a minor impact on the results overall, with the largest impact of them being only 7.5%, and most of them around 2.5%. This suggests that the model is relatively stable with respect to this hyperparameter, and the prediction performance is not significantly affected by the model parameters within the tested range.

10 Conclusion

In our work, we developed a comprehensive model to analyze and predict medal counts of 2028 Los Angeles Olympic Games, mainly focusing on historical performance, host country, and elite coaching. We quantified the "Great Coach Effect" and demonstrated its significant influence on medal counts. Our analysis revealed that investing in top-tier coaching talent can yield substantial performance improvements. Additionally, we examined the phenomenon of host countries experiencing a peak in medal counts during the hosting year, using Z-score analysis to quantify this effect. Our findings may provide valuable insights for national Olympic committees to strategically allocate resources and enhance their chances of success in future Olympic Games.

References

- [1] Ball, D. W. (1972). Olympic Games competition: Structural correlates of national success. *International Journal of Comparative Sociology*, 15, 186–200.
- [2] Bernard, A. B., & Busse, M. R. (2004). Who wins the Olympic Games: Economic resources and medal totals. *Review of Economics and Statistics*, 86, 414–417.

- [3] Bian, X. (2005). Predicting Olympic Medal Counts: The Effects of Economic Development on Olympic Performance. *The Park Place Economist*, 13, 37-44.
- [4] Forrest, D., Sanz, I., & Tena, J. D. (2010). Forecasting national team medal totals at the Summer Olympic Games. *International Journal of Forecasting*, 26, 576–588.
- [5] Schlembach, C., Schmidt, S. L., Schreyer, D., & Wunderlich, L. (2022). Forecasting the Olympic medal distribution – A socioeconomic machine learning model. *Technological Forecasting & Social Change*, 175, 121314.
- [6] Tcha, M., & Pershin, V. (2003). Reconsidering performance at the Summer Olympics and revealed comparative advantage. *Journal of Sports Economics*, 4, 216–239.
- [7] Bernard A B, Busse M R. Who wins the Olympic Games: Economic resources and medal totals[J]. *Review of economics and statistics*, 2004, 86(1): 413-417.
- [8] Xue, J., & Shen, B. (2020). A novel swarm intelligence optimization approach: sparrow search algorithm. *Systems Science & Control Engineering*, 8(1), 22-34.
- [9] Xue, J., & Shen, B. (2023). Dung beetle optimizer: A new meta-heuristic algorithm for global optimization. *The Journal of Supercomputing*, 79(7), 7305-7336.
- [10] Mirjalili, S. (2016). SCA: a sine cosine algorithm for solving optimization problems. *Knowledge-Based Systems*, 96, 120-133.
- [11] Steinbrunn, M., Moerkotte, G., & Kemper, A. (1997). Heuristic and randomized optimization for the join ordering problem. *The VLDB Journal*, 6, 191-208.
- [12] Zhan, Z. H., Zhang, J., Li, Y., & Chung, H. S. H. (2009). Adaptive particle swarm optimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(6), 1362-1381.
- [13] Hashim, F. A., & Hussien, A. G. (2022). Snake Optimizer: A novel meta-heuristic optimization algorithm. *Knowledge-Based Systems*, 242, 108320.
- [14] Trojovský, P., & Dehghani, M. (2022). Pelican optimization algorithm: A novel nature-inspired algorithm for engineering applications. *Sensors*, 22(3), 855.
- [15] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. *Advances in Engineering Software*, 69, 46-61.

Report on Use of AI

1. Github Copilot

Query1: make one elaborate plot out of this data: 1. add scatter points of year-country count 2. for the last 8 years [1996-2024], add the linear regression line onto the plot.

Output: I'll help you create a plot with scatter points and a linear regression line. Here's the solution step by step:

1. Import required libraries (numpy, pandas, matplotlib)
2. Read the CSV data
3. Create scatter plot for all points
4. Calculate and plot linear regression for the last 8 data points
5. Add labels and styling

Create a visualization script that plots the data with scatter points and regression line. [In main.py]

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5 from sklearn.linear_model import LinearRegression
6
7 # Read the CSV file
8 df = pd.read_csv('src/plot_1_2/summerOly_medal_country_count.csv')
9
10 # Set style
11 sns.set_style('whitegrid')
12 sns.set_palette("husl")
13
14 # Create figure
15 plt.figure(figsize=(12, 8))
16
17 # Plot all data points
18 sns.scatterplot(data=df, x='Year', y='Countries having earned medals',
19                  color='blue', alpha=0.6, label='Historical data')
20
21 # Get last 8 data points for regression
22 last_8 = df.tail(8)
23 X = last_8['Year'].values.reshape(-1, 1)
24 y = last_8['Countries having earned medals'].values
25
26 # Perform Linear regression
27 reg = LinearRegression().fit(X, y)
28 y_pred = reg.predict(X)
29
30 # Create extended X range for prediction
31 X_extended = np.array([[x] for x in range(X.min()-4, 2033)]) # Extend ←
   ↘before and after
```

```
32 y_extended = reg.predict(X_extended)
33
34 # Plot extended regression line
35 sns.lineplot(x=X_extended.flatten(), y=y_extended, color='red', ←
    ↪linestyle='--',
36             label=f'Linear trend (1996–2024)\nR2 = {reg.score(X, y):.4f}')
37
38 # Add 2028 prediction point
39 prediction_2028 = reg.predict([[2028]])[0]
40 plt.scatter(2028, prediction_2028, color='red', s=35, zorder=5,
41             label=f'2028 Prediction: {prediction_2028:.0f}')
42
43 # Customize the plot
44 plt.title('Cumulative Medal-Winning Countries by Olympic Year',
45            fontsize=18, pad=15)
46 plt.xlabel('Year', fontsize=12)
47 plt.ylabel('Number of Countries with Medals', fontsize=12)
48 plt.grid(True, alpha=0.3)
49 plt.legend(fontsize=10)
50
51 # Add text showing the regression equation
52 slope = reg.coef_[0]
53 intercept = reg.intercept_
54 equation = f'y = {slope:.2f}x + {intercept:+.1f}'
55
56 # Calculate middle point of the regression line for text placement
57 mid_x = (X_extended.min() + X_extended.max()) / 2
58 mid_y = reg.predict([[mid_x]])[0]
59
60 # Place text horizontally above the regression line
61 plt.text(mid_x - 10, mid_y + 10, equation,
62           fontsize=18,
63           verticalalignment='bottom',
64           horizontalalignment='center',
65           bbox=dict(facecolor='white', alpha=0.8, edgecolor='none'))
66
67 plt.tight_layout()
68 plt.show()
```

This script will create a plot showing:

- 1.Blue scatter points for all historical data
- 2.Red dashed line showing the linear regression for 1996-2024
- 3.Grid, labels, and legend
- 4.The regression equation and R² value
- 5.Modern styling with seaborn
- 6.The regression line will help visualize the trend in the number of countries earning medals in recent Olympics.

2. Github Copilot

Query1: Insert Team_or_not and Team_size as another two columns in the data frame.
Output: Sure, I can help you with that. Here's the modified data frame:

```
1 Event, Team_or_not, Team_size
2 10m Air Pistol Men, false, 1
3 10m Air Pistol Mixed Team, true, 2
4 10m Air Pistol Women, false, 1
5 10m Air Rifle Men, false, 1
6 10m Air Rifle Mixed Team, true, 2
7 10m Air Rifle Women, false, 1
8 25m Pistol Women, false, 1
9 25m Rapid Fire Pistol Men, false, 1
10 4 x 400m Relay Mixed, true, 4
11 4 x 400m Relay Mixed Team, true, 4
12 50m Rifle 3 Positions Men, false, 1
13 50m Rifle 3 Positions Women, false, 1
14 Aeronautics Mixed Aeronautics, false, 1
15 Alpinism Mixed Alpinism, false, 1
16 "Archery Men's Au Chapelet, 33 metres", false, 1
17 "Archery Men's Au Chapelet, 50 metres", false, 1
18 "Archery Men's Au Cordon Dore, 33 metres", false, 1
19 "Archery Men's Au Cordon Dore, 50 metres", false, 1
20 Archery Men's Championnat Du Monde, false, 1
21 Archery Men's Continental Style, false, 1
22 Archery Men's Double American Round, false, 1
23 Archery Men's Double York Round, false, 1
24 Archery Men's Individual, false, 1
25 "Archery Men's Pole Archery, Large Birds, Individual", false, 1
26 "Archery Men's Pole Archery, Large Birds, Team", true, 4
27 "Archery Men's Pole Archery, Small Birds, Individual", false, 1
28 "Archery Men's Pole Archery, Small Birds, Team", true, 4
29 Archery Men's Sur La Perche a La Herse, false, 1
30 Archery Men's Sur La Perche a La Pyramide, false, 1
31 "Archery Men's Target Archery, 28 metres, Individual", false, 1
32 "Archery Men's Target Archery, 28 metres, Team", true, 4
33 "Archery Men's Target Archery, 33 metres, Individual", false, 1
34 "Archery Men's Target Archery, 33 metres, Team", true, 4
35 "Archery Men's Target Archery, 50 metres, Individual", false, 1
36 "Archery Men's Target Archery, 50 metres, Team", true, 4
37 Archery Men's Team, true, 4
38 Archery Men's Team Round, true, 4
39 Archery Men's Unknown Event, false, 1
40 Archery Women's Double Columbia Round, false, 1
41 Archery Women's Double National Round, false, 1
42 Archery Women's Individual, false, 1
43 ...
44 "Wrestling Men's Super-Heavyweight, Greco-Roman", false, 1
45 "Wrestling Men's Unlimited Class, Greco-Roman", false, 1
46 "Wrestling Men's Welterweight, Freestyle", false, 1
```

```
47 "Wrestling Men's Welterweight, Greco-Roman", false, 1
48 "Wrestling Women's Featherweight, Freestyle", false, 1
49 "Wrestling Women's Flyweight, Freestyle", false, 1
50 "Wrestling Women's Heavyweight, Freestyle", false, 1
51 "Wrestling Women's Light-Heavyweight, Freestyle", false, 1
52 "Wrestling Women's Lightweight, Freestyle", false, 1
53 "Wrestling Women's Middleweight, Freestyle", false, 1
```