CDS504: Business Data Analytics

Group Project

Team 8

Final Report

Project Title: Data Analysis and Prediction for the Paris Olympics

Team Leader: ZHU Jun 5523710

Member: ZHAO Yuhao 5528423

LIN Lixin 5586011

OU Lefeng 5509881

LAW Hoi 5503318

Abstract

This report aims to explore the relationship between athletes' age, nationality and medal awards, as well as the performance and medal distribution of each country at the Olympic Games, through an in-depth analysis of the Paris 2024 data set. Through the use of a variety of data analysis techniques, we identify key issues and propose targeted strategies and recommendations to optimize athlete selection, training and participation strategies, and provide data support for managers and policymakers to develop more science-based sports development strategies. At the same time, we also used random forests, linear regression and decision trees to predict the number of medals of the host country in the next Olympic Games.

1 Introduction

As a global sports event, the Olympic Games is not only a platform for countries to showcase their sports strength, but also an important occasion for cultural exchanges. By analysing the Olympic data set, this project aims to uncover the relationship between athletes' age and medal awards, analyze differences in performance across different sports, and explore the host country's performance at the Games. These analyses will provide the scientific basis for national sports delegations to develop more effective athlete selection and training plans, as well as sports development strategies.

Research Methods:

The following techniques and methods were used for data analysis:

- Group comparison: Compare medals in different age groups by grouping athletes according to their age.
- Visualization charts: Use bar charts and pie charts to show the distribution of medals by country.
- Comparative analysis: Analyze the medals won by different countries in different sports.
- Trend analysis: show the trend of the number of medals in each country by drawing a trend graph.
- Random Forest: Predicts the likelihood of an athlete winning a medal and identifies the key factors that affect medal acquisition.
- Decision Tree and Linear Regression: Predict future medal distribution and analyze the factors affecting the number of medals.

2 Data Types and Analytics

By filtering the dataset, the following three tables are selected:

Medals_total:

Nominal Data：Country,

Numerical value: Golden Medals, Silver Medals, Bronze Medals, Total

Medals:

Nominal Data：medal_type, gender, discipline, country

Athletes:

Nominal Data：country, discipline, birth_country, birth_place

Numerical value: birth_date

Total_medal_1896_2024:

Nominal Data：edition, country,

Numerical value: Gold Medal, Silver Medal, Bronze Medal, Total

3 Preliminary Analysis:

(1) Relationship between age and medals:

Preliminary analysis shows that most medal winners are concentrated in specific age groups, such as 24-29 years old, indicating that this age group may be the athletes' competitive peak.

age	求和项:总计	求和项:Silver	求和项:Gold	求和项:Bronze
14-19	73	30	19	24
19-24	571	195	183	193
24-29	965	307	308	350
29-34	522	161	180	181
34-39	144	52	50	42
39-44	27	5	10	12
44-49	7	2	1	4
49-54	2	2		
54-59	4	2	1	1
总计	**2315**	**756**	**752**	**807**

(2) Nationality and medal distribution:

Countries at the top of the medal table dominate the sports powerhouses, but there are significant differences in the performance of countries in different sports.(Top 10 are developed country(expect China), and Top 4 are Permanent Five (expect Russia))

nationality_long	Bronze Medal	Gold Medal	Silver Meda	总计
United States of America	97	134	101	332
France	39	53	95	187
People's Republic of China	40	71	57	168
Great Britain	80	40	42	162
Australia	45	33	45	123
Netherlands	26	66	25	117
Germany	38	25	50	113
Italy	28	31	29	88
Spain	36	40	7	83
Japan	24	27	31	82

4 Data Visualization with Tableau

4.1 Medals by Countries

With figure1 we can conclude the following:

USA and China's medal dominance. The United States and China are tied for first place in the number of gold medals, demonstrating the significant dominance of both countries in competitive sports. However, China's medal tally is the same as that of the United States, but slightly lower in the number of silvers and bronzes, which may indicate that the country has invested heavily in the development of its top athletes.

This chart provides visual data to support analyses of countries‘ performance in competitive sports and can be used to further explore how factors such as countries' sports policies, athlete training systems, and sports cultures influence medal attainment.

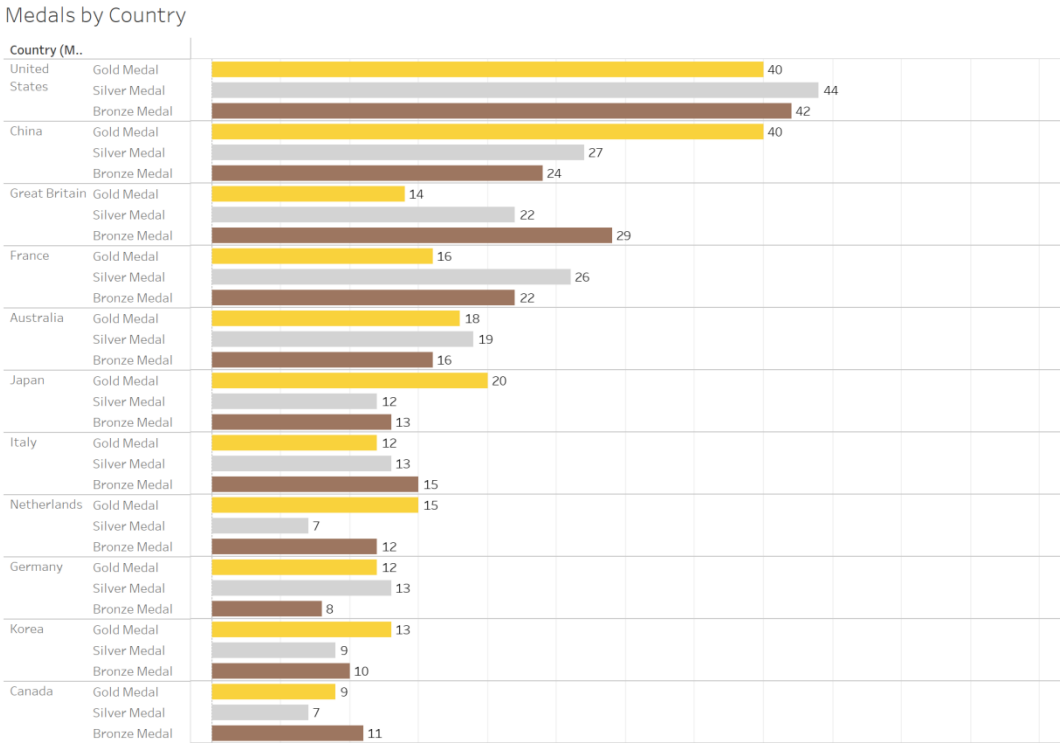


Figure 1: Medals by Countries

4.2 Medals Distribution by World

This world map shows the distribution of medals won by countries around the globe in a particular sporting event or series. The map uses different shades of color to indicate the number of medals, with darker colors indicating more medals won by that country. The map is analyzed and summarized below:

(3) Leading Country

The United States and China are shown as the darkest colors on the map, indicating that these two countries are well ahead in terms of the number of medals.

(4) Medal Distribution

The distribution of medals shows that some countries have a significant advantage in competitive sports, which may be related to the country's sports policy, resource allocation, and athlete training system.

(5) The relationship between sporting prowess and economic prowess

Some countries with stronger economies, such as the United States, China, Japan and Germany, also perform well in sports competitions, which may be related to the country's investment in and support for sports.

(6) Diversity and balanced development

The map shows the distribution of competitive strength in sport globally, which serves as a reference for international sports organizations and governments to develop sport around the world in a more balanced way.

This map can be used as a powerful tool for analyzing the competitive strength of sport and the allocation of sport resources globally, helping to formulate more effective sport development strategies.

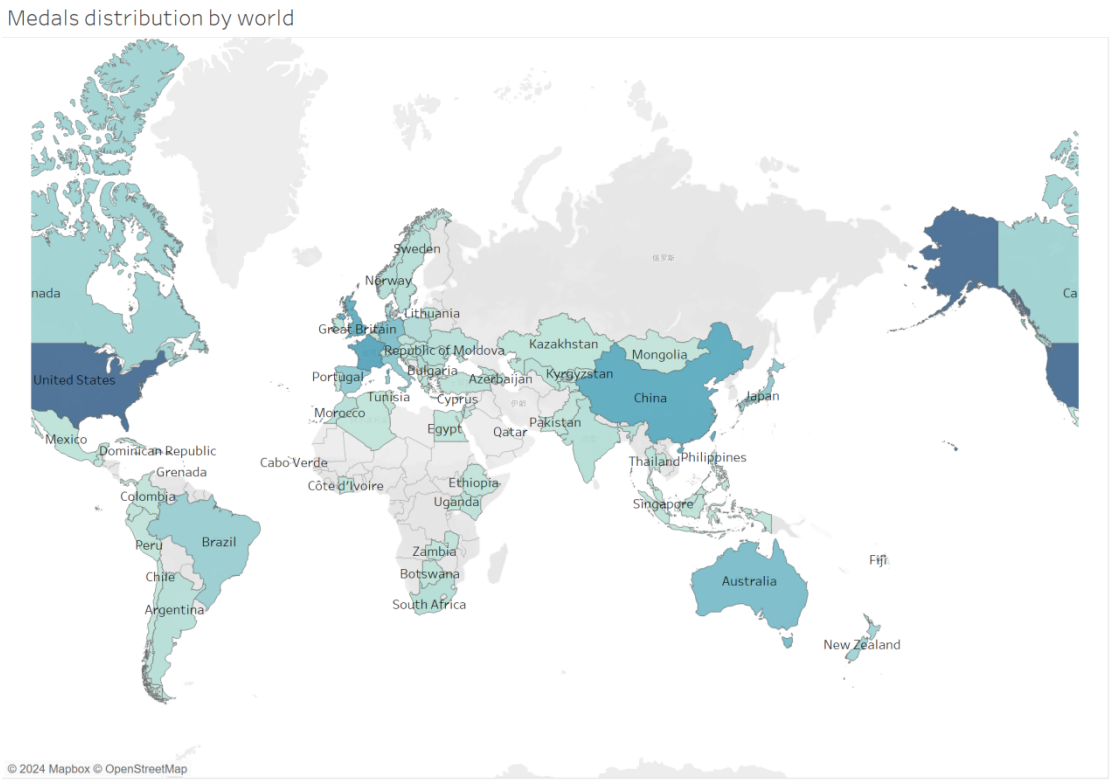


Figure 2: Medals Distribution by World

4.3 TotalMedals by AgeGroup and Gender

This formula is used to calculate a person's age from the date of birth to the current date (the date returned by the TODAY() function), and adjusts the age based on the results to ensure accuracy.

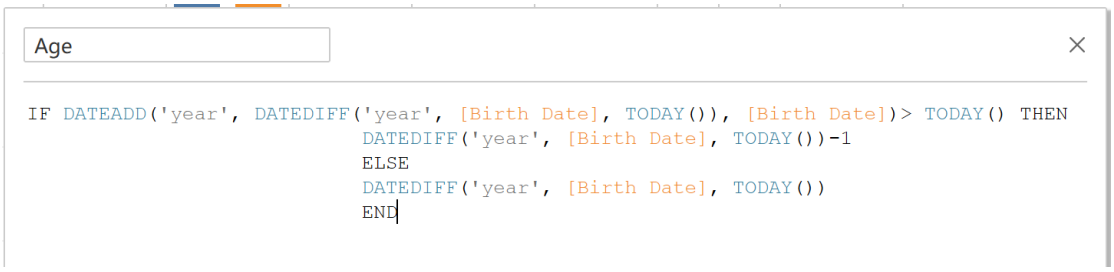


Figure 4: Age calculation formula

Figure 5 is a bar chart showing the total number of medals won by athletes of different age groups and genders.

(1) Distribution by age group:

The charts are divided by age group into 10-15, 15-20, 20-25, 25-30, 30-35, 35-40, 40-45, 45-50, 50-55 and 55+.

(2) Gender distribution:

Within each age group, the number of medals is divided by gender into male (blue) and female (orange).

(3) Number of medals:

The 20-25 and 25-30 age groups had the highest number of medals, especially for male athletes. Male athletes in the 25-30 age group won the highest number of medals with 462. The number of medals won by female athletes in the 20-25 age group was also relatively high at 350.

(4) Trend analysis

The number of medals peaks in the 20-30 age group, which may be related to the prime competitive age of athletes. There is a significant decrease in the number of medals for those over 35 years of age, which may be related to the retirement age of the athletes.

(5) Extreme age groups

The very low number of medals in the 10-15 and 55+ age groups may be related to the low number of participants in competitive sport in these age groups.

This chart can be used to analyze the performance of athletes of different ages and genders in sports competitions.

This bar chart can be used to analyze the performance of athletes of different ages and genders in competitive sports, as well as possible age and gender differences.

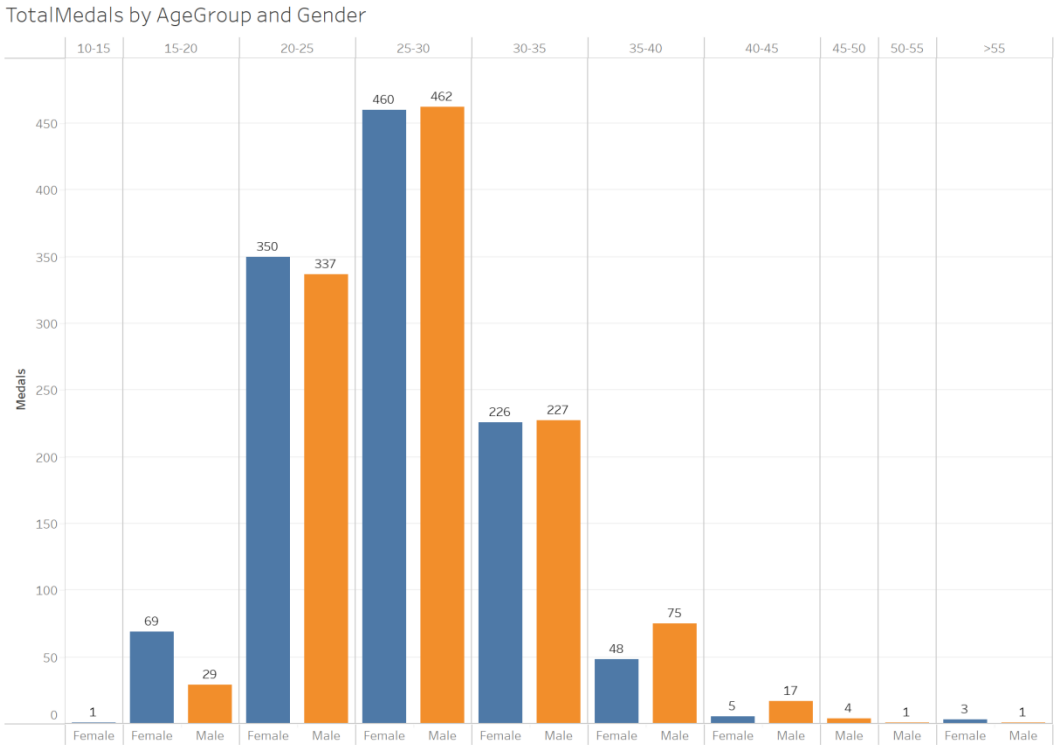


Figure 5: Total Medals by Age Group and Gender

4.4 Each Event Medals by Age Group

This bar chart contains statistics on the number of medals won by different age groups in various sports. The document lists several sports, including but not limited to 3x3 basketball, archery, artistic gymnastics, and synchronized swimming, and etc. There are also age groupings. Data is filtered by age group, sporting event. Gradient green was used to visualize differences in the data across age groups. Used to analyze the performance of different age groups in various sporting disciplines. This

document may be used to analyze the performance of different age groups in a variety of sports, and at the same time allow for targeted training based on the performance of different age groups in different sports.

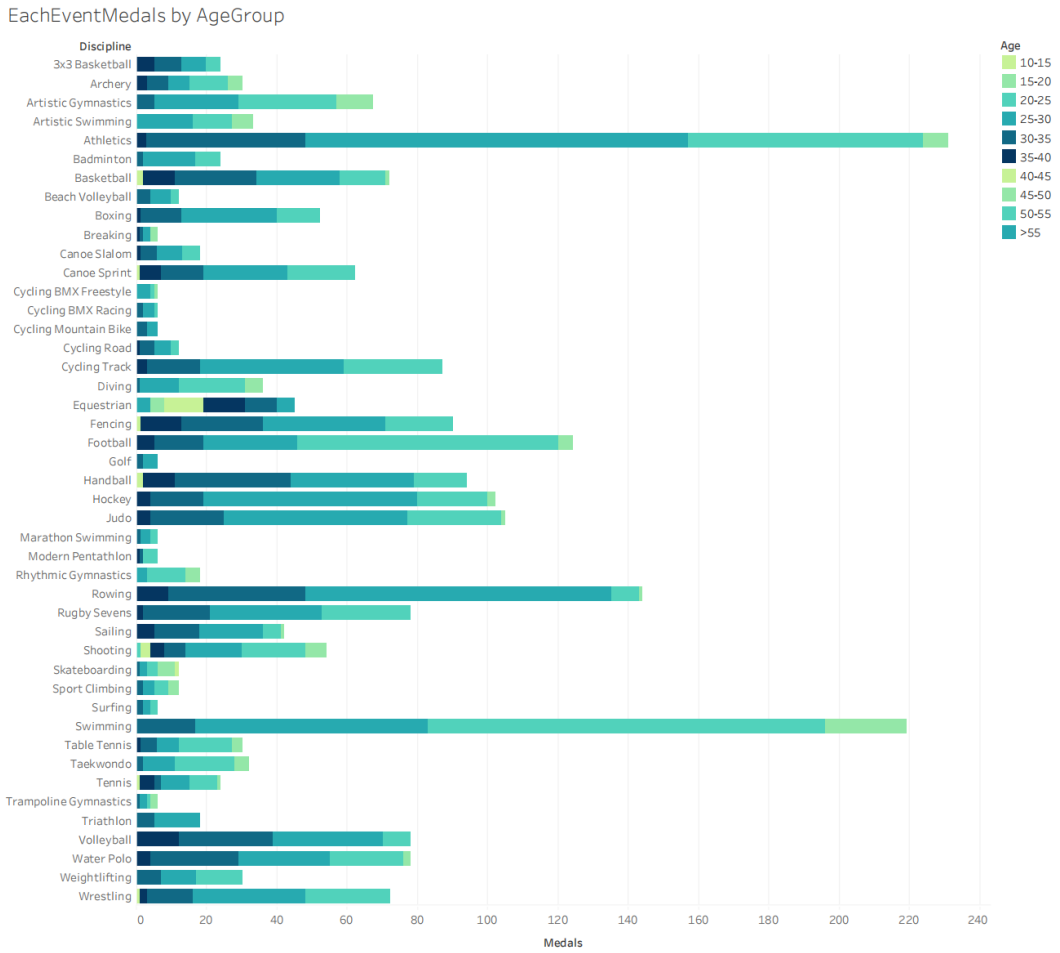


Figure 6: Each Event Medals by Age Group

4.5 Medals by Province in China

(1) Leading Province:

Zhejiang Province (ZHE JIANG) and Sichuan Province (SI CHUAN) tied for first place in the medal count with 20 medals each. Guangdong Province (GUANG DONG) followed with 18 medals.

(2) Medium-Performing Provinces:

Shanghai (SHANG HAI) has 8 medals. Liaoning (LIAO NING) has seven medals. Hebei Province (HE BEI) and Beijing Municipality (BEI JING) have six medals each.

Chinese Special Administrative Region (SAR):

Taiwan Province of China (TAI PEI) has 8 medals. Hong Kong Special Administrative Region (HKSAR) has 4 medals.

This chart can be used to analyze the performance of Chinese provinces in competitive sports and possible regional differences.

This chart can be used to analyze the performance of Chinese provinces in sports competitions.

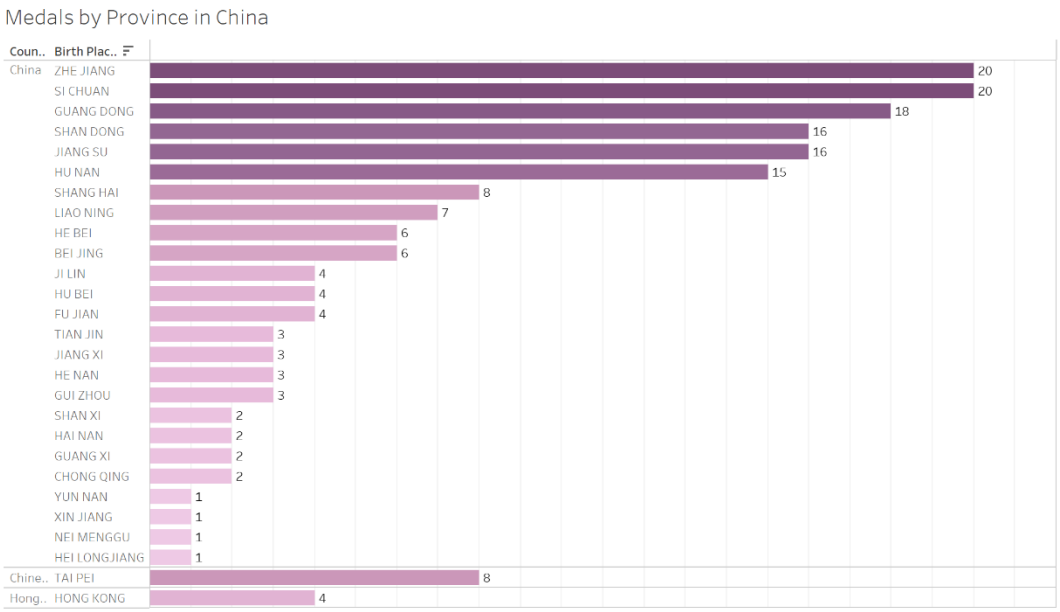


Figure 7: Medals by Province and Discipline in China

4.6 Medals by China Province and Discipline

This chart shows the number of medals won by different provinces in China in various sports disciplines. Some provinces excelled in specific disciplines, such as Zhejiang Province's dominance in swimming and Sichuan Province's performance in hockey and artistic gymnastics. The distribution of medals may be related to the allocation of sports resources, sports culture, and athlete training system in each province. Some provinces may have traditional strengths in certain programmes, which may be related to historical, geographical, social and other factors. It can be used as a reference for sport policy makers, coaches and athletes to help them better understand the current state of competition and develop appropriate strategies.

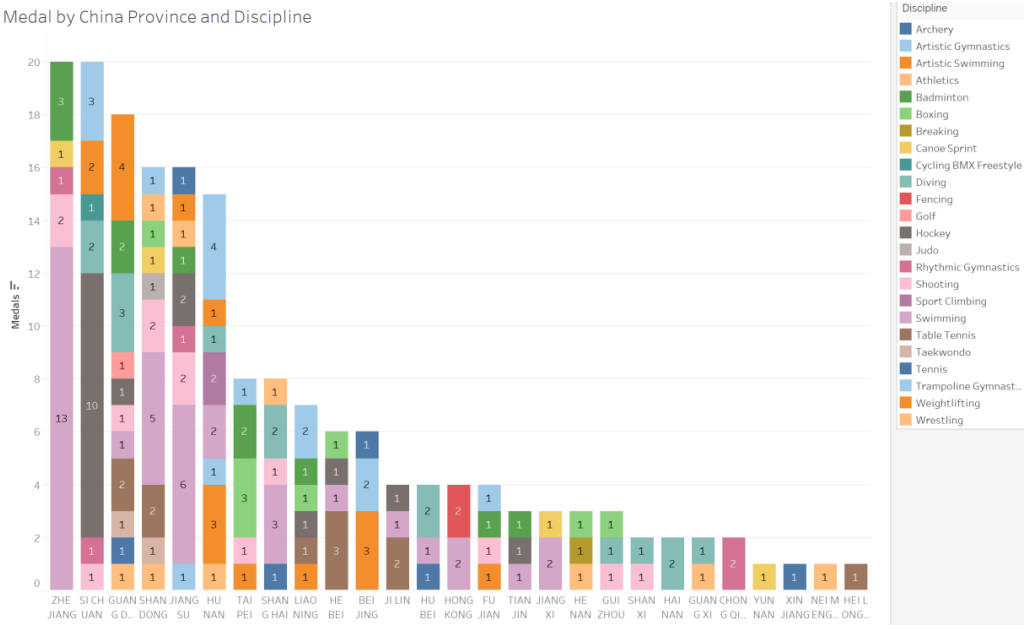


Figure 8: Medal by Province and Displine in China

4.7 Medals(single/team) by Province in China Map

Eastern coastal provinces such as Zhejiang, Jiangsu, Shandong and Guangdong are shown as dark colors on the map, indicating that these regions have excelled in sports competitions and won more medals. Central provinces such as Hunan, Hubei and Henan are colored lighter on the map, which could mean that these regions have relatively fewer medals in competitive sports. Western provinces such as Sichuan, Shaanxi and Gansu are also colored darker on the map, showing that these regions also have some strength in sports competitions. Northeastern provinces such as Liaoning and Jilin have lighter colors on the map, possibly indicating that these regions do not win many medals in competitive sports. The map shows that the provinces with stronger sports competitiveness are mostly concentrated in the more economically developed eastern coastal regions, which may be related to the sports resources, facilities, coaching level, and athlete selection and training mechanisms in these regions.

This map can be used as a powerful tool for analyzing the competitive strength of sports and the distribution of sports resources across China's provinces, helping to formulate more effective sports development strategies.

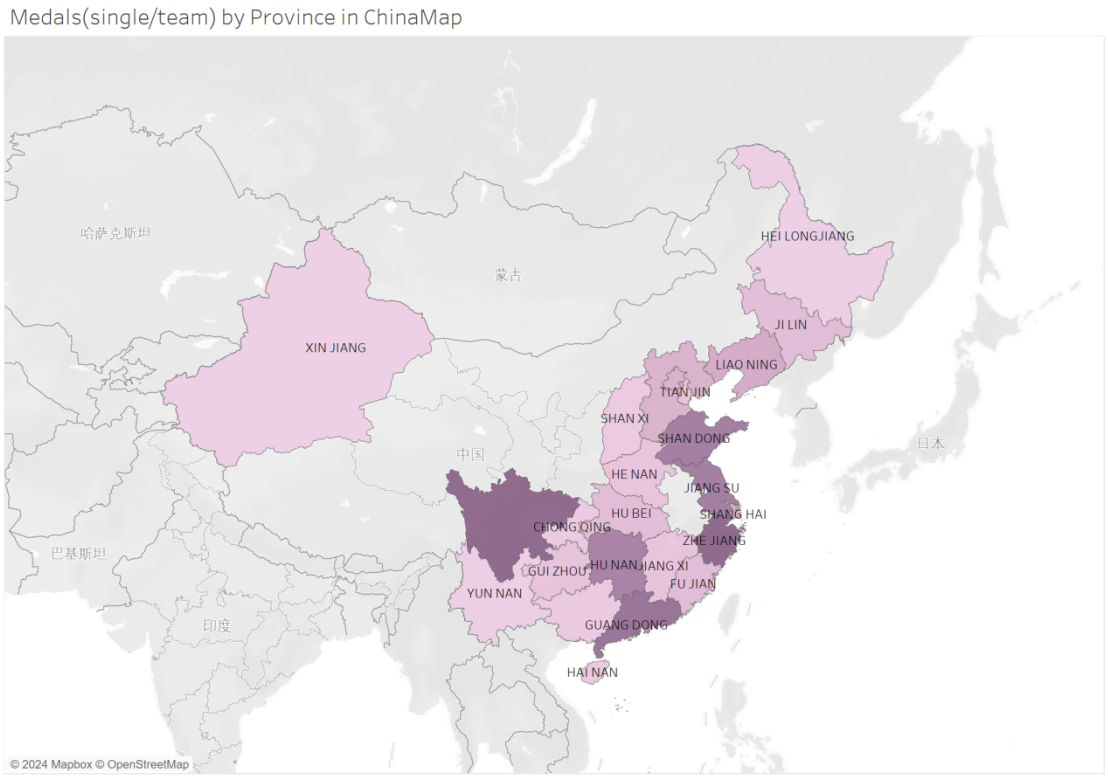


Figure 10: Medals(single/team) by Province in China Map

4.8 History Total Medals by Country

This is a line graph showing the trend in the total number of medals (top graph) and gold medals (bottom graph) won by different countries at the Olympic Games between 2000 and 2024. The chart is analysed and summarised below:

(1) China

The total number of medals peaked at 100 in 2008 and has fluctuated since then, but has remained at a high level. The number of gold medals peaked at 48 in 2008 and has declined since then, but remains a major contender for gold.

(2) Great Britian

The total number of medals peaked in 2012 and 2016, which may be related to being the host of the Olympics. The number of gold medals peaked at 29 in 2012 and has declined since.

(3) United States

The total number of medals peaked in 2004 and 2012 and has fluctuated since then, but has remained high overall. The number of gold medals peaked at 39 in 2004 and has declined since then, but remains a major contender for gold.

The number of medals won by each country is influenced by a number of factors, including sports policy, economic inputs, and the athlete development system. Countries that host the Olympic Games, such as Greece, Great Britain and China, have seen a significant increase in the number of medals won in their host years. China's historic success at the 2008 Beijing Olympics has been followed by some volatility, but overall performance remains strong. Fluctuations in the number of medals may be related to national sports development strategies, athlete performance and changes in the international competitive landscape of sport. These data can inform national sports policy makers, helping them to better understand their country's position in international sports competition and to develop appropriate strategies to raise the level of competition.

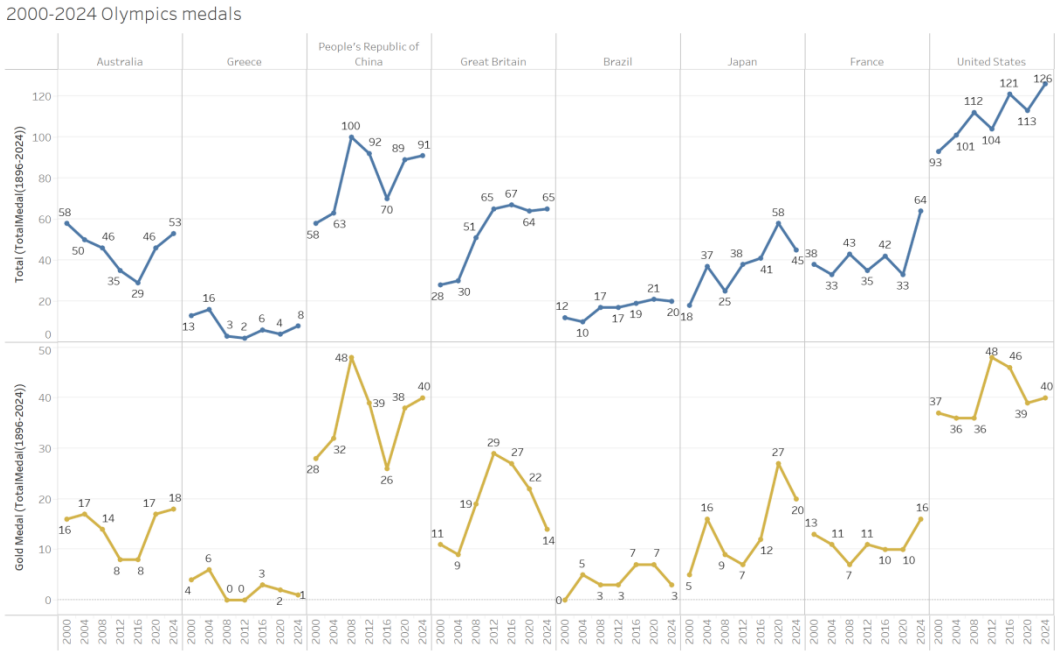


Figure 10: 2000-2024 Olympics Medals

4.9 Dashboard

The dashboard of figure 11 provides a comprehensive view of Olympic medal distribution from 2000 to 2024, including a world map, medal statistics and national medal trends.

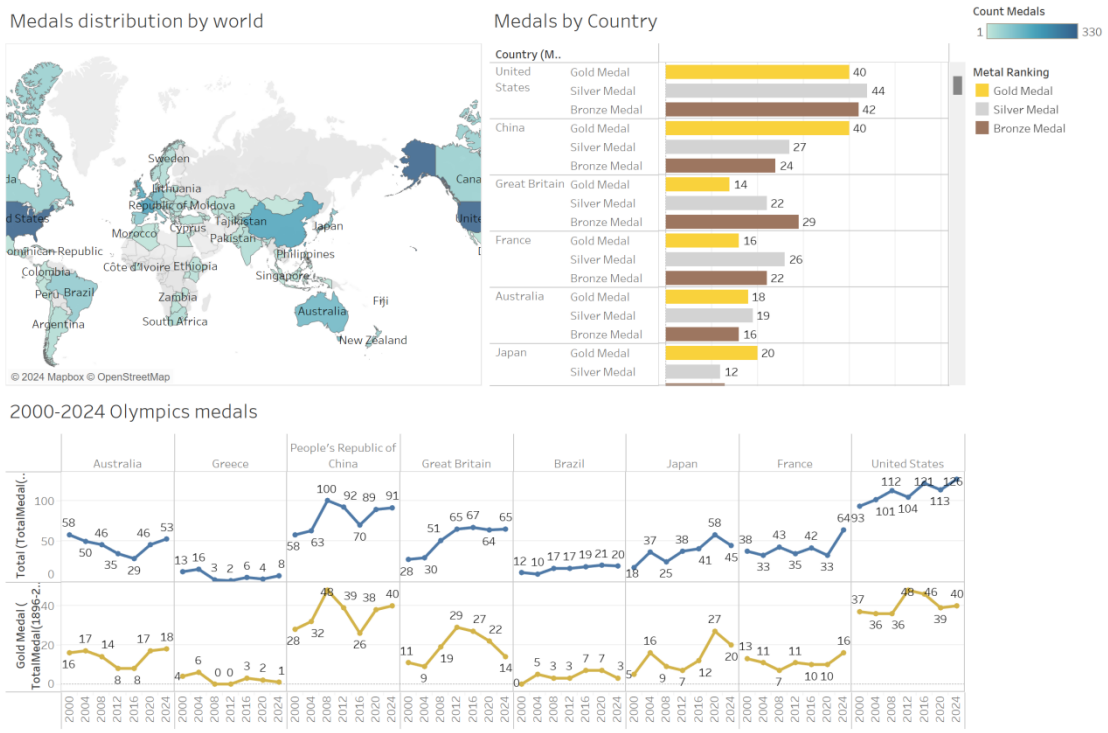


Figure 11: World Medal Distribution Dashboard

This dashboard of figure 12 provides a clear view of the distribution of medals across China's provinces in different sports.

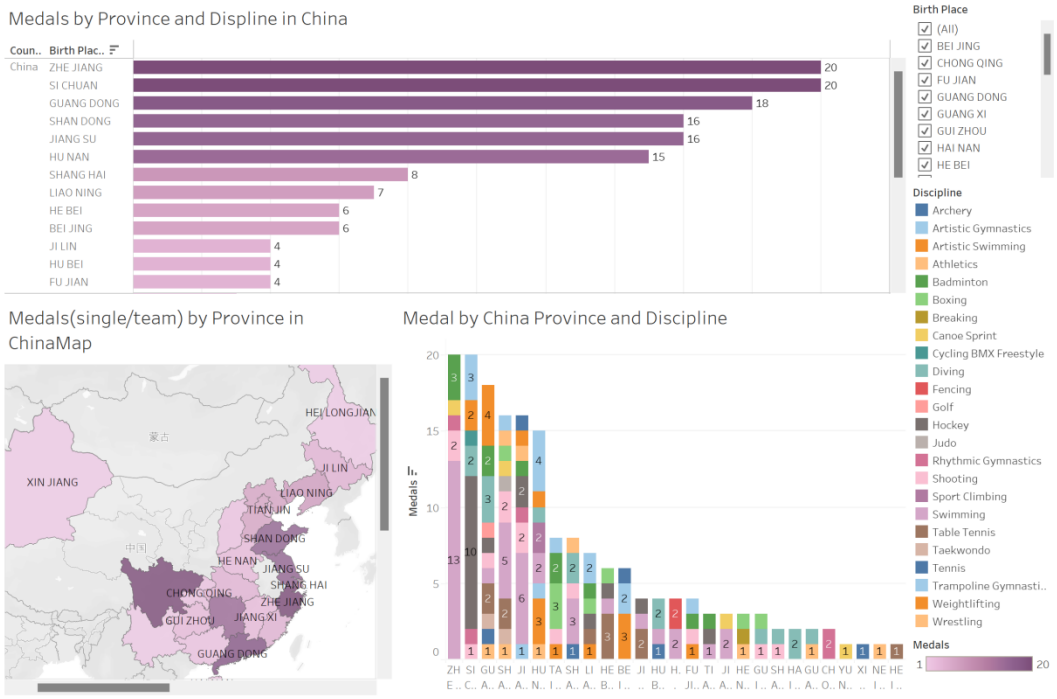


Figure 12: China Province Distribution Dashboard

This dashboard of Figure 13 consists of two main sections showing the total number of medals by age

group and gender, and the distribution of medals by age group in each sport.

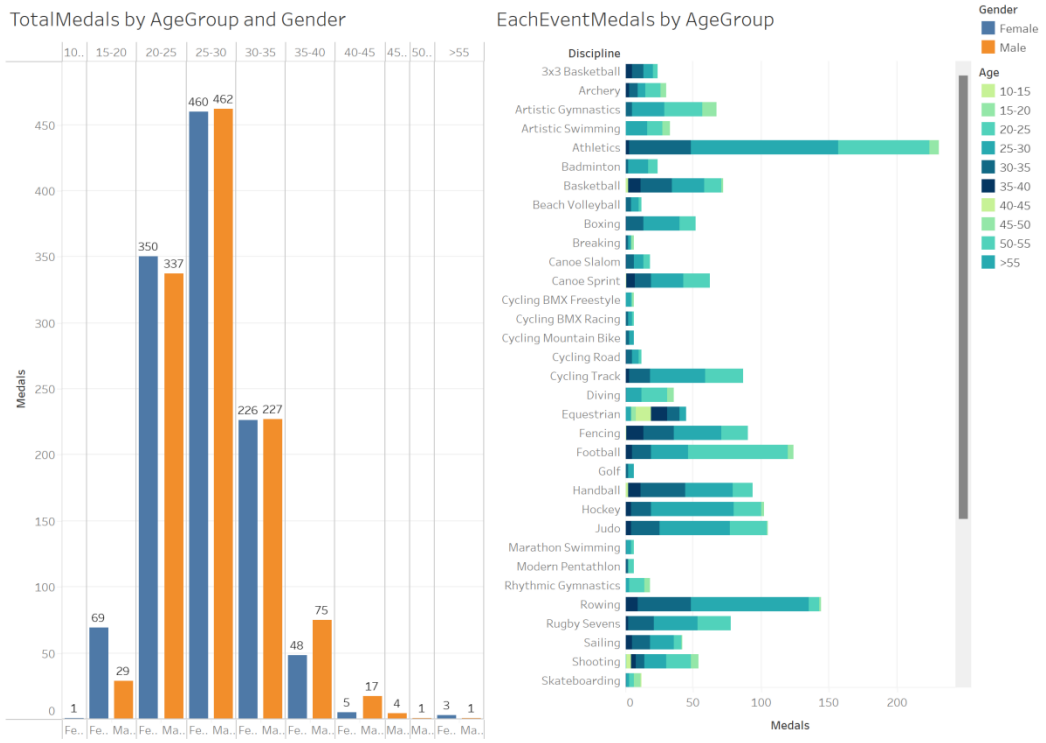


Figure 13: Age Medal Distribution

With the dashboard above, you can select the item you want to see in detail, and the dashboard will highlight the data you want to see.

5 Data Prediction Part

The goal of this study is to predict the total medal counts for key countries in the 2028 Los Angeles Olympic Games based on historical data from 1896 to 2024. The countries mainly included in this analysis are: United States (USA), China (CHN), France (FRA), Australia (AUS), Japan (JPN), Brazil (BRA)

5.1 Data and Feature Engineering

The analysis relies on publicly available historical Olympic medal data from 1896 to 2024, which includes each country's total medal count for every Olympic Games. To ensure robust predictions, advanced feature engineering was applied to extract meaningful insights from the data, including:

- Lag Features: Metrics such as the total medal count in the previous Olympics (prev_total) and the total count from two Olympics ago (prev_2_total).
- Trend Features: Metrics such as the average medal count over the past three Olympics (avg_last_3), capturing long-term performance trends and consistency.
- Temporal Features: The year of each Olympics (year), allowing the model to account for changes

over time in medal distribution.

These features were carefully crafted to capture the historical and dynamic nature of Olympic medal performance, providing the models with high-quality inputs for accurate prediction.

5.2 Methodology and Models

To ensure a comprehensive analysis and reliable predictions, three different machine learning models were employed:

(1) Decision Tree Regressor

- A non-parametric model that recursively partitions the data space to capture non-linear relationships.
- Decision trees are highly interpretable, making it easier to identify which features play the most significant roles in medal prediction.

(2) Random Forest Regressor:

- An ensemble learning method that builds multiple decision trees and averages their outputs to improve robustness and accuracy.
- Random forests reduce the risk of overfitting and provide feature importance metrics to evaluate the contribution of each feature.

(3) Linear Regression:

- A baseline model that assumes a linear relationship between features and the target (medal counts).
- While linear regression is less capable of capturing non-linear trends, it serves as a valuable reference point for comparing the performance of more complex models.

In order to fully explore whether the relationship between the data of countries will have an impact on the result of the outcome prediction, so in the data prediction part, our overall data prediction is divided into two ways, that is, the number of medals of all countries from 1896 to 2024 as the data reference of the model prediction and the number of medals of only one country from 1896 to 2024 as the data reference of the model prediction.

Then, in order to determine whether the model used is accurate or not, in addition to the Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2), etc., in the data prediction process, we will delete the data of the specified year that already exists and use the previous data to predict the data of that year again, and then use the predicted results to predict the data of that year again, and then use the predicted results to predict the data of that year again, and then use the predicted results to predict the data of that year. For example, in this project, we filtered out the data from 2024 and only kept the data before 2024 to predict the number of medals for the country in 2024, and in this way we compared the predicted data from 2024 with the actual data from 2024 to determine the accuracy of the model predictions.

5.3 Model Training and Evaluation

To ensure the reliability of the predictions, the data was split into training and testing sets. Training data included all historical data from 1896 to 2020. The following steps were carried out:

(1) Feature Selection and Engineering:

Lagged and trend features were extracted to provide comprehensive historical context and performance indicators.

(2) Model Training and Hyperparameter Tuning:

Each model was trained using the training data, and hyperparameters (e.g., the number of trees in Random Forest, maximum depth of Decision Trees) were optimized to achieve the best performance.

(3) Model Performance Evaluation:

Metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 were used to evaluate each model's accuracy and reliability.

(4) Prediction:

The best-performing models were used to predict the 2028 medal counts.

5.4 Data Prediction Implementation

(1) Random Forest Model Prediction by Single Country

Compared real data in 2024 and predict data in 2024:

- The mean square error (MSE): 89.05 and The Root Mean Square Error (RMSE): 9.44 indicate that the model predictions deviated from the actual values by an average of about 9 medals.
- The Mean Absolute Error (MAE): 7.39, indicating that the model predictions deviated from the actual values by an average of about 7 medals each time.
- The coefficient of determination (R^2) is 0.86, indicating that the model explains approximately 86 per cent of the variation in the data, which means that the model fits the historical data well.

```
The historical data for the last 3 years:
 year  Total
 2012  104.0
 2016  121.0
 2020  113.0

Medal prediction for USA in 2024:
The predicted number of medals: 113

Model performance on historical data:
Mean Squared Error (MSE): 89.05
Root Mean Squared Error (RMSE): 9.44
Mean Absolute Error (MAE): 7.39
Coefficient of Determination (R2): 0.86
```

Predict data in 2028:

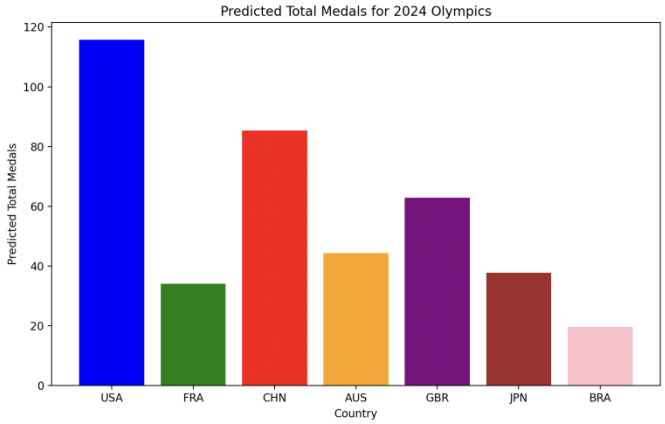
- Mean Square Error (MSE): 93.56 and the Root Mean Square Error (RMSE): 9.67, indicating that the average deviation from the model's predictions was approximately 9.67 medals.
- Mean Absolute Error (MAE): 7.24, indicating that the absolute value of the prediction bias averaged about 7.24 medals.
- The Coefficient of Determination (R^2): 0.86, indicating that the model explained 86 per cent of the variation in the historical data, which is a good fit.

```
The historical data for the last 3 years:
 year  Total
 2016  121.0
 2020  113.0
 2024  126.0

Medal prediction for USA in 2028:
The predicted number of medals: 118

Model performance on historical data:
Mean Squared Error (MSE): 93.56
Root Mean Squared Error (RMSE): 9.67
Mean Absolute Error (MAE): 7.24
Coefficient of Determination (R2): 0.86
```

(2) Random Forest Model Prediction by All Countries

Compared real data in 2024 and predict data in 2024:

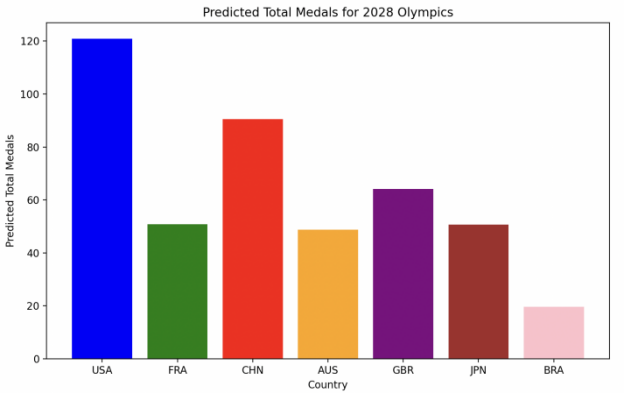
- Mean Absolute Error (MAE): 4.27, which means that the predicted results differed from the actual value by about 4 medals on average.
- Root Mean Square Error (RMSE): 8.50, indicating that the range of prediction deviations is small.
- Coefficient of determination (R^2): 0.82, indicating that the model explains 82% of the variation in the historical data, which is a good fit.




```
Mean Absolute Error (MAE): 4.2737174721189595
Root Mean Squared Error (RMSE): 8.495804978369222
Coefficient of Determination (R2): 0.8153847112642223
Predicted total medals for USA in 2024: 115.72
Predicted total medals for FRA in 2024: 34.01
Predicted total medals for CHN in 2024: 85.28
Predicted total medals for AUS in 2024: 44.31
Predicted total medals for GBR in 2024: 62.92
Predicted total medals for JPN in 2024: 37.68
Predicted total medals for BRA in 2024: 19.53
```

Predict data in 2028:

- Mean Absolute Error (MAE): 4.66, with an average difference of about 4.66 medals between the predicted and actual values.
- Root Mean Square Error (RMSE): 14.25, indicating a certain predictive volatility and a wide margin of error.
- Coefficient of determination (R^2): 0.64, the model only explains about 64% of the variation in the historical data.

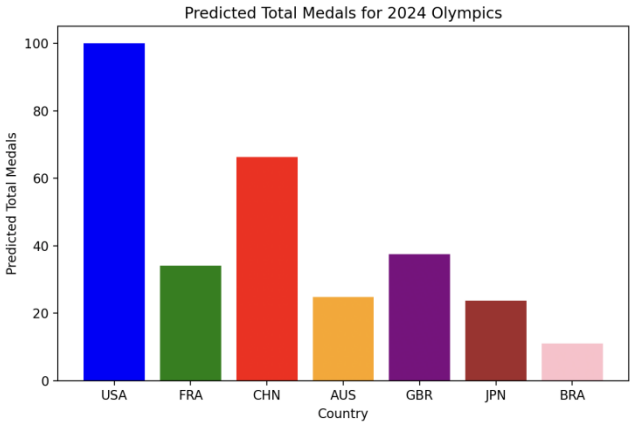



```
Mean Absolute Error (MAE): 4.6606620209059235
Root Mean Squared Error (RMSE): 14.249069423738124
Coefficient of Determination (R2): 0.6350106939660585
Predicted total medals for USA in 2028: 120.87
Predicted total medals for FRA in 2028: 50.88
Predicted total medals for CHN in 2028: 90.52
Predicted total medals for AUS in 2028: 48.70
Predicted total medals for GBR in 2028: 64.13
Predicted total medals for JPN in 2028: 50.64
Predicted total medals for BRA in 2028: 19.61
```

(3) Linear Regression Model Prediction by All Countries

Compared real data in 2024 and predict data in 2024:

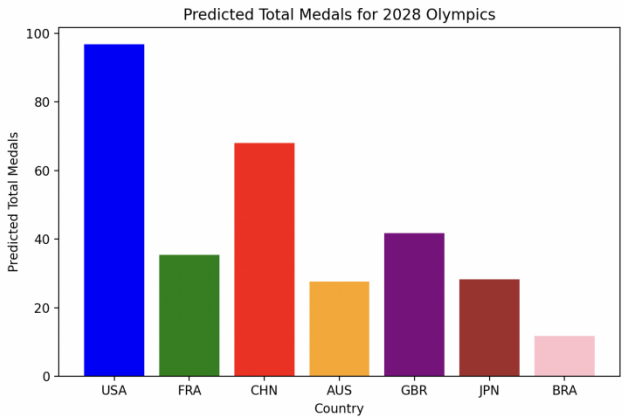
- Mean Absolute Error (MAE): 5.97, indicating that the average difference between the predicted value and the actual value is about 6 medals.
- Root Mean Square Error (RMSE): 10.67, indicating a wide range of error.
- Coefficient of determination (R^2): 0.71, the model explains about 71% of the variation in the data, which is a good fit.




```
Mean Absolute Error (MAE): 5.966301869821674
Root Mean Squared Error (RMSE): 10.673792495877834
Coefficient of Determination (R2): 0.7085955993917666
Predicted total medals for USA in 2024: 100.04
Predicted total medals for FRA in 2024: 34.08
Predicted total medals for CHN in 2024: 66.25
Predicted total medals for AUS in 2024: 24.81
Predicted total medals for GBR in 2024: 37.52
Predicted total medals for JPN in 2024: 23.73
Predicted total medals for BRA in 2024: 10.98
```

Predict data in 2028:

- Mean Absolute Error (MAE): 5.90, the average difference between the predicted value and the actual value is about 5.9 medals.
- Root Mean Square Error (RMSE): 13.58, indicating a wide range of error.
- Coefficient of determination (R^2): 0.67, the model explains about 67% of the variation in the data and the fit is fair.



```
Mean Absolute Error (MAE): 5.895192894803215
Root Mean Squared Error (RMSE): 13.57952753974394
Coefficient of Determination (R2): 0.6685054019820609
Predicted total medals for USA in 2028: 96.89
Predicted total medals for FRA in 2028: 35.37
Predicted total medals for CHN in 2028: 68.01
Predicted total medals for AUS in 2028: 27.58
Predicted total medals for GBR in 2028: 41.70
Predicted total medals for JPN in 2028: 28.26
Predicted total medals for BRA in 2028: 11.75
```

(4) Linear Regression Model Prediction by Single Country

Compared real data in 2024 and predict data in 2024:

- Mean Square Error (MSE): 364.33, which is high error.
- Root Mean Square Error (RMSE): 19.09, indicating that the average deviation of the predicted value from the actual value is about 19 medals.
- Mean Absolute Error (MAE): 12.49, which means that the average absolute error in the forecast is 12.49 medals.
- Coefficient of determination (R^2): 0.43, the model explains only about 43 per cent of the variation in the data, which is a poor fit.

```
The historical data for the last 3 years:
 year  Total
 2012  104.0
 2016  121.0
 2020  113.0

Medal prediction for USA in 2024:
The predicted number of medals: 117

Model performance on historical data:
Mean Squared Error (MSE): 364.33
Root Mean Squared Error (RMSE): 19.09
Mean Absolute Error (MAE): 12.49
Coefficient of Determination (R2): 0.43
```

Predict data in 2028:

- Mean Square Error (MSE): 353.07, a large error.
- Root Mean Square Error (RMSE): 18.79, the average deviation of the predicted value from the actual value is about 18.79 medals.
- Mean Absolute Error (MAE): 12.42, the average absolute error of the predicted values is 12.42 medals.
- Coefficient of determination (R^2): 0.46, the model explains only about 46% of the variation in the data, which is a poor fit.

```
The historical data for the last 3 years:
 year  Total
 2016  121.0
 2020  113.0
 2024  126.0

Medal prediction for USA in 2028:
The predicted number of medals: 120

Model performance on historical data:
Mean Squared Error (MSE): 353.07
Root Mean Squared Error (RMSE): 18.79
Mean Absolute Error (MAE): 12.42
Coefficient of Determination (R2): 0.46
```

(5) Decision Tree Model Prediction by Single Country

Compared real data in 2024 and predict data in 2024:

- Mean Square Error (MSE): 2.83, very small error.
- Coefficient of Determination (R^2): 1.00, the model fits the historical data perfectly, which indicates that its predictive ability is very strong.

```
The historical data for the last 3 years:
 year  Total
 2012  104.0
 2016  121.0
 2020  113.0

Medal prediction for USA in 2024:
The predicted number of medals: 121
Mean Squared Error (MSE): 2.83
Coefficient of Determination (R2): 1.00
Root Mean Squared Error (RMSE): 1.68
Mean Absolute Error (MAE): 0.93
```

Predict data in 2028:

- Mean Square Error (MSE): 8.38, the error is small, indicating that the predicted values are close to the actual values.
- Coefficient of determination (R^2): 0.99, the model explains 99% of the variation in the data and the fit is very good.

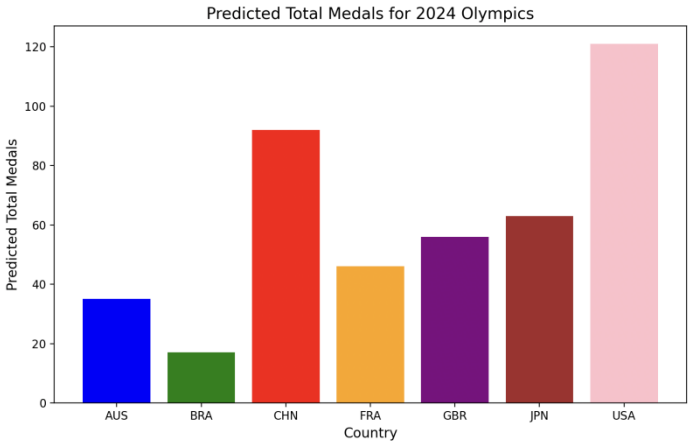
```
The historical data for the last 3 years:
 year  Total
 2016  121.0
 2020  113.0
 2024  126.0

Medal prediction for USA in 2028:
The predicted number of medals: 113
Mean Squared Error (MSE): 8.38
Coefficient of Determination (R2): 0.99
Root Mean Squared Error (RMSE): 2.90
Mean Absolute Error (MAE): 1.81
```

(6) Decision Tree Prediction by All Countries

Compared real data in 2024 and predict data in 2024:

- Mean Square Error (MSE): 0.15, very small error.

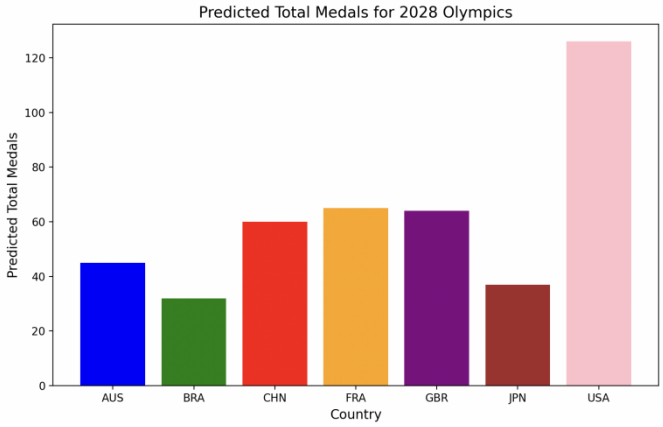
- Root Mean Square Error (RMSE): 0.39, the prediction is very close to the actual value.
- Mean Absolute Error (MAE): 0.10, very low average error.
- Coefficient of Determination (R^2): 1.00, the model fits the historical data perfectly and performs very well.




```
Predicted medal totals for selected countries in 2024:
         predicted_total
country
AUS                35.0
BRA                17.0
CHN                92.0
FRA                46.0
GBR                56.0
JPN                63.0
USA               121.0

Model performance on historical data:
Mean Squared Error (MSE): 0.15
Root Mean Squared Error (RMSE): 0.39
Mean Absolute Error (MAE): 0.10
Coefficient of Determination (R2): 1.00
```

Predict data in 2028:

- Mean Square Error (MSE): 0.15, which is very small.
- Root Mean Square Error (RMSE): 0.39, indicating that the predicted value is almost the same as the actual value.
- Mean Absolute Error (MAE): 0.10, the average error is very low.
- Coefficient of determination (R^2): 1.00, the model fits the historical data perfectly.



```
Predicted medal totals for selected countries in 2028:
         predicted_total
country
AUS                45.0
BRA                32.0
CHN                60.0
FRA                65.0
GBR                64.0
JPN                37.0
USA               126.0

Model performance on historical data:
Mean Squared Error (MSE): 0.15
Root Mean Squared Error (RMSE): 0.39
Mean Absolute Error (MAE): 0.10
Coefficient of Determination (R2): 1.00
```

5.6 Summary of Data Prediction

According to the resultant parameters given by all the methods above, it can be seen that perhaps the data predicted using the Decision Tree Model turned out to be the most accurate, with almost no error, if we look at the accuracy of the data alone, but it can be observed that the value of R^2 is almost equal to 1, and that both the MSE and the MAE are close to being equal to 0, which implies that the model is actually overfitting the features of the training data, the Instead of learning more generalised laws. In addition, the decision tree model itself is prone to overfitting the model may perform perfectly on the training data but may fail when predicting future data. Therefore, combining all the parameter outputs for analysis, in this data prediction, we believe that the random forest model worked the best due to the fact that there was no overfitting and no excessive deviation from the actual data.

5.7 Special Case Analytics

Regarding the prediction of the number of medals for the 2028 Olympic Games in Los Angeles, we can refer to the results given by random forest model. But there is a point worth mentioning is that we can observe that the host country of 2020, Japan, and the host country of 2024, Paris, in fact, there is a certain difference between the actual data and the predicted data of both of them in that year, and it can be clearly seen that the actual data will basically be much higher than the predicted data. Because our predictions are based on past data, it is actually an exception for the host country, meaning that the host country will win more medals in the year it hosts the Olympics than it did in the past, so it is reasonable that there will be a deviation between the predicted results and the actual results. The following are the results predicted by Random Forest Model.

	HOST Olympics	Previous Olympics	Average before	Difference vs. Average	growth rate compared previous
Sydney(2000)	41	27	13.85	27.15	51.9%
Athens(2004)	16	13	5.68	10.32	23.1%
Beijing(2008)	100	63	47.5	52.5	58.7%

London(2012)	51	30	27.2	23.8	70%
Rio(2016)	19	17	6.11	12.89	11.8%
Tokyo(2020)	58	41	21	37	41.5%
Paris(2024)	64	33	27.6	36.4	93.9%

(1) France Case:

- Actual Data**:** Total number of medals is 64 in 2024
- Predicted Data**:** Total number of medals is 33 to 39 in 2024

```
The historical data for the last 3 years:
 year  Total
 2012   35.0
 2016   42.0
 2020   33.0

Medal prediction for FRA in 2024:
The predicted number of medals: 36

Model performance on historical data:
Mean Squared Error (MSE): 18.96
Root Mean Squared Error (RMSE): 4.35
Mean Absolute Error (MAE): 3.38
Coefficient of Determination (R2): 0.86
```

(2) Japan Case:

- Actual Data**:** Total number of medals is 58 in 2020
- Predicted Data**:** Total number of medals is 35 to 39 in 2020

```
The historical data for the last 3 years:
 year  Total
 2008   25.0
 2012   38.0
 2016   41.0

Medal prediction for JPN in 2020:
The predicted number of medals: 37

Model performance on historical data:
Mean Squared Error (MSE): 11.10
Root Mean Squared Error (RMSE): 3.33
Mean Absolute Error (MAE): 2.73
Coefficient of Determination (R2): 0.85
```

Thus, it is reasonable to conclude that because the next host country of the 2028 Olympics is the United States, the total number of medals won by the United States will be higher than the range of 111 to 125 or 116 to 124 medals predicted by the Random Forest model based on the predict by single country and predict by all country methods, respectively. According to predict by single country and predict by all country, the MAE values of 7.24 and 4.66 are given, and the approximate number of medals will in the range of 120 to 130 or more than 130.

6 Conclusion

Based on the statistical analysis of the 2024 Paris Olympic dataset, we've identified several key findings regarding athlete performance and medal distribution. The majority of medal winners are

concentrated in the 24-29 age group, indicating that this may be the peak competitive period for athletes. The medal tally is dominated by sports powerhouses, with the United States and China leading in overall medal counts, likely due to their comprehensive sports policies, resource allocation, and athlete training systems. Additionally, there is a correlation between a country's economic strength and its sports performance, with nations like the U.S., China, Japan, and Germany excelling in both areas. Regional disparities in sports performance within China are also noted, potentially linked to differences in sports resource distribution, culture, and training systems across provinces.

From a data analytics perspective, the report reveals that the number of medals won by countries at the Olympics is influenced by a complex interplay of factors, including sports policies, economic investments, and athlete development systems. Host countries tend to win more medals during their hosting years, as seen with Greece, the UK, and China. Predictive modeling plays a crucial role in forecasting future medal counts, with the random forest model outperforming others by providing predictions with minimal deviation from actual data and without overfitting, making it the most reliable model for such predictions. These insights underscore the importance of data-driven strategies in optimizing athlete selection, training, and participation in future Olympic Games.

Improvement

Dataset Expansion:

- Add more historical data to enhance trend analysis and the accuracy of predictive models.
- Introduce more relevant variables, such as athletes' training backgrounds, injury histories, and psychological states.

Model Optimization:

- Tune parameters of machine learning models like Random Forest to improve the precision and generalization of predictions.
- Attempt ensemble learning methods, such as Boosting or Stacking, to further enhance model performance.
- Further explore feature selection and transformation techniques to extract more meaningful information.
- Create interaction features or polynomial features to capture complex relationships between variables.

Result Validation:

Use techniques such as cross-validation to assess the stability and reliability of models.

Reference

[1] Schmidtke, J., & Schmidt, K. (2006). Data management and data base implementation for GMO monitoring. *Journal Für Verbraucherschutz Und Lebensmittelsicherheit*, *1*(S1), 92–94. https://doi.org/10.1007/s00003-006-0096-0

[2] Ordonez, C., Varghese, R., Phan, N., & Macyna, W. (2024). Growing a FLOWER: Building a Diagram Unifying Flow and ER Notation for Data Science. *Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics*, 1–8. https://doi.org/10.1145/3665939.3665958

[3] Tao, M., Nawaz, M. Z., Nawaz, S., Butt, A. H., & Ahmad, H. (2018). Users' acceptance of innovative mobile hotel booking trends: UK vs. PRC. *Information Technology & Tourism*, *20*(1–4), 9–36. https://doi.org/10.1007/s40558-018-0123-x

[4] Chen, P. P.-S. (1977). The entity-relationship model: a basis for the enterprise view of data. *Proceedings of the June 13-16, 1977, National Computer Conference*, 77–84. https://doi.org/10.1145/1499402.1499421

[5] Weldon, J.-L. (1975). Review of "An introduction to database systems" by C. J. Date. Addison-Wesley Publishing Co. 1975. SIGMOD Record, 7(3–4), 53–54. https://doi.org/10.1145/984403.984407

[6] Hayden, R. W. (2010). A Review of: "Now You See It: Simple Visualization Techniques for Quantitative Analysis, by S. C. Few,": Oakland, CA: Analytics Press, 2009, ISBN 0-9706019-8-0, xi + 327 pp., $45 [Review of A Review of: "Now You See It: Simple Visualization Techniques for Quantitative Analysis, by S. C. Few,": Oakland, CA: Analytics Press, 2009, ISBN 0-9706019-8-0, xi + 327 pp., $45]. Journal of Biopharmaceutical Statistics, 20(3), 701–702. Taylor & Francis Group. https://doi.org/10.1080/10543401003641225

[7] Sun, N., Yang, X., & Liu, Y. (2020). TableQA: a Large-Scale Chinese Text-to-SQL Dataset for Table-Aware SQL Generation. arXiv.Org.

[8] HAERDER, T., & REUTER, A. (1983). Principles of transaction-oriented database recovery. ACM Computing Surveys, 15(4), 287–317. https://doi.org/10.1145/289.291